SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

GLOBAL UNIQUENESS IN GEOMETRIC SINGULAR PERTURBATION THEORY*

NEIL FENICHEL[†]

Abstract. Estimates are developed for the order of contact of center manifolds. It follows from these estimates that global center manifolds have arbitrarily order of contact, even if they are not differentiable. These results give the geometric justification for a large class of uniqueness results in singular perturbation theory.

The purpose of this paper is to present a global uniqueness result for order of contact of center manifolds. This result is the geometric form of the uniqueness of asymptotic expansions in singular perturbation theory.

Center manifolds are invariant manifolds characterized by neutral growth conditions. In [F1], global center manifolds appear as the basic construct in geometric singular perturbation theory. In singular perturbation theory the characterization in terms of neutral growth means that center manifolds are filled with orbits which have no transition layers.

Following [F1], we use global center manifolds to study the geometry of singular perturbation theory. The existence of the center manifolds reflects the presence of distinct time scales. The global center manifolds of interest are not unique and may not be smooth, but they have uniqueness and smoothness properties which are useful for computations and estimates. It is shown in [F1] that certain invariant subsets of global center manifolds are unique, and that over those invariant sets certain formal derivatives are unique. Under appropriate conditions there are center manifolds with any desired finite degree of smoothness, and the unique formal derivatives are ordinary derivatives of the nonunique smooth realizations. In [F1] it is shown that the derivatives of center manifolds and related structures are the coefficients of asymptotic expansions of the inner solutions and outer corrections of singular perturbation theory. It therefore is desirable to have a better understanding of how the unique asymptotic expansions are derived from nonunique geometric structures.

The main technical result of this paper is a contraction inequality from which we derive an estimate for the order of contact of center manifolds. Order of contact is the geometric structure corresponding to asymptotic expansions in analysis, so our estimate is a powerful uniqueness theorem for asymptotic expansions.

1. Uniqueness results for center manifolds. The simplest uniqueness properties of center manifolds are local and algebraic. If the graph of a center manifold of a C^r system is represented near an equilibrium point p as the graph of a C^r function, then all derivatives of the function up to order r are uniquely determined at p. The local algebraic uniqueness of center manifolds has been known for some time, and has been used in applications in bifurcation theory and in singular perturbation theory. See, for example, [F2], Ruelle and Takens [R-T], Lanford [L], Hassard and Wan [H-W], [F3], and [F4]. The method appears to be much older than the references. See Segre [S1]. Wan proved a complete local algebraic uniqueness theorem for invariant manifolds [W]. As a special case, the local algebraic uniqueness of center manifolds is completely resolved.

* Received by the editors December 23, 1981, and in revised form April 27, 1983.

[†] Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5.

Sijbrand [S2] proved a strong local uniqueness theorem for center manifolds. By a detailed analysis which involves the Jordan canonical form of the linearization of a system, rather than just the eigenvalues of the linearization, he estimated the transcendentally small terms in the distance between center manifolds. Sijbrand's result is much stronger than the local algebraic results.

[F1, Thm. 9.1] includes a global uniqueness result which asserts that any two C^r center manifolds agree to order r over specified invariant sets. The invariant sets include all equilibrium points and also many nonequilibrium points, so this theorem is much stronger than the local algebraic results. See, for example the computation of asymptotic expansions in [F1] for the Flatto–Levinson theorem [F-L]. The uniqueness result of [F1] is almost an algebraic uniqueness theorem, in the sense that it is proved by showing that the derivatives of a center manifold satisfy equations which have unique solutions. By passing to an infinite dimensional space of sections of a derived tangent bundle, one could regard the global uniqueness theorem as a local uniqueness theorem in an infinite dimensional space. The method of proof is analytic, however, and the proof cannot be reduced to algebra in a finite dimensional setting.

Subject only to some natural restrictions, there always exist global center manifolds with any desired degree of smoothness. In applications to singular perturbation theory this usually means that it is enough to compute formally using a sufficiently smooth center manifold. The selection of a smooth center manifold is not satisfying, however, as there generally is no fundamental reason to select one center manifold rather than another. When a unique computation depends on an arbitrary selection, an explanation is required.

In this paper we prove a global uniqueness theorem which requires only mild regularity of the center manifolds. We show that center manifolds have contact of specified order over specified invariant sets, even if the center manifolds are not sufficiently smooth for the results of [F1] to be applicable. Even in the local case, the estimate which underlies our proof is interesting, as it gives good estimates for the distances between nonunique center manifolds. In the local case, however, the results of Sijbrand are stronger.

2. Asymptotic expansions and order of contact of functions. We recall the usual definition of asymptotic expansion of functions. This definition is local and geometric. Functions $f(x)$ and $g(x)$ defined for x near 0 agree to order N at 0 if the graphs of f and g have contact of order N at 0. This definition is equivalent to the estimate

$$|f(x)-g(x)|/|x|^N \to 0 \quad \text{as } |x|\to 0.$$

A function $f(x)$ defined near 0 has an asymptotic expansion to order N at 0 if there is a polynomial which agrees with f to order N at 0, or equivalently if there is a polynomial whose graph has contact of order N with the graph of $f(x)$ at 0. Note that the preceding definitions do not require any differentiability. For sufficiently smooth functions the local geometric definition of asymptotic expansion is equivalent to a local algebraic definition. Smooth functions $f(x)$ and $g(x)$ have the same asymptotic expansion to order N at 0 if and only if they have the same derivatives to order N at 0.

3. An estimate for order of contact of center manifolds. Consider a system of differential equations with an equilibrium point at the origin,

$$\begin{aligned} x' &= f(x,y) = ax + \cdots, \\ y' &= g(x,y) = by + \cdots, \end{aligned} \tag{1}$$

where $x \in R^m$, $y \in R^n$, the eigenvalues of a lie on the imaginary axis, and the eigenvalues of b lie in the left half plane. A center manifold for (1) at the origin is any invariant manifold tangent to the x-axis at the origin.

Let $C: y=u(x)$ and $D: y=v(x)$ be two center manifolds for (1) at the origin. That is, let C and D be invariant manifolds for (1) which are tangent to the x-axis at the origin. To make our estimate as general as possible, we do not require that u and v be differentiable; instead we require that u and v be uniformly Lipschitz continuous near the origin.

Let

$$S(r)=\{x: \|x\| \leq r\}.$$

Let

$$d(r)=\sup|u(x)-v(x)|, \qquad x \in S(r).$$

To estimate the order of contact of C and D we will estimate $d(r)$ as $r \to 0$. The following lemma contains our main estimate. It asserts that $d(r)$ decays faster than r as $r \to 0$.

LEMMA. *There is a constant $\lambda < 1$, such that for any κ with $\lambda < \kappa < 1$, there exists r^0 such that*

$$d(\kappa r) \leq \lambda d(r) \tag{2}$$

for all r, $0 \leq r \leq r^0$.

COROLLARY. *$d(r)$ vanishes to arbitrarily high order as $r \to 0$.*

Proof of the corollary. It is sufficient to show that

$$d(r)/r^{\log\lambda/\log\kappa} \tag{3}$$

is bounded as $r \to 0$, because $\log\lambda/\log\kappa$ can be made arbitrarily large by taking κ sufficiently close to 1. Fix a small value r^0, and apply (2) repeatedly with $r=r^0$, κr^0, $\kappa^2 r^0$, and so on. Combining the first j such inequalities yields

$$d(\kappa^j r^0) \leq \lambda^j d(r^0). \tag{4}$$

We now interpolate (4). For any $r \leq r^0$, select j such that

$$\kappa^{j+1} \leq r/r^0 \leq \kappa^j.$$

Then $d(r) \leq d(\kappa^j r^0) \leq \lambda^j d(r^0)$ and

$$j+1 \geq \log(r/r^0)/\log\kappa$$

so

$$d(r) \leq \lambda^{-1} d(r^0)(r/r^0)^{\log\lambda/\log\kappa}. \tag{5}$$

By taking κ close to 1 for fixed λ, the exponent in (5) may be made arbitrarily large. This completes the proof of the corollary.

Proof of the lemma. Let (F, G) denote the time 1 map of (1), the solution of (1) for 1 time unit:

$$X=F(x,y)=Ax+\phi(x,y), \qquad Y=G(x,y)=By+\psi(x,y), \tag{6}$$

where the eigenvalues of A are on the unit circle, the eigenvalues of B are in the left half plane, and ϕ and ψ denote higher order terms at the origin. Note that A and B are just exponentials of a and b. Select any λ which is less than 1 but greater than the real parts of all of the eigenvalues of B. Let κ be any number in the interval $\lambda<\kappa<1$. By a linear coordinate transformation in a small neighborhood of the origin we can make $\|A^{-1}\|$ arbitrarily close to 1, and we can make $\|B\|$ strictly less than λ. Then by restricting to a possibly smaller neighborhood of the origin, we can make the norms of ϕ and ψ, together with their first derivatives, less than a small pre-assigned value ε. In addition, we can make the Lipschitz constants of u and v uniformly bounded by 1. In our choice of ε and $\|A^{-1}\|$ we require

$$\kappa<(\|A^{-1}\|-2\varepsilon)^{-1} \quad \text{and} \quad \lambda>\|B\|+4\varepsilon.$$

The center manifolds C and D are invariant under the flow (1) and hence also under the map (6). Points of the form $(x,u(x))$ are mapped to points of the form $(X,u(X))$, and points of the form $(x_0,v(x_0))$ are mapped to points of the form $(X,v(X))$. Hence we have the invariance conditions

$$X=F(x,u(x)), \qquad u(X)=G(x,u(x))$$

and

$$X=F(x_0,v(x_0)), \qquad v(X)=G(x_0,v(x_0))$$

for the graphs of u and v, respectively. Note that $u(0)=0$, $v(0)=0$, $\phi(0,0)=0$, and $\psi(0,0)=0$.

Let r be given, and let $(X,u(X))$ be a point in C with $\|X\|\le\kappa r$. Then we can solve for a unique point $(x,u(x))$ in C, the pre-image of $(X,u(X))$ under (6), with $\|x\|\le r$. Similarly, given $(X,v(X))$ in D we can solve for a pre-image $(x_0,v(x_0))$ in D. This is just a Lipschitz version of the inverse function theorem, and is proved using a contraction mapping. We find

$$x+\phi(x,u(x))=x_0+\phi(x_0,v(x_0)),$$

so

$$\|x-x_0\|\le2\varepsilon\|x-x_0\|+\varepsilon\|u(x)-v(x)\|$$

and hence $\|x-x_0\|\le2\varepsilon d(r)$ if ε is sufficiently small. Also,

$$u(X)-v(X)=Bu(x)-Bv(x_0)+\psi(x,u(x))-\psi(x_0,v(x_0)),$$

so

$$\|u(X)-v(X)\|\le\|B\|(\|u(x)-v(x_0)\|) \\ +\varepsilon(\|x-x_0\|+\|u(x)-v(x_0)\|).$$

But

$$\begin{aligned}\|u(x)-v(x_0)\|&\le\|u(x)-v(x)\|+\|v(x)-v(x_0)\|\\ &\le d(r)+\|x-x_0\|\\ &\le(1+2\varepsilon)d(r),\end{aligned}$$

so $\|u(X)-v(X)\|\le\lambda d(r)$ if ε is sufficiently small. Taking the supremum over $\|X\|\le\kappa r$ yields $d(\kappa r)\le\lambda d(r)$. This completes the proof of the lemma.

4. Extensions of the estimate. The estimates of the previous section were derived only in the simplest case, for a center manifold at an equilibrium point where the normal flow is attracting. We now sketch a number of extensions. All are based on the same geometric construction.

i. Suppose the matrix a of (1) has eigenvalues on the imaginary axis and also in the right half plane. Then A^{-1} has eigenvalues on and inside the unit circle. The norm of A^{-1} may be chosen arbitrarily close to 1, and so κ also may be chosen arbitrarily close to 1. With these remarks, the derivation of the estimate of the lemma is unchanged. Hence the estimate and the theorem are applicable to a center unstable manifold at an equilibrium point.

ii. By reversing the flow of time, we may apply the same estimate to the case in which a has eigenvalues on the imaginary axis and also possibly in the left half plane, and b has eigenvalues in the right half plane. Thus the estimate is applicable to a center manifold at an equilibrium point at which the normal flow is repelling, or to a center stable manifold at an equilibrium point.

iii. Consider (1) again, and suppose that the eigenvalues of a lie on the imaginary axis and possibly in the right half plane, and that the eigenvalues of b lie in the left half plane. Recall from [F1] that a set V is called negatively invariant under a flow $p \to p \cdot t$ if V is carried into itself under the flow for backward time. This means that the flow for forward time carries V to a set which contains V. The estimates of [F1, Thm. 9.1] show that in a sufficiently small neighborhood of an equilibrium point, any compact negatively invariant set must be contained in every center unstable manifold. Let C and D be center manifolds in a small neighborhood of the origin, as in the previous section. Let V be a small compact negatively invariant set, so that V is contained in both C and D. Then the points in V can be represented either in the form $(x,u(x))$ or $(x,v(x))$, where x ranges over a small set S near the origin in R^m. Let

$$S(r,V)=\{x_0:\rho(x_0,S)<r\},$$

where ρ is the usual Euclidean distance. Define

$$d(r,V)=\sup\{\|u(x)-v(x)\|\}, \qquad x\in S(r,V).$$

Then the estimate (2) goes through unchanged, with $d(r,V)$ in place of $d(r)$, and $\log\lambda/\log\kappa$ may be made arbitrarily small by taking V sufficiently small.

iv. By working in local coordinate patches, the same construction may be carried over to the global center stable manifolds and center unstable manifolds of [F1]. It follows that the center stable manifolds and center unstable manifolds have the expected order of contact, even if they do not satisfy the differentiability hypotheses of [F1, Thm. 9.1(iv)].

REFERENCES

[F1] N. FENICHEL, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.

[F2] ———, *the orbit structure of the Hopf bifurcation problem*, J. Differential Equations, 17 (1975), pp. 308–328.

[F3] ———, *Oscillatory bifurcations in singular perturbation theory*, I, this Journal, 14 (1983), pp. 861–867.

[F4] ———, *Oscillatory bifurcations in singular perturbation theory*, II, this Journal, 14 (1983), pp. 868–874.

[F5] ______, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 23 (1974), pp. 1109–1137.

[F-L] L. Flatto and N. Levinson, *Periodic solutions of singularly perturbed systems*, J. Rational Mech. Anal., 4 (1955), pp. 943–950.

[H-W] B. Hassard and Y.-H. Wan, *Bifurcation formulas derived from center manifold theory*, J. Math. Anal. Appl., 63 (1978), pp. 297–312.

[L] O. Lanford, *Bifurcations of periodic solutions into invariant tori: the work of Ruelle and Takens*, in Nonlinear Problems in the Physical Sciences and Biology, Lecture Notes in Mathematics 322, Springer-Verlag, Berlin, 1973.

[R-T] D. Ruelle and F. Takens, *On the nature of turbulence*, Commun. Math. Phys. 20 (1971), pp. 167–192.

[S] B. Segré, *Some properties of Differentiable Varieties and Transformations*, Springer-Verlag, Berlin, 1957.

[S] J. Sijbrand, *Studies in non-linear stability and bifurcation theory*, thesis, Utrecht, 1981.

[W] Y.-H. Wan, *On the uniqueness of invariant manifolds*, J. Differential Equations, 24 (1977), pp. 268–273.

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

002

AN ASYMPTOTIC DECOMPOSITION METHOD APPLIED TO MULTI-TURNING POINT PROBLEMS*

H. GINGOLD†

Abstract. An asymptotic decomposition technique is developed. It is designed and used for 2 by 2 first order singularly perturbed linear differential systems. A new set of decoupled linear integral equations is introduced in the process of the asymptotic analysis. Its usefulness is demonstrated with multi-turning point problems. An adiabatic theorem in quantum mechanics is proved in a general case of degenerate energy levels.

AMS-MOS subject classifications (1980). Primary 34E20, Secondary 34E15

Key words. singularly perturbed, turning point, asymptotic expansions

1. Introduction. In this paper a new asymptotic decomposition method is introduced. The method is capable of handling second order linear differential equations and first order 2 by 2 singular linear differential systems. The method is able to produce a "fine" asymptotic decomposition due to the fact that the entries of a simplifying transformation are shown to satisfy a *new set of decoupled linear integral equations*. This is in *contrast* to a method developed by Sibuya [14], favored also by Wasow [17], in which entries of the simplifying transformation are shown to satisfy a *nonlinear* Riccati differential equation. The advantages of a linear integral equation over a nonlinear integral equation are many. The "global" existence of a solution as well as other properties are more transparent. Many properties follow easily from the corresponding resolvent series of a linear integral equation. The same cannot be claimed for a nonlinear Riccati equation. In particular, the calculation of the coefficients in the asymptotic expansions of a solution may become more laborious.

Examination of Theorem 4.6 in §4 reveals "that higher order asymptotic terms" in a solution of an integral equation are a by-product of repeated integrations in the resolvent series. Thus, unlike Sibuya [14] and Wasow [18], we do not need to duplicate those calculations by finding "higher order asymptotic terms" *directly* from the Riccati differential equation. In order to have a unified theory for second order singular differential equations as well as for singular 2 by 2 first order differential systems, we adopted matrix formulation, a favorite of many authors. See e.g. Wasow [17], Sibuya [14]. We believe that anything that can be done by the Liouville transformation (see, e.g. Olver [7, Ch. 6]) for second order singular differential equations can also be done by a suitable matrix transformation. The latter method has the potential of being applied to n by n first order singular linear differential systems. This point, however, will be demonstrated elsewhere.

Though the Borel–Ritt theorem is nice (see e.g. Wasow [17, p. 41]), we do not feel it is an indispensable tool in the theory of asymptotic expansions of solutions of linear differential equations. In fact, the use of the Borel–Ritt theorem (see Sibuya [14], Wasow [17]) restricts the theory to applications in cases where only an *infinite* asymptotic expansion in *power series* is possible. Unfortunately, this may not be the rule but

* Received by the editors May 19, 1983, and in revised form December 28, 1983.

† Department of Mathematics, West Virginia University, College of Arts and Sciences, Morgantown, West Virginia 26506.

rather the exception in many applied mathematics problems (see e.g. Van Dyke [15, Chap. III]). The Borel–Ritt theorem introduces a new function which is not explicitly given but becomes a part of the desired solution. Thus, for the purpose of *approximation* of a solution, one may be losing accuracy. Therefore, we avoided the use of the Borel–Ritt theorem.

The main purposes of this paper are threefold. The first is to present the decoupling of the entries of the simplifying transformation by obtaining four basic integral decomposing equations. The second is to demonstrate their usefulness by tackling singularly perturbed problems with multi-turning points on an interval $[a,b]$. As a consequence, we are able to free from restrictions, the type and number of turning points occurring in a singularly perturbed 2 by 2 system discussed by Wasow [16]. The third purpose is to obtain a proof for the adiabatic approximation theorem in quantum mechanics for a general degenerate case. Friedrichs [2] discussed a *special* degenerate case which restricted the type and number of turning points occurring in that theorem.

The order of the article will be as follows. Following the discussion, we introduce a few notations and conventions. In §2 we introduce some matrix identities which will produce, in §3, the decoupled integral decomposing equations needed for asymptotic decomposition. In §4, we prove an asymptotic decomposition theorem. We define when an integral can be put on a "zero uniform scale on $[a,b]$". This is utilized in the proof of the asymptotic decomposition theorem with "multi-turning points". The method of proof demonstrates that it *is not always essential* to reduce the investigation of a singularly perturbed linear differential system with turning points to another one with a coefficient matrix which has distinct eigenvalues. Section 5 shows how the method of stationary phase (see e.g. Olver [7, p. 96]) could be generalized in a certain manner. This is used in integrals which can be put on a "zero uniform scale". Sections 4 and 5 provide the necessary extensions to the results in Wasow [16]. Section 6 combines a well-known theorem of Rellich [9] with our method to provide uniform approximations of solutions of "Hamiltonian systems" occurring in quantum mechanics. Finally, in §6, we use Friedrichs' setting from [2] to prove an adiabatic theorem in quantum mechanics. For a source of 2 by 2 first order differential systems occurring in physics, the reader may consult Feynman [1, §§9, 10, 11].

We now introduce a few notations and conventions.

Notation 1.1. Let $V(\tau)$ be a 2 by 2 matrix function. Assume its entries $v_{jk}(\tau)$, $k,j=1,2$, are Riemann or Lebesgue integrable on $[a,b]$. The symbol

$$\int^{x} V(\tau)\,d\tau \tag{1.1}$$

represents a 2 by 2 matrix function on $[a,b]$ with entries

$$\int_{\alpha_{kj}}^{x} v_{kj}(\tau)\,d\tau, \qquad k,j=1,2. \tag{1.2}$$

The lower limits of integration α_{kj} are certain numbers which belong to $[a,b]$.

Assumption 1.2. Unless otherwise stated, given a matrix differential equation

$$Y'=AY, \tag{1.3}$$

or

$$Z'=AZ-ZB+M, \tag{1.4}$$

on an interval $[a,b]$, we assume that:

i) All matrices involved are 2 by 2 matrices, denoted in general by Roman or Greek capital letters.

ii) All entries of A, B, M, are piecewise smooth functions on $[a,b]$ for each fixed value of parameters $\mu=\varepsilon^{-1}$. The parameter μ varies in a set

$$\mu \geqq \mu_0 > 0, \quad (\text{or } 0<\varepsilon \leqq \varepsilon_0), \quad \mu_0 \text{ (or } \varepsilon_0) \text{ a fixed positive number.} \tag{1.5}$$

iii) Given an integral equation

$$Z=M(x)+\int^x A(x,t)Z(t)B(x,t)\,dt \tag{1.6}$$

for the unknown matrix $Z(x)$, we assume $M(x)$ is a continuous matrix function of x on $[a,b]$ for each fixed value of a parameter μ. We assume for each fixed x,μ, that the entries of $A(x,t)$, $B(x,t)$ are piecewise smooth functions on $[a,b]$.

2. Some identities. In this section we derive a few needed identities. We summarize them in the following lemma.

LEMMA 2.1. *Assume the function $\psi(x)$ and the entries of $\Omega(x)$, $R(x)$ are piecewise smooth functions on $[a,b]$ for each fixed value of a parameter μ, $\mu \geqq \mu_0$. Let the differential equation*

$$Y'=(\psi\Omega+R)Y \tag{2.1}$$

be taken into

$$W'=\psi\Omega W, \tag{2.2}$$

by the transformation

$$Y=(I+P)W. \tag{2.3}$$

I is the 2 by 2 identity matrix. Then the matrix function P satisfies the differential equation

$$P'=\psi\Omega P-P\psi\Omega+RP+R. \tag{2.4}$$

Denote by $\Phi(x,s)$ the matrix solution of the initial value problem

$$\Phi'=\psi\Omega\Phi, \quad \Phi(s,s)=I, \quad s\in[a,b]. \tag{2.5}$$

Assume that $P(x)$ is a piecewise smooth matrix function which solves the integral equation

$$P=P_0+FP, \tag{2.6}$$

with

$$P_0:=\int^x \Phi(x,s)R(s)\Phi^{-1}(x,s)\,ds=FI \tag{2.7}$$

where F is an integral operator defined by

$$FP:=\int^x \Phi(x,s)R(s)P(s)\Phi^{-1}(x,s)\,ds. \tag{2.8}$$

Then P is a piecewise smooth solution of the differential equation (2.4).

Let

$$X=(x_{kj}), \qquad k,j=1,2, \tag{2.9}$$

and denote

$$D[X]:=\begin{bmatrix} x_{11} & 0 \\ 0 & x_{22} \end{bmatrix}, \tag{2.10}$$

$$OF[X]:=\begin{bmatrix} 0 & x_{12} \\ x_{21} & 0 \end{bmatrix}. \tag{2.11}$$

Denote

$$\tilde{R}(x,\tau_1):=\Phi(x,\tau_1)R(\tau_1)\Phi^{-1}(x,\tau_1). \tag{2.12}$$

Then the integral equation (2.6) *can be rewritten as with*

$$P_0(x)=\int^x \tilde{R}(x,\tau_1)\,d\tau_1, \tag{2.13}$$

and

$$FP=\int^x \tilde{R}(x,\tau_1)\Phi(x,\tau_1)P(\tau_1)\Phi^{-1}(x,\tau_1)\,d\tau_1. \tag{2.14}$$

In addition, assume that

$$\Omega:=\operatorname{diag}\{\lambda_1(x),\lambda_2(x)\} \tag{2.15}$$

and that

$$D[R]=0, \qquad R=\begin{bmatrix} 0 & r_{12}(x) \\ r_{21}(x) & 0 \end{bmatrix}. \tag{2.16}$$

Then

$$\tilde{R}(x,\tau)=\begin{bmatrix} 0 & r_{12}(\tau)\exp\int_\tau^x \psi(\lambda_1-\lambda_2)\,ds \\ r_{21}(\tau)\exp\int_\tau^x \psi(\lambda_2-\lambda_1)\,ds & 0 \end{bmatrix}, \tag{2.17}$$

and therefore

$$D[\tilde{R}]=0. \tag{2.18}$$

$D[P]$ *and* $OF[P]$, *respectively, satisfy the integral equations*

(2.19)

$$D[P]=\int^x \tilde{R}(x,\tau_1)\left[\int^{\tau_1} \tilde{R}(x,\tau_2)\,d\tau_2\right]d\tau_1+\int^x \tilde{R}(x,\tau_1)\left[\int^{\tau_1} \tilde{R}(x,\tau_2)D[P]\,d\tau_2\right]d\tau_1,$$

(2.20)

$$OF[P]=\int^x \tilde{R}(x,\tau_1)\,d\tau_1+\int^x \tilde{R}(x,\tau_2)\left[\int^{\tau_1}\tilde{R}(x,\tau_2)\Phi(x,\tau_2)OF[P]\Phi^{-1}(x,\tau_2)\,d\tau_2\right]d\tau_1.$$

Proof. The integral equation (2.6) follows from (2.4) by use of a well-known method. See e.g. Wasow [17, p. 169]. The identities (2.13) and (2.14) follow in an obvious manner. If P is a solution of (2.6), then it is also a solution of

$$P=FI+F^2I+F^2P, \qquad P_0=FI. \tag{2.21}$$

Substitute in (2.21)

$$P=D[P]+OF[P] \tag{2.22}$$

and utilize the following fact: If the matrices X, Y, are such that

$$D[X]=0, \qquad D[Y]=0, \tag{2.23}$$

then

$$OF[XY]=OF[XY]=0. \tag{2.24}$$

If

$$D[X]=0, \qquad OF[Y]=0, \tag{2.25}$$

then

$$D[XY]=D[YX]=0. \tag{2.26}$$

Therefore (2.21) splits into two independent equations, (2.19), (2.20). Actually each element of P satisfies an integral equation which does not involve the other three elements of P. The integral equations for $D[P]$ and $OF[P]$ do not reveal this. Our next aim is to prove this point and to find the corresponding integral equations. This is accomplished in the following section.

3. Decoupled decomposing equations.

We will use the following:

Notation 3.1. Let

$$e(x,\tau):=\exp\int_{\tau}^{x}\psi(\lambda_1-\lambda_2)\,ds, \tag{3.1}$$

and let

$$e^{-1}(x,\tau)=e(\tau,x)=\exp\int_{\tau}^{x}\psi(\lambda_2-\lambda_1)\,ds. \tag{3.2}$$

We proceed to the next lemma.

LEMMA 3.2. *Let p_{vk}, $v,k=1,2$, be the elements of P. Let the assumptions of Lemma 2.1 hold. Then p_{vk} satisfy the following integral equations*:

$$\begin{aligned} p_{11}(x)&=\int_{\alpha_{11}}^{x}\int_{\alpha_{21}}^{\tau_1} r_{12}(\tau_1)r_{21}(\tau_2)e(\tau_2,\tau_1)\,d\tau_2\,d\tau_1 \\ &\quad+\int_{\alpha_{11}}^{x}\int_{\alpha_{21}}^{\tau_1} r_{12}(\tau_1)r_{21}(\tau_2)e(\tau_2,\tau_1)p_{11}(\tau_2)\,d\tau_2\,d\tau_1, \end{aligned} \tag{3.3}$$

$$\begin{aligned} p_{22}(x)&=\int_{\alpha_{22}}^{x}\int_{\alpha_{12}}^{\tau_1} r_{21}(\tau_1)r_{12}(\tau_2)e(\tau_1,\tau_2)\,d\tau_2\,d\tau_1 \\ &\quad+\int_{\alpha_{22}}^{x}\int_{\alpha_{12}}^{\tau_1} r_{21}(\tau_1)r_{12}(\tau_2)e(\tau_1,\tau_2)p_{22}(\tau_2)\,d\tau_2\,d\tau_1, \end{aligned} \tag{3.4}$$

$$\begin{aligned} p_{12}(x)&=\int_{\alpha_{12}}^{x} r_{12}(\tau_1)e(x,\tau_1)\,d\tau_1+ \\ &\quad+\int_{\alpha_{12}}^{x}\int_{\alpha_{22}}^{\tau_1} r_{12}(\tau_1)r_{21}(\tau_2)e(x,\tau_1)p_{12}(\tau_2)\,d\tau_2\,d\tau_1, \end{aligned} \tag{3.5}$$

$$(3.6)\qquad p_{21}(x)=\int_{\alpha_{21}}^{x} r_{21}(\tau_1)e(\tau_1,x)\,d\tau_1 + \int_{\alpha_{21}}^{x}\int_{\alpha_{11}}^{\tau_1} r_{21}(\tau_1)r_{12}(\tau_2)e(\tau_1,x)p_{21}(\tau_2)\,d\tau_2\,d\tau_1.$$

Choose in (3.3)–(3.6)

$$(3.7)\qquad \alpha_{11}=\alpha_{21},\qquad \alpha_{22}=\alpha_{12}.$$

Then $p_{vk}(x)$ are solutions of Volterra integral equations

$$(3.8)\qquad p_{vk}(x)=n_{vk}(x)+\int_{\alpha_{vk}}^{x} K_{vk}(x,\tau_2)p_{vk}(\tau_2)\,d\tau_2,\qquad v,k=1,2,$$

with

$$(3.9)\qquad K_{11}(x,\tau_2):=r_{21}(\tau_2)\int_{\tau_2}^{x} r_{12}(\tau_1)e(\tau_2,\tau_1)\,d\tau_1,$$

$$(3.10)\qquad n_{11}(x,\alpha_{21}):=\int_{\alpha_{21}}^{x} K_{11}(x,\tau_2)\,d\tau_2,$$

$$(3.11)\qquad K_{22}(x,\tau_2):=r_{12}(\tau_2)\int_{\tau_2}^{x} r_{21}(\tau_1)e(\tau_1,\tau_2)\,d\tau_1,$$

$$(3.12)\qquad n_{22}(x,\alpha_{12}):=\int_{\alpha_{12}}^{x} K_{22}(x,\tau_2)\,d\tau_2,$$

$$(3.13)\qquad K_{12}(x,\tau_2):=r_{21}(\tau_2)\int_{\tau_2}^{x} r_{12}(\tau_1)e(x,\tau_1)\,d\tau_1,$$

$$(3.14)\qquad n_{12}(x,\alpha_{12}):=\int_{\alpha_{12}}^{x} r_{12}(\tau_1)e(x,\tau_1)\,d\tau_1,$$

$$(3.15)\qquad K_{21}(x,\tau_2):=r_{12}(\tau_2)\int_{\tau_2}^{x} r_{21}(\tau_1)e(\tau_1,x)\,d\tau_1,$$

$$(3.16)\qquad n_{21}(x,\alpha_{21})=\int_{\alpha_{21}}^{x} r_{21}(\tau_1)e(\tau_1,x)\,d\tau_1.$$

Moreover, let p_{vk}, $v,k=1,2$, satisfy the integral equations (3.3)–(3.6) *or the integral equations* (3.8) *with* (3.9)–(3.16). *Then, the matrix*

$$(3.17)\qquad P=(p_{vk}),\qquad v,k=1,2$$

satisfies the matrix integral equation (2.6).

Proof. We will need a few identities for the elements of the matrices which appear in (3.3)–(3.6). Let

$$(3.18)\qquad J_1:=\int^{x}\tilde{R}(x,\tau_1)\int^{\tau_1}\tilde{R}(x,\tau_2)\,d\tau_2\,d\tau_1.$$

Then

(3.19)

$$J_1=\begin{bmatrix}\int_{\alpha_{11}}^{x}\int_{\alpha_{21}}^{\tau_1} r_{12}(\tau_1)r_{21}(\tau_2)e(\tau_2,\tau_1)\,d\tau_2\,d\tau_1 & 0\\ 0 & \int_{\alpha_{22}}^{x}\int_{\alpha_{12}}^{\tau_1} r_{21}(\tau_1)r_{12}(\tau_2)e(\tau_1,\tau_2)\,d\tau_2\,d\tau_1\end{bmatrix}.$$

Denote

$$J_2 := \int^x \tilde{R}(x,\tau_1)\left[\int^{\tau_1} \tilde{R}(x,\tau_2) D[P]\, d\tau_2\right] d\tau_1. \tag{3.20}$$

By repeating a calculation similar to the one above, we obtain

(3.21)

$$J_2 := \begin{bmatrix} \displaystyle\int_{\alpha_{11}}^{x} \int_{\alpha_{21}}^{\tau_1} r_{12}(\tau_1) r_{21}(\tau_2) e(\tau_2,\tau_1) p_{11}(\tau_2)\, d\tau_2\, d\tau_1 & 0 \\ 0 & \displaystyle\int_{\alpha_{22}}^{x} \int_{\alpha_{12}}^{\tau_1} r_{21}(\tau_1) r_{12}(\tau_2) e(\tau_1,\tau_2) p_{22}(\tau_2)\, d\tau_2\, d\tau_1 \end{bmatrix}.$$

Since

$$D[P] = J_1 + J_2, \tag{3.22}$$

we obtain (3.3), (3.4). In order to find equations for p_{12}, p_{21}, we first compute terms in (2.20). Let

$$J_3 := \Phi(x,\tau_2) OF[P] \Phi^{-1}(x,\tau_2). \tag{3.23}$$

Then

$$J_3 = \begin{bmatrix} 0 & e(x,\tau_2) p_{12}(\tau_2) \\ e(\tau_2,x) p_{21}(\tau_2) & 0 \end{bmatrix}. \tag{3.24}$$

Put

$$J_4 := \begin{bmatrix} \displaystyle\int_{\alpha_{11}}^{\tau_1} r_{12}(\tau_2) p_{21}(\tau_2)\, d\tau_2 & 0 \\ 0 & \displaystyle\int_{\alpha_{22}}^{\tau_1} r_{21}(\tau_2) p_{12}(\tau_2)\, d\tau_2 \end{bmatrix}. \tag{3.25}$$

Let

$$J_5 := \int^x \tilde{R}(x,\tau_1) J_4\, d\tau_1. \tag{3.26}$$

Then

(3.27)

$$J_5 = \begin{bmatrix} 0 & \displaystyle\int_{\alpha_{12}}^{x} \int_{\alpha_{22}}^{\tau_1} r_{12}(\tau_1) r_{21}(\tau_2) e(x,\tau_1) p_{12}(\tau_2)\, d\tau_2\, d\tau_1 \\ \displaystyle\int_{\alpha_{21}}^{x} \int_{\alpha_{11}}^{\tau_1} r_{21}(\tau_1) r_{12}(\tau_2) e(\tau_1,x) p_{21}(\tau_2)\, d\tau_2\, d\tau_1 & 0 \end{bmatrix}.$$

Therefore by

$$OF[P] = \int^x \tilde{R}(x,\tau_1)\, d\tau_1 + J_5 \tag{3.28}$$

we obtain (3.5), (3.6). Inverting the order of integration in (3.3)–(3.6) together with (3.7) leads to (3.8) with (3.9)–(3.16).

Reversing the order of operations in the steps which led to the set of integral equations above implies the last conclusion of the lemma.

It is worth noting that because of Lemma 3.2 we are able to reduce a matrix integral equation to *four decoupled* equations for its entries.

Moreover, this special splitting involves a *double integral* in (3.3)–(3.6) so that the resulting kernels in (3.8) are expected to show "desirable qualities of increased smoothness".

It does not appear to be possible to obtain a *similar* decoupling for an n by n linear differential system. However, the insight that will be gained by this *specific* decoupling for $n=2$ and the use of the scheme (2.21) for $n>2$ will be shown to be very useful.

It has long been recognized that the problem of diagonalizing a system of 2 by 2 linear differential equations can be reduced to solving a certain system of uncoupled integral equations. This can be done in a variety of manners. However we deal here with a *singular* differential system. The *type* of decoupling could be *crucial* for the ultimate goal of a successful asymptotic decomposition.

We now proceed to the next section.

4. Multi-turning points with piecewise smooth coefficients. The previous decomposition equations are able to handle a variety of turning point problems. Consider the singularly perturbed system

$$Y'=\begin{bmatrix} i\mu\lambda_1(x) & r_{12}(x) \\ r_{21}(x) & i\mu\lambda_2(x) \end{bmatrix} Y, \qquad Y'=\frac{d}{dx}, \tag{4.1}$$

with $\lambda_1(x)$, $\lambda_2(x)$, real functions defined on $[a,b]$, and $r_{12}(x)$, $r_{21}(x)$, piecewise smooth functions on $[a,b]$.

We will be concerned with the asymptotic behaviour of a *fundamental solution* of (4.1) to be denoted by Y,

$$Y=Y(x,\mu) \quad \text{as } \mu\to+\infty. \tag{4.2}$$

First we introduce the following assumption.

Assumption 4.1. Let $q(\tau)\in C^1(a,b), p(\tau)\in C^2(a,b)$.

i) Assume there exist continuous functions $g_L(\mu)$, $g_R(\mu)$ defined for $\mu\geqq\mu_0$ such that

$$a\leqq g_L(\mu)\leqq g_R(\mu)\leqq b, \quad \lim_{\mu\to\infty} g_L(\mu)=a, \quad \lim_{\mu\to\infty} g_R(\mu)=b. \tag{4.3}$$

ii) The quantities

$$J_L(a,g_L(\mu)):=\sup_{a\leqq x\leqq g_L(\mu)}\left|\int_a^x q(\tau)\exp i\mu p(\tau)\right|, \tag{4.4}$$

$$J_R(g_R(\mu),b):=\sup_{g_R(\mu)\leqq x\leqq b}\left|\int_x^b q(\tau)(\exp i\mu p(\tau))\,d\tau\right|, \tag{4.5}$$

satisfy

$$\lim_{\mu\to\infty}\left[J_L(a,g_L(\mu))+J_R(g_R(\mu),b)\right]=0. \tag{4.6}$$

iii) With

$$J=J(g_L(\mu),g_R(\mu)) \tag{4.7}$$

$$:=\sup_{g_L(\mu)\leqq s\leqq z\leqq g_R(\mu)}\left[\left|\frac{q(z)}{p'(z)}\right|+\left|\frac{q(s)}{p'(s)}\right|+\int_s^z\left|\frac{d}{d\tau}\frac{q(\tau)}{p'(\tau)}\right|d\tau\right]$$

we have that

$$\lim_{\mu\to\infty} \mu^{-1} J(g_L(\mu), g_R(\mu)) = 0. \tag{4.8}$$

Next we introduce

DEFINITION 4.2. We say that the integrals

$$I_R = I_R(x,b,\mu) := \int_x^b q(\tau)(\exp i\mu p(\tau))\,d\tau, \tag{4.9}$$

$$I_L = I_L(a,x,\mu) := \int_a^x q(\tau)[\exp i\mu p(\tau)]\,d\tau \tag{4.10}$$

are on a *zero uniform scale on* $[a,b]$ if assumption 4.1 holds. We will also need the following assumption.

Assumption 4.3. The interval $[a,b]$ can be written as a union of intervals

$$[a,b] = \bigcup_{j=1}^{N} [\alpha_j, \alpha_{j+1}], \qquad \alpha_1 = a, \quad \alpha_{N+1} = b, \tag{4.11}$$

N an integer, such that on each interval $[\alpha_j, \alpha_{j+1}]$, $j = 1, \cdots, N$, the integrals

$$I_R = \int_x^{\alpha_{j+1}} r_{vk}(\tau)(\exp \pm i\mu p(\tau))\,d\tau, \qquad v,k, = 1,2, \quad v \neq k, \tag{4.12}$$

$$I_L = \int_{\alpha_j}^{x} r_{vk}(\tau)(\exp \pm i\mu p(\tau))\,d\tau, \qquad v,k, = 1,2, \quad v \neq k, \tag{4.13}$$

can be put on a zero uniform scale, the mapping $p(\tau)$ is given in (4.12), (4.13) by

$$p(\tau) := \int^{\tau} (\lambda_1(t) - \lambda_2(t))\,dt. \tag{4.14}$$

Moreover, for $v = 1$, $k = 2$ the sign in (4.12), (4.13) is taken to be minus and for $v = 2$, $k = 1$ the sign in (4.12), (4.13) is taken to be plus.

Let Assumption 4.3 hold. For each subinterval $[\alpha_j, \alpha_{j+1}]$, $j = 1, \cdots, N$ denote the quantities g_L, g_R, J, a, b, (defined in Assumption 4.1) by g_{Lj}, g_{Rj}, J_j, α_j, α_{j+1}, respectively. Put

$$g_{12}(\mu) := \sum_{j=1}^{N} \left[J_{Lj}(\alpha_j, g_{Lj}(\mu)) + \mu^{-1} J_j(g_{Lj}(\mu), g_{Rj}(\mu)) + J_{Rj}(g_{Rj}(\mu), \alpha_{j+1}) \right]. \tag{4.15}$$

Define $g_{21}(\mu)$ in a similar manner. It can be easily verified that

$$|n_{12}(x,a)| \leqq g_{12}(\mu), \tag{4.16}$$

and

$$|n_{21}(x,a)| \leqq g_{21}(\mu). \tag{4.17}$$

By (3.13), (3.15) we have

$$|K_{12}(x,\tau_2)| \leqq |r_{21}(\tau_2)| g_{12}(\mu), \tag{4.18}$$

$$|K_{21}(x,\tau_2)| \leqq |r_{12}(\tau_2)| g_{21}(\mu). \tag{4.19}$$

Also,

$$|K_{11}(x,\tau_2)| \leqq |r_{21}(\tau_2)| g_{12}(\mu), \tag{4.20}$$

$$|K_{22}(x,\tau_2)| \leqq |r_{12}(\tau_1)| g_{21}(\mu), \tag{4.21}$$

$$|n_{11}(x,a)| \leqq g_{12}(\mu) \int_a^x |r_{21}(\tau_2)|\, d\tau_2, \tag{4.22}$$

$$|n_{22}(x,a)| \leqq g_{21}(\mu) \int_a^x |r_{12}(\tau_1)|\, d\tau_1. \tag{4.23}$$

Let Assumptions 4.1 and 4.3 hold.

We would like to verify if each of the integral equations in (3.8) possesses a solution

$$p_{\nu k}(x) = p_{\nu k}(x,\mu), \qquad \nu, k = 1,2, \tag{4.24}$$

with the following properties.

For each fixed μ, $p_{\nu k}(x,\mu)$ is a piecewise smooth function of x and

$$\lim_{\mu \to +\infty} p_{\nu k}(x,\mu) = 0 \tag{4.25}$$

uniformly for $a \leqq x \leqq b$. With (3.8), it turns out that

$$|p_{12}(x)| \leqq g_{12}(\mu) + \left(\int_a^b |r_{21}(\tau_2)|\, d\tau_2 \right) g_{12}(\mu) \|p_{12}\|, \tag{4.26}$$

where $\|p_{12}\|$ is to be interpreted as follows. If $f(x)$ is a mapping on $[a,b]$, then $\|f\|$ is defined by

$$\|f\| = \operatorname*{Sup}_{a \leqq x \leqq b} |f(x)|. \tag{4.27}$$

Therefore, if

$$\hat{g}_{12} := g_{12}(\mu) \int_a^b |r_{21}(\tau)|\, d\tau < 1, \tag{4.28}$$

the integral equation (3.8) with $v=1$, $k=2$ possesses a unique solution $p_{12}(x,\mu)$ such that

$$\|p_{12}(x,\mu)\| \leqq \frac{g_{12}(\mu)}{1-\hat{g}_{12}(\mu)}. \tag{4.29}$$

Similarly, with $\nu=2$, $k=1$, (3.8) possesses a solution $p_{21}(x,\mu)$ such that

$$\|p_{21}(x,\mu)\| \leqq \frac{g_{21}(\mu)}{1-\hat{g}_{21}(\mu)}, \tag{4.30}$$

if

$$\hat{g}_{21}(\mu) := g_{21}(\mu) \int_a^b |r_{12}(\tau)|\, d\tau < 1. \tag{4.31}$$

Also from (3.8), with $\nu=k=1$, it turns out that if (4.28) holds then (3.8) possesses a solution $p_{11}(x,\mu)$ such that

$$\|p_{11}(x,\mu)\| \leqq \frac{\hat{g}_{12}(\mu)}{1-\hat{g}_{12}(\mu)}. \tag{4.32}$$

Similarly, if (4.31) holds, then (3.8) with $\nu=k=2$ possesses a solution $p_{22}(x,\mu)$ with

$$\|p_{22}(x,\mu)\| \leqq \frac{\hat{g}_{21}(\mu)}{1-\hat{g}_{21}(\mu)}. \tag{4.33}$$

It can easily be verified that if $r_{12}(x)$, $r_{21}(x)$, $\lambda_1(x)$, $\lambda_2(x)$ are piecewise smooth on $[a,b]$ then the solutions of (3.8) are continuous with piecewise continuous derivatives with respect to x.

The previous discussion shows that in order to obtain

$$\lim_{\mu\to\infty} p_{\nu k}(x,\mu)=0, \qquad \nu,k=1,2, \tag{4.34}$$

uniformly for $a \leqq x \leqq b$ we also need the following assumption.

Assumption 4.4. The functions $r_{12}(\tau)$, $r_{21}(\tau)$ satisfy on $[a,b]$

$$\int_a^b \left(|r_{12}(\tau)|+|r_{21}(\tau)|\right) d\tau < \infty, \tag{4.35}$$

$$\hat{g}_{12}(\mu)<1, \quad \hat{g}_{21}(\mu)<1, \quad \mu \geqq \mu_0. \tag{4.36}$$

However, (4.34) may not be the only information that we would like to have about $p_{\nu k}(x,\mu)$. Actually (4.34) tells us that a first approximation to $p_{\nu k}(x,\mu)$ may be taken to be 0.

In order to obtain "higher order terms" of uniformly valid approximations, we deviate from a practice in the literature (see e.g. Wasow [17, Chap. IV, VII]). Instead of calculating "higher order terms of asymptotic expansions" from the differential equation for

$$P=(p_{\nu k}), \qquad \nu,k=1,2, \tag{4.37}$$

we demonstrate how to extract higher order terms from the integral equations (3.8).

To this end, we take a closer look at the infinite resolvent series of a Voltera integral equation. We rewrite (3.8)

$$p_{mk}(x)=n_{mk}(x)+\int_a^x K_{mk}(x,\tau_2)\,p_{mk}(\tau_2)\,d\tau_2, \qquad m,k=1,2. \tag{4.38}$$

It is well known that if the series s

$$S := \sum_{\nu=0}^{\infty} T^{\nu} n_{mk} \tag{4.39}$$

with

$$T^0 n_{mk}=n_{mk}(x), \tag{4.40}$$

and

(4.41)

$$T^\nu n_{mk}=\int_a^x K_{mk}(x,t_1)\int_a^{t_1} K_{mk}(t_1,t_2)\cdots\int_a^{t_{\nu-1}} K_{mk}(t_{\nu-1},t_\nu)n_{mk}(t_\nu)\,dt_\nu dt_{\nu-1}\cdots dt_1,$$

$$\nu=1,2,\cdots,$$

are convergent, then

$$p_{mk}(x)=S \tag{4.42}$$

is a solution of (4.38).

It is easily shown that for $m=1$, $k=2$ or for $m=2$, $k=1$ we obtain with (4.18)–(4.21) and with (4.41),

$$|T^\nu n_{mk}|\leq g_{mk}(\mu)[\hat{g}_{mk}(\mu)]^\nu=g_{mk}(\mu)]^{\nu+1}\left[\int_a^b |r_{km}(t)|\,dt\right]^\nu,\qquad \nu=0,1,\cdots. \tag{4.43}$$

Therefore, each expression $T^\nu n_{mk}$ is a "generalized asymptotic term of order ν" in a "generalized asymptotic expansion" of $p_{mk}(x)$. Similarly, it is easily shown that

$$|T^\nu n_{11}|\leq\left[g_{12}(\mu)\int_a^b |r_{21}(\tau)|\,dt\right]^{\nu+1}=[\hat{g}_{12}(\mu)]^{\nu+1}, \tag{4.44}$$

and

$$|T^\nu n_{22}|\leq\left[g_{21}(\mu)\int_a^b |r_{21}(t)|\,dt\right]^{\nu+1}=[\hat{g}_{21}(\mu)]^{\nu+1}. \tag{4.45}$$

Thus, we obtain higher order terms in "generalized expansions" of $p_{11}(x,\mu)$ and $p_{22}(x,\mu)$. Let us adopt a definition for these circumstances.

DEFINITION 4.5. We say that

$$S:=\sum_{\nu=0}^{\infty} f_\nu(\mu) \tag{4.46}$$

is a converging generalized power type expansion of $f(\mu)$ in the neighborhood of $\mu=\infty$ if

i)

$$f(\mu)=\sum_{\nu=0}^{\infty} f_\nu(\mu); \tag{4.47}$$

ii) there exists a sequence of positive gauge functions $g_\nu(\mu)$ such that

$$|f_\nu(\mu)|\leq g_\nu(\mu),\qquad \mu>\mu_0, \tag{4.48}$$

$$\lim_{\mu\to\infty}\frac{g_{\nu+1}(\mu)}{g_\nu(\mu)}=0, \tag{4.49}$$

with

$$g_\nu(\mu)=[\phi(\mu)]^\nu,\qquad \nu=1,2,\cdots. \tag{4.50}$$

A definition of a generalized asymptotic expansion can be found e.g. in Olver [7, p. 25].

However, we demanded here a strict additional requirement of convergence of the series in (4.47).

The following theorem has been established.

THEOREM 4.6. *With the notation of this section and with Assumptions* 4.1, 4.3 *and* 4.4, *the differential system* (4.1) *possesses a fundamental solution*

$$Y(x,\mu)=(I+P)\operatorname{diag}\left\{\exp i\mu\int_a^x \lambda_1(t)\,dt, \exp i\mu\int_a^x \lambda_2(t)\,dt\right\} \tag{4.51}$$

for $\mu\geqq\mu_0$, *and* $a\leqq x\leqq b$. *We have*

$$P=(p_{mk}(x,\mu)), \qquad p_{mk}(a,\mu)=0, \qquad m,k=1,2. \tag{4.52}$$

The entries $p_{mk}(x,\mu)$ *are continuous with piecewise continuous derivatives on* $[a,b]$ *for each fixed* $\mu\geqq\mu_0$. *Moreover, the entries* $p_{mk}(x,\mu)$ *possess convergent generalized power type expansions*

$$P_{mk}(x,\mu)=\sum_{\nu=0}^{\infty} T^{\nu}n_{mk}. \tag{4.53}$$

The convergence in (4.53) *is absolute and uniform with respect to* $x\in[a,b]$. *The terms* $T^{\nu}n_{mk}$ *are subject to the inequalities* (4.43), *and the terms* $p_{mk}(x,\mu)$ *are subject to the inequalities* (4.29), (4.30), (4.32), (4.33).

Proof. The theorem is a consequence of the previous discussion.

Theorem 4.6 provides a differential proof and an extension to the results given in Wasow [16].

There is a widespread belief in the theory of asymptotic expansions which roughly states that "asymptotic expansions are preferable to (slowly) converging expansions". This statement is true only in a right context. For example, an approximation to $\exp(-\mu)$ for $\mu>0$, μ large, by the use of a power series expansion about $\mu=0$ may be a practical handicap. An asymptotic expansion may turn out to be far more superior. However, in this setting formula (4.53) combined with (4.29)–(4.33) point out the following. Though the entries of the matrix P in (4.51) are given by an absolutely converging series, each entry of P has a "converging generalized power type expansion" according to Definition 4.5.

Therefore, the "order" of accuracy provided by each term $T^{\nu}n_{mk}$ in (4.53) is no worse then the order of accuracy which may be provided by the ν term in a generalized asymptotic expansion of p_{mk}. In this case I have not come across anything to disprove the following statement. "The provided approximation to p_{mk} given by p_{mk}^n:

$$p_{mk}^n := \sum_{\nu=0}^{\nu=n} T^{\nu}n_{mk},$$

may be better then an asymptotic approximation of order $(n+1)$."

It is also worth noticing that it is possible to use integration by parts or some other method to extract from p_{mk}^n an asymptotic expansion of order $(N+1)$ in cases where it is available. This could be done without using direct substitution in the differential system (4.1) or (2.4). The series in (4.53) are of a double nature. They are absolutely convergent and at the same time they provide a generalized asymptotic expansion.

We now turn to an example which clarifies the relation between the method of stationary phase and our definition of an integral which is on a zero uniform scale on $[a,b]$.

5. An example. Adopt the notation of §4 and consider on $[0,1]$

$$I_L = I_L(0,x,\mu) = \int_0^x \tau^{\delta-1}(\exp - i\mu\tau^{m+1})\, d\tau \tag{5.1}$$

with

$$\delta > 0, \qquad m+1 > 0. \tag{5.2}$$

If

$$\delta - 1 - m \geqq 0, \tag{5.3}$$

then by integration by parts, it turns out that

$$I_L(0,x,\mu) \leqq \mu^{-1}K \tag{5.4}$$

for some constant K. This corresponds to the case where I_L is on zero uniform scale on $[0,1]$ with

$$g_L(\mu) \equiv 0, \quad g_R(\mu) = 1, \quad |J(0,1)| \leqq \mu^{-1}K. \tag{5.5}$$

However, for

$$\delta - 1 - m < 0, \tag{5.6}$$

we find with

$$g_L(\mu) = \left(\frac{1}{\mu}\right)^l, \quad g_R(\mu) \equiv 1, \quad 0 < l < (m+1-\delta)^{-1}, \quad \mu > 1, \tag{5.7}$$

that

$$J_L(0, g_L(\mu)) \leqq \frac{1}{\delta}\left(\frac{1}{\mu}\right)^{\delta l}, \qquad J_R(1,1) \equiv 0. \tag{5.8}$$

(The reason for specifying l as above will be apparent later.) Let

$$\tilde{H} = \left[\frac{x^{(\delta-1-m)}}{(m+1)} + \frac{(1/\mu)^{(\delta-1-m)l}}{(m+1)} + \int_{(1/\mu)^l}^x \left|\frac{(\tau^{\delta-1-m})'}{(m+1)}\right| d\tau\right]. \tag{5.9}$$

For

$$0 < \left(\frac{1}{\mu}\right)^l \leqq x \leqq 1 \tag{5.10}$$

we have

$$(\delta - 1 - m)\ln x \leqq (\delta - 1 - m) l \ln\left(\frac{1}{\mu}\right) \tag{5.11}$$

or

$$x^{\delta-1-m} \leqq \left(\frac{1}{\mu}\right)^{l(\delta-1-m)}. \tag{5.12}$$

Therefore,

$$\tilde{H}=\frac{1}{(m+1)}\left[x^{\delta-1-m}+\left(\frac{1}{\mu}\right)^{(\delta-1-m)l}+\left(\frac{1}{\mu}\right)^{(\delta-1-m)l}-x^{\delta-1-m}\right] \tag{5.13}$$

and

$$\mu^{-1}J(g_L(\mu),1)\leqq\mu^{-1}\tilde{H}=\frac{2}{(m+1)\mu^{1-l(m+1-\delta)}}. \tag{5.14}$$

Thus, $I_L(0,x,\mu)$ can be put on a zero uniform scale on [0, 1].

We notice that with

$$l=(m+1)^{-1} \tag{5.15}$$

we obtain that

$$I_L=O\left(\mu^{-\delta(m+1)^{-1}}\right) \tag{5.16}$$

holds uniformly for $0\leqq x\leqq 1$ in complete agreement with the method of stationary phase. (See e.g. Olver [7], pp. 98–108.) Consider now a differential system (4.1) on [0, 2] with

$$r_{12}(t)=r_{21}(t)=\begin{cases}t^{\delta-1}, & \delta>0,\quad 0\leqq t<1,\\ 0, & \qquad\quad 1<t\leqq 2,\end{cases} \tag{5.17}$$

$$\lambda_1(t)=-\lambda_2(t)=\frac{(m+1)t^m}{2},\qquad 0\leqq t\leqq 2,\quad m+1-\delta>0. \tag{5.18}$$

Then, by using the notations of §4, and utilizing (5.1)–(5.15) we have

$$N=2,\quad g_{L1}(\mu)=\left(\frac{1}{\mu}\right)^l,\quad g_{R1}(\mu)=1,\quad \mu>1, \tag{5.19}$$

$$J_{L1}(0,g_{L1}(\mu))\leqq\frac{1}{\delta}\left(\frac{1}{\mu}\right)^{\delta l},\quad J_{R1}(1,1)\equiv 0,\quad 0<l<(m+1-\delta)^{-1}, \tag{5.20}$$

$$\mu^{-1}J_1(q_L(\mu),1)=\frac{2}{(m+1)\mu^{1-l(m+1-\delta)}}, \tag{5.21}$$

$$g_{L2}(\mu)=1,\qquad g_{R2}(\mu)=2, \tag{5.22}$$

$$J_{L2}(1,1)=0=J_{R2}(2,2), \tag{5.23}$$

$$J_2(1,1)=0. \tag{5.24}$$

Therefore,

$$g_{21}(\mu)=g_{12}(\mu)=\frac{1}{\delta}\left(\frac{1}{\mu}\right)^{\delta l}+\frac{2}{(m+1)\mu^{1-l(m+1-\delta)}},\qquad 0<l<(m+1-\delta)^{-1}, \tag{5.25}$$

$$\int_0^2|r_{12}(t)|\,dt=\int_0^2|r_{21}(t)|\,dt=\int_0^1 t^{\delta-1}dt=\frac{1}{\delta}. \tag{5.26}$$

If we choose l in accordance with (5.15), we end up with

$$g_{12}(\mu)=\left(\delta^{-1}+2(m+1)^{-1}\right)\left(\frac{1}{\mu}\right)^{\delta(m+1)^{-1}}. \tag{5.27}$$

Consequently, for some $\mu_1>0$ we have for $\mu\geqq\mu_1$,

$$\hat{g}_{12}(\mu)=\hat{g}_{21}(\mu)=\delta^{-1}\left(\delta^{-1}+2(m+1)^{-1}\right)\left(\frac{1}{\mu}\right)^{\delta(m+1)^{-1}}<1. \tag{5.28}$$

Let

$$\mu_0=\max\{1,\mu_1\}. \tag{5.29}$$

It can be easily verified that we may choose

$$\mu_1=\left[\delta^{-1}\left(\delta^{-1}+2(m+1)^{-1}\right)\right]^{\delta^{-1}(m+1)}+\delta_1, \qquad \delta_1>0. \tag{5.30}$$

The differential system (4.1) with entries specified by (5.17), (5.18) possesses a fundamental solution

$$Y(x,\mu)=(I+P)\operatorname{diag}\{0.5i\mu x^{m+1},-0.5i\mu x^{m+1}\}, \qquad \mu\geqq\mu_0. \tag{5.31}$$

The desired properties of P are described in Theorem 4.6.

We turn to the next section.

6. Applications to quantum mechanics. Consider the differential system

$$i\varepsilon Y'=H(\tau)Y, \qquad Y'=\frac{d}{d\tau}, \tag{6.1}$$

with ε a small positive parameter. We need the following

Assumption 6.1. $H(\tau)$ is an analytic 2 by 2 Hermitian matrix function of the real variable τ, $a\leqq\tau\leqq b$. The case $b-a=\infty$ is not excluded.

By use of a theorem of Rellich [9] it turns out that there exists a unitary transformation $U(\tau)$ with the following properties: $U(\tau)$ is an analytic function of τ in $[a,b]$ such that

$$U^{-1}(\tau)H(\tau)U(\tau)=\tilde{\Omega}(\tau), \quad U^{-1}(\tau)=U^*(\tau), \quad \tilde{\Omega}(\tau)=\operatorname{diag}\{\lambda_1(\tau),\lambda_2(\tau)\}. \tag{6.2}$$

$\lambda_1(\tau)$, $\lambda_2(\tau)$, the eigenvalues of $H(\tau)$, are real analytic functions of τ in $[a,b]$. It can be easily verified that the transformation

$$Y=U(\tau)Z \tag{6.3}$$

takes the differential system (6.1) into

$$i\varepsilon Z'=\left[\tilde{\Omega}(\tau)-i\varepsilon U^*(\tau)U'(\tau)\right]Z. \tag{6.4}$$

Moreover, if we demand that the transformation

$$Z=(I+P)W \tag{6.5}$$

takes the differential system (6.4) into

$$i\varepsilon W'=\Omega(\tau)W, \tag{6.6}$$

then $(I+P)$ must satisfy the differential system

$$i\varepsilon(I+P)' = (\tilde{\Omega}(\tau) - i\varepsilon U^*(\tau)U'(\tau))(I+P) - (I+P)\Omega(\tau). \tag{6.7}$$

$\Omega(\tau)$ is to be specified in the sequel. Two situations may occur.

Case I.

$$\lambda(\tau) :\equiv \tilde{\lambda}_1(\tau) \equiv \tilde{\lambda}_2(\tau), \qquad a \leqq \tau \leqq b. \tag{6.8}$$

Then, we notice that in (6.7) we have

$$(I+P)' = -U^*(\tau)U'(\tau)(I+P), \qquad \text{with } \Omega := \tilde{\Omega}, \tag{6.9}$$

since

$$\tilde{\Omega}(\tau)(I+P) \equiv (I+P)\Omega(\tau). \tag{6.10}$$

With the initial value

$$(I+P)(a) = I \tag{6.11}$$

a solution of (6.9) is obtained which is an analytic invertible matrix function of τ on $[a,b]$. (This solution is independent of ε.)

Therefore, a fundamental solution of (6.1) is given by

$$Y(\tau) = U(\tau)(I+P(\tau))\left(\exp\left(-i\varepsilon^{-1}\int_a^\tau \lambda(s)\,ds\right)I\right). \tag{6.12}$$

Case II. The identity (6.8) does not hold. Therefore, there is a *finite number* of points $t_1, \cdots, t_n$ such that

$$a \leqq t_1 < t_2 < \cdots < t_N \leqq b, \tag{6.13}$$

$$\tilde{\lambda}_1(\tau) - \tilde{\lambda}_2(\tau) = (\tau - t_j)^{m_j} h_j(\tau), \qquad j = 1, \cdots, N. \tag{6.14}$$

m_j are some integers and

$$h_j(\tau) \neq 0 \tag{6.15}$$

for τ in the neighborhood of t_j. Let us now conform to the notation in the integral equations (3.8)–(3.16). By $v_{\nu k}$ we denote the elements of U^*U'.

$$(v_{\nu k}) := U^*U', \qquad \nu, k = 1,2. \tag{6.16}$$

In (3.8) we pick

$$\alpha_{\nu k} = a, \qquad \nu, k = 1,2. \tag{6.17}$$

Then we have

(6.18)

$$n_{12}(\tau,a) = \int_a^\tau -v_{12}(\tau_1)\left\{\exp\left(-\int_{\tau_1}^\tau i\varepsilon^{-1}[\tilde{\lambda}_1(s) - \tilde{\lambda}_2(s) - i\varepsilon(v_{11}(s) - v_{22}(s))]\right)ds\right\}d\tau$$

$$= \int_a^\tau \tilde{r}_{12}(\tau_1)\left(\exp\int_{\tau_1}^\tau -i\varepsilon^{-1}(\tilde{\lambda}_1(s) - \tilde{\lambda}_2(s)\,ds)\right)d\tau_1$$

with

$$(6.19)\qquad \tilde{r}_{12}(\tau_1) := -v_{12}(\tau_1)\exp\int_{\tau}^{\tau_1}(v_{11}(s)-v_{22}(s))\,ds.$$

Similarly,

$$(6.20)\qquad n_{21}(\tau,a)=\int_a^{\tau}\tilde{r}_{21}(\tau_1)\left(\exp\int_{\tau}^{\tau_1}-i\varepsilon^{-1}(\tilde{\lambda}_1(s)-\tilde{\lambda}_2(s))\,ds\right)d\tau_1$$

with

$$(6.21)\qquad \tilde{r}_{21}(\tau_1) := -v_{21}(\tau_1)\exp\int_{\tau_1}^{\tau}(v_{11}(s)-v_{22}(s))\,ds.$$

In agreement with that notation we define

(6.22)

$$\Omega(\tau) := \operatorname{diag}\{\lambda_1,\lambda_2\},\quad \lambda_1=\tilde{\lambda}_1(\tau)-i\varepsilon v_{11}(\tau),\quad \lambda_2=\tilde{\lambda}_2(\tau)-i\varepsilon v_{22}(\tau),$$

(6.23)

$$R(\tau) := \begin{bmatrix} 0 & -i\varepsilon v_{12}(\tau) \\ -i\varepsilon v_{21}(\tau) & 0 \end{bmatrix},\quad r_{12}=-v_{12}(\tau),\quad r_{21}(\tau)=-v_{21}(\tau)$$

$$(6.24)\qquad \psi^{-1}=i\varepsilon.$$

We plug ψ, λ_1, λ_2, r_{12}, r_{21}, into (3.8) and further rearrange the terms.

There result for $p_{\nu k}$, $\nu,k=1,2$, four equations similar to the original ones (3.8). However, from now on, $e(x,\tau)$ in (3.1) is defined by

$$(6.25)\qquad e(x,\tau) := \exp\left(-i\varepsilon^{-1}\int_{\tau}^{x}(\tilde{\lambda}_1(s)-\tilde{\lambda}_2(s))\,ds\right),$$

and $r_{12}(\tau)$, $r_{21}(\tau)$ in (3.8) are replaced by $\tilde{r}_{12}(\tau)$, $\tilde{r}_{21}(\tau)$, defined by (6.19), (6.21), respectively. The previous discussion leads us to the following theorem.

THEOREM 6.2. *With the previous notation and with Assumption* 6.1 *the differential system* (6.1) *possesses a fundamental solution on* $[a,b]$ ($b-a=\infty$ *is not excluded*) *as follows*: *In Case* I

$$(6.26)\qquad Y(\tau)=U(\tau)(I+P(\tau))\left(\exp\left(-i\varepsilon^{-1}\int_a^{\tau}\lambda(s)\,ds\right)I\right),\quad P(a)=0,\quad \varepsilon>0,$$

$P(\tau)$ *continuous on* $[a,b]$. *In Case* II *there exists* $\varepsilon_0>0$, *such that for* $0<\varepsilon<\varepsilon_0$

$$(6.27)\qquad Y(\tau)=U(\tau)(I+P(\tau))\left(\exp\left(-i\varepsilon^{-1}\int_a^{\tau}\Omega(s)\,ds\right)\right),\qquad P(a)=0,$$

$$(6.28)\qquad \|P(\tau)\|\leqq K\varepsilon^{d}\leqq K\varepsilon_0^{d}<1,$$

uniformly for $a\leqq\tau\leqq b$. *K and d are positive constants independent of* τ,ε. *Moreover, if* $b-a<\infty$, *then*

$$(6.29)\qquad d=\min_j\left\{(m_j+1)^{-1},1\right\},\qquad j=1,\cdots,N.$$

Proof. We first show that $(I+P(\tau))$ is well defined in the neighborhood of $\tau=b$ if $b=\infty$. By use of Assumption 6.1 it turns out that

$$(6.30)\qquad \|U^*U'\|=O(\tau^{-2}),\qquad \tau\to\infty.$$

Therefore (4.35) holds for $\tilde{r}_{12}(\tau)$, $\tilde{r}_{21}(\tau)$. For Case II with $\mu=\varepsilon^{-1}$, we proceed to show that for the system (6.4) all conditions of Theorem 4.6 are fulfilled. Due to (6.14), by repeating the arguments of §5 one can show that Assumption 4.3 is satisfied. The integrals (4.12), (4.13) are being defined in this case with $r_{\nu k}$ replaced by $\tilde{r}_{\nu k}$ $\nu\neq k$, $\nu,k=1,2$, given by (6.19), (6.21) and with

$$(6.31)\qquad p(\tau):=\int^{\tau}(\tilde{\lambda}_1(t)-\tilde{\lambda}_2(t))\,dt.$$

Moreover, there exists a finite number of appropriate subintervals $[\alpha_n,\alpha_{n+1}]$ such that

$$[a,b]=\bigcup_{n=1}^{n=N_1}[\alpha_n,\alpha_{n+1}].$$

N_1, is some natural number, $N_1>N$, where M so given by (6.13). On each subinterval $[\alpha_n,\alpha_{n+1}]$

$$\lim_{\varepsilon\to 0^+}\left(|n_{12}(x,a)|+|n_{21}(x,a)|\right)=0$$

uniformly for $\alpha_n\leqq x\leqq\alpha_{n+1}$, $n=1,\cdots,N_1$. This follows by use of techniques given in §§4 and 5. The problem then boils down to the evaluation of integrals of the form (5.1). We may choose $\delta=1$ and l as one of the numbers $(m_j+1)^{-1}, j=1,\cdots,N$, if $|t_j|<\infty$.

$$(6.32)\qquad g_{12}(\mu)\leqq K\mu^{-d}=K\varepsilon^d,\qquad 0<\varepsilon\leqq\varepsilon_0<1.$$

Without loss of generality, we may also assume that we have

$$(6.33)\qquad g_{21}(\mu)\leqq K\varepsilon^d,\qquad 0<\varepsilon\leqq\varepsilon_0<1.$$

Without loss of generality, we may assume that

$$(6.34)\qquad \hat{g}_{12}(\mu):=g_{12}(\mu)\int_a^b|\tilde{r}_{21}(\tau_2)|\,d\tau_2\leqq K\varepsilon^d\left(\int_a^b|r_{21}(\tau)|\,d\tau\right)<1,$$

and

$$(6.35)\qquad \hat{g}_{21}(\mu):=g_{21}(\mu)\int_a^b|\tilde{r}_{12}(\tau)|\,d\tau\leqq K\varepsilon^d\left(\int_a^b|r_{12}(\tau)|\,d\tau\right)<1$$

for $0<\varepsilon\leqq\varepsilon_0<1$. By virtue of (6.30) this is so even if an end point of $[a,b]$ is infinite. Thus,

$$(6.36)\qquad \|P(\tau)\|\leqq\hat{K}\varepsilon^d<1$$

for some constant $\hat{K}$, and $0<\varepsilon\leqq\varepsilon_0<1$, uniformly for $a\leqq\tau\leqq b$. The results follows.

The adiabatic approximation theorem (see e.g. Messiah [6, Chap. XVII]) attracted a considerable amount of attention. It has been rigorously proven by Kato [4] for a general Hamiltonian with noncoalescing eigenvalues. Friedrichs [2], [3] discussed the theorem. For a 2 by 2 Hamiltonian with a *special type* of coalescing zeros Friedrichs proved the adiabatic approximation theorem. We intend now to give a proof of the above theorem for a 2 by 2 Hamiltonian with very general type of degeneracy of its ("energy levels") eigenvalues.

We will use the setting used by Friedrichs [2]. Accordingly, the differential system (6.1) is subject to assumption 6.1 with $0 \leqq \tau \leqq 1$ and ε is a "stretching" variable. Given the initial value problem

$$i\varepsilon Y' = H(\tau)Y, \qquad Y(0) = U(0). \tag{6.37}$$

The proof of an adiabatic theorem which follows is accomplished by showing that uniformly for $0 < \tau \leqq 1$

$$H(\tau)Y(\tau) \sim Y(\tau)\tilde{\Omega}(\tau) \quad \text{as } \varepsilon \to 0. \tag{6.38}$$

The interpretation is that "if a Hamiltonian system started to evolve from an initial 'eigenstate', then *asymptotically* it will continue to evolve into the same 'eigenstate'." An "eigenstate" of the Hamiltonian $H(\tau)$ is to be identified with an eigenvector of $H(\tau)$.

To prove (6.38) we invoke Theorem 6.2 with $[a,b]=[0,1]$ and we use (6.26) or (6.27). If $H(\tau)$ possess identical eigenvalues then by (6.26) we have for the left-hand side of (6.38)

$$H(\tau)Y(\tau) = H(\tau)U(\tau)(I+P(\tau))\left(\exp\left(-i\varepsilon^{-1}\int^{\tau}\lambda(s)\,ds\right)I\right). \tag{6.39}$$

Since

$$\tilde{\Omega}(\tau) = \lambda(\tau)I \tag{6.40}$$

then

$$H(\tau)U(\tau) = \lambda(\tau)U(\tau). \tag{6.41}$$

For the right-hand side of (6.38) we have

$$Y(\tau)\tilde{\Omega}(\tau) = U(\tau)(I+P(\tau))\left(\exp\left(-i\varepsilon^{-1}\int^{\tau}\lambda(s)\,ds\right)I\right)\lambda(\tau). \tag{6.42}$$

Thus in the case of identical eigenvalues, (6.38) is actually an identity.

If $H(\tau)$ does not possess identical eigenvalues, then by (6.27) we have

$$H(\tau)Y(\tau) = H(\tau)U(\tau)(I+P(\tau))\left(\exp\left(-i\varepsilon^{-1}\int^{\tau}\Omega(s)\,ds\right)\right). \tag{6.43}$$

Consequently, by the asymptotic nature of P we have

$$H(\tau)Y(\tau) \sim H(\tau)U(\tau)\left(\exp\left(-i\varepsilon^{-1}\int_0^{\tau}\Omega(s)\,ds\right)\right). \tag{6.44}$$

By (6.2) we conclude that

$$H(\tau)Y(\tau) \sim U(\tau)\tilde{\Omega}(\tau)\left(\exp\left(-i\varepsilon^{-1}\int_0^{\tau}\Omega(s)\,ds\right)\right). \tag{6.45}$$

Since the diagonal matrices $\tilde{\Omega}(s)$ and $\Omega(s)$ commute, we conclude that (6.38) holds.

7. Concluding remarks. The analysis presented in this work demonstrates that it is possible to obtain "global results" with multi-turning point problems.

Uniformly valid asymptotic formulas can be obtained in a full neighborhood of a turning point on the real line.

It is not necessary to always use "lateral connection formulas".

"Central connection formulas" may be superior in certain important applications. Formula (6.27) may be considered a special case of a central connection formula. Therefore, methods related to the construction of central connection formulas should be considered important. See Wasow [19] for "connection problems".

Acknowledgment. Acknowledgment is due to Professor A. Levine from the Physics Department with whom I had useful clarifying discussions.

REFERENCES

[1] R. P. Feynman, R. B. Leighton, and M. Sands [1965], *The Feynman Lectures on Physics, Vol.* 3, *Quantum Mechanics*, Addison-Wesley, Reading, MA, §§9, 10, 11.

[2] K. O. Friedrichs [1953], *Special Topics in Analysis (Lecture Notes)*, New York Univ., New York.

[3] ______ [1955], *On the adiabatic theorem in quantum theory*, Report IMM.NYU-218, New York Univ., New York.

[4] T. Kato [1950], *On the adiabatic theorem of quantum mechanics*, J. Phys. Soc. Japan, 5, pp. 435–439.

[5] ______ [1976], *Perturbation Theory for Linear Operators*, second ed., Springer-Verlag, Berlin, Heidelberg, New York.

[6] A. Messiah [1961], *Quantum Mechanics, Vol* II, Interscience, New York.

[7] F. W. J. Olver [1974], *Asymptotics and Special Functions*, Academic Press, New York.

[8] R. E. O'Malley, Jr. [1974], *Introduction to Singular Perturbations*, Academic Press, New York and London.

[9] F. Rellich [1936], *Störungstheorie der Spektralzerlegung*, I, *Mitteilung*, Math. Ann., 113, pp. 600–619.

[10] ______ [1937], *Störungstheorie der Spektralzerlegung*, II, Math. Ann., 113, pp. 667–685.

[11] ______ [1937], *Störungstheorie der Spektralzerlegung*, III, Math. Ann., 116, pp. 555–570.

[12] ______ [1939], *Störungstheorie der Spektralzerlegung*, IV, Math. Ann., 117, pp. 356–382.

[13] ______ [1940], *Störungstheorie der Spektralzerlegung*, V, Math. Ann., 118, pp. 462–484.

[14] Y. Sibuya [1975], *Global Theory of Second Order Linear Ordinary Differential Equation with a Polynomials Coefficient*, Mathematics Studies 18, North-Holland, Amsterdam.

[15] M. Van Dyke [1964], *Perturbation Methods in Fluid Mechanics*, Academic Press, London.

[16] W. Wasow [1960], *A turning point problem for a system of two linear differential equations*, J. Math. Phys., 38, pp. 257–278.

[17] ______ [1965], *Asymptotic Expansions for Ordinary Differential Equations*, Wiley (Interscience) New York.

[18] ______ [1973], *Adiabatic invariance of a simple oscillator*, this Journal, 4, pp. 78–88.

[19] ______ [1968], *Connection problems for asymptotic series*, Bull. Amer. Math. Soc., 74, pp. 831–853.

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

FINITE DETERMINATION OF BIFURCATION PROBLEMS*

PETER B. PERCELL† AND PETER N. BROWN†

Abstract. A C^0 theory of finite determination of bifurcation problems is presented in this paper which supplements a corresponding C^∞ theory of Golubitsky and Schaeffer. Finite determination of both bifurcation diagrams and stability properties of branches is considered. C^0 finite determination of bifurcation diagrams is shown to follow from an analytic-geometric nondegeneracy condition which is modelled on a criterion of Kuo, rather than an algebraic condition of the type found in the C^∞ theory. The class of "quasi-homogeneous" bifurcation problems, which contains bifurcation problems previously studied by McLeod and Sattinger and Landman and Rosenblat using more classical methods, is introduced and shown to admit a simplified and computable nondegeneracy condition which suffices to ensure finite determination of the bifurcation diagram. The results of the C^0 theory are compared with those of the C^∞ theory and are found to be a distinct improvement in some cases.

Two different notions of equivalence of bifurcation problems are used in the results on finite determination of bifurcation diagrams. Contact equivalence is used primarily because it appears in the C^∞ theory. BD equivalence (i.e. existence of an ambient, parameter-preserving homeomorphism of bifurcation diagrams) is a simpler and more fundamental concept of equivalence. Furthermore, it permits the possibility that each coordinate function of a bifurcation problem may have its own "order of determination".

1. Introduction. Here a *bifurcation problem* is considered to be a family of maps

$$G(\cdot,\lambda): \mathbb{R}^n \to \mathbb{R}^p$$

parameterized by $\lambda \in \mathbb{R}^l$ such that $G(0,0)=0$, or, more compactly, a map

$$G: (\mathbb{R}^n \times \mathbb{R}^l, 0) \to (\mathbb{R}^p, 0).$$

(It is assumed that infinite dimensional bifurcation problems have already been reduced to finite dimensional ones by some device such as the Lyapunov–Schmidt reduction.) The first feature of interest in the present context is the variation of the set of zeros of $G(\cdot,\lambda)$ with the parameter λ; the set $G^{-1}(0)$ is called the *bifurcation diagram*. Second, if $p=n$ and $G(\cdot,\lambda)$ is thought of as a family of vector fields, then the zeros of $G(\cdot,\lambda)$ are steady state solutions of the differential equations $\dot{x}=G(x,\lambda)$ whose asymptotic stability properties are of interest. Roughly, two bifurcation problems are equivalent if their bifurcation diagrams are locally homeomorphic in a neighborhood of the origin, and a bifurcation problem G is finitely determined if some Taylor polynomial of G has the property that every other bifurcation problem with that Taylor polynomial is equivalent to G. The theory developed in this paper is entirely local in the sense that bifurcation problems are considered only on small neighborhoods of the origin.

By the implicit function theorem, in order for the bifurcation diagram of G to actually bifurcate at the origin it is necessary that

$$\operatorname{rank}[G_x(0,0)] < p,$$

where G_x denotes the partial derivative of G with respect to the variable $x \in \mathbb{R}^n$. On the other hand, again by the implicit function theorem, even when there is true bifurcation at the origin, the bifurcation diagram of G is well behaved away from the origin if

$$\operatorname{rank}[G_x(u)] = p \quad \text{whenever } G(u)=0 \text{ and } u \equiv (x,\lambda) \neq 0. \tag{1.1}$$

* Received by the editors February 23, 1982, and in revised form September 14, 1983.

† Mathematics Department, University of Houston, Central Campus, Houston, Texas 77004.

This leads one to propose (1.1) as the basic nondegeneracy condition for bifurcation diagrams. In fact, §§4 and 5 are devoted to consideration of two rather general classes of bifurcation problems ("quasi-homogeneous" and real analytic) for which (1.1) is sufficient to guarantee C^0 finite determination of the bifurcation diagram.

However, (1.1) alone is clearly not always adequate as a criterion for finite determination. For example,

$$G(x,\lambda)=\begin{cases} e^{-(1/x^2)}(x-\lambda), & x\neq 0,\\ 0, & x=0,\end{cases}$$

is a C^∞ bifurcation problem which satisfies (1.1), but all its Taylor polynomials are identically zero. A somewhat more delicate inadequacy is found in the example

$$H(x,\lambda)=x^2-\lambda^2 x.$$

H satisfies (1.1) and is equal to its Taylor polynomial of degree 3, so one might hope that every bifurcation problem whose Taylor polynomial of degree 3 equals H is equivalent to H. But this is not true because

$$x^2-\lambda^2x+\lambda^4=(x-\lambda^2/2)^2+\tfrac{3}{4}\lambda^4$$

is zero only at the origin, while H is zero on the curves $x=0$ and $x=\lambda^2$. (It will follow from Theorem 4.1 that the Taylor polynomial of degree 4 for H does qualitatively determine bifurcation diagrams for all perturbations of order greater than 4.)

Our full nondegeneracy condition and criterion for C^0 finite determination of bifurcation diagrams, which contains and refines condition (1.1), will be stated in §3. It is a version of a condition of Kuo [3]. Buchner, Marsden and Schecter [7] have obtained results similar to those proved below using a blowing-up construction and techniques from algebraic geometry. There, however, no special emphasis is placed on requiring the equivalence to preserve the value of the bifurcation parameter λ. The bifurcation problems $x^3-\lambda x=0$ and $x\lambda=0$ would be considered equivalent in [7], but not here. However, if one constructs the theory presented here without the above restriction on λ, then the two theories are essentially the same. The interested reader should compare the results in §§2 and 3 below with those obtained in [7, §1].

In the case $p=n$, the basic nondegeneracy condition for stability properties of bifurcating branches is simply that no eigenvalue of $G_x(u)$ be purely imaginary when $G(u)=0$ and $u\neq 0$.

The rest of the paper is organized as follows. Section 2 contains necessary definitions, notation and minor technical preliminaries. Section 3 is devoted to the statement and proof of our basic result. In §4 the class of "quasi-homogeneous" bifurcation problems is introduced and used to show that our basic result contains classical results such as those of McLeod and Sattinger [6]. Section 5 deals with real analytic bifurction problems. In §6 our C^0 theory of finite determination of bifurcation diagrams is compared with the C^∞ theory of Golubitsky and Schaeffer [1]. Finally, in §7 several results on finite determination of stability properties of bifurcating branches are presented.

2. Preliminaries. We begin by presenting notation and concepts needed from linear algebra. The standard inner product on a Euclidean space $\mathbb{R}^n$ is denoted by $\langle\cdot,\cdot\rangle$ and the associated norm by $\|\cdot\|$. $\mathscr{L}(\mathbb{R}^m,\mathbb{R}^p)$ denotes the set of all linear maps from $\mathbb{R}^m$ to $\mathbb{R}^p$ and is identified with the collection of $p\times m$ matrices; $GL(\mathbb{R}^p)$ denotes the set of invertible linear maps from $\mathbb{R}^p$ to itself. For $A \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^p)$, let

$$d(A) = \inf\{\|\alpha^t A\| : \alpha \in \mathbb{R}^p, \|\alpha\| = 1\},$$

with the convention that all vectors are columns and the superscript "t" is used to denote a transpose. Clearly $d(A) > 0$ if and only if $\operatorname{rank} A = p$. When $A \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^p)$ and $\operatorname{rank} A = p$, let A^+ denote the *pseudo-inverse of A* which in this situation is given by

$$A^+ = A^t(AA^t)^{-1}.$$

Obviously $AA^+ = I$, where I is the identity map of $\mathbb{R}^p$. Also, if $B \in GL(\mathbb{R}^p)$, then it can verified by a short computation that

$$A^+B^{-1} = (BA)^+. \tag{2.1}$$

Furthermore, when $A \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^p)$ and $\operatorname{rank} A = p$,

$$d(A) = \|A^+\|^{-1} \tag{2.2}$$

since $[d(A)]^2$ equals the smallest eigenvalue of AA^t and $\|A^+\|^2$ equals the largest eigenvalue of $(A^+)^t A^+ = (AA^t)^{-1}$. For $A \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^n)$, let $\sigma(A)$ denote the spectrum of A, and let $\operatorname{Ind}(A)$ denote the *index of A* which is the triple (a,b,c) with a, b and c the number of eigenvalues of A with positive, negative and zero real part, respectively.

The basic notation connected with nonlinear functions which we shall need is the following. The notation

$$F\colon (\mathbb{R}^m, 0) \to (\mathbb{R}^p, 0)$$

means that F is a map from $\mathbb{R}^m$ to $\mathbb{R}^p$ whose domain contains a neighborhood of the origin and that $F(0) = 0$. If f and g are real-valued functions defined on a neighborhood of $0 \in \mathbb{R}^m$, then

$$f(u) = o(g(u))$$

means that

$$[f(u)/g(u)] \to 0 \quad \text{as } u \to 0,$$

while

$$f(u) = O(g(u))$$

means that $[f(u)/g(u)]$ is merely bounded. For $G\colon \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k} \to \mathbb{R}^p$, G_{x_i} denotes the partial derivative of G with respect to $x_i \in \mathbb{R}^{n_i}$ and G' denotes the total derivative of G.

Next we give precise definitions for various notions of equivalence between bifurcation problems. Let

$$G, \tilde{G}\colon (\mathbb{R}^n \times \mathbb{R}^l, 0) \to (\mathbb{R}^p, 0)$$

be continuous maps. We say that G and $\tilde{G}$ are $[C^r]$ BD *equivalent* (BD for bifurcation diagram) if there are a neighborhood $V \subset \mathbb{R}^n \times \mathbb{R}^l$ of the origin and a map

$$\phi\colon (V, 0) \to (\mathbb{R}^n \times \mathbb{R}^l, 0)$$

of the form

$$\phi(x,\lambda)=(\bar{\phi}(x,\lambda),\lambda) \tag{2.3}$$

which is a homeomorphism [C^r diffeomorphism] onto its image such that

$$\phi(G^{-1}(0)\cap V)=\tilde{G}^{-1}(0)\cap\phi(V).$$

We call G and $\tilde{G}$ [C^r] *contact equivalent* if there exist a neighborhood V of the origin in $\mathbb{R}^n\times\mathbb{R}^l$, a map

$$\phi\colon (V,0)\to(\mathbb{R}^n\times\mathbb{R}^l,0)$$

which is a [C^r] BD equivalence between G and $\tilde{G}$ and a continuous [C^r] map

$$\tau\colon V\to GL(\mathbb{R}^p)$$

such that

$$G(u)=\tau(u)\cdot(\tilde{G}(\phi(u))), \qquad u\equiv(x,\lambda)\in V.$$

Now, suppose $p=n$ and let

$$G,\ \tilde{G}\colon (\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^n,0)$$

be C^1 maps. We say that G and $\tilde{G}$ are [C^r] *$\mathscr{S}$ equivalent* ($\mathscr{S}$ for stability) if there exists a [C^r] BD equivalence ϕ with domain V between G and $\tilde{G}$ such that

$$\operatorname{Ind}(G_x(u))=\operatorname{Ind}(\tilde{G}_x(\phi(u))) \quad \text{when } u\in(G^{-1}(0)-\{0\})\cap V.$$

Note that, because of the form (2.3) imposed on ϕ, these equivalences can be thought of as equivalences between two families of maps parameterized by λ.

The following "multi-exponent" notation is introduced in order to be able to recognize that each coordinate function of a bifurcation problem G may have its own "order of determination". For $\rho>0$ and $\nu\in\mathbb{R}^m$, let

$$\rho^\nu=\operatorname{diag}(\rho^{\nu_1},\cdots,\rho^{\nu_m}),$$

where the right-hand side is the $m\times m$ diagonal matrix with the diagonal entries $\rho^{\nu_1},\cdots,\rho^{\nu_m}$. In this context we call $\nu\in\mathbb{R}^m$ a *multi-exponent*. For $\nu\in\mathbb{R}^m$ a multi-exponent, let

$$|\nu|=\max\{|\nu_i|: i=1,\cdots,m\}.$$

We call a multi-exponent $\nu\in\mathbb{R}^m$ a *constant exponent* when $\nu_1=\cdots=\nu_m$ and then we identify it with a single real number also denoted by ν. For example, in the expression $\rho^{1-\nu}$, where $\nu\in\mathbb{R}^m$, the 1 represents $(1,\cdots,1)\in\mathbb{R}^m$.

Now we are in a position to rigorously define the finite determination concepts. If

$$G\colon (\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^p,0)$$

is a C^2 bifurcation problem and $\nu\in\mathbb{R}$ is a multi-exponent, we say that G is [C^r] BD, *respectively contact or $\mathscr{S}$, ν-determined* if G and $G+P$ are [C^r] BD, respectively contact or $\mathscr{S}$, equivalent for every $C^2[C^{\max(r,2)}]$ perturbation

$$P\colon (\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^p,0)$$

such that, with $u \equiv (x,\lambda) \in \mathbb{R}^n \times \mathbb{R}^l$ and $\rho = \|u\|$,

$$\|\rho^{-\nu} P(u)\| = o(1) \quad \text{and} \quad \|\rho^{1-\nu} P_x(u)\| = o(1).$$

(Note that if $r \geqq |\nu|$, then requiring that $\|\rho^{1-\nu} P_x(u)\| = o(1)$ is redundant.) A bifurcation problem is called $[C^r]$ BD (*respectively contact or* $\mathcal{S}$, *finitely determined*) if it is $[C^r]$ BD (respectively contact or $\mathcal{S}$, ν-determined) for some multi-exponent $\nu \in \mathbb{R}^p$.

The last concept we shall need is that of a *horn neighborhood* of a bifurcation diagram (compare with [3]). For a bifurcation problem

$$G\colon (\mathbb{R}^n \times \mathbb{R}^l, 0) \to (\mathbb{R}^p, 0),$$

a multi-exponent $\nu \in \mathbb{R}^p$ and a constant $\delta > 0$, let

$$H(G,\nu,\delta) = \left\{ u \in \mathbb{R}^n \times \mathbb{R}^l \colon \|\rho^{-\nu} G(u)\| < \delta, \text{ where } \rho = \|u\| > 0 \right\}.$$

3. Finite determination of the bifurcation diagram. This section contains the fundamental result of the paper which is a precise statement of our nondegeneracy condition for bifurcation problems together with a proof that it refines the crude nondegeneracy condition (1.1) sufficiently to ensure BD ν-determination and contact $|\nu|$-determination.

DEFINITION. Let G: $(\mathbb{R}^n \times \mathbb{R}^l, 0) \to (\mathbb{R}^p, 0)$ be a C^1 map and $\nu \in \mathbb{R}^p$ be a multi-exponent. We say that G is ND(ν) (*nondegenerate of order* ν) if there exist $\varepsilon > 0$, $\delta > 0$ and a neighborhood U of the origin in $\mathbb{R}^n \times \mathbb{R}^l$ for which, with $\rho = \|u\|$,

$$d\left(\rho^{1-\nu} G_x(u)\right) \geqq \varepsilon \quad \text{whenever } u \equiv (x,\lambda) \in H(G,\nu,\delta) \cap U. \tag{3.1}$$

Remark. If G is ND(ν), then G is also ND($|\nu|$). To see this, note that we may assume that diam $U < 1$. Then clearly

$$H(G,|\nu|,\delta) \cap U \subset H(G,\nu,\delta) \cap U.$$

Furthermore,

$$d\left(\rho^{1-|\nu|} G_x(u)\right) \geqq d\left(\rho^{1-\nu} G_x(u)\right),$$

because, for $\alpha \in \mathbb{R}^p$ a unit vector,

$$\begin{aligned} \left\|\alpha^t \rho^{1-|\nu|} G_x(u)\right\| &= \left\|\left(\rho^{\nu-|\nu|}\alpha\right)^t \rho^{1-\nu} G_x(u)\right\| \\ &\geqq \left\|\rho^{\nu-|\nu|}\alpha\right\| \cdot d\left(\rho^{1-\nu} G_x(u)\right) \\ &\geqq d\left(\rho^{1-\nu} G_x(u)\right). \end{aligned}$$

THEOREM 3.1. *Suppose* G: $(\mathbb{R}^n \times \mathbb{R}^l, 0) \to (\mathbb{R}^p, 0)$ *is a* C^2 *map and* $\nu \in \mathbb{R}^p$ *is a multi-exponent such that* G *is* ND(ν). *Then* G *is* BD ν-*determined. Furthermore, if* ν *is a constant exponent, then* G *is contact* ν-*determined.* (*In particular, by the remark,* G *is contact* $|\nu|$-*determined even when* ν *is not a constant exponent.*)

Proof. Suppose P: $(\mathbb{R}^n \times \mathbb{R}^l, 0) \to (\mathbb{R}^p, 0)$ is a C^2 map such that

$$\|\rho^{-\nu} P(u)\| = o(1) \quad \text{and} \quad \|\rho^{1-\nu} P_x(u)\| = o(1). \tag{3.2}$$

We must construct a BD equivalence ϕ between G and $\tilde{G} \equiv G + P$, and with the added assumption that ν is a constant exponent improve it to a contact equivalence. Throughout the construction of ϕ, W is a neighborhood of the origin in $\mathbb{R}^n \times \mathbb{R}^l$ which we shall assume is shrunk as necessary and $J \equiv [0,1]$ is the unit interval in $\mathbb{R}$.

Let F: $\mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R} \to \mathbb{R}^p$ be defined by

$$F(u,s) = G(u) + sP(u), \qquad u \in \mathbb{R}^n \times \mathbb{R}^l, \quad s \in \mathbb{R}.$$

The homeomorphism ϕ will be found by following integral curves of a vector field ζ on $\mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}$ from $s=0$ to $s=1$, where $\zeta(u,s)$ is such that for $(u,s) \in W \times J$,

$$(3.3) \qquad F'(u,s) \cdot \zeta(u,s) = 0 \quad \text{when } F(u,s) = 0,$$

$$(3.4) \qquad \langle \zeta(u,s), \Lambda \rangle = 0, \qquad \Lambda \in \{0\} \times \mathbb{R}^l \times \{0\} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R},$$

$$(3.5) \qquad \langle \zeta(u,s), \theta \rangle = 1, \qquad \theta = (0,0,1)^t \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}.$$

Note that (3.3) means that ζ is tangent to $F^{-1}(0)$, (3.4) means that the λ coordinate is constant along integral curves of ζ and (3.5) means that the s coordinate equals time along interval curves of ζ. For $W \subset U$ sufficiently small,

$$(3.6) \qquad d\left(\rho^{1-\nu} F_x(u,s)\right) \geqq \varepsilon/2 \quad \text{for } (u,s) \in (H(G,\nu,\delta) \cap W) \times J,$$

because if $\alpha \in \mathbb{R}^p$ and $\|\alpha\| = 1$, then

$$\begin{aligned}\left\|\alpha^t \rho^{1-\nu} F_x(u,s)\right\| &= \left\|\alpha^t \rho^{1-\nu} G_x(u) + s\alpha^t \rho^{1-\nu} P_x(u)\right\| \\ &\geqq d\left(\rho^{1-\nu} G_x(u)\right) - \left\|\rho^{1-\nu} P_x(u)\right\| \\ &\geqq \varepsilon - o(1),\end{aligned}$$

when $(u,s) \in (H(G,\nu,\delta) \cap U) \times J$, by (3.1) and (3.2). In particular, $F_x(u,s)$ has rank$p$ when $(u,s) \in (H(G,\nu,\delta) \cap W) \times J$. Let

$$\bar{\zeta}\colon (H(G,\nu,\delta) \cap W) \times J \to \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}$$

be the vector field defined by

$$\begin{aligned}\bar{\zeta}(u,s) &= \left(-\left[(F_x(u,s))^+ (F_s(u,s))\right]^t, 0, 1\right)^t \\ &= \left(-\left[(F_x(u,s))^+ (P(u))\right]^t, 0, 1\right)^t.\end{aligned}$$

Then clearly

$$(3.7) \qquad F'(u,s) \cdot \bar{\zeta}(u,s) = 0 \quad \text{for } (u,s) \in (H(G,\nu,\delta) \cap W) \times J$$

and $\bar{\zeta}$ satisfies (3.4) and (3.5). Furthermore, by (3.2) and (3.6),

$$(3.8) \qquad \|\bar{\zeta}(u,s) - \theta\| = o(\|u\|) \quad \text{uniformly in } s \in J,$$

since, by (2.1) and (2.2),

$$\begin{aligned}\|\bar{\zeta}(u,s)-\theta\| &= \|(F_x(u,s))^+(P(u))\|\\ &= \|(F_x(u,s))^+(\rho^{\nu-1})(\rho^{1-\nu}P(u))\|\\ &= \|(\rho^{1-\nu}F_x(u,s))^+(\rho^{1-\nu}P(u))\|\\ &\leqq \|(\rho^{1-\nu}F_x(u,s))^+\|\cdot\|\rho^{1-\nu}P(u)\|\\ &= \|\rho^{1-\nu}P(u)\|/d(\rho^{1-\nu}F_x(u,s)).\end{aligned}$$

Let $\bar{\chi}$: $\mathbb{R}^p\to\mathbb{R}$ be a C^∞ "bump function" such that $0\leqq\bar{\chi}(y)\leqq 1$ for all $y\in\mathbb{R}^p$, $\bar{\chi}(y)=1$ if $\|y\|\leqq 1/2$ and $\bar{\chi}(y)=0$ if $\|y\|\geqq 1$, and then let

$$\chi(u)=\bar{\chi}(\rho^{-\nu}G(u)/\delta)\quad\text{when } u\neq 0.$$

Then $0\leqq\chi(u)\leqq 1$ for all $u\neq 0$, $\chi(u)=1$ if $u\in H(G,\nu,\delta/2)$, $\chi(u)=0$ if $u\notin H(G,\nu,\delta)$ and χ is C^2. Finally, let ζ be the vector field on $W\times J$ defined by

$$\zeta(u,s)=\begin{cases}\chi(u)\bar{\zeta}(u,s)+(1-\chi(u))\theta, & u\neq 0,\\ \theta, & u=0.\end{cases}$$

Clearly ζ is C^1 on $(W-\{0\})\times J$ and, by (3.8),

$$\|\zeta(u,s)-\theta\|=o(\|u\|)\quad\text{uniformly in } s\in J, \tag{3.9}$$

so ζ is differentiable, but not necessarily C^1, on all of $W\times J$. By (3.7) and the properties of χ,

$$F'(u,s)\cdot\zeta(u,s)=0\quad\text{when } (u,s)\in(H(G,\nu,\delta/2)\cap W)\times J. \tag{3.10}$$

Thus, by choosing W small enough, we can ensure that ζ satisfies (3.3), since if $F(u,s)=0$ and $u\neq 0$, then

$$\rho^{-\nu}(G(u)+sP(u))=0$$

so

$$\|\rho^{-\nu}G(u)\|\leqq\|\rho^{-\nu}P(u)\|=o(1).$$

It is obvious that ζ satisfies (3.4) and (3.5).

In order to be able to define ϕ by following integral curves of ζ, we must check uniqueness. Integral curves through points not on the s-axis are unique because ζ is C^1 on $(W-\{0\})\times J$. By (3.9) there exists $\bar{\eta}>0$ such that

$$\|\zeta(u,s)-\theta\|\leqq\bar{\eta}\|u\|\quad\text{when } (u,s)\in W\times J,$$

so if

$$t\mapsto(u(t),s(t)) \tag{3.11}$$

is an integral curve of $\zeta|(W-\{0\})\times J$, then

$$\|u'(t)\|=\|\zeta(u(t),s(t))-\theta\|\leqq\bar{\eta}\|u(t)\|.$$

Therefore

$$\pm\frac{d}{dt}\left(\|u(t)\|^2\right)=\pm 2\langle u'(t),u(t)\rangle \leqq \|u'(t)\|^2+\|u(t)\|^2 \leqq (\bar{\eta}^2+1)\|u(t)\|^2,$$

in other words

$$-(\bar{\eta}^2+1)\leqq\left[\frac{d}{dt}\left(\|u(t)\|^2\right)\right]\Big/\|u(t)\|^2\leqq\bar{\eta}^2+1.$$

Thus, for $t_0<t_1$ in the domain of the curve (3.11),

$$-(\bar{\eta}^2+1)(t_1-t_0)\leqq\ln\left(\|u(t_1)\|^2/\|u(t_0)\|^2\right)\leqq(\bar{\eta}^2+1)(t_1-t_0),$$

so

$$\exp[-(\bar{\eta}^2+1)(t_1-t_0)]\leqq\|u(t_1)\|^2/\|u(t_0)\|^2\leqq\exp[(\bar{\eta}^2+1)(t_1-t_0)],$$

hence, with $\eta=\exp[(\bar{\eta}^2+1)/2]$,

$$\eta^{-1}\|u(t_0)\|\leqq\|u(t_1)\|\leqq\eta\|u(t_0)\|. \tag{3.12}$$

This implies that the s-axis itself is the only trajectory of ζ through each of its own points.

Now, using (3.12), choose $V\subset W$ to be a neighborhood of the origin small enough so that if

$$\psi\colon V\times J\to W\times J$$

is determined by letting $t\mapsto\psi(u,t)$ be the integral curve of ζ which passes through $(u,0)$ when $t=0$, then ψ is defined on all of $V\times J$ and ψ is a homeomorphism onto a neighborhood of $\{0\}\times J\subset W\times J$. Note that, by (3.5), ψ has the form

$$\psi(u,s)=(\bar{\psi}(u,s),s)\in(\mathbb{R}^n\times\mathbb{R}^l)\times\mathbb{R}$$

and, by (3.12),

$$\eta^{-1}\|u\|\leqq\|\bar{\psi}(u,s)\|\leqq\eta\|u\|\quad\text{when }(u,s)\in V\times J. \tag{3.13}$$

At last, we define $\phi\colon V\to W$ by letting $\phi(u)=\bar{\psi}(u,1)$. Then ϕ is clearly a homeomorphism onto its image and, by (3.3) and (3.4), it is a BD equivalence between $G=F(\cdot,0)$ and $\tilde{G}=F(\cdot,1)$.

Before upgrading the equivalence between G and $\tilde{G}$ to a contact equivalence, we must study the effect of ϕ in greater detail. In particular, letting

$$P_\phi(u)=\tilde{G}(\phi(u))-G(u)\quad\text{for }u\in V,$$

we shall need to know that

$$\|\rho^{-\nu}P_\phi(u)\|=o(1) \tag{3.14}$$

and that, for γ and V sufficiently small,

$$P_\phi(u)=0\quad\text{when }u\in H(G,\nu,\gamma)\cap V. \tag{3.15}$$

For $(u,s)\in V\times J$, let

$$Q_\psi(u,s)=F(\psi(u,s))-G(u)$$

and

$$\sigma=\|\bar{\psi}(u,s)\|.$$

Then

$$\begin{aligned}\frac{\partial}{\partial s}\left[Q_\psi(u,s)\right]&=F'(\psi(u,s))\cdot\zeta(\psi(u,s))\\&=\left[1-\chi(\bar{\psi}(u,s))\right]P(\bar{\psi}(u,s)),\end{aligned}$$

so, since $Q_\psi(\mathrm{u},0)=0$,

$$\begin{aligned}\|\rho^{-\nu}Q_\psi(u,t)\|&\leqq\int_0^t\left\|\rho^{-\nu}\frac{\partial}{\partial s}\left[Q_\psi(u,s)\right]\right\|ds\\&\leqq\int_0^t\|\rho^{-\nu}P(\bar{\psi}(u,s))\|\,ds\\&\leqq\int_0^t\|(\sigma/\rho)^\nu\|\cdot\|\sigma^{-\nu}P(\bar{\psi}(u,s))\|\,ds.\end{aligned}$$

Thus, by (3.2) and (3.13),

$$\|\rho^{-\nu}Q_\psi(u,s)\|=o(1)\quad\text{uniformly in } s\in J. \tag{3.16}$$

This proves (3.14) because $Q_\psi(u,1)=P_\phi(u)$. Since

$$F(\psi(u,s))=G(\bar{\psi}(u,s))+sP(\bar{\psi}(u,s)),$$

it follows from (3.2), (3.13) and (3.16) that if $u\in H(G,\nu,\gamma)\cap V$, then

$$\begin{aligned}\|\sigma^{-\nu}G(\bar{\psi}(u,s))\|&=\|\sigma^{-\nu}\left[F(\psi(u,s))-sP(\bar{\psi}(u,s))\right]\|\\&\leqq\|(\rho/\sigma)^\nu\|\left\{\|\rho^{-\nu}G(u)\|+\|\rho^{-\nu}Q_\psi(u,s)\|\right\}+\|\sigma^{-\nu}P(\bar{\psi}(u,s))\|\\&=\|(\rho/\sigma)^\nu\|\cdot\|\rho^{-\nu}G(u)\|+o(1)\\&<\eta^{|\nu|}\gamma+o(1).\end{aligned}$$

uniformly for $s\in J$. Therefore, for γ and V sufficiently small,

$$\psi(u,s)\in(H(G,\nu,\delta/2)\cap W)\times J\quad\text{when } (u,s)\in(H(G,\nu,\gamma)\cap V)\times J. \tag{3.17}$$

This proves (3.15) because, by (3.10), F is constant on trajectories of ζ which remain in $(H(G,\nu,\delta/2)\cap W)\times J$, so

$$P_\phi(u)=F(\psi(u,1))-F(\psi(u,0))=0\quad\text{when } u\in H(G,\nu,\gamma)\cap V.$$

Finally we construct a matrix valued map τ which completes a contact equivalence. For $u\in V$ and $v\in\mathbb{R}^p$, let

$$\theta\colon V\to\mathscr{L}(\mathbb{R}^p,\mathbb{R}^p)$$

be defined by

$$\theta(u)(v)=\begin{cases}\left(\langle G(u),v\rangle/\|G(u)\|^2\right)P_\phi(u), & G(u)\neq 0,\\ 0, & G(u)=0.\end{cases} \tag{3.18}$$

When $G(u)\neq 0$ an $P_\phi(u)\neq 0$, $\theta(u)$ is just a rank one linear transformation designed so that

$$\theta(u)(G(u))=P_\phi(u).$$

By (3.15) and (3.18), $\theta(u)=0$ for u in $H(G,\nu,\gamma)\cap V$, which is a neighborhood of $[G|(V-\{0\})]^{-1}(0)$, so θ is continuous on $V-\{0\}$. By (3.14), (3.18) and the extra hypothesis that ν is a constant exponent,

$$\|\theta(u)\|=\|P_\phi(u)\|/\|G(u)\|\leqq o(\rho^\nu)/\gamma\rho^\nu=o(1),$$

for $u\in V-H(G,\nu,\gamma)$, so θ is also continuous at the origin. (This is the only place the extra hypothesis is used in the proof.) Since θ is continuous on V and $\theta(0)=0$, we can shrink V so that

$$I+\theta(u)\in GL(\mathbb{R}^p) \quad \text{when } u\in V,$$

where I is the identity map of $\mathbb{R}^p$. Let

$$\tau(u)=(I+\theta(u))^{-1} \quad \text{for } u\in V.$$

Then

$$\tau\colon V\to GL(\mathbb{R}^p)$$

is clearly continuous and

$$G(u)=\tau(u)(\tilde{G}(\phi(u))),$$

since

$$(I+\theta(u))(G(u))=G(u)+P_\phi(u)=\tilde{G}(\phi(u)). \qquad \square$$

4. Quasi-homogeneous bifurcation problems. In this section we show that for "quasi-homogeneous" polynomial bifurcation problems the basic nondegeneracy condition (1.1), checked only at points on the unit ball, is sufficient to ensure finite determination with the order ν of determination depending only on the degree of quasi-homogeneity. This result is a generalization of results of McLeod and Sattinger [6, §§3 and 6] and Landman and Rosenblat [4, §§4, 6 and 7]. It is interesting to note that (4.3) below contains in a unified form the three apparently different sets of nondegeneracy conditions listed as hypotheses in [6, Thms. 3.1 and 6.1] and [4, Thms. 4.1 and 4.4].

DEFINITION. Let G: $(\mathbb{R}^m,0)\to(\mathbb{R}^p,0)$ be a map and let $\mu\in\mathbb{R}^m$ and $\nu\in\mathbb{R}^p$ be multi-exponents such that

$$\min\{\mu_i\colon 1\leqq i\leqq m\}=1. \tag{4.1}$$

We say that G is *quasi-homogeneous of degree* (μ,ν) if

$$G(\sigma^\mu u)=\sigma^\nu G(u) \quad \text{for } u\in\mathbb{R}^m \text{ and } \sigma>0. \tag{4.2}$$

Remark. Note that the above definition could have been formulated with (4.1) replaced by the weaker condition that $\min\{\mu_i\}>0$. However, $\min\{\mu_i\}\geqq 1$ is necessary for the next theorem and whenever G is quasi-homogeneous of degree (μ,ν) with (4.1) replaced by $\min\{\mu_i\}>0$, one can rescale by setting $\sigma=\bar{\sigma}^{\varepsilon}$ for $\varepsilon=\min\{\mu_i\}$ so that both (4.1) and (4.2) hold with μ and ν replaced by $\bar{\mu}=\mu/\varepsilon$ and $\bar{\nu}=\nu/\varepsilon$. The advantage of the definition as it has been formulated is that it minimizes the order of determination in the next theorem.

THEOREM 4.1. *Suppose G: $(\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^p,0)$ is a polynomial mapping and $\mu\in\mathbb{R}^n\times\mathbb{R}^l$ and $\nu\in\mathbb{R}^p$ are multi-exponents such that G is quasi-homogeneous of degree (μ,ν). If*

$$\text{(4.3)}\qquad \operatorname{rank}[G_x(v)]=p \quad \textit{whenever } G(v)=0 \textit{ and } \|v\|=1,$$

then G is ND(ν), *so G is* BD *ν-determined and contact $|\nu|$-determined.*

Proof. By (4.3) and compactness of the unit sphere in $\mathbb{R}^n\times\mathbb{R}^l$, there exists $\varepsilon>0$ and $\delta>0$ such that

$$\text{(4.4)}\qquad d(G_x(v))\geqq\varepsilon \quad \text{whenever } \|G(v)\|<\delta \text{ and } \|v\|=1.$$

We shall show that G is ND(ν) by proving that if U is the open unit ball in $\mathbb{R}^n\times\mathbb{R}^l$, then G,ν,ε,δ and U satisfy (3.1).

Suppose $u\in H(G,\nu,\delta)\cap U$. Since $0<\rho\equiv\|u\|<1$, there exists a $\sigma\in\mathbb{R}$ such that

$$\text{(4.5)}\qquad 0<\sigma<1 \quad\text{and}\quad \|\sigma^{-\mu}u\|=1.$$

Let

$$\text{(4.6)}\qquad v=\sigma^{-\mu}u.$$

Then

$$\text{(4.7)}\qquad \rho\leqq\sigma$$

since, by (4.1), (4.5) and (4.6),

$$\rho/\sigma=\|u\|/\sigma=\|\sigma^{\mu-1}v\|\leqq\|v\|=1.$$

Note that, by (4.1),

$$\text{(4.8)}\qquad \min\{\nu_j\colon 1\leqq j\leqq p\}\geqq 1,$$

since $G(0)=0$, and, by (4.2),

$$\text{(4.9)}\qquad G_x(\sigma^{\mu}v)=\sigma^{\nu}G_x(v)\sigma^{-\xi},$$

where $\mu\equiv(\xi,\eta)\in\mathbb{R}^n\times\mathbb{R}^l$. Then $\|G(v)\|<\delta$ since, by (4.2), (4.6), (4.7) and (4.8),

$$\|G(v)\|=\|G(\sigma^{-\mu}u)\|=\|\sigma^{-\nu}G(u)\|\leqq\|\rho^{-\nu}G(u)\|<\delta,$$

so, by (4.4), $d(G_x(v))\geqq\varepsilon$. Thus, for α a unit vector in $\mathbb{R}^p$ and

$$\beta=(\sigma/\rho)^{\nu-1}\alpha/\|(\sigma/\rho)^{\nu-1}\alpha\|,$$

by (4.1), (4.5), (4.6), (4.7), (4.8) and (4.9),

$$\begin{aligned}\|\alpha^t\rho^{1-\nu}G_x(u)\| &= \|\alpha^t\rho^{1-\nu}G_x(\sigma^\mu v)\|\\ &= \|\alpha^t\rho^{1-\nu}\sigma^\nu G_x(v)\sigma^{-\xi}\|\\ &= \|(\sigma/\rho)^{\nu-1}\alpha\|\cdot\|\beta^t G_x(v)\sigma^{1-\xi}\|\\ &\geqq \|\beta^t G_x(\nu)\|\\ &\geqq d(G_x(v))\geqq\varepsilon,\end{aligned}$$

so $d(\rho^{1-\nu}G_x(u))\geqq\varepsilon$. □

Example 4.2. Let G: $(\mathbb{R}^2\times\mathbb{R},0)\to(\mathbb{R}^2,0)$ be defined by

$$G(x,\lambda)=(x_2+\lambda x_1,x_1^2+\lambda x_2)^t.$$

It is easy to check that, with $\mu=(2,3,1)$ and $\nu=(3,4)$, G satisfies the hypotheses of the last theorem. Therefore G is BD (3,4)-determined and contact 4-determined. This G is a sample of the type of bifurcation problem treated by Landman and Rosenblat [4, §6]. □

The two examples in §6 are samples of the type of bifurcation problem considered by McLeod and Sattinger [6].

5. Real analytic bifurcation problems. The result of this section is that for real analytic bifurcation problems the basic nondegeneracy condition (1.1) is indeed the key to contact finite determination, so the role of the extra detail in ND(ν) is simply to pick out an order ν of contact determination. The idea of using the Lojasiewicz inequality [5, p. 59] to prove such a result is well known (e.g. [2] and [3]).

THEOREM 5.1. *If* G: $(\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^p,0)$ *is a real analytic map such that*

$$\text{(5.1)}\qquad \operatorname{rank}[G_x(u)]=p\quad\textit{whenever } G(u)=0 \textit{ and } u\neq 0,$$

then G is contact finitely determined.

Proof. Consider the real analytic function

$$f(u)=\|G(u)\|^2+\det[(G_x(u))(G_x(u))^t].$$

It follows immediately from (5.1) that $f(u)$ can vanish only when $u=0$. Therefore if U is a closed ball centered at the origin and contained in the domain of f, then there exist constants $\alpha>0$ and $\gamma>0$ such that

$$\|G(u)\|^2+\det[(G_x(u))(G_x(u))^t]\geqq\gamma\|u\|^\alpha\quad\text{for } u\in U,$$

by the Lojasiewicz inequality. Thus

$$\det[(G_x(u))(G_x(u))^t]\geqq(\gamma/2)\|u\|^\alpha\quad\text{when } u\in H(G,\alpha/2,\sqrt{\gamma/2})\cap U.$$

But $[d(G_x(u))]^2$ is just the smallest eigenvalue of $(G_x(u))(G_x(u))^t$ and $\|G_x(u)\|^2$ is its largest eigenvalue, so

$$[d(G_x(u))]^2\geqq\det[(G_x(u))(G_x(u))^t]/\|G_x(u)\|^{2(p-1)}.$$

Hence if $\varepsilon=(\sqrt{\gamma/2}\,\inf_{u\in U}\|G_x(u)\|^{1-p})$, $\nu=\alpha/2+1$ and $\delta=\sqrt{\gamma/2}$, then

$$d(G_x(u))\geqq\varepsilon\|u\|^{\nu-1}\quad\text{when } u\in H(G,\nu,\delta)\cap U.$$

The conclusion now follows from Theorem 3.1. □

6. Comparison with C^∞ finite determination results. In this section we shall compare the results of the previous sections with a slightly improved version of the result on finite determination obtained by Golubitsky and Schaeffer in [1]. For G: $(\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^p,0)$ a C^∞ map, we let $\tilde{T}G$ denote the set of all maps H: $\mathbb{R}^n\times\mathbb{R}^l\to\mathbb{R}^p$ of the form

$$H(u)=T(u)\cdot G(u)+G_x(u)\cdot R(u),$$

where

$$T\colon \mathbb{R}^n\times\mathbb{R}^l\to\mathscr{L}(\mathbb{R}^p,\mathbb{R}^p)\quad\text{and}\quad R\colon(\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^n,0)\tag{6.1}$$

are C^∞ maps. This differs from the $\tilde{T}G$ of [1] in that we require that $R(0)=0$. With this notation, the C^∞ finite determination result of [1] can be rewritten in the following somewhat strengthened form.

THEOREM 6.1. *If G: $(\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^p,0)$ is a C^∞ map and ν is a positive integer such that*

$$Q\in\tilde{T}G\quad\textit{whenever } Q\in C^\infty(\mathbb{R}^n\times\mathbb{R}^l,\mathbb{R}^p)\textit{ and } Q(u)=O\left(\|u\|^\nu\right),\tag{6.2}$$

then G is C^∞ contact ν-determined.

The content of the next theorem is simply the unsurprising result that condition (6.2) which is strong enough to ensure C^∞ contact ν-determination is stronger than our condition which ensures C^0 contact ν-determination.

THEOREM 6.2. *If G: $(\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^p,0)$ is a C^∞ map and ν is a positive integer such that* (6.2) *is satisfied, then G is* ND(ν).

Proof. We shall give a proof by contradiction. Suppose G is not ND(ν). Then there are nonzero sequences $\{u\}\subset\mathbb{R}^n\times\mathbb{R}^l$ and $\{\alpha\}\subset\mathbb{R}^p$ such that

$$\rho=\|u\|\to0,\qquad\|\alpha\|=1,$$

$$\rho^{-\nu}G(u)\to0,\tag{6.3}$$

$$\alpha^t\rho^{1-\nu}G_x(u)\to0.\tag{6.4}$$

We may assume that there is a unit vector $\bar{\alpha}\in\mathbb{R}^p$ such that $\alpha\to\bar{\alpha}$. At this point the cases ν even and ν odd must be handled separately.

If ν is even, let $Q(u)=\|u\|^\nu\bar{\alpha}$. Then, by (6.2), there exist C^∞ maps T and R as in (6.1) such that

$$Q(u)=T(u)\cdot G(u)+G_x(u)\cdot R(u).\tag{6.5}$$

When multiplied on the left by $\rho^{-\nu}\alpha^t$, (6.5) becomes

$$\alpha^t\bar{\alpha}=\left[\alpha^tT(u)\right]\cdot\left[\rho^{-\nu}G(u)\right]+\left[\alpha^t\rho^{1-\nu}G_x(u)\right]\cdot\left[\rho^{-1}R(u)\right].\tag{6.6}$$

But $\alpha^t\bar{\alpha}\to1$ while, by (6.3) and (6.4), the right-hand side of (6.6) converges to 0 (because $\rho^{-1}R(u)$ is bounded since $R(0)=0$). This is a contradiction.

If ν is odd, let $Q(u)=\|u\|^{\nu+1}\bar{\alpha}$, so that Q is again C^∞. Then each component of $Q(u)$ is a summation of polynomials in the components of u, and hence $Q(u)$ has the basic form

$$Q(u)=\Big(\sum_k u_{j_k}P_k(u)\Big)\bar{\alpha}=\sum_k u_{j_k}(P_k(u)\bar{\alpha}),$$

where $j_k\in\{1,\cdots,n+l\}$ for each k, $P_k\in C^\infty(\mathbb{R}^n\times\mathbb{R}^l,\mathbb{R}^p)$, and $P_k(u)=O(\|u\|^\nu)$. Therefore, by (6.2) $P_k(u)\bar{\alpha}\in\tilde{T}G$ for each k, and so there exist C^∞ maps T_k and R_k as in (6.1) such that

$$P_k(u)\bar{\alpha}=T_k(u)\cdot G(u)+G_x(u)\cdot R_k(u).$$

Hence,

$$\begin{aligned}(6.7)\qquad Q(u)&=\Big(\sum_k u_{j_k}T_k(u)\Big)\cdot G(u)+G_x(u)\cdot\Big(\sum_k u_{j_k}R_k(u)\Big)\\&=T(u)\cdot G(u)+G_x(u)\cdot R(u),\end{aligned}$$

where T and R are C^∞ maps such that $T(u)=0(\|u\|)$ and $R(u)=0(\|u\|^2)$. Multiplying (6.7) on the left by $\rho^{-(\nu+1)}\alpha^t$ this time leads quickly to a contradiction as in the first case. □

In [1], the main emphasis of the C^∞ theory developed there is concerned with finding a universal unfolding of a bifurcation problem. However, if one is only interested in equivalence of bifurcation diagrams, then condition (6.2) and the notion of C^∞ contact finite determination are sometimes unnecessarily restrictive, as the following simple examples indicate.

Example 6.3. Consider G: $(\mathbb{R}^2\times\mathbb{R},0)\to(\mathbb{R}^2,0)$ defined by

$$G(x,\lambda)=(x_1^2+x_2^2+\lambda x_1,\lambda x_2)^t.$$

It is easy to verify that G satisfies the hypotheses of Theorem 4.1, with $\mu=(1,1,1)$ and $\nu=(2,2)$, so G is C^0 contact 2-determined. However, G is not C^∞ contact finitely determined. In particular, the result of §5 for real analytic bifurcation problems is false with a conclusion of C^∞ contact finite determination. To see that G is not C^∞ contact ν-determined for any ν, note that if $H\in\tilde{T}G$ has the form $H(u)=(0,h(x))^t$, then $h(x)=\bar{h}(x)\|x\|^2$. (This can be shown by finding an appropriately simple set of generators for $\tilde{T}G$.) Thus $(0,x_1^\nu)^t$ is not in $\tilde{T}G$ for any integer ν, so (6.2) cannot be satisfied for any ν. But for general G it can be shown that if G is C^∞ contact ν-determined, then (6.2) holds with ν replaced by $\nu+1$. □

Example 6.4. Next consider H: $(\mathbb{R}^2\times\mathbb{R},0)\to(\mathbb{R}^2,0)$ defined by

$$H(x,\lambda)=(x_1^3-\lambda x_1,x_2^3-\lambda x_2)^t.$$

Again it is easy to check that H satisfies the hypotheses of Theorem 4.1, this time with $\mu=(1,1,2)$ and $\nu=(3,3)$, so H is C^0 contact 3-determined. Here the shortcoming of the C^∞ theory is that H is only C^∞ contact 4-determined, so the degree of C^∞ contact determination is higher than is necessary for the purpose of capturing the qualitative structure of the bifurcation diagram. To see that H is not C^∞ contact 3-determined, note that, for example, $(x_1^2x_2^2,0)^t$ is not in $\tilde{T}G$, so a perturbation of H by this fourth order term gives an $\tilde{H}$ which is not C^∞ contact equivalent to H. □

7. Results on $\mathscr{S}$ finite determination (for $p=n$). It is clear that nondegeneracy alone is insufficient to ensure $\mathscr{S}$ finite determination because there is nothing in our nondegeneracy conditions to prevent eigenvalues of $G_x(u)$ from being purely imaginary. In this section we shall consider several cases in which $\mathscr{S}$ finite determination follows from a nondegeneracy condition combined with a condition which keeps the spectrum of $G_x(u)$ away from the imaginary axis

$$\mathscr{I}=\{z\in\mathbb{C}: \operatorname{Re} z=0\}.$$

First we present a lemma which gives a useful formulation of what is needed to prove $\mathscr{S}$ ν-determination when G is ND(ν).

LEMMA 7.1. *Suppose G: $(\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^n,0)$ is a C^2 map and $\nu\in\mathbb{R}^n$ is a multi-exponent such that G is* ND(ν). *If there exist $\delta>0$, $\gamma>0$ and a neighborhood U of the origin in $\mathbb{R}^n\times\mathbb{R}^l$ such that, with $E\in\mathscr{L}(\mathbb{R}^n,\mathbb{R}^n)$ and $\rho=\|u\|$,*

$$\sigma(G_x(u)+E)\cap\mathscr{I}=\varnothing \quad \text{when } u\in H(G,\nu,\delta)\cap U \text{ and } \|\rho^{1-\nu}E\|<\gamma, \tag{7.1}$$

then G is $\mathscr{S}\nu$-determined.

Proof. Let P: $(\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^n,0)$ be a C^2 map such that

$$\|\rho^{-\nu}P(u)\|=o(1) \quad\text{and}\quad \|\rho^{1-\nu}P_x(u)\|=o(1).$$

Then there exists a neighborhood $W\subset U$ of the origin for which

$$\|\rho^{1-\nu}sP_x(u)\|<\gamma \quad \text{when } (u,s)\in W\times J, \tag{7.2}$$

where $J\equiv[0,1]$. Note that we may assume, by making them smaller if necessary, that the δ and W here have all the properties of those in the proof of Theorem 3.1. Therefore, by the proof of Theorem 3.1, there exists a BD equivalence ϕ: $V\to W$ between G and $\tilde{G}\equiv G+P$. Furthermore, with

$$F(u,s)=G(u)+sP(u) \quad\text{and}\quad \psi\colon V\times J\to W\times J$$

as in the proof of Theorem 3.1,

$$s\mapsto F_x(\psi(u,s)), \qquad s\in J,$$

is a continuous path in $\mathscr{L}(\mathbb{R}^n,\mathbb{R}^n)$ from $G_x(u)$ to $\tilde{G}_x(\phi(u))$ when $u\in V$. But, by (7.1) and (7.2),

$$\sigma(F_x(\psi(u,s)))\cap\mathscr{I}=\varnothing \quad \text{when } (u,s)\in\left[(G^{-1}(0)-\{0\})\cap V\right]\times J,$$

since if w is defined by $(w,s)=\psi(u,s)$, then

$$F_x(\psi(u,s))=G_x(w)+sP_x(w)$$

and, by (3.17),

$$w\in H(G,\nu,\delta/2)\cap W\subset H(G,\nu,\delta)\cap U.$$

Thus ϕ is actually an $\mathscr{S}$equivalence, i.e.,

$$\operatorname{Ind}(G_x(u))=\operatorname{Ind}(\tilde{G}_x(\phi(u))) \quad \text{when } u\in(G^{-1}(0)-\{0\})\cap V,$$

because $\sigma(F_x(\psi(u,s)))$ depends continuously on s (so eigenvalues of $F_x(\psi(u,s))$ cannot migrate across $\mathscr{I}$ as s varies without lying on $\mathscr{I}$ for some $s\in J$). □

The next lemma is just a minor technical observation which will be used in some of our proofs later.

LEMMA 7.2. *If* G: $(\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^n,0)$ *is* ND(ν) *and* ε, δ *and* U *are such that* (3.1) *holds, then, for* $E\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)$,

$$\|G_x(u)^{-1}E\|<1 \quad \textit{whenever } u\in H(G,\nu,\delta)\cap U \textit{ and } \|\rho^{1-\nu}E\|<\varepsilon.$$

Proof.

$$\begin{aligned}\|G_x(u)^{-1}E\|&=\|(\rho^{1-\nu}G_x(u))^{-1}(\rho^{1-\nu}E)\|\\&\leqq\|(\rho^{1-\nu}G_x(u))^{-1}\|\cdot\|\rho^{1-\nu}E\|\\&=\|\rho^{1-\nu}E\|/d(\rho^{1-\nu}G_x(u)).\end{aligned}$$ □

A very simple and natural condition which keeps the spectrum of a nonsingular matrix off $\mathcal{I}$ is that the matrix be symmetric. The following lemma will be used to prove our first $\mathcal{S}\nu$-determination result, in which the condition added to nondegeneracy is just that $G_x(u)$ be symmetric. This result covers the important class of gradient vector fields.

LEMMA 7.3. *Let* $A\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)$. *If there exists* $Q\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)$ *such that*

$$QA+A^tQ \textit{ is positive definite}, \tag{7.3}$$

then

$$\sigma(A)\cap\mathcal{I}=\varnothing.$$

Proof. We shall prove the contrapositive. Suppose that $\sigma(A)\cap\mathcal{I}\neq\varnothing$. Then there exist $\beta\in\mathbb{R}$ and a nonzero $v\in\mathbb{C}^n$ such that

$$Av=i\beta v,$$

which is equivalent to

$$\bar{v}^tA^t=-i\beta\bar{v}^t,$$

where $\bar{v}$ is the complex conjugate of v. Thus, for all $Q\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)$,

$$\bar{v}^t(QA+A^tQ)v=i\beta\bar{v}^tQv-i\beta\bar{v}^tQv=0.$$

It follows that there is no Q for which (7.3) holds. □

THEOREM 7.4. *Suppose* G: $(\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^n,0)$ *is a* C^2 *map and* $\nu\in\mathbb{R}^n$ *is a multi-exponent such that* G *is* ND(ν). *If* $G_x(u)$ *is symmetric for all* $u\in\mathbb{R}^n\times\mathbb{R}^l$, *i.e.* $G(\cdot,\lambda)$ *is a gradient vector field on* $\mathbb{R}^n$ *for each* $\lambda\in\mathbb{R}^l$, *then* G *is* $\mathcal{S}\nu$*-determined.*

Proof. Since G is ND(ν), there exist $\varepsilon>0$, $\delta>0$ and U such that (3.1) holds. We shall show that the theorem is a consequence of Lemma 7.1 by applying Lemma 7.3 to

$$A=G_x(u)+E \quad \text{when } u\in H(G,\nu,\delta)\cap U \text{ and } \|\rho^{1-\nu}E\|<\varepsilon$$

in order to verify that (7.1) holds if $\gamma=\varepsilon$. It suffices to show that if $Q=G_x(u)^{-1}$, then (7.3) holds. Note that, since $G_x(u)$ is symmetric,

$$\begin{aligned}QA+A^tQ&=G_x(u)^{-1}[G_x(u)+E]+[G_x(u)+E^t]G_x(u)^{-1}\\&=2I+\left[G_x(u)^{-1}E+\left(G_x(u)^{-1}E\right)^t\right].\end{aligned}$$

Thus $QA+A^tQ$ is symmetric. Furthermore, by Lemma 7.2,

$$\left\|G_x(u)^{-1}E+\left(G_x(u)^{-1}E\right)^t\right\|\leqq 2\left\|G_x(u)^{-1}E\right\|<2.$$

Therefore $QA+A^tQ$ is positive definite. □

For nonsingular real 2×2 matrices the trace can be used to keep eigenvalues off $\mathscr{I}$; if $A\in\mathscr{L}(\mathbb{R}^2,\mathbb{R}^2)$ is nonsingular, then $\sigma(A)\cap\mathscr{I}\neq\varnothing$ if and only if $\det A>0$ and $\operatorname{tr} A=0$. In the next theorem a nondegeneracy condition on the trace is used to control $\sigma(G_x(u))$.

THEOREM 7.5. *Suppose G: $(\mathbb{R}^2\times\mathbb{R}^l,0)\to(\mathbb{R}^2,0)$ is a C^2 map and ν is a constant exponent such that G is* ND(ν). *If there exist $\varepsilon>0$, $\delta>0$ and a neighborhood U of the origin in $\mathbb{R}^2\times\mathbb{R}^l$ such that*

$$(7.4)\qquad |\operatorname{tr}(G_x(u))|\geqq\varepsilon\|u\|^{\nu-1}\quad \textit{when } u\in H(G,\nu,\delta)\cap U \textit{ and } \det(G_x(u))>0,$$

then G is $\mathscr{S}\nu$-determined.

Proof. We may assume that (3.1) holds with the same ε,δ and U. The theorem will be proved by showing that (7.1) holds if $\gamma=\varepsilon/2$. Supppose that

$$u\in H(G,\nu,\delta)\cap U\quad\text{and}\quad \|E\|<(\varepsilon/2)\|u\|^{\nu-1}.$$

Note that

$$\operatorname{sgn}(\det(G_x(u)+E))=\operatorname{sgn}(\det(G_x(u)))$$

because

$$G_x(u)+E=G_x(u)\left(I+G_x(u)^{-1}E\right)$$

and, by Lemma 7.2,

$$\det\left(I+G_x(u)^{-1}E\right)>0.$$

Therefore, if $\det(G_x(u))<0$, then $\det(G_x(u)+E)<0$ so the two eigenvalues of $G_x(u)+E$ are real and of opposite sign. On the other hand, if $\det(G_x(u))>0$, then $\det(G_x(u)+E)>0$, so $G_x(u)+E$ can have purely imaginary eigenvalues only if $\operatorname{tr}(G_x(u)+E)=0$. But

$$\operatorname{tr}(G_x(u)+E)=\operatorname{tr}(G_x(u))+\operatorname{tr}E=\operatorname{tr}(G_x(u))[1+(\operatorname{tr}E)/\operatorname{tr}(G_x(u))]$$

which cannot be zero because, by (7.4), $\operatorname{tr}(G_x(u))\neq 0$ and

$$|(\operatorname{tr}E)/\operatorname{tr}(G_x(u))|\leqq 2\|E\|/|\operatorname{tr}(G_x(u))|<1.$$

Thus, whatever the sign of $\det(G_x(u))$ happens to be,

$$\sigma(G_x(u)+E)\cap\mathscr{I}=\varnothing. \qquad \square$$

For a restricted class of quasi-homogeneous bifurcation problems there is a straightforward and computable condition which is sufficient to guarantee $\mathscr{S}$ finite determination. This class includes our examples in §4 and §6, all the problems considered by McLeod and Sattinger [6] and some of the problems considered by Landman and Rosenblat [4].

THEOREM 7.6. *Suppose G: $(\mathbb{R}^n\times\mathbb{R}^l,0)\to(\mathbb{R}^n,0)$ is a polynomial mapping and $\mu\equiv(\xi,\eta)\in\mathbb{R}^n\times\mathbb{R}^l$ and $\nu\in\mathbb{R}^n$ are multi-exponents such that G is quasi-homogeneous of degree (μ,ν). If*

$$(7.5)\qquad \nu-\xi \text{ is a constant exponent}$$

and

$$(7.6)\qquad \sigma(G_x(v))\cap\mathscr{I}=\varnothing \quad \text{whenever } G(v)=0 \text{ and } \|v\|=1,$$

then G is $\mathscr{S}\nu$-determined.

Proof. Clearly (7.6) implies (4.3), so, by Theorem 4.1, G is ND(ν). Since $\sigma(G_x(v)+B)$ depends continuously on $v\in\mathbb{R}^n\times\mathbb{R}^l$ and $B\in\mathscr{L}(\mathbb{R}^n,\mathbb{R}^n)$, $\mathscr{I}$ is closed and the set of zeros of G on the unit sphere in $\mathbb{R}^n\times\mathbb{R}^l$ is compact, there exists $\delta>0$ and $\gamma>0$ such that

$$(7.7)\qquad \sigma(G_x(v)+B)\cap\mathscr{I}=\varnothing \quad \text{whenever } \|G(v)\|<\delta,\ \|v\|=1 \text{ and } \|B\|<\gamma.$$

Let U be the open unit ball in $\mathbb{R}^n\times\mathbb{R}^l$. Once again we shall prove the theorem by showing that (7.1) holds.

Suppose that $u\in H(G,\nu,\delta)\cap U$ and $\|\rho^{1-\nu}E\|<\gamma$. Let $\sigma\in\mathbb{R}$ and $v\in\mathbb{R}^n\times\mathbb{R}^l$ be defined by (4.5) and (4.6). Then, as in the proof of Theorem 4.1, (4.7), (4.8) and (4.9) hold and $\|G(v)\|<\delta$. But, by (4.6) and (4.9),

$$G_x(u)+E=G_x(\sigma^{\mu}v)+E=\sigma^{\nu}G_x(v)\sigma^{-\xi}+E=\sigma^{\nu}\left[G_x(v)+\sigma^{-\nu}E\sigma^{\xi}\right]\sigma^{-\xi}$$
$$\sim\sigma^{\nu-\xi}\left[G_x(v)+\sigma^{-\nu}E\sigma^{\xi}\right],$$

where " ~ " means "is similar to". Thus, by (7.5),

$$\sigma(G_x(u)+E)=k\sigma(G_x(v)+B),$$

where $k=\sigma^{|\nu-\xi|}$ and $B=\sigma^{-\nu}E\sigma^{\xi}$. Therefore, by (7.7),

$$\sigma(G_x(u)+E)\cap\mathscr{I}=\varnothing,$$

since, by (4.1), (4.5), (4.7) and (4.8),

$$\begin{aligned}\|\sigma^{-\nu}E\sigma^{\xi}\|&=\|\sigma^{1-\nu}E\sigma^{\xi-1}\|\le\|\sigma^{1-\nu}E\|\\&=\left\|(\rho/\sigma)^{\nu-1}(\rho^{1-\nu}E)\right\|\\&\le\left\|(\rho/\sigma)^{\nu-1}\right\|\cdot\|\rho^{1-\nu}E\|\\&\le\|\rho^{1-\nu}E\|<\gamma.\end{aligned}$$

□

Naively, one might expect, for a C^2 bifurcation problem G and a constant exponent ν for which G is ND(ν), that an apparently strong nondegeneracy condition on the real parts of the eigenvalues of $G_x(u)$ such as "there exist $\varepsilon>0$, $\delta>0$ and a neighborhood U of the origin such that

$$\min\{|\mathrm{Re}\lambda|:\lambda\in\sigma(G_x(u))\}\geqq\varepsilon\|u\|^{\nu-1}\quad\text{when } u\in H(G,\nu,\delta)\cap U\text{"}$$

would be enough to make it easy to prove that G is $\mathscr{S}\nu$-determined. Unfortunately, the eigenvalues of a matrix can be extremely sensitive to perturbations of the matrix, especially if the Jordan form of the matrix is not diagonal. Thus, if one uses the known perturbation theory for eigenvalues of a general matrix together with Lemma 7.1, all one can conclude from this strong condition keeping $\sigma(G_x(u))$ away from $\mathscr{I}$ is the disappointing result that G is $\mathscr{S}$ finitely determined with an order of determination possibly much larger than ν. However, it is our feeling that it should be possible to obtain a better result from essentially the same hypotheses with a deeper study of the significance of the assumption that G is ND(ν).

REFERENCES

[1] M. Golubitsky and D. Schaeffer, *A theory for imperfect bifurcation via singularity theory*, Comm. Pure Appl. Math., 32 (1979), pp. 21–98.

[2] N. H. Kuiper, *C^1-equivalence of functions near isolated critical points*, Symposium on Infinite Dimensional Topology, No. 69, Princeton, Univ. Press, Princeton, NJ 1972.

[3] T. C. Kuo, *Characterizations of v-sufficiency of jets*, Topology, 11 (1972), pp. 115–131.

[4] K. A. Landman and S. Rosenblat, *Bifurcation from a multiple eigenvalue and stability of solutions*, SIAM J. Appl. Math., 34 (1978), pp. 743–759.

[5] B. Malgrange, *Ideals of Differentiable Functions*, Oxford Univ. Press, Cambridge 1966.

[6] J. B. McLeod and D. H. Sattinger, *Loss of stability and bifurcation at a double eigenvalue*, J. Funct. Anal., 14 (1973), pp. 62–84.

[7] M. Buchner, J. Marsden and S. Schecter, *Applications of the blowing-up construction and algebraic geometry to bifurcation problems*, J. Differential Equations, 48 (1983), pp. 404–433.

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

004

RESTRICTED QUADRATIC FORMS AND THEIR APPLICATION TO BIFURCATION AND STABILITY IN CONSTRAINED VARIATIONAL PRINCIPLES*

JOHN H. MADDOCKS†

Abstract. The subjects of this investigation are the abstract properties and applications of *restricted quadratic forms*. The first part of the presentation resolves the following question: if L is a self-adjoint linear operator mapping a Hilbert space H into itself, and S is a subspace of H, when is the quadratic form $\langle u, Lu\rangle$ positive for any nonzero $u \in S$? In the second part of the presentation, restricted quadratic forms are further examined in the specific context of constrained variational principles; and the general theory is applied to obtain information on stability and bifurcation. Two examples are then solved: one is finite-dimensional and of an illustrative nature; the other is a longstanding problem in elasticity concerning the stability of a buckled rod. In addition to being a valuable analytical tool for isoperimetric problems in the calculus of variations, the tests described are amenable to numerical treatment.

1. Introduction. This presentation has two main parts. The first describes conditions determining whether a quadratic form is positive when restricted to a given subspace of its domain of definition. The second motivates the inquiry by a description of how the discussed property arises naturally in applications. In particular, bifurcation and stability phenomena are considered within the context of constrained variational principles.

The question of restricted positivity is precisely formulated and answered in §2, where the main development is given in the form of a sequence of lemmas leading to Theorems 1 and 2 and a corollary. The relevance of §2 to bifurcation and stability in constrained variational problems is explained in §3. Two types of bifurcation are identified dependent upon whether the constraints play an active or passive role. In §§4 and 5, the theory of §2 is applied to two examples of the general type introduced in §3. The problem of §4 is set in three dimensions and is used to illustrate the underlying geometrical nature of Theorem 2. The example worked in §5 describes applications to the isoperimetric calculus of variations.

The remainder of this section describes connexions between this paper and previous work. The development of §2 parallels that of M. R. Hestenes as it appears in Hestenes (1951), and is reported in Gregory (1980, §2.2). The presentation here differs in that results are described in terms of a self-adjoint linear operator associated with a quadratic form, whereas Hestenes discusses the quadratic form *per se*. Consideration of the operator formulation allows an explicit statement of theorems that is not possible in terms of quadratic forms. In particular, the characterization of *relative nullity* given in Theorem 1 seems to be new; it is this result that allows Theorem 2 and its corollary to be applied in specific problems.

Results on quadratic forms are of two main types, namely conjugate or focal point theorems, and index theorems. This paper is concerned with index theorems. Bolza (1904) obtained a theory of constrained conjugate points within the specific context of the theory of the isoperimetric calculus of variations, but the tests entailed are complex.

*Received by the editors July 28, 1982, and in revised form July 2, 1983. This research was supported by the National Science Foundation, the Office of Naval Research, the Office of Army Research, and the Air Force Office of Scientific Research.

†Institute for Mathematics and Its Applications, Minneapolis, Minnesota 55455. This research was conducted while the author was in the Mathematics Department, Stanford University, Stanford, California 94305.

Other results of index type in the isoperimetric calculus of variations were obtained by Birkhoff and Hestenes (1935). They considered the following problem: given an unconstrained problem in the calculus of variations and an extremal that is not a minimum, can a finite set of isoperimetric constraints be added to the problem so as to make the extremal into a constrained minimum? In a sense this problem is inverse to the one treated here; we consider variational problems subject to given constraints, and determine which extremals are constrained minima. Hestenes (1951) and Gregory (1980), describe both focal point and index theorems, but only the former theory appears to have been applied to any extent (see Gregory (1980) for further details). The method of intermediate problems for eigenvalues (see, for example, Weinstein and Stenger (1972)) can also be regarded as a focal point theory for constrained quadratic forms.

The problem of primary concern here, as in the works cited above, is the determination of the properties of a quadratic form that is defined on an infinite-dimensional space, but is restricted to a subspace. For applications the case in which the subspace has (in a certain sense) finite codimension is of particular interest. Although the results of §2 are not dependent on this delimitation, we present a proof of Theorem 2 in the special case which is considerably simpler than that given in the comparable result of Hestenes.

The results of Hestenes are not well known, even in the highly special case of the whole space being finite-dimensional. For example, the results of Morse (1971a, b), (1973, p. 172) are easy consequences of the work of Hestenes (1951, Theorem 15.2). The specialization of Theorem 2 to finite dimensions actually extends the results of Morse. See also Cottle (1974), and Bellman (1960, §5) for discussion of the finite-dimensional problem.

The following two references are given for completeness; they discuss restricted quadratic forms but have no direct bearing on this work: Bognár (1974), Uhlig (1979).

2. Quadratic forms. In this section we consider quadratic forms $Q(u)$ of the type

$$Q(u)=\langle u,Lu\rangle, \qquad u\in\mathcal{D}\subset\mathcal{H}$$

where $\mathcal{H}$ is a real Hilbert space with inner-product $\langle\cdot,\cdot\rangle$, $\mathcal{D}$ is a dense subspace of $\mathcal{H}$, and L is a linear operator from $\mathcal{D}$ to $\mathcal{H}$. The following three hypotheses on L are made throughout:

H1. L is self-adjoint and Fredholm.

H2. L has a finite number σ^- of orthonormal eigenvectors ξ_i^-, $i=1,\cdots,\sigma^-$, corresponding to negative eigenvalues.

H3. L is positive on the orthogonal complement of span $\{\xi_i^-\}\oplus\ker L$.

Here $\ker L$ denotes the null space of L. The $\{\xi_i^-\}$ will be referred to as the *negative* eigenvectors. By the statement that L is Fredholm it is meant that the equation

$$Lu=h$$

has a solution if and only if h is orthogonal to $\ker L$. By the statement that L or Q is positive on a set $\mathcal{S}\in\mathcal{D}$ (as in H3, for example), it is meant that

$$Q(u)=\langle u,Lu\rangle>0 \quad \forall u\in\mathcal{S}, \quad u\neq 0. \tag{2.1}$$

L or Q being nonnegative is defined similarly.

It should be understood that in H3, and in the sequel, the *orthogonal complement* of a set means the orthogonal complement with respect to $\langle\cdot,\cdot\rangle$, but in the set $\mathcal{D}$; that is, the orthogonal complement of a set $\mathcal{S}\subset\mathcal{D}$ is defined by

$$\mathcal{S}^{\perp}=\{u\in\mathcal{D}:\ \langle u,v\rangle=0,\forall v\in\mathcal{S}\}. \tag{2.2}$$

In either the case of $\mathcal{H}$ being of finite dimension, or $\mathcal{H}$ being infinite-dimensional and L being bounded, the dense subspace $\mathcal{D}$ can be identified with $\mathcal{H}$. If L is unbounded this is not possible. In this last case the quadratic form $Q(u)$ can actually be defined on a larger space than the domain of definition of L. The theory is standard, but not relevant to the results presented here; further details can be obtained from Kato (1976, Chap. VI), for example.

The following notion of orthogonality will be useful.

DEFINITION 1. Two vectors u_1, $u_2\in\mathcal{D}$ are termed *L-orthogonal* if

$$\langle u_1,Lu_2\rangle=0.$$

Remarks. (a) As L is self-adjoint the relation is symmetric.

(b) Any vector is L-orthogonal to any element of $\ker L$.

(c) If one of the vectors involved is an eigenvector not in $\ker L$, then orthogonality and L-orthogonality are equivalent.

(d) The concept of L-orthogonality extends to sets in the obvious way.

To motivate the treatment adopted below, consider the following theorem, which is due to Hestenes (Gregory (1980, p. 62, Thm. 1)):

THEOREM (Hestenes). *Let $Q(u)$ be a quadratic form on a Hilbert space $\mathcal{H}$. Then there exist three subspaces $\mathcal{H}_-$, $\mathcal{H}_0$, and $\mathcal{H}_+$ such that* (a) $\mathcal{H}=\mathcal{H}_-\oplus\mathcal{H}_0\oplus\mathcal{H}_+$, *the sum being direct;* (b) *the three subspaces are mutually orthogonal and Q-orthogonal;* (c) *if the zero vector is excepted, $Q(u)$ is negative on $\mathcal{H}_-$, zero on $\mathcal{H}_0$, and positive on $\mathcal{H}_+$.*

Remarks.

(i) The definition of Q-orthogonality is directly analogous to Definition 1 of L-orthogonality.

(ii) The theorem is intuitive whenever $Q(u)$ can be written in the form $\langle u,Lu\rangle$ with L satisfying H1, H2 and H3; for $\mathcal{H}_-$ is the span of the negative eigenvectors of L, $\mathcal{H}_0$ is the kernel of L, and $\mathcal{H}_+$ is the orthogonal complement of $\mathcal{H}_-$ and $\mathcal{H}_0$.

(iii) Notice that $\mathcal{H}_-$ does *not* contain all vectors u that make $Q(u)$ negative. However, $\mathcal{H}_-$ is maximal in the sense that if u is orthogonal to $\mathcal{H}_-$, then $Q(u)$ is nonnegative.

(iv) Any closed subspace $\mathcal{B}$ of $\mathcal{H}$ forms another Hilbert space, so $\mathcal{B}$ can itself be decomposed in the above manner. Remark (ii) is again relevant if the self-adjoint operator L is replaced by the usual restriction of L to $\mathcal{B}$, which operator is also self-adjoint.

The above theorem will not be used directly, but the first essential idea is that a "maximal" negative subspace $\mathcal{B}_-$ can be associated with any given subspace $\mathcal{B}$. In particular, if $\mathcal{B}_-$ can be shown to have dimension zero, then Q is nonnegative on $\mathcal{B}$. Because of Remark (iv) above, the tests to be presented also provide information about the spectrum of the restriction of the operator L to $\mathcal{B}$, but this viewpoint will not be stressed in the sequel.

The second essential idea, detailed in Theorem 2, is that the sizes of the maximal negative and nonpositive subspaces of a subspace $\mathcal{B}$ are intimately connected with the sizes of the corresponding subspaces in the L-orthogonal complement of $\mathcal{B}$. Thus knowledge of $Q(u)$ on one subspace provides information about $Q(u)$ on the L-orthogonal complement. We shall be particularly concerned with a case arising in many applications, namely one subspace having finite dimension.

The simple result stated in Lemma 1 will be repeatedly exploited in the sequel.

LEMMA 1. *Any subspace of $\mathcal{V}$ with finite dimension n has a mutually L-orthogonal basis. That is, there is a basis $\{u_i, i=1,\cdots,n\}$ such that*

$$\langle u_i, Lu_j\rangle = 0, \qquad i \neq j.$$

Proof. Consider any basis $\{v_i, i=1,\cdots,n\}$. Since L is self-adjoint, the $n\times n$ matrix $A=\{a_{ij}\}$, where

$$a_{ij}=\langle v_i, Lv_j\rangle,$$

is symmetric. Therefore there exists an *orthogonal* $n\times n$ matrix $P=\{p_{ij}\}$ such that P^TAP is diagonal. The set

$$\left\{u_i\colon u_i=\sum_j p_{ji}v_j\right\}$$

is a basis because P is nonsingular. Also, this basis is mutually L-orthogonal by construction. □

The next definition generalizes the properties of span $\{\xi_i^-\}$ relevant in the study of constrained quadratic forms.

DEFINITION 2. For any subspace $\mathcal{S}\subset\mathcal{V}$, a *maximal negative subspace* of $\mathcal{S}$, denoted $\mathcal{M}(\mathcal{S})$, satisfies the following two properties:

Negativity: $\forall$ nonzero $u\in\mathcal{M}$, $Q(u)<0$, and

Maximality: $\forall v\in\mathcal{S}$ that are L-orthogonal to $\mathcal{M}$, $Q(u)\geq 0$.

Remarks. (a) $\mathcal{M}$ need not contain all vectors u that make $Q(u)<0$.

(b) The maximality condition given is equivalent to:

If $v\in\mathcal{S}$ and $v\notin\mathcal{M}(s)$, then $\mathcal{M}\oplus\text{span}\{v\}$ is not a negative subspace. The proof of equivalence is by contradiction, and is straightforward once $\mathcal{M}$ is shown to have finite dimension (vide infra), and Lemma 1 is invoked.

(c) Consideration of the example described in §4 shows that maximal negative subspaces need not be unique. However, we do have the following two lemmas.

LEMMA 2. span$\{\xi_i^-\}$ *is a maximal negative subspace of* $\mathcal{V}$.

Proof. Negativity is trivial. Maximality is also clear; for if

$$\langle w, L\xi_i^-\rangle = 0 = \lambda_i\langle w, \xi_i^-\rangle, \qquad i=1,\cdots,\sigma^-,$$

then H3 implies that $\langle w, Lw\rangle$ is positive. □

LEMMA 3. *Each maximal negative subspace of $\mathcal{S}\subseteq\mathcal{V}$ has the same finite dimension, denoted $d^-[\mathcal{S}]$.*

Proof. Note that if n is greater than m, then any subspace of dimension n contains a vector L-orthogonal to any subspace of dimension m. Hypothesis H3 then implies that any negative subspace has a dimension less than σ^-. Similarly, the existence of two maximal negative subspaces of $\mathcal{S}$ with different dimensions contradicts the maximality of one of the subspaces. □

Remark. The nonnegative integer $d^-[\mathcal{S}]$ is known as the *signature* or index of the quadratic form Q on the subspace $\mathcal{S}$. For completeness we give the following characterization theorem (Hestenes (1951, p. 547)).

THEOREM (Hestenes). *The index of $Q(u)$ on $\mathcal{S}$ is given by either*:

(i) *the dimension of a maximal subspace of $\mathcal{S}$ on which $Q(u)<0$, $u\neq 0$;*

(ii) *the least integer k such that $Q(u)\geq 0$ on the Q-orthogonal complement of a subspace of $\mathcal{S}$ with dimension k;*

(iii) *the least integer k such that $Q(u)\geq 0$ on the orthogonal complement of a subspace of $\mathbb{S}$ with dimension k*;

(iv) *the least integer k such that there exist k linear forms $K_1(u),\cdots,K_k(u)$ such that $Q(u)\geq 0$ whenever $K_\alpha(u)=0$ $(\alpha=1,\cdots,k)$.*

The usual definition of index is (i); by remark (b) above this is equivalent to Definition 2.

DEFINITION 3. For any subspace $\mathbb{S}\subseteq\mathfrak{D}$, a *maximal nonpositive subspace* of $\mathbb{S}$, denoted $\mathfrak{N}\mathfrak{P}(\mathbb{S})$, satisfies the following three properties:

Nonpositivity: $\forall u\in\mathfrak{N}\mathfrak{P}(\mathbb{S})$, $Q(u)\leq 0$,

Maximality: If $v\in\mathbb{S}$ satisfies $v\notin\mathfrak{N}\mathfrak{P}(\mathbb{S})\oplus\ker L$ and if v is L-orthogonal to $\mathfrak{N}\mathfrak{P}(\mathbb{S})$, then

$$Q(v)>0.$$

Thirdly,

$$\mathfrak{N}\mathfrak{P}(\mathbb{S})\cap\ker L=\{\mathbf{0}\}.$$

Remarks. (a) This definition is directly analogous to Definition 2. The maximality condition cannot apply to elements of $\mathfrak{N}\mathfrak{P}(\mathbb{S})$ because it is possible for an element of $\mathfrak{N}\mathfrak{P}(\mathbb{S})$ to be L-orthogonal to the set. The third condition could be omitted, but it allows us to distinguish between $Q(u)$ vanishing because $u\in\ker L$, and $Q(u)$ vanishing because u is orthogonal to Lu.

(b) The maximality condition is equivalent to: if $v\in\mathbb{S}$ and $v\notin\mathfrak{N}\mathfrak{P}(\mathbb{S})\oplus\ker L$, then $\mathfrak{N}\mathfrak{P}(\mathbb{S})\oplus\operatorname{span}\{v\}$ is not a nonpositive subspace.

(c) The proof of Theorem 1 (vide infra) demonstrates that any maximal negative subspace of $\mathbb{S}$ can be extended to a maximal nonpositive subspace. However, not all maximal nonpositive subspaces can be obtained in this way.

LEMMA 4. *Each maximal nonpositive subspace of $\mathbb{S}$ has the same finite dimension, $d^-(\mathbb{S})+d^0(\mathbb{S})$ say.*

Proof. The proof is analogous to that of Lemma 3. □

Remark. As a maximal negative subspace satisfies all the conditions for a nonpositive subspace except maximality, it is clear that

$$d^0(\mathbb{S})\geq 0.$$

The next two lemmas are not necessary for the succeeding development, but are given to provide some familiarity with $d^-(\mathbb{S})$ and $d^0(\mathbb{S})$. Lemma 5 is actually a particular case of Theorem 1. Lemma 6 provides a constructive method for calculation of $d^-(\mathbb{S})$ and $d^0(\mathbb{S})$ in problems where $\mathbb{S}$ is finite dimensional.

LEMMA 5. $d^0(\mathfrak{D})=0$.

Proof. $\operatorname{span}\{\xi_i^-\}$ is both a maximal negative and a maximal nonpositive subspace of $\mathfrak{D}$. □

LEMMA 6. *Let $\mathbb{S}\subset\mathfrak{D}$ have finite dimension n, let $\{v_i, i=1,\cdots,n\}$ be any basis of $\mathbb{S}$, and let the $n\times n$ matrix W be defined by*

$$W=\{\langle v_i, Lv_j\rangle\}.$$

Then

$$d^-(\mathbb{S})=\textit{number of negative eigenvalues of } W,$$

and

$$d^0(\mathbb{S})=(\textit{multiplicity of zero as an eigenvalue of } W)-\dim(\ker L\cap\mathbb{S}).$$

Proof. The proof of Lemma 1 demonstrates that there is an orthogonal matrix $P=\{p_{ij}\}$ such that the set

$$\left\{u_i\colon u_i=\sum_j p_{ji}v_j\right\}$$

is a mutually L-orthogonal basis for $\mathcal{S}$. Clearly the space

$$\operatorname{span}\{u_i\colon \langle u_i, Lu_i\rangle<0\}$$

is a maximal negative subspace of $\mathcal{S}$ with dimension equal to the number of negative eigenvalues of $\operatorname{diag}\{\langle u_i, Lu_i\rangle\}=P^T\langle v_i, Lv_j\rangle P=P^TWP$. But as the eigenvalues of a matrix are invariant under orthogonal equivalence this gives the required expression for $d^-(\mathcal{S})$. The expression for $d^0(\mathcal{S})$ can be derived similarly because

$$\operatorname{span}\{u_i\colon \langle u_i, Lu_i\rangle\leq 0\}=\mathcal{M}\mathcal{P}(\mathcal{S})\oplus\{\ker L\cap\mathcal{S}\}. \qquad \square$$

Theorem 1 is of interest because it provides an alternative characterization of $d^0(\mathcal{S})$. It is also used in the derivation of Theorem 2.

If $\mathcal{Q}$ is a subspace of $\mathcal{H}$, the preimage of $\mathcal{Q}$ under L will be denoted $L^{-1}(\mathcal{Q})$. Recall that the orthogonal complement $\mathcal{S}^\perp$ of a subspace $\mathcal{S}\subset\mathcal{D}$, connotes the orthogonal complement in $\mathcal{D}$.

THEOREM 1. *Let the subspace* $\mathcal{C}\subseteq\mathcal{D}$, *be closed in* $\mathcal{D}$. *Then*

$$d^0(\mathcal{C})=\dim\{\mathcal{C}\cap L^{-1}(\mathcal{C}^\perp)\cap(\ker L)^\perp\}. \tag{2.3}$$

Proof. Let $\mathcal{G}(\mathcal{C})$ denote the subspace $\mathcal{C}\cap L^{-1}(\mathcal{C}^\perp)\cap(\ker L)^\perp$, and let $\mathcal{M}(\mathcal{C})$ be any maximal negative subspace of $\mathcal{C}$. We prove that $\mathcal{M}(\mathcal{C})\oplus\mathcal{G}(\mathcal{C})$ is a maximal nonpositive subspace of $\mathcal{C}$. Equation (2.3) is a consequence of this fact because the sum of $\mathcal{M}$ and $\mathcal{G}$ is direct, and because

$$d^0(\mathcal{C})=\dim[\mathcal{M}\oplus\mathcal{G}]-\dim\mathcal{M}.$$

To see that the sum is direct note that any $x\in\mathcal{G}$ satisfies $x\in\mathcal{C}$ and $Lx\in\mathcal{C}^\perp$. Consequently $Q(x)$ vanishes and therefore $x\notin\mathcal{M}(\mathcal{C})$. It is also clear that $\ker L\cap(\mathcal{M}\oplus\mathcal{G})=\{0\}$, and that $\mathcal{M}\oplus\mathcal{G}$ is a nonpositive subspace.

It remains to prove that $\mathcal{M}\oplus\mathcal{G}$ is maximal, that is, to demonstrate that

(2.4) If $x\in\mathcal{C}$ is L-orthogonal to $\mathcal{M}\oplus\mathcal{G}$, and $x\notin\mathcal{M}\oplus\mathcal{G}\oplus\ker L$, then $\langle x, Lx\rangle>0$.

The maximality of $\mathcal{M}$ as a negative subspace implies that any such x satisfies $\langle x, Lx\rangle\geq 0$, so we obtain the maximality of $\mathcal{M}\oplus\mathcal{G}$ after reaching a contradiction on the assumption $\langle x, Lx\rangle=0$. Note that because L is Fredholm, and because $\mathcal{C}$ is closed in $\mathcal{D}$, any $x\in\mathcal{C}$ can be written as a sum

$$x=p+q, \text{ where } p\in\mathcal{C}\cap L^{-1}(\mathcal{C}) \text{ and } q\in\mathcal{C}\cap L^{-1}(\mathcal{C}^\perp).$$

Moreover, $q\in\mathcal{G}\oplus\ker L$, so that:

if $x\notin\mathcal{M}\oplus\mathcal{G}\oplus\ker L$, then $p\notin\mathcal{M}\oplus\mathcal{G}\oplus\ker L$;
if x is L-orthogonal to $\mathcal{M}\oplus\mathcal{G}$, so is p;

and, by choice of p and q,

$$\langle x, Lx\rangle=\langle p, Lp\rangle.$$

A contradiction on the assumption that $\exists p \in \mathcal{C} \cap L^{-1}(\mathcal{C})$ satisfying the hypotheses of (2.4) and $\langle p, Lp \rangle = 0$, is therefore sufficient to obtain the desired maximality. Let $\{u_i, i = 1, \cdots, d^-(\mathcal{C})\}$ be a mutually L-orthogonal basis of $\mathfrak{M}(\mathcal{C})$ (the existence of which is guaranteed by Lemma 1). As $p \in L^{-1}(\mathcal{C})$, $Lp \in \mathcal{C}$, so

$$f \equiv Lp - \sum_i \alpha_i u_i,$$

where

$$\alpha_i = \langle Lp, Lu_i \rangle / \langle u_i, Lu_i \rangle$$

is an element of $\mathcal{C}$ that is L-orthogonal to $\mathfrak{M}(\mathcal{C})$. Hence

$$\beta p + f \in \mathcal{C}, \qquad \beta \in \mathbb{R}$$

is L-orthogonal to $\mathfrak{M}(\mathcal{C})$. But

$$\langle \beta p + f, L(\beta p + f) \rangle = \beta^2 \langle p, Lp \rangle + 2\beta \langle Lp, f \rangle + \langle f, Lf \rangle,$$

which equals

$$\beta^2 \langle p, Lp \rangle + 2\beta \langle Lp, Lp \rangle - 2\beta \left\langle Lp, \sum \alpha_i u_i \right\rangle + \langle f, Lf \rangle.$$

But by hypothesis this expression is

$$2\beta \langle Lp, Lp \rangle + \langle f, Lf \rangle. \tag{2.5}$$

Also, $p \notin \ker L$, so that $\langle Lp, Lp \rangle = \|Lp\|^2 > 0$. Therefore, $\beta \in \mathbb{R}$ can be chosen such that (2.5) is negative, contradicting the maximality of $\mathfrak{M}(\mathcal{C})$. □

Remarks (a). Elements of $\mathcal{C} \cap L^{-1}(\mathcal{C}^\perp)$ are termed *Q-transversals of* $\mathcal{C}$ by Hestenes (1951). His definition is that x is a Q-transversal of $\mathcal{C}$ if it is Q-(or L-) orthogonal to the whole of $\mathcal{C}$. Theorem 1 can be viewed as proving that $\mathcal{C} \cap L^{-1}(\mathcal{C}^\perp)$ coincides precisely with the set of all Q-transversals of $\mathcal{C}$. The dimension of $\mathcal{C} \cap L^{-1}(\mathcal{C}^\perp)$ is called the nullity of Q on $\mathcal{C}$. The dimension $d^0[\mathcal{C}]$, is called the relative nullity of Q on $\mathcal{C}$ (Gregory (1980, §2.2)), and characterizes those Q-transversals not in $\ker L$.

(b) For operators L satisfying H1 the subspace $L^{-1}(\mathcal{C})$ coincides with the subspace $(L\mathcal{C}^\perp)^\perp$. Theorems 1 and 2 could be restated accordingly.

The next result relates the properties of the quadratic form $Q(u)$ on the orthogonal complement of a closed subspace $\mathcal{C}$, to the properties of $Q(u)$ on the L-preimage of $\mathcal{C}$.

THEOREM 2. *For any subspace $\mathcal{C} \subseteq \mathfrak{D}$, that is closed in $\mathfrak{D}$,*

$$d^0[\mathcal{C}^\perp] = \dim[\mathcal{C}^\perp \cap L^{-1}(\mathcal{C}) \cap (\ker L)^\perp] = d^0[L^{-1}(\mathcal{C})], \tag{2.6}$$

and

$$d^-[\mathcal{C}^\perp] + d^-[L^{-1}(\mathcal{C})] + d^0[L^{-1}(\mathcal{C})] = \sigma^-. \tag{2.7}$$

Proof. The first equality in (2.7) is an immediate consequence of Theorem 1. The second equality in (2.6) is also implied by Theorem 1 because the Fredholm property of L provides the following identity,

$$\mathcal{C}^\perp \cap L^{-1}(\mathcal{C}) \cap (\ker L)^\perp = L^{-1}(\mathcal{C}) \cap L^{-1}\left[\{L^{-1}(\mathcal{C})\}^\perp\right] \cap (\ker L)^\perp.$$

Equation (2.7) is consequent upon the following argument. Consider the subspace

$$\mathfrak{M}[\mathcal{C}^\perp] \oplus \mathfrak{N}\mathfrak{P}[L^{-1}(C)]. \tag{2.8}$$

The sum is direct; for if $x \in \mathcal{N}[\mathcal{C}^\perp]$, then $x \in \mathcal{C}^\perp$ and $\langle x, Lx \rangle < 0$, which is incompatible with $Lx \in \mathcal{C}$. Notice also that $\mathcal{C}^\perp$ and $L^{-1}(\mathcal{C})$ are mutually L-orthogonal. The subspace (2.8) is therefore a nonpositive subspace of $\mathcal{H}$, for if $u \in \mathcal{N}[\mathcal{C}^\perp]$ and $v \in \mathcal{N}\mathcal{P}[L^{-1}(\mathcal{C})]$, then

$$\langle u+v, L(u+v) \rangle = \langle u, Lu \rangle + \langle v, Lv \rangle \leq 0.$$

It can be shown that (2.8) is actually a maximal nonpositive subspace of $\mathcal{D}$. As (2.8) has dimension $d^-[\mathcal{C}^\perp] + d^-[L^{-1}(\mathcal{C})] + d^0[L^{-1}(\mathcal{C})]$, Lemmas 2, 4 and 5 combined then imply (2.6). The proof is similar to that of Theorem 1 and depends on the following decomposition. Because $\mathcal{C}$ is closed and L is Fredholm, $\mathcal{D}$ can be written as the sum of the four subspaces

$$\{\mathcal{C} \cap L^{-1}(\mathcal{C})\} + \{\mathcal{C} \cap L^{-1}(\mathcal{C}^\perp)\} + \{\mathcal{C}^\perp \cap L^{-1}(\mathcal{C})\} + \{\mathcal{C}^\perp \cap L^{-1}(\mathcal{C}^\perp)\}.$$

Rather than proceed with the proof in general, we present a more illuminating, and simpler proof for the special case arising in the application considered subsequently, namely $\mathcal{C}$ having finite dimension. The above argument is retained to the extent that (2.8) being a nonpositive subspace implies the inequality

$$(2.9) \qquad d^-[\mathcal{C}^\perp] + d^-[L^{-1}(\mathcal{C})] + d^0[L^{-1}(\mathcal{C})] \leq \sigma^-,$$

but maximality arguments are not used to obtain equality. Instead, the opposite inequality is proven by our demonstrating the existence of a negative subspace of $\mathcal{C}^\perp$ with dimension $(\sigma^- - d^-[L^{-1}(\mathcal{C})] - d^0[L^{-1}(\mathcal{C})])$.

First decompose $L^{-1}(\mathcal{C})$ as

$$\mathcal{X} \oplus \ker L,$$

where $\mathcal{X}$ is orthogonal to $\ker L$. As $\mathcal{C}$ is of finite dimension, so is $\mathcal{X}$. Denote the dimension of $\mathcal{X}$ by m, and let $\{\eta_i\}$, $i = 1, \cdots, m$ be an L-orthogonal basis for $\mathcal{X}$ (the existence of which is guaranteed by Lemma 1). According to Lemma 6 we may assume that

$$\langle \eta_i, L\eta_i \rangle \leq 0, \qquad i = 1, \cdots, d^-[L^{-1}(\mathcal{C})] + d^0[L^{-1}(\mathcal{C})].$$

Let η_j^+ denote the remaining $(m - d^-[L^-(\mathcal{C})] - d^0[L^-(\mathcal{C})])$ elements of the basis, which satisfy

$$(2.10) \qquad \langle \eta_j^+, L\eta_j^+ \rangle > 0.$$

Consider the subspace spanned by the η_j^+ and the negative eigenvectors ξ_i^-, and notice that

$$(2.11) \qquad \operatorname{span}(\xi_i^-, \eta_j^+) = \operatorname{span}(\xi_i^-) \oplus \operatorname{span}(\eta_j^+).$$

The sum is direct because any nonzero vector common to both subspaces would make Q simultaneously positive and negative. Accordingly, (2.11) has dimension $(\sigma^- + m - d^-[L^{-1}(\mathcal{C})] - d^0[L^{-1}(\mathcal{C})])$.

Now, by definition, ξ_i^- and η_j are orthogonal to $\ker L$, so in order that an element v of (2.11) be in $\mathcal{C}^\perp$ it need only satisfy the m orthogonality relations

$$(2.12) \qquad \langle v, L\eta_i \rangle = 0, \qquad i = 1, \cdots, m.$$

Consequently, there exists a subspace of (2.11), with dimension $(\sigma^- - d^-[L^{-1}(\mathcal{C})] - d^0[L^{-1}(\mathcal{C})])$, that is contained in $\mathcal{C}^\perp$. This subspace is next shown to be a negative subspace, thus completing the proof of (2.7), and Theorem 2, in the case of $\mathcal{C}$ being of finite dimension.

Let v be an arbitrary element of subspace (2.11) satisfying (2.12). The vector v can be expressed in the form

$$v=\xi+\sum_i \alpha_i \eta_i^+,$$

where $\xi \in \operatorname{span}\{\xi_i^-\}$. By the L-orthogonality of the η_i, equations (2.12) imply that

$$\langle \xi, L\eta_j^+ \rangle = -\alpha_j \langle \eta_j^+, L\eta_j^+ \rangle. \tag{2.13}$$

Furthermore, the L-orthogonality of the η_j^+ implies that

$$\langle v, Lv \rangle = \langle \xi, L\xi \rangle + 2\sum_j \alpha_j \langle \xi, L\eta_j^+ \rangle + \sum_j \alpha_j^2 \langle \eta_j^+, L\eta_j^+ \rangle,$$

which, because of (2.13), can be written as

$$\langle v, Lv \rangle = \langle \xi, L\xi \rangle - \sum_j \alpha_j^2 \langle \eta_j^+, L\eta_j^+ \rangle. \tag{2.14}$$

But $\xi \in \operatorname{span}\{\xi_i^-\}$, so that $\langle \xi, L\xi \rangle < 0$, which combined with (2.10) demonstrates the right-hand side of (2.14) to be negative, as was required. □

Remarks. (a) Gregory (1980, Thm. 16, p. 71) presents a result that has close connexions with Theorem 2, and attributes it to unpublished work of M. R. Hestenes. In that work the results are stated in terms of the index, nullity and relative nullity of the quadratic form Q, and the operator L does not appear explicitly. The two theorems coincide in many cases, but it is not clear that they are always equivalent. Theorem 2 above, certainly has a more explicit form, which is suitable for the applications described in §§4, 5. We believe the simpler proof for the case of $\mathcal{C}$ being of finite dimension, to be completely new.

(b) The apparent asymmetry in (2.7) between $\mathcal{C}^\perp$ and $L^{-1}(\mathcal{C})$ is nebulous because of (2.6): the result applies to any closed subspace and its L-orthogonal complement.

(c) It has nowhere been assumed that $\ker L$ is finite.

The usefulness of the dimensions d^- and d^0 in applications is apparent from the following lemma and corollary.

LEMMA 7. *For any subspace $\mathcal{S} \subseteq \mathcal{H}$,*

(i) *Q is nonnegative on $\mathcal{S}$ iff $d^-[\mathcal{S}]=0$,*

(ii) *Q is positive on $\mathcal{S}$ iff $d^-[\mathcal{S}]=d^0[\mathcal{S}]=\dim[\mathcal{S}\cap\ker L]=0$.*

Proof. Immediate from definitions of $d^-[\mathcal{S}]$ and $d^0[\mathcal{S}]$. □

COROLLARY 1. *For any closed subspace $\mathcal{C} \subseteq \mathcal{H}$,*

(i) *Q is nonnegative on $\mathcal{C}^\perp$ iff $d^-[L^{-1}(\mathcal{C})]+d^0[L^{-1}(\mathcal{C})]=\sigma^-$,*

(ii) *Q is positive on $\mathcal{C}^\perp$ iff $d^-[L^{-1}(\mathcal{C})]=\sigma^-$, and $\dim[\mathcal{C}^\perp\cap\ker L]=0$*

Proof. Immediate from Theorem 2 and Lemma 7. □

The implicit belief motivating Corollary 1 is that in many applications $\mathcal{C}$ will have small dimension, so that Lemma 6 can be practicably applied to determine d^- and d^0 of $L^{-1}(\mathcal{C})$. It is also reasonable to expect that the number σ^- can be estimated by one of the many standard techniques applying to self-adjoint eigenvalue problems. Therefore, if $\ker L$ is also known—for example, if L is nonsingular—then Corollary 1 can be used to ascertain the properties of Q over the (possibly infinite) subspace $\mathcal{C}^\perp$.

3. Bifurcation and stability. We now describe the way in which restricted quadratic forms arise in constrained variational principles. In such problems the solutions sought are *stationary points* or *extremals* of a functional

$$F(u,\lambda)\colon \mathcal{H}\times\mathbb{R}\to\mathbb{R} \tag{3.1}$$

subject to constraints

$$\mathbb{G}(u,\lambda)=\mathbf{0}, \qquad \mathbb{G}\colon \mathcal{H}\times\mathbb{R}\to\mathbb{R}^m. \tag{3.2}$$

Here u is a variable in a real Hilbert space $\mathcal{H}$, and λ is a bifurcation parameter or possibly a vector of bifurcation parameters. The theory of Lagrange undetermined multipliers (see e.g. Hestenes (1966)) demonstrates that, provided certain smoothness assumptions are made, constrained extremals of (3.1) are solutions of the extended gradient system

$$F_u+\boldsymbol{\nu}\cdot\mathbb{G}_u=0, \tag{3.3}$$

$$\mathbb{G}=\mathbf{0}, \tag{3.4}$$

where $\boldsymbol{\nu}$ is an m-vector of Lagrange multipliers, and the subscript u denotes Gateaux differentiation with respect to u. We shall assume that a branch of solutions $(u^*(\lambda),\boldsymbol{\nu}^*(\lambda))$ to (3.3) and (3.4) exists. Two related questions are then addressed: for which λ does the stationary point realize a minimum, and for which λ can other extremals bifurcate from the given solution?

Whether an extremal is a constrained local minimum is typically of interest because this property often coincides with *stability* of the extremal regarded as an equilibrium configuration of an underlying dynamical system. For each λ, an extremal u^* is said to be a *strict constrained local minimum* of (3.1) if there is a neighbourhood U of u^* in $\mathcal{H}$ such that $\forall v\in U$ *satisfying constraints* (3.2), and $u\neq u^*$,

$$F(u)>F(u^*). \tag{3.5}$$

A necessary condition for (3.5) to be satisfied is:

(3.6) If $u\in\mathcal{D}\setminus\{0\}$ satisfies $\langle u,T_i\rangle=0$, $i=1,\cdots,m$, then $\langle u,Lu\rangle>0$.

Here $T_i\in\mathcal{H}$ is defined as the ith component of $G_u(u^*,\lambda)$, L: $\mathcal{H}\to\mathcal{H}$ is the linear self-adjoint operator

$$L(u^*,\boldsymbol{\nu}^*,\lambda)\equiv F_{uu}(u^*,\lambda)+\boldsymbol{\nu}^*\cdot G_{uu}(u^*,\lambda), \tag{3.7}$$

$\mathcal{D}$ denotes the domain of definition of L, and $\langle\cdot,\cdot\rangle$ is the inner-product on $\mathcal{H}$. The connexion between conditions (3.5) and (3.6) is discussed by, for example, Hestenes (1966). In many applications criterion (3.6) is also a sufficient condition for (3.5) to hold; however, this is not always the case. We do not pursue this point further in abstract, but the example of §4 is representative of a category of problems in which condition (3.6) is both a necessary and sufficient test for an extremal to be a strict constrained local minimum.

Condition (3.6) can be rephrased as a requirement that the quadratic form $\langle u,Lu\rangle$ be positive-definite on the subspace $(\text{span}\{T_i\})^\perp$. Then, provided that the operator L satisfies Hypotheses 1 to 3, Corollary 1 gives necessary and sufficient conditions for (3.6) to hold. In particular, Corollary 1 always provides necessary conditions for an extremal to be a strict constrained local minimum, and Corollary 1 is also a sufficient test whenever (3.6) is.

We now turn to the second point of inquiry, namely, when can there be bifurcation from the extremal $(u^*(\lambda), \boldsymbol{\nu}(\lambda))$? Equation (3.3) and (3.4) are of the form

$$\Omega(u,\boldsymbol{\nu},\lambda)=\mathbf{0}, \tag{3.8}$$

where Ω: $\mathcal{H}\times\mathbb{R}^m\times\mathbb{R}\rightarrow\mathcal{H}\times\mathbb{R}^m$ is a *gradient* (or *potential*) map, so that all the theory of bifurcation in variational problems is applicable. In particular, a necessary condition for there to be bifurcation is that the linearization of Ω be singular, and relatively simple additional tests make this condition both necessary and sufficient (see, for example, Rabinowitz (1976) and references therein).

By the statement that the linearization of Ω is singular we mean that there is a nontrivial solution $(\dot{u},\dot{\boldsymbol{\nu}})\in\mathcal{H}\times\mathbb{R}^m$ of the system

$$\{F_{uu}(u^*)+\boldsymbol{\nu}^*_{\cdot}G_{uu}(u^*)\}\dot{u}+\dot{\boldsymbol{\nu}}.\mathbf{G}_u(u^*)=0, \qquad \mathbf{G}_u(u^*)\dot{u}=\mathbf{0}.$$

But, in the notation of (3.6) and (3.7), these equations are

$$L\dot{u}+\sum_{i=1}^{m}\dot{\nu}_iT_i=0, \tag{3.9}$$

$$\langle T_i,\dot{u}\rangle=0, \qquad i=1,\cdots,m, \tag{3.10}$$

whose solutions can be classified in the following way.

We distinguish between two categories of solutions: Type 1 in which $\dot{\boldsymbol{\nu}}=\mathbf{0}$, and Type 2 in which $\dot{\boldsymbol{\nu}}\neq\mathbf{0}$. Now (3.9) implies that $\dot{u}\in L^{-1}(\text{span}\{T_i\})$, and equation (3.10) implies that $\dot{u}\in(\text{span}\{T_i\})^{\perp}$, so the number of linearly independent solutions of Type 1 is

$$\dim\left[\ker L\cap(\text{span}\{T_i\})^{\perp}\right],$$

and the number of linearly independent solutions of Type 2 is

$$\dim\left[(\text{span}\{T_i\})^{\perp}\cap L^{-1}(\text{span}\{T_i\})\cap(\ker L)^{\perp}\right]. \tag{3.11}$$

By Theorem 1, quantity (3.11) equals

$$d^0\left[(\text{span}\{T_i\})^{\perp}\right].$$

A comparison with the results of §2 reveals that—as might have been expected a priori—there is a candidate bifurcation point whenever there is a nontrivial vector u satisfying the constraints that also makes $Q(u)$ vanish. Of greater interest is that the operator L is not singular at a bifurcation point of Type 2, and that L can be singular without there being a candidate bifurcation point. We remark that a Type 2 bifurcation cannot occur unless there is at least one negative eigenvalue of L.

The constraints play different roles in the two types of bifurcation. At bifurcation points of Type 1 the vector of multipliers $\boldsymbol{\nu}$ remains constant to first order along bifurcating branches, and the bifurcating branches are also solutions of the unconstrained bifurcation problem obtained by our retaining equation (3.9) and discarding equation (3.10). At bifurcation points of Type 2 the vector $\boldsymbol{\nu}$ does not remain constant on bifurcating branches, and bifurcating branches do not satisfy the corresponding unconstrained problem. These observations are important in many engineering applications where typically the Lagrange multipliers represent the reactions of boundary supports. These reactions can often be found without explict knowledge of the solution, in which case the problem is said to be *statically determinate*. Only bifurcations of Type

1 are possible in a statically determinate problem, whereas bifurcations of Type 2 are to be expected whenever a problem ceases to be statically determinate. The examples of §§4 and 5 illustrate these points.

4. An example in finite dimensions. The first example we describe is a finite-dimensional problem having its origins in an idealized model for bending of a beam in the plane. The system is illustrated in Fig. 1. Three rigid rods of lengths $1, \mu$ and 1 are connected by two elastic joints whose equal stiffnesses are normalized to unity. The extreme ends of the linkage are free to rotate, but are constrained to lie on a given line. The position of one end-point is fixed, but the other end-point is free to move under the action of a compressive load λ. Any configuration is determined by the angles θ_1, θ_2 and θ_3 between the three rods and the line joining the end-points.

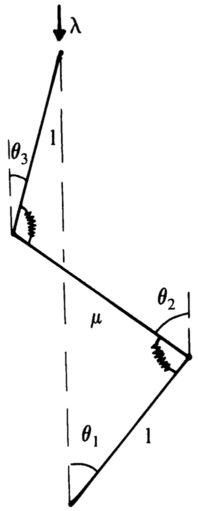


FIG. 1. *The model considered in* §4. *Three rigid rods of lengths* $1, \mu$ *and* 1 *are connected by springs and subjected to a compressive load* λ.

Equilibrium configurations are determined by the requirement that they realize stationary values of the potential energy

$$\tag{4.1} \tfrac{1}{2}(\theta_2-\theta_1)^2+\tfrac{1}{2}(\theta_3-\theta_2)^2+\lambda(\cos\theta_1+\mu\cos\theta_2+\cos\theta_3)$$

subject to the constraint

$$\tag{4.2} \sin\theta_1+\mu\sin\theta_2+\sin\theta_3=0.$$

The first two terms of (4.1) are the energy stored in the springs and the remaining terms correspond to work done by the external load λ. Constraint (4.2) guarantees that the end-points lie on the line $\theta_i=0$.

The equilibrium equations are

$$\tag{4.3} \begin{gathered} \theta_1-\theta_2+\nu\cos\theta_1-\lambda\sin\theta_1=0, \\ -\theta_1+2\theta_2-\theta_3+\nu\mu\cos\theta_2-\lambda\mu\sin\theta_2=0, \\ -\theta_2+\theta_3+\nu\cos\theta_3-\lambda\sin\theta_3=0, \end{gathered}$$

where ν is a Lagrange multiplier. Addition of the three equations (4.3) implies that

$$\nu(\cos\theta_1+\mu\cos\theta_2+\cos\theta_3)=0.$$

Therefore, if the distance between the end-points

$$l \equiv \cos\theta_1 + \mu\cos\theta_2 + \cos\theta_3 \tag{4.4}$$

is nonzero, then the problem is statically determinate, with $\nu = 0$.

We consider "stability" properties of solutions with the form

$$\theta_1 = -\theta_3 = \Theta, \quad \theta_2 = 0, \quad \nu = 0, \tag{4.5}$$

where Θ satisfies

$$\Theta - \lambda\sin\Theta = 0, \quad \text{and} \quad 0 < \Theta < \pi. \tag{4.6}$$

Such a solution exists for each $\lambda > 1$. Observe that on this branch of equilibria l decreases as λ increases. We adopt the following conventions: an extremal is said to be *stable* if condition (3.6) is satisfied, to be *neutrally stable* if condition (3.6) holds only if equality is allowed, and to be *unstable* otherwise.

As anticipated, it can be shown for this particular problem that condition (3.6) is sufficient as well as necessary for a solution of (4.3) to be a strict constrained local minimum. The salient special features of this example are firstly that the space $\mathcal{H}(=\mathbb{R}^3)$ is finite-dimensional, and secondly that (4.1) and (4.2) both have continuous second-order partial derivatives with respect to θ_1, θ_2 and θ_3. Theorems derived, for example, by Hestenes (1975, Chapter 3) then imply the desired sufficiency of (3.6). Consequently, Corollary 1 can be applied to determine precisely when constrained extremals are stable.

In this example, condition (3.6) becomes

$$\mathbf{h}^T L\mathbf{h} = \mathbf{h}^T \begin{bmatrix} (1-\lambda\cos\Theta) & -1 & 0 \\ -1 & (2-\lambda\mu) & -1 \\ 0 & -1 & (1-\lambda\cos\Theta) \end{bmatrix} \mathbf{h} > 0 \tag{4.7}$$

for all nonzero $\mathbf{h} \in \mathbb{R}^3$ such that

$$\mathbf{h}\cdot[\cos\Theta, \mu, \cos\Theta] = 0. \tag{4.8}$$

Here the matrix L is obtained by linearization of (4.3) about the particular solution under consideration. Similarly, (4.8) is the linearization of constraint (4.2). The operator L obviously satisfies H1–H3.

The number of negative eigenvalues of L is easily ascertained from the characteristic polynomial. Details are omitted as the calculation is straightforward once the following two observations are made. Because Θ and λ are related by (4.6) it can be shown (i) that

$$(1-\lambda\cos\Theta) > 0,$$

and (ii) that $\mu^* \in \mathbb{R}$ can be defined by

$$\mu^* = \sup_{0 \le \Theta < \pi} \frac{-2\cos\Theta}{(1-\lambda\cos\Theta)}, \quad \text{whence } \mu^* \simeq 0.48. \tag{4.9}$$

The conclusions finally reached are as follows. If $\mu > \mu^*$, then, for all λ, L has one negative eigenvalue and is nonsingular; that is, in the notation of §2,

$$\sigma^- = 1 \quad \text{and} \quad \ker L = \{\mathbf{0}\}. \tag{4.10}$$

Whereas, if $\mu<\mu^*$, there exist λ_1 and λ_2 such that, for $\lambda_1<\lambda<\lambda_2$,

$$\sigma^- = 0 \quad \text{and} \quad \ker L = \{\mathbf{0}\}; \tag{4.11}$$

for $\lambda=\lambda_1$ or $\lambda=\lambda_2$,

$$\sigma^- = 0 \quad \text{and} \quad \dim[\ker L] = 1; \tag{4.12}$$

and for $\lambda \notin [\lambda_1, \lambda_2]$ (4.10) holds. We remark that the parameter value λ_1 corresponds to the negative eigenvalue crossing through zero, and that λ_2 corresponds to the same eigenvalue becoming negative again. It is also easily shown that the value of l corresponding to any $\lambda \in [\lambda_1, \lambda_2]$ is negative.

The highly exceptional behaviour when $\mu=\mu^*$ deserves mention. In this case there is precisely one value of λ at which the otherwise negative eigenvalue touches zero. It can be shown by explicit construction that at this critical point there occurs multiple bifurcation from a simple eigenvalue. Standard bifurcation results are not contradicted as any transversality hypothesis fails. Although the model here described is similar to that of Bauer, Keller and Reiss (1975), the two examples exhibit intrinsically different behaviours.

Whenever (4.11) holds we have immediately that the extremal is stable. In other cases Corollary 1 is required. Once it is noticed that $\eta = -1/\lambda[1,1,1]$ is a solution of $L\eta = [\cos\Theta, \mu, \cos\Theta]$, and that $\langle \eta, L\eta \rangle = -1/\lambda$, then Lemma 6 can be invoked to obtain the conclusions of the following table:

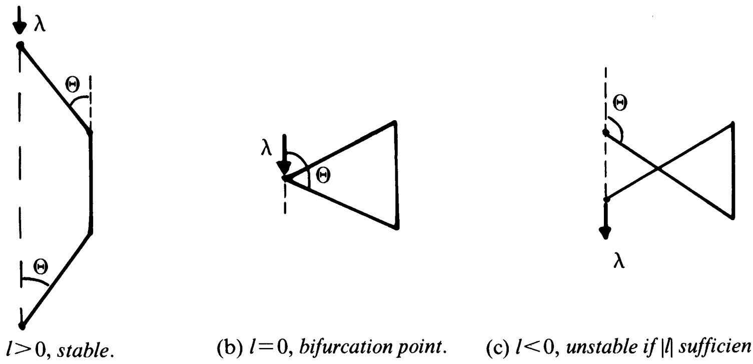
(4.13)

	$d^-[L^{-1}(\mathcal{C})]$,	$d^0[L^{-1}(\mathcal{C})]$
$l>0$	1	0 .
$l=0$	0	1
$l<0$	0	0

Consider the case $\mu>\mu^*$, so that (4.10) holds. When (4.13) is also taken into account, Corollary 1 implies that the extremal is stable when $l>0$, is neutrally stable when $l=0$, and is unstable when $l<0$. Figures 2 (a), (b) and (c) depict equilibria satisfying (4.5) and (4.6) with l positive, zero, and negative respectively. Of course, if

$$\mu>2,$$

all configurations satisfying (4.5) and (4.6) have l positive. Notice that the extremal with l zero is a bifurcation point of Type 2, and corresponds to the system becoming



(a) $l>0$, *stable.* (b) $l=0$, *bifurcation point.* (c) $l<0$, *unstable if* $|l|$ *sufficiently small.*

FIG. 2. *Various equilibria of* (4.3). *The quantity l is the distance (with sign) between the end-points of the linkage. The stability of equilibria with $l<0$ and $|l|$ large depends upon the relative lengths of the middle and end rods.*

statically indeterminate. If $\mu<\mu^*$ the conclusions are unaltered whenever $l\geq 0$. However, extremals with $l<0$ are unstable only when $\lambda\notin[\lambda_1,\lambda_2]$, which is the range on which either (4.11) or (4.12) replaces (4.10). As η exists for all $\lambda\neq 0$, the Fredholm property of L demonstrates the kernel of L to be orthogonal to the constraint, and so the extremals corresponding to $\lambda=\lambda_1$ or $\lambda=\lambda_2$ can be shown to be bifurcation points of Type 1. A partial bifurcation and stability diagram for the case $\mu<\mu^*$ is sketched in Fig. 3. The stabilities of the secondary branches and of the trivial solution are given without justification; they are easily checked.

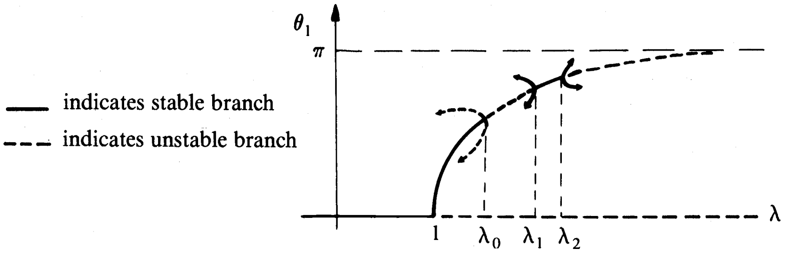


FIG. 3. *Bifurcation diagram for* (4.3), *with* $\mu<\mu^*$. *The diagram is symmetric about the* λ-*axis. Not all solutions are shown. At* λ_0, $l=0$.

The geometry of this simple example helps clarify the theory presented in § 2. Trivially, there exist three orthonormal eigenfunctions of L, and the statements made here can all be verified by expansion in terms of this basis. Assume for the moment that one eigenvalue is negative and that the other two are positive. Then there is a conical surface—as drawn in Fig. 4—defined by

$$Q(\mathbf{u})=\langle\mathbf{u},L\mathbf{u}\rangle=0. \tag{4.14}$$

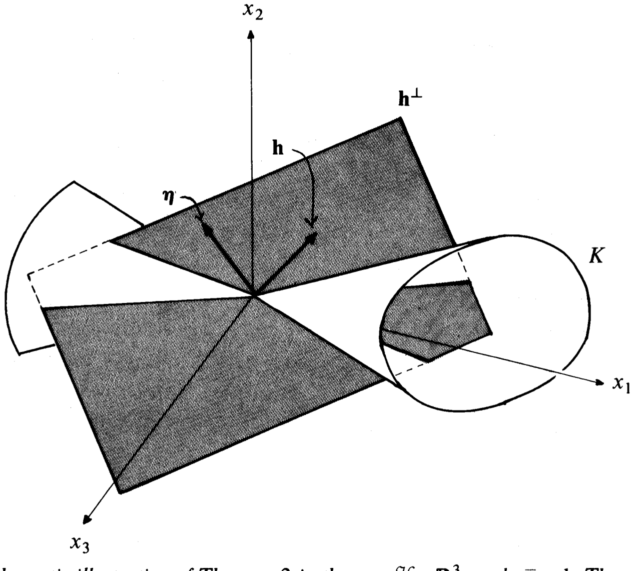


FIG. 4. *A schematic illustration of Theorem* 2 *in the case* $\mathcal{H}=\mathbb{R}^3$, *and* $\sigma^-=1$. *The axes are the eigenvectors of the self-adjoint operator* L; *the cone* K *is defined by the surface* $\langle\mathbf{x},L\mathbf{x}\rangle=0$; *the vector* $\mathbf{h}$ *is arbitrary and could lie inside the cone. Theorem* 2 *states that the plane* $\mathbf{h}^{\perp}$ *and the cone* K *intersect nontrivially if and only if the vector* $\boldsymbol{\eta}$, *such that* $L\boldsymbol{\eta}=\mathbf{h}$, *does not lie in the interior of the cone. Any line in the intersection of* $\mathbf{h}^{\perp}$ *and* K *is a maximal nonpositive subspace of* $\mathbf{h}^{\perp}$.

The closed cone K whose surface is defined by (4.14) has the following properties:

(a) The negative eigenvector is the axis of K.

(b) For any vector $\mathbf{u}$ external to K, $Q(\mathbf{u})$ is positive

(c) For any vector $\mathbf{v}$ internal to K, $Q(\mathbf{v})$ is negative, and span$\{\mathbf{v}\}$ is a maximal negative subspace of $\mathbb{R}^3$. Moreover, there is no nontrivial $\mathbf{w}$ L-orthogonal to $\mathbf{v}$ for which $Q(\mathbf{w})$ vanishes. Consequently, span$\{\mathbf{v}\}$ is also a maximal nonpositive subspace of $\mathbb{R}^3$.

(d) If $\mathbf{x}$ is a vector in the surface of K, then span$\{\mathbf{x}\}$ is a maximal nonpositive subspace of $\mathbb{R}^3$.

The introduction of one linear constraint restricts our attention to a planar subspace of $\mathbb{R}^3$, $\mathbf{h}^\perp$ say. When $\mathbf{h}^\perp$ intersects the interior of K, Q is indefinite on $\mathbf{h}^\perp$; when $\mathbf{h}^\perp$ is tangent to K, Q is nonnegative, but not positive, on $\mathbf{h}^\perp$; and when $\mathbf{h}^\perp$ does not intersect K, Q is positive on $\mathbf{h}^\perp$. However, the subspace $\mathbf{h}^\perp$ can be alternatively defined as the subspace L-orthogonal to $\boldsymbol{\eta}$, where

$$L\boldsymbol{\eta}=\mathbf{h}.$$

By property (c) above, Q is positive on $\mathbf{h}^\perp$ whenever $\boldsymbol{\eta}$ is in the interior of K. By property (d), Q is nonnegative on $\mathbf{h}^\perp$ whenever $\boldsymbol{\eta}$ is in the surface of K. And, although it is not as clear geometrically, Corollary 1 states that Q is indefinite on $\mathbf{h}^\perp$ whenever $\boldsymbol{\eta}\notin K$.

The final point of interest is the behaviour when the negative eigenvalue approaches zero. As this happens, the cone K narrows until the eigenvalue touches zero, at which time the cone degenerates to a line which forms the kernel of L.

5. An example from the calculus of variations. One of the main motivations for this work is the case commonly arising in continuum mechanics where (3.1) and (3.2) take the form of integrals and $\mathcal{H}$ is a space of functions satisfying given boundary conditions. The minimization problem is then of the classic isoperimetric type in the calculus of variations: equation (3.3) corresponds to the weak form of the Euler-Lagrange equation, L is in effect the operator arising in Jacobi's accessory equation, and condition (3.6) is the property characterizing an extremal as a minimum, namely that the second variation be positive on a set of admissible variations satisfying the linearized constraints.

The example presented in this section is another idealization of a buckled rod. We now adopt the model of a continuous inextensible line that resists bending according to a nonlinearly elastic law. The system to be considered (cf. Fig. 5) consists of a uniform rod whose end-points are constrained to lie apart by a specified distance and whose ends are clamped so that the tangents to the rod at its end-points coincide with the line between the ends. Any configuration is determined once the angle θ, defined in Fig. 5a, is specified as a function of arc-length s.

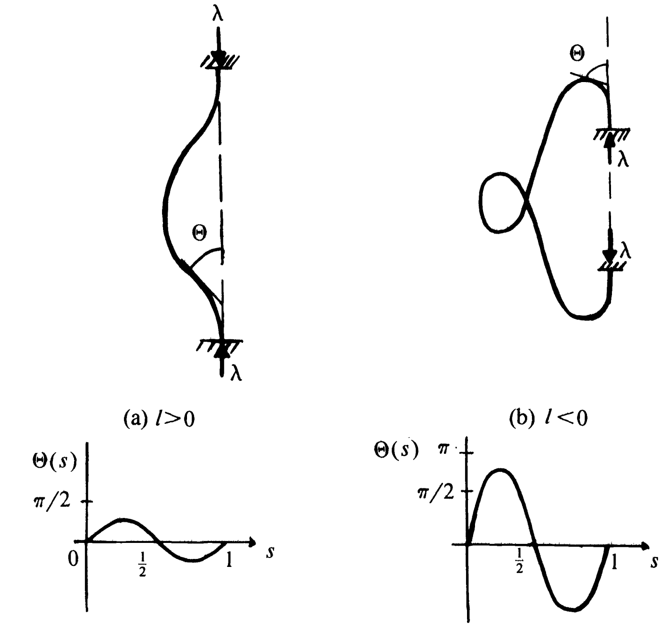
Rather than becoming embroiled in technical aspects of the calculus of variations and the theory of second-order self-adjoint boundary-value problems, we choose to make a formal presentation. A prime on a function denotes differentiation with respect to its argument.

The problem is to minimize

$$\int_0^1 \mathrm{W}(\theta')\,ds \tag{5.1}$$

subject to two constraints

$$\int_0^1 \sin\theta\,ds=0 \tag{5.2}$$



(a) $l>0$

(b) $l<0$

FIG. 5. *Buckled equilibria of a uniform elastic rod. Equilibria with $l>0$ are stable; equilibria with $l<0$ are unstable.*

and

$$\int_0^1 \cos\theta\, ds = l, \qquad |l|<1, \tag{5.3}$$

where $\theta(s)$ satisfies the boundary conditions

$$\theta(0)=\theta(1)=0. \tag{5.4}$$

The function $\mathbf{W}\colon \mathbb{R}\to\mathbb{R}$ measures the stored elastic energy per unit arc-length, and in accord with normal practice $\mathbf{W}$ is assumed to be $\mathbf{C}^2$ (i.e. twice continuously differentiable) and to satisfy

$$\mathbf{W}(x)=\mathbf{W}(-x), \quad \mathbf{W}(0)=0 \quad \text{and} \quad \mathbf{W}''(x)>0. \tag{5.5}$$

A consequence of (5.5) and the smoothness of $\mathbf{W}$, is that

$$\mathbf{W}'(0)=0.$$

We also make an additional convexity assumption, namely that

$$\mathbf{W}''(x)\geq\frac{\mathbf{W}'(x)}{x}. \tag{5.6}$$

Both (5.5) and (5.6) are satisfied in the linearly elastic case, where—after normalization—

$$\mathbf{W}(\theta')=\tfrac{1}{2}\theta'^2.$$

Constrained extremals satisfy the Euler–Lagrange equation

$$-\frac{d}{ds}\mathbf{W}'(\theta')-\lambda\sin\theta+\nu\cos\theta=0, \tag{5.7}$$

where both λ and ν are Lagrange multipliers. Physical considerations show that λ is the compressive force exerted by the boundary supports. We determine stability properties of the branch of solutions $\Theta(s)$ that satisfy: $\Theta(s-\frac{1}{2})$ is an odd function; $|\Theta|<\pi$; Θ has one interior zero; and Θ' has two interior zeros. Two such extremals are depicted in Fig. 5. Because of (5.2), (5.3) and (5.5), integration of (5.7) shows that for such a solution if $l\neq 0$, then $\nu=0$. We also assume that λ increases as l is decreased, so that

$$\frac{d}{d\lambda}l<0. \tag{5.8}$$

This assumption is physically realistic as it means that a greater compressive force is required to obtain a lesser distance between the end-points of the rod. Given constitutive hypothesis (5.6), inequality (5.8) can actually be proven (see Maddocks (1984)).

The problem can be formulated as a minimization problem over the Hilbert space $\mathcal{L}^2$ in the following manner:

A functional f: $\mathcal{L}^2\mapsto\mathbb{R}\cup\{+\infty\}$ is defined by

$$f(\theta)=\begin{cases}\int_0^1 W(\theta')\,ds & \text{if } \theta\in\mathcal{H}_0^1[0,1] \text{ and } W(\theta')\in\mathcal{L}^1(0,1),\\ +\infty & \text{otherwise.}\end{cases}$$

Here the elements of $\mathcal{H}_0^1$ are those functions in the Sobolev space $\mathcal{H}^1$ that vanish weakly at $s=0$ and $s=1$. The existence of weak solutions to the Euler–Lagrange equations for a constrained extremal of f subject to constraints (5.2) and (5.3) can be proved. It can be further shown that these solutions actually satisfy (5.7) which is the strong form of the Euler–Lagrange equations.

It is then necessary that condition (5.9) below be satisfied if an extremal θ is a strict constrained local minimum. Moreover, the theory of §2 can be applied to the quadratic form $\langle u,Lu\rangle$, where $\langle\cdot,\cdot\rangle$ is the $\mathcal{L}^2$-inner-product, and the operator L: $\mathbf{C}_0^2\mapsto\mathcal{L}^2$ is defined in (5.10) below. Trivially the space $\mathbf{C}_0^2[0,1]$ of twice continuously differentiable functions that vanish at 0 and 1 is dense in $\mathcal{L}^2[0,1]$. We remark in passing that the space $\mathbf{C}_0^2$ is sufficiently large as to make the necessary condition (5.9) below relevant in the classic calculus of variations; this is because the strict convexity of $\mathbf{W}$ implies that L has a maximal closed extension defined on the space $\mathcal{H}_0^1[0,1]\cap\mathcal{H}^2[0,1]$, and because the eigenvalues of L and its extension coincide. In the language of the calculus of variations, Legendre's strengthened condition is satisfied, and consideration of "weak" (or smooth) variations is sufficient. The criterion for stability is whether

$$\int_0^1\{\mathbf{W}''(\Theta')u'^2-\lambda\cos\Theta u^2\}\,ds>0$$

(5.9) for those $u(s)\in\mathbf{C}^1$ satisfying

$$\int_0^1\cos\Theta\,u\,ds=0 \quad\text{and}\quad \int_0^1\sin\Theta\,u\,ds=0.$$

In order to apply the theory of §2, condition (5.9) is recast in terms of the operator L, defined by

$$Lu \equiv -\frac{d}{ds}\{\mathbf{W}''(\Theta')u'\} - \lambda\cos\Theta\, u, \tag{5.10}$$

and the formal $\mathfrak{L}^2$-inner-product. Because of the boundary conditions (5.4), condition (5.9) becomes

$$\langle u, Lu\rangle > 0,$$

for any $u \in \mathbf{C}_0^2$ satisfying

$$\langle \cos\Theta, u\rangle = 0$$

and

$$\langle \sin\Theta, u\rangle = 0.$$

The operator L is next shown to be nonsingular and to have one negative eigenvalue. The result employed to reach this conclusion is:

The number of negative eigenvalues of L subject to boundary conditions (5.4), is given by the number of zeros interior to $(0,1)$ of the solution v to the one point boundary value problem

$$Lv = 0, \quad v(0) = 0, \quad v'(0) = 1. \tag{5.11}$$

This result can be obtained from standard Sturm–Liouville comparison theorems.

That L has at least one negative eigenvalue is then a consequence of the Sturm separation theorem. For $\Theta(s)$ satisfies (5.7) with $\nu = 0$, so differentiation with respect to s demonstrates that $L\Theta' = 0$; and by choice of Θ there are two interior zeros of Θ'. That L is nonsingular, and has only one negative eigenvalue is implied by contradictions arising from comparison of solutions to (5.11) having two zeros in $(0,1]$, with the solution

$$t(s) = -\lambda\sin\Theta$$

to

$$-\frac{d}{ds}\left\{\frac{\mathbf{W}'(\Theta)}{\Theta'}t'\right\} - \{\mathbf{W}(\Theta')\Theta' + \lambda\cos\Theta\}t = 0, \qquad t(0) = t(1) = 0.$$

Verification of this identity is straightforward given that Θ satisfies (5.7). The formula arises in the study of stability of planar buckled rods subjected to perturbations out of their plane of deformation (Maddocks (1984)). Comparison techniques can be applied because of assumption (5.6), and because the properties of Θ imply that t vanishes only at zeros of Θ. In the notation of §2 the conclusion is that

$$\sigma^- = 1 \quad \text{and} \quad \ker L = \{\mathbf{0}\}. \tag{5.12}$$

Lemma 6 is now invoked. Note that

$$x = \alpha\Theta' - 1/\lambda \quad \text{and} \quad y = \frac{\partial}{\partial\lambda}\Theta$$

are solutions of

$$Lx = \cos\Theta \quad \text{and} \quad Ly = \sin\Theta,$$

respectively. Here $\alpha \in \mathbb{R}$ can be chosen so that x satisfies boundary conditions (5.4), for $\Theta'(s-\frac{1}{2})$ is an even function. That y is a solution can be seen from differentiation of (5.7) with respect to λ. Boundary conditions (5.4) are satisfied by y because each member of the family of solutions $\Theta(s,\lambda)$ satisfies the boundary conditions. Furthermore,

$$\langle x, Lx \rangle = \int_0 \alpha \cos\Theta \, \Theta' \, ds - 1/\lambda \int_0^1 \cos\Theta \, ds = \alpha \sin\Theta \Big|_0^1 - l/\lambda = -l/\lambda,$$

$$\langle x, Ly \rangle = \langle Lx, y \rangle = \int_0^1 \cos\Theta \, \Theta_\lambda \, ds = \frac{d}{d\lambda} \int_0^1 \sin\Theta \, ds = 0,$$

$$\langle y, Ly \rangle = \int_0^1 \sin\Theta \, \Theta_\lambda \, ds = -\frac{d}{d\lambda} \int_0^1 \cos\Theta \, ds = -\frac{d}{d\lambda} l,$$

and by assumption (5.8)

$$-\frac{d}{d\lambda} l > 0.$$

The matrix $W$of Lemma 6 is therefore diagonal; and the conclusions shown in the table follow immediately.

	$d^-[L^{-1}(\mathcal{C})]$	$d^0[L^{-1}(\mathcal{C})]$
$l>0$	1	0
$l=0$	0	1
$l<0$	0	0

After an application of Corollary 1 it can therefore be concluded that extremals satisfying $l>0$ (such as that illustrated in Fig. 5a) are stable, that the extremal satisfying $l=0$ is a candidate bifurcation point of Type 2, and that extremals satisfying $l<0$ (such as Fig. 5b) are unstable.

These results are directly analogous with those of §4; the rod being uniform corresponds to the parameter μ being one. Were the rod not uniform we could not obtain the instability result, for Θ' would no longer be a solution of the accessory equation. There could therefore be a region corresponding to l negative in which the accessory equation has no negative eigenvalue. Further study of stability in rod problems has been made by Maddocks (1984).

6. Conclusion. The theory developed in §2 comprehensively describes the properties of restricted quadratic forms. The key concepts are maximal negative and maximal nonnegative subspaces; Theorem 2 and Corollary 1 are stated in terms of dimensions of these subspaces, and it is these results that allow new work in applications. One interesting feature revealed in the treatment is that a restricted nonnegative quadratic form can fail to be positive on its subspace because of an element $u \notin \ker L$ with the property that u is orthogonal to Lu. This behaviour cannot occur with an unrestricted quadratic form of the same type. The connexions between the theory of §2, and the results of M. R. Hestenes are described in the introduction.

We remark that whenever the Hilbert space is finite dimensional the operator L has, by necessity, a finite number, σ^+ say, of positive eigenvalues. A maximal positive subspace of dimension d^+ can then be defined, and various results can be added to the theory of §2. For example, if $\mathbb{S}$ is any subspace, then

$$\{d^-[\mathbb{S}^\perp]+d^0[\mathbb{S}^\perp]+d^+[\mathbb{S}^\perp]\}$$
$$+\{d^-[L^{-1}(\mathbb{S})]+d^0[L^{-1}(\mathbb{S})]+d^+[L^{-1}(\mathbb{S})]\}=\sigma^-+\sigma^+.$$

Connexions between the theory of §2 and necessary conditions for bifurcation in constrained variational problems were described in the account of §3. Such connexions are strong; two types of bifurcation can be distinguished, and, according to Corollary 1, bifurcation of either type from a stable extremal necessitates loss of stability. However, it is not immediately clear that a loss of stability implies the presence of bifurcation. General arguments demonstrate that the implication is true; nevertheless an equivalent result, namely that the principle of exchange of stability applies in constrained problems, is not proven here.

The main application of the theory of §2 is probably the use of Lemma 6 and Corollary 1 to determine stability in constrained variational problems. The results here presented combine well with a numerical treatment. The method described in §2 requires only an estimate for the number of eigenvalues, some knowledge of the kernel, and calculation of solutions to a—usually small—number of linear nonhomogeneous problems of the form

$$L\eta_i=T_i, \qquad i=1,\cdots,n. \tag{6.1}$$

Here the T_i are known. Furthermore, a high degree of accuracy is easily obtained because of the following observation. The n-vector $[\eta_i]$ whose elements are the solutions to (6.1) is an extremal of the matrix-valued variational principle

$$V([u_i])=\{-\langle u_i,Lu_j\rangle+\langle u_i,T_j\rangle+\langle u_j,T_i\rangle\}; \tag{6.2}$$

moreover

$$V([\eta_i])=\{\langle\eta_i,T_j\rangle\}=W.$$

The only information required is contained in the matrix W, and if the η_i are calculated to first-order accuracy, then W can be found to second-order accurary by use of (6.2). A direct numerical attack would entail computation of all negative eigenvalues and their eigenfunctions; the example of §4 could be used to demonstrate the saving in labour achieved by our invoking Lemma 6 and Corollary 1.

Although the finite-dimensional example of §4 is included primarily for illustrative purposes, it is also of interest in itself, for the behaviour of the finite-dimensional model supports conjectures about continuous rods. The secondary bifurcation and associated restabilization that occurs, appears to be typical of planar buckling of nonuniform rods, but such an explicit analysis is not possible in the continuous case.

The results of §5 are also of interest in the context of the stability of rods. Similar systems have been discussed by Maddocks (1984), but in that work the rod is subject to one constraint and a prescribed load. The system of §5 arises when the prescribed force is replaced by a prescribed displacement. It transpires that for the example considered, the stabilities in the two versions coincide. This result is of practical importance; in engineering terms dead and hard loadings are equivalent.

As shown in §5 the theory described provides a genuine analytical tool in the isoperimetric calculus of variations. As previously mentioned, there is a prior theory, described by Bolza (1904, Chapter 6), that determines whether a constrained second-variation is positive; but verification of the conditions entailed is much more complex than the test presented in §2. For example, Born (1906) attempted to apply the results of Bolza to determine stability in a simplified version of the example treated in §5, but was unable to reach any conclusion.

Acknowledgments. I wish to thank Dr. A. D. Jepson and Professor J. B. Keller for their help during the preparation of this paper. I am also indebted to Professors S. S. Antman, J. F. G. Auchmuty and T. B. Benjamin for various comments and advice. In addition, I am grateful to a referee who brought certain references to my attention.

REFERENCES

S. S. ANTMAN AND G. ROSENFELD (1978), *Global behavior of buckled states of nonlinearly elastic rods*, SIAM Rev., 20, pp. 513–566.

L. BAUER, H. B. KELLER AND E. L. REISS (1975), *Multiple eigenvalues lead to secondary bifurcation*, SIAM Rev., 17, pp. 101–122.

R. BELLMAN (1970), *Introduction to Matrix Analysis*, 2nd ed., McGraw-Hill, New York.

G. D. BIRKHOFF AND M. R. HESTENES (1935), *Natural isoperimetric conditions in the calculus of variations*, Duke Math. J., 1, pp. 198 ff.

J. BOGNÁR (1974), *Indefinite Inner Product Spaces*, Springer-Verlag, New York.

O. BOLZA (1904), *Lectures on the Calculus of Variations*, Dover, New York.

M. BORN (1906), *Untersuchungen über die Stabilität der elastischen Linie in Ebene und Raum, unter verscheidenen Grenzbedingungen*, Dietrich University, Buchdruckerei, Göttingen.

R. W. COTTLE (1974), *Manifestations of the Schur complement*, Linear Alg. and Appl., 8, pp. 189–211.

M. G. CRANDALL AND P. H. RABINOWITZ (1970), *Nonlinear Sturm-Liouville eigenvalue problems and topological degree*, J. Math. Mech., 19, pp. 1083–1102.

J. GREGORY (1980), *Quadratic Form Theory and Differential Equations*, Academic Press, New York.

M. R. HESTENES (1951), *Applications of the theory of quadratic forms in Hilbert space in the calculus of variations*, Pacific J. Math., 1, pp. 525–581.

______, (1966), *Calculus of Variations and Optimal Control Theory*, John Wiley, New York.

______, (1975), *Optimization Theory, The Finite Dimensional Case*, John Wiley, New York.

T. KATO (1976), *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, New York.

J. H. MADDOCKS (1984), *Stability of nonlinearly elastic rods*, Arch. Rat. Mech. Anal., 85, pp. 311-354.

M. MORSE (1971a), *Subordinate quadratic forms and their complementary forms*, Rev. Roum. Math. Pures. et Appl., XVI, pp. 559–569.

______ (1971b), *Subordinate quadratic forms and their complementary forms*, Proc. Nat. Acad. Sci. USA, 68, p. 579.

______ (1973), *Variational Analysis, Critical Extremals and Sturmian Extensions*, John Wiley, New York.

P. H. RABINOWITZ (1976), *A survey of bifurcation theory*, in Dynamical Systems, Vol. 1, Cesari, Hale & La Salle, eds., Academic Press, New York.

F. UHLIG (1979) *A recurring theorem about pairs of quadratic-forms and extensions*, Linear Algebra and Appl., 25, pp. 219–227.

A. WEINSTEIN AND W. STENGER (1972), *Intermediate Problems for Eigenvalues*, Academic Press, New York.

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

005

STABILITY AND ASYMPTOTIC ESTIMATES IN NONAUTONOMOUS LINEAR DIFFERENTIAL SYSTEMS*

GUSTAF SÖDERLIND[†] AND ROBERT M. M. MATTHEIJ[‡]

Abstract. A new theory is presented, in which a generalized kinematic similarity transformation is used to diagonalize linear differential systems. No matrices of Jordan form are needed. The relation to Lyapunov's classical stability theory is explored, and asymptotic estimates of fundamental solutions are given. Finally, some possible numerical applications of the presented theory are suggested.

1. Introduction. In this paper, we consider linear systems of ordinary differential equations

$$\dot{x}=A(t)x+g(t), \qquad x\in\mathbb{R}^m \tag{1.1}$$

to model the propagation of perturbations in a general nonlinear system

$$\dot{y}=f(t,y), \tag{1.2}$$

where $y\in\mathbb{R}^m$ and f: $\mathbb{R}\times\mathbb{R}^m\to\mathbb{R}^m$. If $z\in\mathbb{R}^m$ satisfies the perturbed equation

$$\dot{z}=f(t,z)+g(t), \tag{1.2$'$}$$

we note that the difference $x=z-y$ satisfies (1.1), where $A(t)$ is the "average Jacobian"

$$A(t)=\int_0^1 J(t,y+\theta x)\,d\theta.$$

Here the $m\times m$ matrix $J(\cdot,\cdot)$ is the partial derivative of f with respect to its second argument. Although a linearization is not necessary in order to establish (1.1), the matrix $A(t)$ depends, by construction, not only on t but also upon x and y. This limits the validity of (1.1) as a model for the error propagation in (1.2), since $A(t)$ may not be uniformly bounded with respect to x. However, with the additional requirement that f satisfies the Lipschitz condition

$$\|f(t,z)-f(t,y)\|\le L\|z-y\| \quad \forall t,y,z$$

one easily shows that

$$\|A(t)\|\le L \quad \forall t. \tag{1.3}$$

We note that the Lipschitz condition can be relaxed; it is sufficient that the condition holds over a convex domain $D\subset\mathbb{R}^m$, i.e. whenever y, $z\in D$.

Under mild conditions, the homogeneous problem $\dot{x}=A(t)x$ has a continuously differentiable fundamental solution matrix Φ, i.e.

$$\dot{\Phi}=A(t)\Phi. \tag{1.4}$$

* Received by the editors May 31, 1983.

[†] Department of Numerical Analysis, Royal Institute of Technology, S-100 44 Stockholm, Sweden.

[‡] Mathematisch Instituut, Katholieke Universiteit, 6525 ED Nijmegen, the Netherlands.

Using this operator, the solution of (1.1) can, in terms of some given initial condition $x(0)$, be written

$$x(t)=\Phi(t)\Phi^{-1}(0)x(0)+\Phi(t)\int_0^t\Phi^{-1}(\tau)g(\tau)\,d\tau. \tag{1.5}$$

We remark that if Φ is a fundamental matrix over the semi-infinite interval $[0,\infty)$, then Φ^{-1} exists on any finite subinterval of $[0,\infty)$.

The object of the paper is to estimate the solution x given by (1.5). In particular, we are interested in asymptotic estimates and stability, i.e. we want to find estimates of $\|x(t)\|$ as well as of $\|\Phi(t)\Phi^{-1}(0)\|$. We will derive these estimates for a monotonic but otherwise unspecified norm. In particular cases we will consider the Hölder norms.

The estimates for global error propagation that we obtain are similar to corresponding results derived by using the logarithmic norm, [6], [8] and [20]. Although the latter estimates are sharp for "short-range" error propagation, our estimates are generally better for large t. Thus, they can be viewed as a complement to the traditional logarithmic norm bounds on the error.

In §2, basic concepts will be introduced and classical results reviewed. In §3 we consider various choices of a fundamental solution. The fundamental solution will then be decomposed into a normalized direction matrix and a size matrix which satisfies a differential equation kinematically similar to (1.1) [11]. We also prove a new diagonalization theorem, demonstrating that any matrix can be brought to diagonal form using a (time-dependent) transformation of Lyapunov type. This result is of fundamental importance since it allows a unified treatment of all linear systems, whether A be constant, defective or time-dependent. It is particularly useful in the latter case, when a Jordan form no longer has a clear meaning. It should be noted that the techniques presented here are of equal importance to initial value problems and boundary value problems.

In §4 we derive the asymptotic error estimates for IVP's by considering the Lyapunov transformation and its adjoint equation. Finally, in §5 we consider some applications of the presented theory.

2. Differential inequalities and logarithmic norms. "Classical" estimates of the solution to (1.1) are obtained from the differential inequality[1]

$$\frac{d}{dt}\|x\|\leq\|A(t)\|\,\|x\|+\|g\|.$$

Due to the Lipschitz constant $\|A\|$ being positive, these estimates are in practice useless, since they fail to provide information about the actual growth or decay rate in (1.1). The situation was greatly improved by the introduction of logarithmic norms [8], [6], [20]. In terms of the logarithmic norm of the matrix A, defined by

$$\mu[A]=\lim_{h\to0+}\frac{\|I+hA\|-1}{h}, \tag{2.1}$$

solutions to (1.1) can be estimated from

$$\frac{d}{dt}\|x\|\leq\mu[A(t)]\,\|x\|+\|g\|. \tag{2.2}$$

More precisely, we can state the following lemma [20].

[1] Since $\|x\|$ may be only piecewise differentiable, the derivative of $\|x\|$ is to be interpreted as a right-hand derivative.

LEMMA 1. *Let $x(t)$ be a solution of $\dot{x}=A(t)x+g(t)$. Then $\|x(t)\|\leq\xi(t)$, where the scalar function ξ satisfies the differential equation*

$$\dot{\xi}=\mu[A(t)]\xi+\|g(t)\| \tag{2.2'}$$

with the initial condition $\xi(0)=\|x(0)\|$.

While the Lipschitz constant is always positive, the logarithmic norm may be negative. This implies that sufficient conditions for classical stability notions can easily be expressed in terms of $\mu[A]$, see e.g. [6]. Instead of going into details, we shall only summarize some useful basic properties of the logarithmic norm that can be found elsewhere in the literature (see [8] and [20]).

We define the spectral abscissa of a matrix A by

$$\alpha[A]=\max_i \operatorname{Re}(\lambda_i) \tag{2.3}$$

where $\lambda_1,\cdots,\lambda_m$ are the eigenvalues of A.

LEMMA 2. *Let A and B be square matrices. Let γ be a nonnegative real number and z be a complex number. Then*

a) $\alpha[A]\leq\mu[A]$;

b) $\mu[\gamma A]=\gamma\mu[A], \gamma\geq 0$;

c) $\mu[A+zI]=\mu[A]+\operatorname{Re}(z)$;

d) $-\|A\|\leq\mu[A]\leq\|A\|$;

e) $\mu[A+B]\leq\mu[A]+\mu[B]$.

Furthermore, if Λ is a diagonal matrix and the norm $\|\cdot\|$ is monotonic, [1], *then*

f) $\mu[\Lambda]=\alpha[\Lambda]$.

LEMMA 3. *Let A be a constant quadratic matrix. Then*

a) $\|e^{At}\|\leq e^{\mu[A]t}$;

b) $\mu[A]=\lim_{h\to 0+}\log\|e^{Ah}\|/h$.

If A depends on t, we can derive nonautonomous counterparts to the statements in Lemma 3:

LEMMA 3'. *Let Φ be a continuously differentiable fundamental solution satisfying* (1.4). *Then*

a) $\|\Phi(t)\Phi^{-1}(\tau)\|\leq\exp\int_\tau^t\mu[A(s)]ds$;

b) $\mu[A(t)]=\lim_{h\to 0+}\log\|\Phi(t+h)\Phi^{-1}(t)\|/h$.

Proof. Part a) follows immediately from (1.5) and Lemma 1. In part b), note that $\Phi(t+h)=\Phi(t)+h\dot{\Phi}(t)+o(h)=(I+hA(t))\Phi(t)+o(h)$. Hence $\Phi(t+h)\Phi^{-1}(t)=I+hA(t)+o(h)$, and, as a consequence of (2.1),

$$\lim_{h\to 0+}\frac{\|\Phi(t+h)\Phi^{-1}(t)\|-1}{h}=\mu[A(t)]. \tag{2.4}$$

We also obtain

$$\frac{d}{dt}\|\Phi(t)\Phi^{-1}(\tau)\|\Big|_{t=\tau}=\mu[A(\tau)]. \tag{2.5}$$

It follows that $\|\Phi(t+h)\Phi^{-1}(t)\|=1+h\mu[A(t)]+o(h)$ as $h\to 0+$. The result then follows by taking logarithms and letting $h\to 0+$. □

The significance of Lemma 3b) and 3'b) is that error bounds obtained by using Lemma 1 are sharp (with respect to the particular choice of norm) for short term error propagation. However, over long intervals, the logarithmic norm may sometimes give gross overestimates. We illustrate these matters by considering the nonautonomous homogeneous equation

$$\dot{x}=A(t)x, \qquad \|x(0)\|=1 \quad \text{(unit initial error)}. \tag{2.6}$$

By Lemma 1,

$$\|x(h)\| \le \xi(h)=\exp \int_0^h \mu[A(s)]\,ds.$$

Since $x(h)=\Phi(h)\Phi^{-1}(0)x(0)$, $\xi(h)$ must clearly majorize $\|\Phi(h)\Phi^{-1}(0)\|$, which is the largest possible value of $\|x(h)\|$ given that $\|x(0)\|=1$. By (2.5), the Maclaurin expansions of $\xi(h)$ and $\|\Phi(h)\Phi^{-1}(0)\|$ agree to first-order terms in h. The minimal margin is therefore

$$\xi(h)-\|\Phi(h)\Phi^{-1}(0)\|=o(h)$$

as $h \to 0+$, showing that Lemma 1 indeed yields sharp results for short-term error propagation.

Estimates of the asymptotic behavior based on the logarithmic norm give useful results only if the vector norm has been chosen with extreme care. Consider, for example, the constant coefficient system

$$\dot{x}=\begin{bmatrix} -1 & 10 \\ 0 & -2 \end{bmatrix} x \tag{2.7}$$

with the matrix exponential

$$e^{At}=\begin{bmatrix} e^{-t} & 10(e^{-t}-e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix}. \tag{2.8}$$

It is immediately clear that for any choice of norm and initial condition, the asymptotic behavior of the solution is $\|x(t)\| \sim e^{-t}$. Yet, if we choose the maximum norm, we find that $\mu_\infty[A]=9$. Thus Lemma 1 yields $\xi(t)=e^{9t}$, whereas

$$\|e^{At}\|_\infty = 11e^{-t}-10e^{-2t}. \tag{2.8$'$}$$

These bounds are illustrated in Fig. 1.

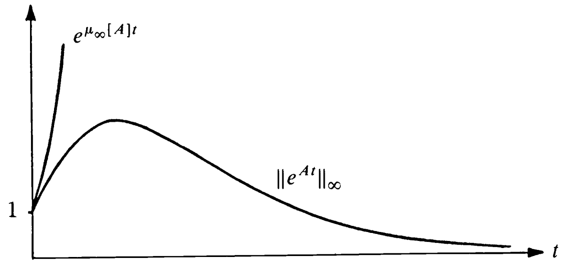


FIG. 1.

In a constant coefficient system it is a fairly straightforward task to construct a norm giving useful asymptotic estimates. Thus if $T^{-1}AT$ is diagonal, we can define

$\|x\|_T=\|T^{-1}x\|_\infty$, from which we can derive $\mu_T[A]=\alpha[A]$. We now generalize this technique to defective and time-dependent systems by using a local nonsingular coordinate transformation

$$x(t)=T(t)y(t). \tag{2.9}$$

We want to estimate the solution in terms of a given *monotonic* norm $\|\cdot\|_p$, the *global norm* of the solution. We define a *local* (time-dependent) *norm* $\|\cdot\|_T$ by

$$\|x(t)\|_T=\|T^{-1}(t)x(t)\|_p=\|y(t)\|_p. \tag{2.10}$$

We first estimate $\|x\|_T$, then these results are transformed back to estimates with respect to the fixed global norm. The following inequalities are readily established:

$$\begin{aligned}&\|x\|_p\leq\|T\|_p\|y\|_p=\|T\|_p\|x\|_T,\\&\|x\|_T\leq\|T^{-1}\|_p\|x\|_p.\end{aligned} \tag{2.11}$$

From the definition of the logarithmic norm (2.1) it follows that

$$\mu_T[A]=\mu_p[T^{-1}AT]. \tag{2.12}$$

When (2.9) is applied to (1.1), we obtain the differential equation

$$\dot{y}=(T^{-1}AT-T^{-1}\dot{T})y+T^{-1}g, \tag{2.13}$$

from which we derive the differential inequality

$$\frac{d}{dt}\|y\|_p\leq\mu_p[T^{-1}AT-T^{-1}\dot{T}]\|y\|_p+\|T^{-1}g\|_p. \tag{2.14}$$

While (2.2) still remains valid for the fixed time-independent global norm, (2.14) clearly shows that for the local norm (cf. (2.12)),

$$\frac{d}{dt}\|x\|_T\leq\mu_T[A-\dot{T}T^{-1}]\|x\|_T+\|g\|_T. \tag{2.14'}$$

Note the term $\dot{T}T^{-1}$, which accounts for the time-dependence of the local norm. Estimates of $\|x\|_T$ can now be obtained by applying Lemma 1 to (2.13), and then transformed back to estimates of $\|x\|_p$ by means of (2.11). We shall see that we can choose the coordinate transformation (2.9) in such a way that $\mu_T[A-\dot{T}T^{-1}]$ is significantly smaller than $\mu_p[A]$, thereby permitting estimates with better asymptotic properties. The price to be paid for this advantage is that we lose sharpness for short-term error propagation.

Before concluding this section, we point out that the following inequality,

$$\frac{d}{dt}\|T\|_p\leq\mu_p[T^{-1}\dot{T}]\|T\|_p=\mu_p[\dot{T}T^{-1}]\|T\|_p, \tag{2.15}$$

which follows directly from the identities $\dot{T}=TT^{-1}\dot{T}=\dot{T}T^{-1}T$, is sometimes useful in deriving the asymptotic estimates.

3. Kinematic eigenvalues and the Lyapunov transformation. Throughout the paper we shall assume that the matrix function $A(t)$ satisfies the following assumptions:

Assumption A1. $A(t)$ is uniformly bounded with respect to t, i.e. $\|A(t)\|_p\leq L$, $\forall t$.

Assumption A2. There exists a continuously differentiable fundamental solution matrix Φ satisfying $\dot{\Phi}=A\Phi$, $\forall t$.

It follows from A1 that no solution to the homogeneous problem $\dot{x}=Ax$ can grow faster than $\exp(Lt)$. Similarly, no homogeneous solution can decay faster than $\exp(-Lt)$. In order to measure the asymptotic behavior of solutions, it is therefore convenient to use the concept of characteristic exponents or type numbers, [4, p. 50], [11] and [17, p. 165].

DEFINITION. The *generalized Euclidean characteristic exponent* of a vector function $f(t)$ is defined by

$$\chi(f)=\overline{\lim_{t\to\infty}}\frac{\log\|f(t)\|_2}{t}. \tag{3.1}$$

If f and g are vector functions and γ is a scalar function of t, then it is clear from the definition that the generalized characteristic exponent satisfies the following rules:

$$\chi(f+g)\le\max(\chi(f),\chi(g)) \tag{3.2}$$

with equality if $\chi(f)\neq\chi(g)$, and

$$\chi(\gamma f)\le\chi(\gamma)+\chi(f). \tag{3.3}$$

DEFINITION. Let $\chi(f)=\chi(g)=\chi_0$. If for some nonzero constants γ_1 and γ_2 we have

$$\chi(\gamma_1 f+\gamma_2 g)<\chi_0,$$

we call f and g *exponentially linearly dependent*. Otherwise they are exponentially linearly independent.

Example. Consider the constant coefficient system

$$\dot{x}=\begin{bmatrix}0&1\\1&0\end{bmatrix}x,$$

with fundamental solutions

$$\Phi=\begin{bmatrix}e^{t}&e^{-t}\\e^{t}&-e^{-t}\end{bmatrix},\qquad\Psi=\begin{bmatrix}\cosh t&\sinh t\\\sinh t&\cosh t\end{bmatrix}.$$

Ψ serves well as a fundamental solution when t is small, but away from the origin its columns rapidly become almost linearly dependent. In fact, they are exponentially linearly dependent. The columns of Φ, on the other hand, are orthogonal for all t. We remark that given a fundamental matrix Φ, one can construct a new fundamental solution by postmultiplying Φ by any constant nonsingular matrix M. It should also be noted that since Φ^{-1} exists for all t, the columns of Φ are linearly independent in such a way that every time-dependent linear combination $\Phi(t)\gamma(t)\neq0$ if $\gamma(t)\neq0$ for all t. This "spatial" linear independence is much stronger than the "functional" linear independence $\phi(t)\gamma\not\equiv0$ for every constant vector γ.

DEFINITION. Let $\Phi=(\phi_1,\phi_2,\cdots,\phi_m)$, where ϕ_j: $\mathbb{R}\to\mathbb{R}^m$, be a fundamental matrix. Φ is said to be *normal* in the sense of Lyapunov [17, p. 169], [4, p. 52] if

$$\sum_{j=1}^{m}\chi(\phi_j)\text{ is minimal.}$$

Clearly, in the previous example, Φ is normal with $\Sigma\chi(\phi_j)=0$, whereas Ψ, with $\Sigma\chi(\psi_j)=2$ is not normal.

In the following analysis we shall be concerned with the following particular choice of Φ:

Assumption A3. Φ is normal, and its columns have been permuted so that Φ can be partitioned columnwise into

$$\Phi=\left[\Phi_1,\Phi_2,\cdots,\Phi_q\right],\qquad 1\leq q\leq m, \tag{3.4}$$

where for every j, each column in the submatrix Φ_j has the same generalized characteristic exponent χ_j. The characteristic exponents are assumed to be arranged in descending order, i.e. $\chi_j>\chi_{j+1}$.

It can be shown that a normal fundamental matrix always exists, and so Assumption A3 is always satisfied for some fundamental matrix. Unless otherwise stated, in the sequel we will deal exclusively with fundamental matrices satisfying A3.

Remark. Observe that if Φ satisfies A3, then so does $\tilde{\Phi}=\Phi M$, where M is a nonsingular block lower triangular matrix partitioned conformally with (3.4), i.e.

$$M=\begin{bmatrix} M_{11} & 0 & \cdots & 0 \\ M_{21} & M_{22} & & \\ \vdots & & \ddots & \vdots \\ & & & 0 \\ M_{q1} & & \cdots & M_{qq}\end{bmatrix},\qquad \det M_{jj}\neq 0. \tag{3.5}$$

Any result induced by A3 therefore remains qualitatively, but not necessarily quantitatively, the same for Φ and $\tilde{\Phi}$.

PROPOSITION 4. *Let Φ satisfy Assumption* A3. *Then the columns within each submatrix Φ_j are exponentially linearly independent.*

Proof. Suppose there is a nonzero vector γ for which $\chi(\Phi_j\gamma)<\chi_j$. Then we can construct a new fundamental matrix $\tilde{\Phi}$ by replacing an arbitrary column in Φ_j by $\Phi_j\gamma$. We then have

$$\sum_{j=1}^{m} x(\tilde{\phi}_j)<\sum_{j=1}^{m}\chi(\phi_j),$$

thus contradicting the assumption that Φ is normal. □

We shall now decompose Φ into a *direction matrix* T and a diagonal *size matrix* D,

$$\Phi=TD \tag{3.6}$$

where $T=(t_1,t_2,\cdots,t_m)$ has columns of unit Euclidean norm,

$$t_j^Tt_j=1. \tag{3.7}$$

Then, since $\phi_j=t_jd_{jj}$, we have

$$d_{jj}=\|\phi_j\|_2. \tag{3.8}$$

The following properties of T and D immediately follow from Assumption A2 and the differentiability of the Euclidean norm:

PROPOSITION 5. *T and D are nonsingular and continuously differentiable.*

Now, since $\dot{\Phi}=A\Phi$, we obtain $\dot{T}D+T\dot{D}=ATD$, or

$$\dot{T}=AT-T\dot{D}D^{-1}.$$

Denote the diagonal matrix $\dot{D}D^{-1}$ by Λ. Then

$$\dot{T}=AT-T\Lambda, \tag{3.9}$$

$$\dot{D}=\Lambda D. \tag{3.10}$$

Thus we have proved the following diagonalization theorem:

THEOREM 6. *For every matrix function $A(t)$ satisfying Assumptions* A1 *and* A2 *there exists a continuously differentiable nonsingular matrix T such that $\Lambda=T^{-1}AT-T^{-1}\dot{T}$ is diagonal. Moreover, under Assumption* A3, *a possible choice of T is given by* (3.6)–(3.8).

COROLLARY 7. *For every matrix function $A(t)$ satisfying Assumptions* A1 *and* A2 *there exists a continuously differentiable nonsingular matrix T such that the differential equation*

$$\dot{x}=A(t)x+g(t) \tag{3.11}$$

is decoupled by the coordinate transformation $x=Ty$ into a system of scalar differential equations

$$\dot{y}=\Lambda(t)y+T^{-1}(t)g(t). \tag{3.12}$$

The fundamental solutions Φ and D, associated with (3.11) *and* (3.12), *respectively, are related by the same transformation, i.e., $\Phi=TD$.*

A coordinate transformation $x=Ty$ is called a Lyapunov transformation, [10, p. 117], under the conditions that

(i) T is uniformly bounded,

(ii) $\dot{T}$ is continuous and uniformly bounded,

(iii) T^{-1} is uniformly bounded.

We shall see that the direction matrix T obtained by the decomposition (3.6) satisfies conditions (i) and (ii) but not always (iii). Thus it is well-known that a defective matrix cannot be transformed to diagonal form with a transformation T satisfying (iii) over a semi-infinite interval. However, there are only two reasons for considering condition (iii). Firstly, T may not become singular for any finite t. By Proposition 5 this cannot occur in our case. Secondly, if any uniform upper bound of T^{-1} appears in some estimate of the solution, it clearly has to be finite. However, such a bound will not be needed in our estimates. Thus (iii) is unnecessarily restrictive, and unless otherwise stated, we shall replace that condition by the weaker requirement

(iii′) T^{-1} exists on every finite interval,

which, according to Proposition 5 is always satisfied. We call the resulting transformation, satisfying (i), (ii) and (iii′), a *generalized Lyapunov transformation*. In particular, we shall refer to (3.9) as a *diagonalizing Lyapunov transformation* (boundedness of $\dot{T}$ will be established in Propositions 9 and 10). The systems (3.11) and (3.12) are said to be *kinematically similar* (cf. [11] and [4, p. 54]), and we call the diagonal elements $\lambda_i(t)$ of $\Lambda(t)$ the *kinematic eigenvalues of A with respect to T*. Let

$$S^T=T^{-1}. \tag{3.13}$$

Then S and T provide the left and right kinematic eigenvalues of A. We remark that T and Λ are *not unique* unless we specify exactly which normal fundamental matrix Φ is to be used for the construction of T. Asymptotic properties, however, are uniquely determined as we shall see in Propositions 11 and 12.

We now illustrate our results by considering the defective system

$$\dot{x}=\begin{bmatrix}-1 & 1\\ 0 & -1\end{bmatrix}x, \qquad t\geq 0, \tag{3.14}$$

which has a fundamental matrix

$$\Phi=\begin{bmatrix}e^{-t} & te^{-t}\\ 0 & e^{-t}\end{bmatrix}.$$

It is easily verified that

$$T=\begin{bmatrix}1 & \dfrac{t}{\sqrt{1+t^2}}\\ 0 & \dfrac{1}{\sqrt{1+t^2}}\end{bmatrix}, \qquad S^T=\begin{bmatrix}1 & -t\\ 0 & \sqrt{1+t^2}\end{bmatrix}, \tag{3.14$'$}$$

corresponding to

$$\Lambda=\begin{bmatrix}-1 & 0\\ 0 & -1+\dfrac{t}{1+t^2}\end{bmatrix}, \qquad D=\begin{bmatrix}e^{-t} & 0\\ 0 & \sqrt{1+t^2}e^{-t}\end{bmatrix}. \tag{3.14$''$}$$

Note that the kinematic eigenvalues are not constant despite the fact that the original system has constant coefficients. This is a consequence of the diagonalizing Lyapunov transformation being time-dependent. Also note that the fundamental matrices Φ and D both exhibit asymptotic growths e^{-t} and te^{-t}.

It is clearly seen that S^T is not uniformly bounded with respect to t. We expressly state that this is not a deficiency of the presented theory. It merely reflects the fact that for *any* choice of fundamental matrix Φ, the space spanned by its columns collapses as $t\rightarrow\infty$. This property is inherited by the kinematic eigensystem T which is aligned with the directions of the linearly independent solutions ϕ_j. Finally, we point out that there are nonautonomous systems with distinct eigenvalues that behave in a similar way. Thus, for instance, the system

$$\dot{x}=\begin{bmatrix}-1 & 1\\ 0 & -1+\dfrac{1}{1+t}\end{bmatrix}x, \qquad t\geq 0$$

has a fundamental matrix

$$\Phi=\begin{bmatrix}e^{-t} & -\dfrac{t^2}{2}e^{-t}\\ 0 & (1+t)e^{-t}\end{bmatrix}.$$

It is clear that for any choice of Φ, S^T will be $O(t)$ as $t\rightarrow\infty$. This "pseudodefective" behavior is due to the two eigenvalues of A approaching a defective pair as $t\rightarrow\infty$.

The kinematic eigenvalues and eigenvectors have a number of interesting properties:

PROPOSITION 8. *The kinematic eigenvalue λ_j is equal to the Rayleigh quotient formed by the corresponding kinematic eigenvector t_j and A, i.e.*

$$\lambda_j=t_j^TAt_j. \tag{3.15}$$

Proof. Differentiating (3.7), we find that $t_j^T\dot{t}_j=0$. By (3.9), $\dot{t}_j=At_j-t_j\lambda_j$. Thus, $0=t_j^TAt_j-\lambda_j$.

PROPOSITION 9. *All kinematic eigenvalues satisfy the inequalities*

$$-L\leq-\mu_2[-A]\leq\lambda_j\leq\mu_2[A]\leq L. \tag{3.16}$$

Proof. For all x with $x^Tx=1$ we have that $-\mu_2[-A]\leq x^TAx\leq\mu_2[A]$. The inequalities then follow from (3.15) and part d) of Lemma 2. □

PROPOSITION 10. $\dot{T}$ *is uniformly bounded with respect to t.*

Proof. The result immediately follows from equation (3.9) and the boundedness of A, T and Λ. □

PROPOSITION 11. *The characteristic exponents are preserved by the diagonalizing Lyapunov transformation, i.e. if* $\Phi=TD$, *then* $\chi(\phi_j)=\chi(d_j)$.

Proof.

$$\chi(\phi_j)=\overline{\lim_{t\to\infty}}\frac{1}{t}\log\|\phi_j\|_2=\overline{\lim_{t\to\infty}}\frac{1}{t}\log d_{jj}=\chi(d_j).$$ □

PROPOSITION 12. $\chi(\phi_j)$ *can be expressed as the "infinite average"*

$$\chi(\phi_j)=\overline{\lim_{t\to\infty}}\frac{1}{t}\int_0^t\lambda_j(s)\,ds. \tag{3.17}$$

Proof. By (3.10), $\dot{d}_j=\lambda_jd_j$. Integrating yields

$$d_j(t)=\exp\int_0^t\lambda_j(s)\,ds\,d_j(0).$$

Hence Proposition 11 gives

$$\chi(d_j)=\overline{\lim_{t\to\infty}}\frac{1}{t}\int_0^t\lambda_j(s)\,ds=\chi(\phi_j).$$ □

DEFINITION. The *kinematic spectral abscissa* of A with respect to T is defined by

$$\alpha_T[A]=\max_j\lambda_j. \tag{3.18}$$

We then have

PROPOSITION 13. *Let* $\|\cdot\|_p$ *be monotonic and let* $\|\cdot\|_T$ *be defined by* (2.10) *where T is a diagonalizing Lyapunov transformation. Then*

$$\alpha_T[A]=\mu_T[A-\dot{T}T^{-1}]=\mu_p[\Lambda].$$

Proof. From (2.12) we obtain $\mu_T[A-\dot{T}T^{-1}]=\mu_p[T^{-1}AT-T^{-1}\dot{T}]=\mu_p[\Lambda]$ by (3.9). Since $\|\cdot\|_p$ is monotonic, Lemma 2f) gives $\mu_p[\Lambda]=\alpha_T[A]$. □

It is clear that $\alpha_T[A]$ has strong implications as to the stability of (1.1). Not only is the kinematic spectral abscissa closely related to the characteristic exponents, but it appears explicitly in (2.14) and (2.14′). Thus we have uniform stability if $\alpha_T[A]\leq0$ and uniform asymptotic stability if $\alpha_T[A]\leq-\alpha<0$ for all t. We note that these results cannot be concluded from corresponding conditions for the spectral abscissa $\alpha[A]$ if the system is nonautonomous. These questions will be further discussed in §4.

An interesting consequence of Theorem 6 is

THEOREM 14 (exponential representation theorem). *Every fundamental solution admits the exponential representation*

$$\Phi(t)\Phi^{-1}(\tau)=T(t)\exp\int_\tau^t\Lambda(s)\,ds S^T(\tau) \tag{3.19}$$

whenever Assumptions A1 *and* A2 *are satisfied.*

Proof. Take M so that ΦM satisfies A3. Then

$$\Phi(t)\Phi^{-1}(\tau)=\Phi(t)MM^{-1}\Phi^{-1}(\tau)=T(t)D(t)D^{-1}(\tau)S^T(\tau).$$

Since $\dot{D}=\Lambda D$, (3.19) follows from

$$D(t)=\exp\int_\tau^t\Lambda(s)\,ds D(\tau), \tag{3.20}$$

and the representation

$$\Phi(t)=T(t)D(t). \tag{3.21}$$

□

Remark. Note that (3.19) is a generalization to the nonautonomous case of the corresponding formula in the diagonalizable constant coefficient case. Thus, if there exists a static similarity transformation that takes A to diagonal form,

$$0=AT-T\Lambda$$

where $T^{-1}=S^T$ and $A=T\Lambda S^T$, then

$$e^{A(t-\tau)}=Te^{\Lambda(t-\tau)}S^T.$$

It is clearly seen that this formula appears as a special case in Theorem 14. In the nonautonomous case, however, it is well known that

$$\Phi(t)\Phi^{-1}(\tau)=\exp\int_\tau^t A(s)\,ds$$

if and only if A commutes with its derivative, i.e. when $\dot{A}A-A\dot{A}=0$. The importance of Theorem 14 is that we indeed still have an exponential representation, even if the commutativity condition is not satisfied. It should be noted that this is made possible by the kinematic similarity transformation to diagonal form, and the Lyapunov-type relation $\Phi=TD$, where the fundamental solution D associated with the decoupled system (3.12) always has an exponential representation (3.20). We finally remark that the kinematic diagonalization is a transformation of global character; the case when A is defective locally requires no special attention and no matrices of Jordan form are needed.

We shall now turn to the question of how the matrix $T^{-1}=S^T$ behaves for increasing t. We have already seen that globally defective or pseudo-defective systems will (in general) cause an $O(t^\beta)$ growth for some power $\beta>0$, due to the inherent structure of the problem. In Theorem 14, however, we would like to avoid any exponential growth of S^T in (3.19), or any exponential linear dependence in the columns of T, so that the exponential behavior is due to Λ only.

Introduce the notation $\phi=\det\Phi$, $\tau=\det T$, $\sigma=\det S^T$ and $\delta=\det D$.

LEMMA 15 [4. p. 53]. $\Sigma\chi(\phi_j)\geq\chi(\phi)\geq-\chi(\phi^{-1})$.

Proof. Since $1=\phi\phi^{-1}$, we have $0\leq\chi(\phi)+\chi(\phi^{-1})$ from which the last inequality follows. For the first inequality, note that $\phi=\tau\delta\Rightarrow\chi(\phi)\leq\chi(\tau)+\chi(\delta)$. However, the normalization of the columns of T gives $|\tau|\leq 1 \Rightarrow \chi(\tau)\leq 0$. Since $\delta=\Pi\|\phi_j\|_2$, we find that $\chi(\delta)\leq\Sigma\chi(\phi_j)$, thus completing the proof. □

In order to show that S^T does not grow exponentially, we have to show that $\chi(\sigma)\leq 0$. Since $\sigma\tau=1$, we obtain

$$0\leq\chi(\tau)+\chi(\sigma). \tag{3.22}$$

However, $\chi(\tau)\leq 0$, and it follows that S^T does not grow exponentially if and only if $\chi(\tau)=\chi(\sigma)=0$.

DEFINITION. Under Assumption A1, the system $\dot{x}=Ax$ is said to be *regular* if there exists at least one fundamental solution Φ satisfying

$$\sum\chi(\phi_j)=-\chi(\phi^{-1}). \tag{3.23}$$

It is clear that a fundamental solution satisfying (3.23) must be normal, and, without loss of generality, we may assume that it has the form described in Assumption A3. We now have

THEOREM 16. *Let the system $\dot{x}=Ax$ be regular and let T be a diagonalizing Lyapunov transformation with inverse S^T. Let $\tau=\det T$ and $\sigma=\det S^T$. Then $\chi(\tau)=\chi(\sigma)=0$. In other words: the columns of T are exponentially linearly independent and S^T does not grow exponentially as t increases.*

Proof. Note that $\sigma=\phi^{-1}\delta=\phi^{-1}\Pi\|\phi_j\|_2$. Hence

$$\chi(\sigma)\leq\chi(\phi^{-1})+\sum\chi(\phi_j). \tag{3.24}$$

Since the system is regular, (3.23) gives $\chi(\sigma)\leq 0$. (3.22) together with $\chi(\tau)\leq 0$ then yields $\chi(\tau)=\chi(\sigma)=0$. □

PROPOSITION 17. *The system $\dot{x}=Ax$ is regular only if*

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t \operatorname{tr}A(s)ds \tag{3.25}$$

exists.

Proof. It is well known [5, p. 67] that ϕ satisfies the differential equation $\dot{\phi}=\operatorname{tr}A(t)\phi$. Hence

$$\chi(\phi)=\overline{\lim_{t\to\infty}}\,\frac{1}{t}\int_0^t \operatorname{tr}A(s)\,ds.$$

Similarly, $\psi=\phi^{-1}$ satisfies the adjoint equation $\dot{\psi}=-\operatorname{tr}A^T(t)\psi$, from which we derive

$$\chi(\phi^{-1})=\overline{\lim_{t\to\infty}}\,\frac{1}{t}\int_0^t -\operatorname{tr}A(s)\,ds=-\underline{\lim_{t\to\infty}}\,\frac{1}{t}\int_0^t \operatorname{tr}A(s)\,ds.$$

Since a regular system has $\chi(\phi)=-\chi(\phi^{-1})$, the existence of the limit (3.25) follows. □

PROPOSITION 17′. *The system $\dot{y}=\Lambda y$ is regular if and only if*

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t \operatorname{tr}\Lambda(s)\,ds \tag{3.26}$$

exists.

Proof. The "if" part follows from the decoupled structure of the system $\dot{y}=\Lambda y$. It is clearly seen that a scalar system $\dot{d}=\lambda d$ is regular if and only if $\chi(d)=-\chi(d^{-1})$. □

THEOREM 18. *If $\dot{x}=Ax$ is regular, then so is the transformed system $\dot{y}=\Lambda y$, and*

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t \operatorname{tr}A(s)-\operatorname{tr}\Lambda(s)\,ds=0. \tag{3.27}$$

Proof. From $\delta^{-1}=\phi^{-1}\tau$ and Theorem 16 it follows that $\chi(\delta^{-1})\leq\chi(\phi^{-1})$. Hence, if A is regular,

$$\chi(\delta)\leq\sum\chi(\phi_j)=\chi(\phi)=-\chi(\phi^{-1})\leq-\chi(\delta^{-1}).$$

By Lemma 15 we must have $\chi(\delta)\geq-\chi(\delta^{-1})$, and so the regularity of the transformed system follows. Thus we have $\chi(\phi)=\chi(\delta)$, where (cf. Proposition 12) $\chi(\phi)$ and $\chi(\delta)$ are given by the limits (3.25) and (3.26) respectively. Alternatively, (3.27) may be derived from the two adjoint differential equations

$$\dot{\tau}=(\operatorname{tr}A-\operatorname{tr}\Lambda)\tau,\qquad \dot{\sigma}=(\operatorname{tr}\Lambda-\operatorname{tr}A)\sigma$$

together with $\chi(\tau)=\chi(\sigma)=0$. □

Remark. Note that $\operatorname{tr}A$ is equal to the sum of the eigenvalues of A. Thus (3.27) states that the kinematic eigenvalues of A are "close" to the eigenvalues in the infinite average.

If A is permitted to grow exponentially, it is simple to construct problems where S^T grows exponentially, see e.g. [7, p. 12]. However, under Assumption A1, we have found that regularity is a sufficient condition for S^T to grow at most at a polynomial rate. Necessary conditions are still an open question, and at present we are not aware of any system where S^T does grow exponentially. Indeed, in Lyapunov's classical example of an irregular system, [4, pp. 53–54], we actually have a uniformly bounded S^T. One should note, however, that the class of regular systems is very wide. Thus, for instance, all systems with constant or periodic coefficients fall in this class. Irregular systems have fundamental solutions containing elements with a quite odd behavior, e.g. like $\exp(t\sin\log t)$.

4. Asymptotic estimates and condition numbers. We shall derive asymptotic estimates by applying the theory of §3 to the differential inequalities in §2. We begin by giving an estimate for $\|\Phi(t)\Phi^{-1}(0)\|_p$, i.e. we consider the homogeneous problem (2.6).

LEMMA 19. *Let $\|\cdot\|_p$ and $\|\cdot\|_q$ be dual Hölder norms (i.e. $1/p+1/q=1$) and assume that A is a rank one matrix, $A=uv^T$. Then*

$$\|A\|_p=\|u\|_p\|v\|_q. \tag{4.1}$$

Proof.

$$\|A\|_p=\sup_{\|x\|_p=1}\|Ax\|_p=\sup_{\|x\|_p=1}\|uv^Tx\|_p=\|u\|_p\sup_{\|x\|_p=1}|v^Tx|.$$

By Hölder's inequality, $|v^Tx|\leq\|v\|_q\|x\|_p$ with equality for some x. Hence $\|A\|_p=\|u\|_p\|v\|_q$. □

THEOREM 20. *Let $\|\cdot\|_p$ and $\|\cdot\|_q$ be dual Hölder norms. In addition to Assumptions* A1 *and* A2, *assume that $\chi(\phi_1)>\chi(\phi_j)$ for $j\geq2$. Then, as $t\to\infty$,*

$$\|\Phi(t)\Phi^{-1}(0)\|_p\approx\|t_1(t)\|_p\|s_1^T(0)\|_q\exp\int_0^t\lambda_1(s)\,ds, \tag{4.2}$$

where t_1 and s_1^T are the first column and row, respectively, of the matrices T and S^T.

Proof. From the exponential representation in Theorem 14, we see that $\Phi(t)\Phi^{-1}(0)$ can be written as a sum of rank one matrices,

$$\Phi(t)\Phi^{-1}(0)=\sum_{j=1}^{m} t_j(t)s_j^T(0)\exp\int_0^t\lambda_j(s)\,ds. \tag{4.3}$$

If $\chi(\phi_1)>\chi(\phi_j)$ for $j\geq 2$, then by Proposition 12, the terms 2 through m will be exponentially small compared to the first term of the sum in the right-hand side of (4.3). Hence, for large t,

$$\Phi(t)\Phi^{-1}(0)\approx t_1(t)s_1^T(0)\exp\int_0^t\lambda_1(s)\,ds,$$

and the result follows by application of Lemma 19. □

Remark. The most important application of Theorem 20 is to systems satisfying

$$\lambda_1(t)>\lambda_j(t), \qquad t\geq 0, \quad 2\leq j\leq m. \tag{4.4}$$

We then have $\chi(\phi_1)>\chi(\phi_j)$ for $j\geq 2$ if and only if (4.4) holds uniformly with respect to t. It is worth noting, however, that Theorem 20 and its proof remain valid for systems satisfying the weaker requirement

$$\lim_{t\to\infty}\int_0^t\lambda_1(s)-\lambda_j(s)\,ds=+\infty, \qquad 2\leq j\leq m, \tag{4.5}$$

although the dominated terms may no longer be exponentially small. Thus (4.4) does not have to hold uniformly in t, and Theorem 20 can also be applied in the defective case. Also note that if (4.4)–(4.5) hold, then the asymptotic behavior is determined (sharply) by the kinematic spectral abscissa, $\alpha_T[A]=\lambda_1$. Finally, note that s_1^T is only evaluated at $t=0$ in the estimate (4.2), showing that a uniform upper bound of S^T is not needed.

We now illustrate Theorem 20 by returning to the problem (2.7). A can be brought to diagonal form by a static similarity transformation $0=AT-T\Lambda$, with

$$\lambda_1=-1, \quad t_1=\begin{bmatrix}1\\0\end{bmatrix}, \quad s_1^T=(1 \quad 10).$$

Thus (4.2) yields the asymptotic estimate

$$\|e^{At}\|_\infty\approx\|t_1\|_\infty\|s_1^T\|_1e^{\lambda_1 t}=11e^{-t}$$

in agreement with (2.8′). For the defective system (3.14), we obtain kinematic eigenvalues (3.14″) that (after permutation) satisfy (4.5). The kinematic spectral abscissa is $-1+t/(1+t^2)$, corresponding to the left kinematic eigenvector $(0 \ \sqrt{1+t^2})$ appearing in the second row of S^T in (3.14′). Evaluating this vector at $t=0$, we obtain (0 1). Hence, for the Euclidean norm we have

$$\|e^{At}\|_2\approx 1\cdot 1\cdot\sqrt{1+t^2}\,e^{-t}=\sqrt{1+t^2}\,e^{-t}\approx te^{-t},$$

a result which is asymptotically sharp for large t.

Next, we turn to estimates of $\|x\|_p$. If $x=Ty$ is a diagonalizing Lyapunov transformation, then by Corollary 7 we have

$$\dot{y}=\Lambda y+S^Tg. \tag{4.6}$$

$\|y\|_p$ can be estimated from (2.14) or by application of Lemma 1, i.e. $\|y\|_p \le \eta$, where

$$\dot{\eta} = \alpha_T[A]\eta + \|S^T g\|_p \tag{4.7}$$

with the initial condition $\eta(0) = \|y(0)\|_p$. Integration of (4.7) yields

$$\|y(t)\|_p \le \|y(0)\|_p \exp\int_0^t \alpha_T[A(s)]\,ds + \int_0^t \left\{\exp\int_\tau^t \alpha_T[A(s)]\,ds\right\} \|S^T(\tau)g(\tau)\|_p\,d\tau. \tag{4.8}$$

Using (2.11) to transform this estimate into an estimate for $\|x(t)\|_p$, we get

$$\|x(t)\|_p \le \|T(t)\|_p \|S^T(0)x(0)\|_p \exp\int_0^t \alpha_T[A(s)]\,ds + \|T(t)\|_p \int_0^t \left\{\exp\int_\tau^t \alpha_T[A(s)]\,ds\right\} \|S^T(\tau)g(\tau)\|_p\,d\tau \tag{4.9}$$

whereas direct application of Lemma 1 to (1.1) gives

$$\|x(t)\|_p \le \|x(0)\|_p \exp\int_0^t \mu_p[A(s)]\,ds + \int_0^t \left\{\exp\int_\tau^t \mu_p[A(s)]\,ds\right\} \|g(\tau)\|_p\,d\tau. \tag{4.9'}$$

Note that in (4.9), $S^T(\tau)g(\tau)$ is the kinematic spectral projection of $g(\tau)$ onto the local coordinate system at time τ, having the columns of $T(\tau)$ as basis vectors. By applying (4.9) to the homogeneous case, we readily establish the following (usually cruder) alternative to the result of Theorem 20,

$$\|\Phi(t)\Phi^{-1}(0)\|_p \le \|T(t)\|_p \|S^T(0)\|_p \exp\int_0^t \alpha_T[A(s)]\,ds \tag{4.10}$$

whereas (4.9′) or Lemma 3′a) yields

$$\|\Phi(t)\Phi^{-1}(0)\|_p \le \exp\int_0^t \mu_p[A(s)]\,ds. \tag{4.10'}$$

Note that the bound (4.10) holds without the special assumptions of Theorem 20 or restrictions like (4.4)–(4.5). Although for a given norm it is usually superior to (4.10′) for asymptotic purposes, it should be clear that (4.10) does not necessarily give the optimal exponential behavior unless some restrictions of the mentioned type are imposed. Thus,

$$\chi\left(\|\Phi(t)\Phi^{-1}(0)\|_p\right) \sim \max_i \chi(\phi_i) = \max_i \overline{\lim_{t\to\infty}} \frac{1}{t}\int_0^t \lambda_i(s)\,ds \le \overline{\lim_{t\to\infty}} \frac{1}{t}\int_0^t \max_i \lambda_i(s)\,ds = \overline{\lim_{t\to\infty}} \frac{1}{t}\int_0^t \alpha_T[A(s)]\,ds.$$

The advantage, however, is that $\alpha_T[A]$ is, in principle, a computable quantity; the characteristic exponents, on the other hand, are in practice virtually impossible to compute.

As for the stability properties of the homogeneous problem, we can state the following theorems. We leave the first theorem without proof since the uniform boundedness of T implies that the bounds (4.10) and (4.10′) have the same generic structure, and the corresponding results are well known in the case of the logarithmic norm, [6, p. 59].

THEOREM 21. *Let $\alpha_T[A]$ be the kinematic spectral abscissa of A with respect to the diagonalizing Lyapunov transformation T. Then the zero solution of $\dot{x}=A(t)x$ is*

a) *stable if* $\overline{\lim}_{t\to\infty}\int_0^t\alpha_T[A(s)]\,ds<\infty$;

b) *asymptotically stable if* $\lim_{t\to\infty}\int_0^t\alpha_T[A(s)]\,ds=-\infty$;

c) *uniformly stable if* $\alpha_T[A(t)]\leq 0$ *for* $t\geq 0$;

d) *uniformly asymptotically stable if* $\alpha_T[A(t)]\leq -\alpha<0$ *for* $t\geq 0$.

THEOREM 22. *Let T be a diagonalizing Lyapunov transformation, and assume that $\alpha_T[A]\leq 0$ for all $t\geq 0$. Then the quadratic form $x^T(TT^T)^{-1}x$ is a Lyapunov function for the system $\dot{x}=A(t)x$ provided that T^{-1} is uniformly bounded for $t\geq 0$.*

Proof. Note that $x=Ty$ gives $\dot{y}=\Lambda(t)y$. Define

$$V(y)=y^Ty=\|y\|_2^2.$$

By assumption, $\alpha_T[A]=\mu_2[\Lambda]\leq 0$, implying that $V(y)$ is a Lyapunov function for the y-system, i.e. $\dot{V}\leq 0$. Transforming back, $y=T^{-1}x$ now gives

$$V(y)=x^TT^{-T}T^{-1}x=x^T(TT^T)^{-1}x,$$

and the theorem is proved. □

Remark. A time-dependent function $V(t,x)$ is a Lyapunov function if $\dot{V}\leq 0$ along the solution under consideration, *and* if there are positive definite time-invariant functions $U(x)$, $W(x)$ such that $U(x)\leq V(t,x)\leq W(x)$. Therefore, we have to require that T^{-1} be uniformly bounded in *this* application.

Quantitatively, (4.10) is superior to (4.10′) for t large enough to make

$$\|T(t)\|_p\|S^T(0)\|_p\exp\int_0^t\alpha_T[A(s)]-\mu_p[A(s)]\,ds\leq 1$$

or, equivalently, when

$$\log\|T(t)\|_p\|S^T(0)\|_p\leq\int_0^t\mu_p[A(s)]-\alpha_T[A(s)]\,ds. \tag{4.11}$$

In the constant coefficient diagonalizable case, (4.11) reduces to

$$\log\kappa_p[T]\leq\big(\mu_p[A]-\alpha[A]\big)t$$

where $\kappa_p[T]$ is the condition number of the eigenvector matrix t with respect to $\|\cdot\|_p$. In §2 we saw that the logarithmic norm gives sharp estimates initially, but in this case (4.10) is preferable for $t\geq t^*$, where

$$t^*=\frac{\log\kappa_p[T]}{\mu_p[A]-\alpha[A]}.$$

The quantity in the denominator is called the logarithmic inefficiency of the norm $\|\cdot\|_p$, [20]. In the general case we may, because of the normalization of the columns of T, think of $\|s_j^T\|_2$ as a condition number of the corresponding column t_j in T. It follows that $\|S^T\|_p$ is an indication of the local conditioning of the Lyapunov transformation. *We therefore suggest that the matrix* (3.5) *be taken to minimize $\kappa_p[T]$ in a suitable way.*

The significance of this is clearly seen in the problem

$$\dot{x}=\begin{bmatrix}1 & 0\\ 0 & 1-\dfrac{1}{t+1}\end{bmatrix}x, \qquad t\geq 0,$$

which has fundamental solutions

$$\Phi=\begin{bmatrix}e^t & 0\\ 0 & \dfrac{1}{t+1}e^t\end{bmatrix}, \qquad \Psi=\begin{bmatrix}e^t & e^t\\ 0 & \dfrac{1}{t+1}e^t\end{bmatrix}.$$

Although both matrices satisfy Assumption A3, the first one gives $T=S^T=I$, whereas the latter choice yields a matrix S^T which grows like $O(t)$ as $t\to\infty$.

It is possible to derive differential inequalities where the condition number does appear explicitly. Indeed, in a closely related context, albeit with somewhat different aims, it has been proposed by Dahlquist (private communication) to consider the quantity $\xi=\|T\|_p\|y\|_p$ directly. Thus if $x=Ty$, then $\|x\|_p\leq\xi$. Upon differentiation of ξ, one obtains

$$\dot{\xi}=\|y\|_p\frac{d}{dt}\|T\|_p+\|T\|_p\frac{d}{dt}\|y\|_p.$$

The derivatives are, as usual, interpreted as right-hand derivatives. Using (2.15) for the first term and (2.14) for the second, we find

$$\dot{\xi}\leq\left(\mu_p[T^{-1}AT-T^{-1}\dot{T}]+\mu_p[T^{-1}\dot{T}]\right)\xi+\kappa_p[T]\|g\|_p \tag{4.12}$$

with the initial condition $\xi(0)=\kappa_p[T(0)]\|x(0)\|_p$. Thus, if T is a diagonalizing Lyapunov transformation,

$$\dot{\xi}\leq\left(\alpha_T[A]+\mu_p[T^{-1}\dot{T}]\right)\xi+\kappa_p[T]\|g\|_p. \tag{4.12$'$}$$

In general, the term $\mu_p[T^{-1}\dot{T}]$ will prevent us from obtaining estimates with the desired exponential behavior, and so (4.12) is best suited for transformations other than the diagonalizing Lyapunov transformation considered in this paper. Comparing (4.12′) and (4.7), we see that in both cases, a small $\kappa_p[T]$ is needed in order to avoid a too large amplification of the forcing term $g(t)$ when it is projected onto the columns of T. In the case of a diagonalizing Lyapunov transformation suitably chosen to minimize $\kappa_p[T]$, a large or growing $\kappa_p[T]$ (such as in the defective case) merely reflects an inherent "ill-conditioning" of the differential system that is inevitable. We repeat that this is not a consequence of our transformation technique; the equations (3.11) and (3.12) appearing in Corollary 7 are completely equivalent. Thus nothing can be gained by forcing S^T to be uniformly bounded at the price of transforming the system to a rather artificial time-dependent Jordan form. Instead we suggest that one interpret $\kappa_p[T]$ as a condition number indicating how well one can distinguish different homogeneous solutions asymptotically as $t\to\infty$. Formally, one may impose conditions that would define T and Λ uniquely, but at present it is not clear what additional properties the "best possible" diagonalizing Lyapunov transformation should possess. Knowing that any transformation of this type does give the optimal exponential behavior in terms of the generalized characteristic exponents, we leave this question open.

5. Applications. In this final section, we shall hint at some possible areas of application of the presented theory. First, we shall briefly discuss how the kinematic spectral abscissa can be computed.

The practical computation of $\alpha_T[A]$ is based on the matrix differential equation (3.9). Note that the structure of (3.9) admits a convenient incorporation of a shift. Thus, if

$$\tilde{A}=A+\beta I, \qquad \tilde{\Lambda}=\Lambda+\beta I, \tag{5.1}$$

then

$$\dot{T}=\tilde{A}T-T\tilde{\Lambda}. \tag{5.2}$$

In order to compute $\alpha_T[A]$, we have to compute the maximum kinematic eigenvalue and the corresponding kinematic eigenvector (λ_1 and t_1 say). By (5.2) these quantities satisfy

$$\dot{t}_1=(\tilde{A}-\tilde{\lambda}_1 I)t_1 \tag{5.3}$$

together with the normalization requirement (3.7), i.e.

$$t_1^T t_1=1. \tag{5.4}$$

An approximation to $\lambda_1=\tilde{\lambda}_1-\beta$ is obtained by discretizing (5.3), for instance by the backward Euler method, in which case one gets

$$t_{1,n+1}-t_{1,n}=h(\tilde{A}_{n+1}-\tilde{\lambda}_{1,n+1}I)t_{1,n+1}. \tag{5.5}$$

Here h is the time-step, and the subscript n indicates an approximation at time t_n. Next, we rearrange (5.5) to obtain

$$h\tilde{\lambda}_{1,n+1}t_{1,n+1}=-(I-h\tilde{A}_{n+1})t_{1,n+1}+t_{1,n}.$$

This formula is the basis for the iteration

$$h\tilde{\lambda}_{1,n+1}^{k+1}t_{1,n+1}^{k+1}=-(I-h\tilde{A}_{n+1})t_{1,n+1}^{k}+t_{1,n}. \tag{5.6}$$

In each iteration, $\tilde{\lambda}_{1,n+1}^{k+1}$ and $t_{1,n+1}^{k+1}$ are defined by imposing (5.4). Under mild assumptions, the iteration converges with an appropriate shift β. Note that (5.6) is essentially a power iteration, but with an inhomogeneous term taking the time-dependence of the kinematic eigenvector into account. Several other discretizations and iterations are also possible. This is currently being investigated and will be reported elsewhere. Finally, we point out that the mentioned technique is good only for the computation of approximations to the dominant kinematic eigenpair, i.e., we cannot compute the whole T matrix this way. However, this is not the purpose of our analysis. Moreover, if a full transformation were to be computed, one might expect some numerical difficulties. As an alternative, one may consider the possibility of using orthogonal transformation matrices. Thus, if

$$\Phi=QR \tag{5.7}$$

is the QR factorization of a fundamental matrix Φ, it is easily verified that

$$\dot{Q}=AQ-QU, \qquad \dot{R}=UR, \tag{5.8}$$

where R and U are upper triangular matrices. (5.8) can then be regarded as a kinematic Schur form of the matrix A, but note that the diagonal elements of U are, in general, not equal to our kinematic eigenvalues. We remark that the existence of the transformation (5.8) is a classical result, cf. [4, p. 54].

Because of the strong relation between $\alpha_T[A]$ and the stability of the system, one of the most important numerical applications of the presented theory is to monitor the mathematical stability of the problem when it is solved numerically. It is well known, [9] and [14], that frequently used numerical methods for solving stiff initial value problems may sometimes produce erroneous results. This is a consequence of the difference between mathematical stability on the one hand and numerical stability on the other hand. Thus most methods for stiff problems are numerically stable also in large portions of the right half-plane. As a result, the numerical method sometimes follows a (mathematicially) unstable particular solution without ever detecting this instability. However, by numerically computing $\alpha_T[A]$ along the approximate particular solution, such instabilities are easily detected. We therefore propose that $\alpha_T[A]$ be computed so as to implement a stability check in stiff codes that would increase their reliability. Also note that once $\alpha_T[A]$ is computed, it is simple to find an approximation to the integral

$$\int_0^t \alpha_T[A(s)]\,ds$$

which gives information about the global stability properties, cf. Theorem 21.

A second possible application is to estimate the global truncation error in the numerical integration. Instead of solving a variational equation of the type (3.11), we consider a kinematically similar system (3.12). If $\alpha_T[A]=\lambda_1$, then

$$\dot{y}_1=\alpha_T[A]y_1+s_1^Tg \tag{5.9}$$

defines the asymptotically dominant component in the y-system. Since

$$x=\sum_{j=1}^{m} t_j y_j,$$

we see that t_1y_1 is the global error component in the direction with the least damping in the x-system. This component can be estimated from the scalar equation

$$\dot{\eta}=\alpha_T[A]\eta+|s_1^Tg|. \tag{5.10}$$

If the global error components in the other directions can be neglected, then

$$\|x\|_p\sim\eta\|t_1\|_p$$

and, in particular, $\|x\|_2\sim\eta$. It is clear, however, that η is neither a bound nor an estimate of the global error, but rather a "global error indicator" which is not very robust. In some preliminary computations we have, nevertheless, obtained some fairly reasonable results by solving

$$\dot{\eta}=\alpha_T[A]\eta+\|s_1^T\|_q\|g\|_p \tag{5.10$'$}$$

instead of (5.10). The norms are dual Hölder norms, and (5.10′) corresponds to projecting the local error vector g entirely onto the t_1 direction. While (5.10′) still does not give more than an indication of the global error, it is more robust and only involves quantities associated with λ_1 (s_1 satisfies the adjoint of (5.3)). To obtain a global error bound, we need $\max_i\|s_i^T\|_q$ or $\|S^T\|_p$, and at present we have not been able to compute these quantities (cf. (4.7)).

We would finally like to give an example showing that the presented theory is also useful for problems not satisfying Assumption A1, i.e. when the matrix A is no longer uniformly bounded with respect to t or other parameters. The application of our technique to such problems can be justified, although the theory is considerably more complicated. Thus the limits defining the characteristic exponents may not exist, and it is also questionable whether one may consider a linear equation (1.1) as a model for the error propagation in a nonlinear equation (1.2).

We consider the following very simple turning point problem

$$\varepsilon u''+2u'+0\cdot u=0 \tag{5.11}$$

over the interval $(-\infty,\infty)$. We are especially interested in how well it is possible to distinguish the two linearly independent solutions of (5.11). In particular, we would like to compute the condition number of the transformation matrix T as $t\to-\infty$, $+\infty$ and at the turning point $t=0$. To this end, we rewrite (5.11) as the first-order system

$$\begin{bmatrix} u' \\ \sqrt{\pi\varepsilon}\,u'' \end{bmatrix}=\begin{bmatrix} 0 & 1/\sqrt{\pi\varepsilon} \\ 0 & -2t/\varepsilon \end{bmatrix}\begin{bmatrix} u \\ \sqrt{\pi\varepsilon}\,u' \end{bmatrix}. \tag{5.12}$$

Introduce

$$E(t)=e^{-t^2/\varepsilon},\qquad I(t)=\frac{1}{\sqrt{\pi\varepsilon}}\int_{-\infty}^{t}E(\tau)\,d\tau. \tag{5.13}$$

The factor $\sqrt{\pi\varepsilon}$ appears in (5.13) to normalize I so that $I(\infty)=1$. It is now easily verified that (5.12) has a fundamental solution matrix

$$\Phi=\begin{bmatrix} 1 & I \\ 0 & E \end{bmatrix}. \tag{5.14}$$

Since $E\to0$ as $t\to\infty$, we see that the T matrix associated with (5.14) has a rapidly growing inverse S^T. We therefore try to improve this behavior by considering another fundamental matrix ΦM. Indeed, for

$$M=\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

we obtain

$$\Phi M=\begin{bmatrix} 1-I & I \\ -E & E \end{bmatrix}, \tag{5.14'}$$

a fundamental solution having much better properties. Thus, when (5.14′) is decomposed into its direction matrix T and size matrix D, we get

$$T=\begin{bmatrix} \dfrac{1-I}{\sqrt{(1-I)^2+E^2}} & \dfrac{I}{\sqrt{I^2+E^2}} \\ \dfrac{-E}{\sqrt{(1-I)^2+E^2}} & \dfrac{E}{\sqrt{I^2+E^2}} \end{bmatrix} \tag{5.15}$$

and

$$D=\begin{bmatrix}\sqrt{(1-I)^2+E^2} & 0\\ 0 & \sqrt{I^2+E^2}\end{bmatrix}. \tag{5.16}$$

After some limit calculations, one obtains Table 1, valid uniformly with respect to ε.

TABLE 1

t	$-\infty$	0	∞
T	$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$	$\frac{1}{\sqrt{5}}\begin{bmatrix}1 & 1\\ -2 & 2\end{bmatrix}$	$\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}$
S^T	$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$	$\frac{\sqrt{5}}{4}\begin{bmatrix}2 & -1\\ 2 & 1\end{bmatrix}$	$\begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix}$
$\kappa_\infty[T]$	1	3	1

The size functions d_{11} and d_{22} in (5.16) are illustrated in Fig. 2.

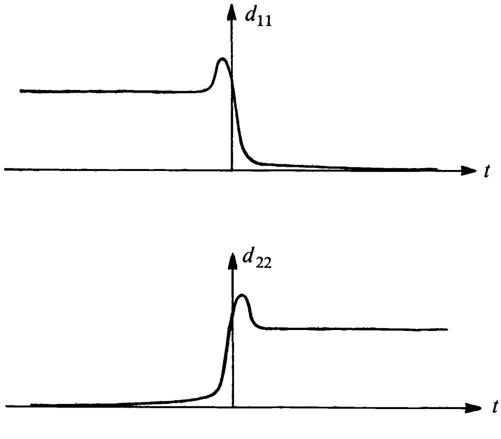


FIG. 2

In Fig. 2 we see that our turning point problem has a dichotomy, [7], in a very general sense. From the table it is clear that the direction matrix T is well behaved. At the turning point this local coordinate system rapidly flips an angle of $\pi/2$ from one orthogonal system to another. This change takes place quicker as $\varepsilon\to 0$, and consequently $\dot{T}$ is not uniformly bounded with respect to ε. The kinematic eigenvalues with respect to T are $0(\varepsilon^{-1/2})$ in a neighborhood of the turning point $t=0$.

We remark that the special scaling with respect to ε used to derive the first-order system (5.12) is necessary to obtain these good results (an alternative would be to choose a linear combination M that depends in a nonuniform way on ε). Our interpretation of this is that a proper minimization of $\kappa[T]$ implies a proper scaling with respect to the perturbation parameter.

We shall now demonstrate the importance of directional well-conditioning for boundary value problems, and apply some results from [15] to this turning point problem. Consider the BVP

$$\dot{x}=Ax+g, \qquad a\le t\le b, \tag{5.17}$$

$$M_a x(a)+M_b x(b)=c, \tag{5.18}$$

where c is a vector and M_a, M_b are square matrices which are normalized such that $\max(\|M_a\|_p, \|M_b\|_p)=1$. It was shown in [15] that the sensitivity of the solution x with respect to the boundary condition (5.18) can be quantified by the following *condition number* ($\|\cdot\|_p$ is a Hölder norm)

$$CN_p := \max\|\Phi(t)Q^{-1}\|_p, \tag{5.19}$$

where

$$Q=M_a\Phi(a)+M_b\Phi(b). \tag{5.20}$$

One should realize that CN_p is independent of the actual choice for Φ and that CN_p does not have to be greater than 1. In order to have a more workable quantity, it was suggested in [15] to use the following estimate for CN_p,

$$\gamma_p=\|Q^{-1}\|_p. \tag{5.21}$$

This estimate is meaningful only if we make the following (not restrictive) assumption. Let $\Phi(t)=T(t)D(t)$ be such that

(i) $\max_{a\le t\le b}\|d_{jj}(t)\|_p=\max_{a\le s\le b}\|d_l(s)\|_p$, $\forall j, l$;

(ii) $\max_{a\le t\le b}\|\Phi(t)\|_p=1$.

We obtain

THEOREM 23.

$$\frac{1}{\max_{a\le t\le b}\|S^T(t)\|_p}\,\frac{1}{\max_{a\le t\le b}\|T(t)\|_p}\left(\frac{1}{n}\right)^{1/p}\gamma_p\le CN_p\le\gamma_p.$$

Proof. The second inequality is an immediate consequence of our normalization assumption and the fact that $\|\Phi(t)Q^{-1}\|_p\le\|\Phi(t)\|_p\|Q^{-1}\|_p$. To show the first inequality, let z be a maximizing vector of Q^{-1}, i.e. $\|Q^{-1}z\|_p=\gamma_p\|z\|_p$. Define $y := Q^{-1}z$; then

$$\|\Phi(t)Q^{-1}z\|_p=\|T(t)D(t)y\|_p\ge glb_p(T(t))\|\phi(t)y\|_p. \tag{5.22}$$

Now we have

$$\max_t\|D(t)y\|_p\ge\max_t|d_{ii}(t)|\cdot\|y\|_\infty\ge\max_t|d_{ii}(t)|\left(\frac{1}{n}\right)^{1/p}\|y\|_\infty \tag{5.23}$$

(where i is arbitrary). Finally, from $D(t)=T^{-1}(t)\Phi(t)$, we derive

$$\max_t|d_{ii}(t)|\ge\max_t\left\{\mathrm{glb}_p(T^{-1}(t))\|\Phi(t)\|_p\right\}. \tag{5.24}$$

Substituting (5.24) into (5.23) and this into (5.22) where we now take the max over all t yields

$$CN_p\ge\left[\min_t\mathrm{glb}_p(T^{-1}(t))\right]\left[\min_t\mathrm{glb}_p(T(t))\right]\left(\frac{1}{n}\right)^{1/p}\gamma_p\frac{\|z\|_p}{\|z\|_p},$$

since $\mathrm{glb}_p(T^{-1})=1/\|T\|_p$ and $\mathrm{glb}_p(T)=1/\|T^{-1}\|_p$ the result immediately follows. □

Note that whereas CN_p is independent of the choice of Φ, γ_p is not. It appears that γ_p is a sharper estimate the less "skew" the direction matrix T is. For a useful estimate γ_p, we therefore have to choose a Φ such that the basis solutions have fairly well separated directions (if this is at all possible). The preceding turning point problem provides a nice example to demonstrate this. To this end, let $[a,b]=[-1,1]$ and assume that ε is sufficiently small to let the asymptotic behavior ($t \to \pm\infty$) be valid already for $t \to \pm 1$. As boundary condition we consider

$$(5.25) \qquad \begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}x(-1)+\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}x(1)=\begin{pmatrix}-1\\ 1\end{pmatrix}$$

(here $x=(u, \sqrt{\pi\varepsilon}\,u')^T$, cf. (5.12)).

Choosing Φ according to (5.14) results in

$$(5.26) \qquad Q\approx\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}+\begin{pmatrix}0 & 0\\ 1 & 1\end{pmatrix}=\begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix}.$$

Hence $\gamma_\infty \approx 2$. Although this bound seems small, it may not be a sharp estimate for CN_∞, for as we can see from (5.14) we have $\max_t \|S(t)\|_\infty \sim e^{1/\varepsilon}$, implying that the lower bound in Theorem 23 tends to zero as $\varepsilon \to 0$. On the other hand, if we choose ΦM as in (5.14′), we obtain

$$(5.26) \qquad Q\approx\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}+\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}.$$

Hence $\gamma_\infty \approx 1$. Moreover, it follows from (5.15) (see also Table 1) that $\max_t \|S^T(t)\|_\infty \cdot \max_t \|T(t)\|_\infty \approx 3$. Hence

$$(5.27) \qquad \frac{1}{3} \lesssim CN_\infty \lesssim 1.$$

Acknowledgments. Much of this work was carried out while the authors were visitors at the Mathematical Sciences Department of Rensselaer Polytechnic Institute, Troy, New York. We are grateful to Professor R. E. O'Malley, Jr. for providing excellent working conditions during this time. The financial support given by the Catholic University of Nijmegen, when the first author visited the second, is also gratefully acknowledged.

REFERENCES

[1] F. Bauer, J. Stoer, and C. Witzgall, *Absolute and monotonic norms*, Numer. Math., 3 (1961), pp. 257–264.

[2] R. Bellman, *Stability Theory of Differential Equations*, McGraw-Hill, New York, 1965.

[3] R. Bulirsch and J. Stoer, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[4] L. Cesari, *Asymptotic Behavior and Stability Problems in Ordinary Differential Equations*, Academic Press, New York, 1963.

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

[6] W. A. Coppel, *Stability and Asymptotic Behavior of Differential Equations*, Heath, Boston, 1965.

[7] ______, *Dichotomies in Stability Theory*, Springer-Verlag, New York, 1978.

[8] G. Dahlquist, *Stability and error bounds in the numerical integration of ordinary differential equations*, Trans. Royal Inst. of Tech., Stockholm, 1959.

[9] G. Dahlquist, L. Edsberg, G. Sköllermo and G. Söderlind, *Are the numerical methods and software satisfactory for chemical kinematics*? in Numerical Integration of Differential Equations and Large Linear Systems, J. Hinze, ed., Proceedings, Bielefeld 1980, Springer-Verlag, New York, 1982.

[10] F. R. Gantmacher, *The Theory of Matrices*, Vol. 2, Chelsea, New York, 1959.

[11] E. J. Harris and J. F. Miles, *Stability of Linear Systems: Some Aspects of Kinematic Similarity*, Academic Press, New York, 1980.

[12] B. Kågström, *Bounds and perturbations bounds for the matrix exponential*, BIT, 17 (1977), pp. 39–57.

[13] H. B. KELLER AND J. B. KELLER, *Exponential-like solution of systems of linear ordinary differential equations*, J. Soc. Ind. Appl. Math, 10 (1962), pp. 246–259.

[14] B. LINDBERG, *A dangerous property of methods for stiff differential equations*, BIT, 14 (1974), pp. 430–436.

[15] R. M. M. MATTHEIJ, *The conditioning of linear boundary value problems*, SIAM J. Numer. Anal., 19 (1982), pp. 963–978.

[16] ———, *Estimates for the errors in the solutions of linear boundary value problems, due to perturbations*, Computing, 27 (1981), pp. 299–318.

[17] V. V. NEMYTSKIJ AND V. V. STEPANOV, *Qualitative Theory of Differential Equations*, Princeton Univ. Press, Princeton, NJ, 1960.

[18] A. VAN DER SLUIS, *Estimating the solutions of slowly varying differential equations*, Preprint 152, Dept. Mathematics, Univ. Utrecht, 1980.

[19] G. SÖDERLIND, *On nonlinear difference and differential equations*, BIT, 24 (1984), to appear.

[20] T. STRÖM, *On logarithmic norms*, SIAM J. Numer. Anal., 2 (1975), pp. 741–753.

[21] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, Wiley Interscience, New York, 1965.

[22] J. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

006

SEPARATION OF VARIABLES FOR THE HAMILTON–JACOBI EQUATION ON COMPLEX PROJECTIVE SPACES*

C. P. BOYER†, E. G. KALNINS‡ AND P. WINTERNITZ§

Abstract. The additive separation of variables in the Hamilton–Jacobi equation and the multiplicative separation of variables in the Laplace–Beltrami equation are studied for the complex projective space $\mathbb{C}P^n$ considered as a Riemannian Einstein space with the standard Fubini–Study metric. The isometry group of $\mathbb{C}P^n$ is $SU(n+1)$ and its Cartan subgroup is used to generate n ignorable variables (variables not figuring in the metric tensor). A one-to-one correspondence is established between separable coordinate systems on S^n and separable systems with n ignorable variables on $\mathbb{C}P^n$. The separable coordinates in $\mathbb{C}P^n$ are characterized by $2n$ integrals of motion in involution: n of them are elements of the Cartan subalgebra of $SU(n+1)$ and the remaining n are linear combinations of the Casimir operators of $n(n+1)/2$ different $su(2)$ subalgebras of $su(n+1)$. Each system of $2n$ integrals of motion in involution, and hence each separable system of coordinates on CP^n, thus provides a completely integrable Hamiltonian system. For $n=2$ it is shown that only two separable systems on CP^2 exist, both nonorthogonal with two ignorable variables, coming from spherical and elliptic coordinates on S^2, respectively.

1. Introduction. The purpose of this paper is to discuss the problem of separation of variables for the Hamilton–Jacobi equation

$$g^{ij}S_{x^i}S_{x^j}=E \tag{1.1}$$

on a complex projective space $\mathbb{C}P^n$ with respect to the standard Fubini–Study metric. Indeed we will prove that every separable coordinate system on $\mathbb{C}P^n$ with n ignorable coordinates comes from a separable coordinate system on the real projective space $\mathbb{R}P^n$ or equivalently on the real sphere S^n. Conversely, every orthogonal separable coordinate system on S^n induces a nonorthogonal separable coordinate system on $\mathbb{C}P^n$ with n ignorable coordinates. Moreover, we prove that for $\mathbb{C}P^2$ these are all possible separable coordinates.

Also of interest is the separation of variables problem for the Laplace–Beltrami equation

$$\begin{aligned}&\Delta\psi=E\psi,\\ &\Delta=g^{-1/2}\frac{\partial}{\partial x^i}g^{1/2}g^{ij}\frac{\partial}{\partial x^j},\qquad g=\det(g_{ij}).\end{aligned} \tag{1.2}$$

Since $\mathbb{C}P^n$ (with the standard metric) is a Riemannian Einstein space, the separation of variables problem for (1.2) is equivalent [1] to that for (1.1).

Recall [1], [2] that on any pseudo-Riemannian manifold M a local coordinate system $\{x^i\}$ is said to be a *separable coordinate system* for (1.1) if it is possible to find a solution W of (1.1) depending on n-parameters $\lambda_1,\cdots,\lambda_n$ satisfying

$$W=\sum_{i=1}^{n}W_i(x^i,\lambda_1,\cdots,\lambda_n),\qquad \det\left(\frac{\partial^2 W}{\partial x^i\partial\lambda_j}\right)\neq 0 \tag{1.3}$$

* Received by the editors February 11, 1983, and in revised form July 6, 1983.

† Department of Mathematics, Harvard University, Cambridge Massachusetts 02138; on sabbatical leave from IIMAS, UNAM, Mexico D.F., Mexico. Present address: Department of Mathematics and Computer Science, Clarkson College of Technology, Potsdam, New York, 13676.

‡ Mathematics Department, University of Waikato, Hamilton, New Zealand.

§ Centre de recherche de mathématiques apliquées, Université de Montréal, Montréal, Québec, Canada H3C 3J7. The research of this author was supported in part by the Natural Sciences and Engineering Research Council of Canada and by the "Fonds FCAC pour l'aide et le soutien à la recherche".

where the function W_i is independent of x^j for $i \neq j$. This is sometimes referred to as *additive separation* as opposed to *multiplicative separation* [3]–[5] of (1.2).

There are two natural types of coordinates [1]–[5]: (1) Ignorable coordinates x^α for which $\partial_{x^\alpha} g_{ij} = 0$. (2) Essential coordinates x^a for which $\partial_{x^a} g_{ij} \neq 0$. Ignorable coordinates are naturally associated with abelian infinitesimal isometries [6]. In terms of these two types of variables additive separation can be characterized by the equations

$$(1.4) \qquad \begin{aligned} &\partial_{x^a x^b} S = 0 \quad && \text{for } a \neq b, \\ &\partial_{x^a x^\alpha} S = 0 \quad && \text{for all } a, \alpha, \\ &\partial_{x^\alpha x^\beta} S = 0 \quad && \text{for all } \alpha, \beta, \end{aligned}$$

where a, b indicate essential coordinates and α, β indicate ignorable coordinates. The pseudogroup $\mathcal{P}$ of coordinate transformations which leaves invariant this system of equations is given by

$$(1.5) \qquad x'^a = X^a(x^a), \qquad x'^\alpha = \sum_\beta A^\alpha_\beta x^\beta + \sum_a f^\alpha_a(x^a)$$

where $\det(A^\alpha_\beta) \neq 0$. Such transformations thus preserve separation; hence we say that two coordinate systems $\{x'^i\}$ and $\{x^i\}$ are *equivalent* if they can be related by a transformation in $\mathcal{P}$. By abuse of language a separable coordinate system will mean an equivalence class of separable coordinate systems.

On a Riemannian manifold every separable coordinate system $\{x^i\}$ can be brought to the canonical form [2]

$$(1.6) \qquad (g^{ij}) = \begin{bmatrix} \delta^{ab} H_a^{-2} & 0 \\ 0 & g^{\alpha\beta} \end{bmatrix}$$

by a member of $\mathcal{P}$, where δ^{ab} is the Kronecker delta and the functions H_a and $g^{\alpha\beta}$ are specified as follows: (1) The quadratic form $Q = \Sigma_a H_a^2 (dx^a)^2$ is in *Stäckel form*, i.e. H_a satisfies

$$\partial_{x^j x^k} \ln H_i^2 - \partial_{x^j} \ln H_i^2 \partial_{x^k} \ln H_i^2 + \partial_{x^j} \ln H_i^2 \partial_{x^k} \ln H_j^2 + \partial_{x^k} \ln H_i^2 \partial_{x^j} \ln H_k^2 = 0$$

for $j \neq k$. (2) Each function $g^{\alpha\beta}$ is a *Stäckel multiplier*, i.e. the quadratic form $g^{\alpha\beta} Q$ is also in Stäckel form for all α, β. The subpseudogroup $\mathcal{P}_C$ which preserves canonical forms is given by the transformations (1.5) with f^α_a constant. All our coordinate systems will be in canonical form.

Suppose G acts on M as a group of isometries with action $\phi: G \times M \to M$. Then if $\{x^i\}$ is a coordinate system about $p \in M$, then $\phi^*\{x^i\}$ is a coordinate system about $\phi^{-1}(p)$. Furthermore, if $\{x^i\}$ is a separable, so is $\phi^*\{x^i\}$. In fact $\{x^i\}$ and $\phi^*\{x^i\}$ are conjugate and we deal with conjugacy classes under G. Thus separable coordinate systems are classified up to equivalence under $\mathcal{P} \times G$.

It is classical [7] that associated with every orthogonal separable coordinate system there is an n-dimensional vector space of locally defined second order contravariant symmetric C^∞ tensor fields $A_1, \cdots, A_n$ one of which is the metric itself and which mutually commute with respect to the induced Lie bracket [3]. Recently [1,2] this has been extended to nonorthogonal separable coordinate systems where both first and second order tensor fields must be allowed. Furthermore, practical criteria have been given to determine precisely which tensor fields give rise to separation of variables [2], [8]. Up to now this is purely local; however, we will say that a separable coordinate system $\{x^i\}$ is *globally admissible* if the locally defined tensor fields $A_1,\cdots,A_n$ extend to global C^∞ tensor fields on M.

For the purpose of separation of variables it is convenient to consider the cotangent bundle $T^*(M)$ with its canonical symplectic structure [9]. Let $S^p(M)$ denote the vector space of pth-order symmetric contravariant C^∞ tensor fields on M and let $S(M)$ denote the direct sum $S(M)=\oplus_p S^p(M)$. Every tensor field in $S(M)$ defines a unique C^∞ function on $T^*(M)$ which is a polynomial in the canonical coordinates p_i associated with a coordinate system $\{x^i\}$ on M and vice-versa. The Lie bracket operation in $S(M)$ goes over to the Poisson bracket operation in $C^\infty(T^*(M))$. The contravariant metric g goes over to the Hamiltonian H_0 for geodesic motion or what we call the *free Hamiltonian*, and tensor fields which commute with H_0 become *constants of the motion*. Constants of the motion which commute under Poisson bracket are said to be in involution. Since constants of the motion will be globally defined functions in $C^\infty(T^*(M))$ if and only if the corresponding tensor fields are globally defined on M, a globally admissible separable coordinate system gives a completely integrable Hamiltonian system [9], [10].

Now let the Lie group G act on M as isometries and denote by $\mathcal{G}$ the Lie algebra of G. Let $\mathcal{U}(\mathcal{G})$ denote the universal enveloping algebra [11] of $\mathcal{G}$. $U(\mathcal{G})$ has a canonical filtration $\cdots\supset\mathcal{U}^{(2)}\supset\mathcal{U}^{(1)}\supset\mathcal{U}^0$. Denote by $\operatorname{gr}\mathcal{U}$ its associated graded algebra and by $\mathcal{U}_p=\mathcal{U}^{(p)}/\mathcal{U}^{(p-1)}$ the elements of degree p. $\mathcal{U}_p$ is naturally identified with the symmetric tensors $S^p(\mathcal{G})$. The Lie algebra homomorphism of $\mathcal{G}$ into the C^∞ vector fields on M induces a homomorphism of $\operatorname{gr}\mathcal{U}$ into $S(M)$. We will be particularly interested in $\mathcal{U}_2\to S^2(M)$. Recall [12] that a separable coordinate system $\{x^i\}$ is *class one* if the corresponding tensor fields $A_1,\cdots,A_n$ lie in the image of $\mathcal{U}_2(\mathcal{G})$, and *class two* otherwise. It follows that every class one coordinate system is globally admissible. It is known [8] that all coordinate systems on the n-sphere S^n are class one and hence globally admissible. Similarly since the isometry group $SO(n+1)$ of S^n passes to the quotient—real projective space $\mathbb{R}P^n$, all coordinate systems on $\mathbb{R}P^n$ are class one and globally admissible.

An example of a space with coordinate systems which are not globally admissible is given by the torus T_n obtained as the quotient space of $\mathbb{R}^n$ by the integer lattice. Here the tensor fields associated with, for example, spherical coordinates on $\mathbb{R}^n$, do not pass to globally defined tensor fields on T_n and thus spherical coordinates are not globally admissible on T_n. Likewise $SO(n)$ does not define global isometries on T_n.

There are two motivations for our work. First a study of separable coordinate systems on Riemannian manifolds with a large group G of isometries leads through the intimate connection [12], [13] between these coordinates and second order elements of $\mathcal{U}(\mathcal{G})$ to an algebraic understanding of special function theory. Our article constitutes a first step towards this understanding in spaces of nonconstant curvature.

Second, $\mathbb{C}P^n$ is of considerable interest in physics. For $n=2$ $\mathbb{C}P^2$ has recently been used as a model for a gravitational instanton [14], [15].

2. The geometry of complex projective spaces. Let us begin by considering the complex manifold $\mathbb{C}^{n+1}$ and the standard complex coordinates $\{\omega^\mu\}$ $\mu=1,\cdots,n+1$. On $\mathbb{C}^{n+1}$ there is a flat hermitian metric $\tilde{h}$ given in local coordinates by

$$\tilde{h}=\sum_{\mu=1}^{n+1} d\omega^\mu d\bar{\omega}^\mu. \tag{2.1}$$

The real part $\tilde{g}=\operatorname{Re}\tilde{h}$ of $\tilde{h}$ is just the standard flat Riemannian metric on $\mathbb{C}^{n+1}\sim\mathbb{R}^{2n+2}$ and the imaginary part $\operatorname{Im}\tilde{h}$ of $\tilde{h}$ is just the standard Kählerian 2-form on $\mathbb{C}^{n+1}$.

Consider the sphere S^{2n+1} in $\mathbb{C}^{n+1}$ defined by

$$\sum_{\mu} |\omega^{\mu}|^2 = 1. \tag{2.2}$$

If ω is the complex vector in $\mathbb{C}^{n+1}$ whose components are ω^{μ}, and we identify the tangent space to $\mathbb{C}^{n+1}$ at ω with $\mathbb{C}^{n+1}$ itself, then the tangent space $T_{\omega}(S^{2n+1})$ to S^{2n+1} at ω can be identified with the set $\{\xi \in T_{\omega}(\mathbb{C}^{n+1})\colon \tilde{g}(\omega,\xi)=0\}$. Moreover, the flat Riemannian metric $\tilde{g}$ on $\mathbb{R}^{2n+2}$ pulls back to the standard metric on S^{2n+1}.

Recall that the complex projective plane $\mathbb{C}P^n$ is the set of complex lines through the origin $\{0\}$ in $\mathbb{C}^{n+1}$. Two points $z, z' \in \mathbb{C}^{n+1} - \{0\}$ are equivalent if there is a $\lambda \in \mathbb{C} - \{0\}$ such that $z' = \lambda z$. Then $\mathbb{C}P^n$ is the quotient manifold $\mathbb{C}^{n+1} - \{0\} \xrightarrow{p} \mathbb{C}P^n$. Every complex line through 0 intersects the sphere S^{2n+1} in a great circle. These circles can be obtained as the orbits of the free circle group action on S^{2n+1} by

$$\omega \to e^{i\theta}\omega \tag{2.3}$$

and the space of orbits is just $\mathbb{C}P^n$. This gives the well-known Hopf fibration

$$S^1 \to S^{2n+1} \xrightarrow{\pi} \mathbb{C}P^n \tag{2.4}$$

giving S^{2n+1} as a principal bundle over $\mathbb{C}P^n$ with group $U(1) \sim S^1$.

Just as the tangent space to the sphere at a point can be determined by the condition $\tilde{g}(\omega,\xi)=0$, the tangent space to $\mathbb{C}P^n$ at a point $[\omega] \in \mathbb{C}P^n$ (here $[\omega]$ denotes the equivalence class determined by $\omega \in \mathbb{C}^{n+1} - \{0\}$) can be identified with the set

$$\{\xi \in T_{\omega}(\mathbb{C}^{n+1} - \{0\}) : \tilde{h}(\omega,\xi) = 0\} \tag{2.5}$$

Alternatively this is the set of $\xi \in T_{\omega}(S^{2n+1})$ such that

$$\tilde{g}(i\omega,\xi) = 0.$$

But $i\omega$ is a vector tangent to the great circle determined as the intersection of the projective line $[\omega]$ with S^{2n+1}. Thus the tangent space $T_{[\omega]}(\mathbb{C}P^n)$ is precisely the set of vectors tangent to S^{2n+1} and orthogonal to the great circle determined by $[\omega]$.

We can now put a metric on $\mathbb{C}P^n$ by requiring that the distance between two points on $\mathbb{C}P^n$ be measured by the corresponding distance between two great circles on S^{2n+1}. That is, for any tangent vectors $\xi_i \in T_{\omega}(S^{2n+1})$, we put

$$h(\pi_*\xi_1, \pi_*\xi_2) = \tilde{h}(\xi_1^{\perp}, \xi_2^{\perp}) \tag{2.6}$$

where $\xi^{\perp}$ is the component of ξ which is hermitian orthogonal to ω. It is easily verified that (2.6) depends only on $\pi_*\xi_i$, $i=1,2$. In local coordinates if we put

$$\omega_{n+1} = \left(1+|z|^2\right)^{-1/2}, \qquad \omega_i = z_i\left(1+|z|^2\right)^{-1/2} \tag{2.7}$$

where $|z|^2 = |z_1|^2 + \cdots + |z_n|^2$, we obtain the usual Fubini–Study metric [16] on $\mathbb{C}P^n$, viz.

$$h = \left(1+|z|^2\right)^{-2}\left[|dz|^2\left(1+|z|^2\right) - |\bar{z}\cdot dz|^2\right] \tag{2.8}$$

where we are employing standard vector notation.

Consider again the Hopf fibration (2.4). The isometry group of S^{2n+1} (with the usual metric) is the orthogonal group $O(2n+2)$. A necessary condition for an isometry

ϕ of S^{2n+1} to project to an isometry on $\mathbb{C}P^n$ is that ϕ lie in the centralizer $C(U(1))$ in $O(2n+2)$. But a straightforward computation using local coordinates ω on S^{2n+1} shows that

$$C(U(1)) \simeq U(n+1) \simeq U(1) \times SU(n+1).$$

As is well known $SU(n+1)$ is the isometry group for the metric (2.8) on $\mathbb{C}P^n$. (In fact an effective action is given only by the group $SU(n+1)/\mathbb{Z}_{n+1}$, but we will usually suppress the $\mathbb{Z}_{n+1}$ and consider $SU(n+1)$ as the isometry group of $\mathbb{C}P^n$.) We will be interested in the maximal torus $T^{n+1} \subset U(n+1) \subset O(2n+2)$. On the coordinates ω the action of T^{n+1} is given by

$$\omega^i \to e^{i\theta_i}\omega^i, \tag{2.9}$$

$i=1,\cdots,n+1$. Notice that T^{n+1} is a maximal torus both for $U(n+1)$ and $O(2n+2)$.

Let us introduce polyspherical coordinates on S^{2n+1} by writing

$$\omega^i = s^i e^{i\alpha_i}, \qquad 0 < s_i < \infty, \quad 0 < \alpha_i < 2\pi. \tag{2.10}$$

The surface defined by $\alpha_i = 0$ is an n-sphere S^n given by

$$\sum_{i=0}^{n} (s^i)^2 = 1. \tag{2.11}$$

Furthermore, we recover the whole coordinate domain by the action of T^{n+1} on S^n. Now consider the circle group action given by (2.3). Its induced action on S^n is the discrete group $\mathbb{Z}_2$. Thus the Hopf fibration of S^{2n+1} induces the fibration of S^n over the real projective space $\mathbb{Z}_2 \to S^n \to \mathbb{R}P^n$ and we get the commutative diagram

$$\begin{tikzcd} \mathbb{Z}_2 \arrow[r] \arrow[d] & S^1 \arrow[d] \\ S^n \arrow[r, "i"] \arrow[d, "\tilde{\pi}"'] & S^{2n+1} \arrow[d, "\pi"] \\ \mathbb{R}P^n \arrow[r, "\hat{i}"] & \mathbb{C}P^n \end{tikzcd} \tag{2.12}$$

This diagram is fundamental in understanding the underlying geometry. The Fubini–Study metric on $\mathbb{C}P^n$ pulls back under $(i \circ \pi)^*$ to the standard metric on S^n. However, by (2.6) h pulls back under π not to the standard metric on S^{2n+1} but to a degenerate symmetric two form on S^{2n+1} whose null space consists precisely of those tangent vectors on S^{2n+1} that are tangent to the great circles that are the orbits of S^1 in (2.12). This degenerate two form then pulls back under i to the standard metric on S^n.

3. Hamiltonian systems. In this section we discuss the relation between the free Hamiltonian on $\mathbb{C}P^n$ and a singular Hamiltonian on $\mathbb{R}P^n$ with a certain inverse square potential. This relation is an example of a general procedure in classical mechanics known as reduction of the phase space. Although this procedure is classical, it has only recently been understood in its proper context in the work of Marsden and Weinstein [17] and Kazhdan, Kostant, and Sternberg [18]. In the latter work this technique was used to obtain completely integrable Hamiltonian systems. Our interest is the classical method of reduction of the phase space by ignorable coordinates and then using separation of variables in the reduced system to give certain completely integrable Hamiltonian systems.

We briefly outline the reduction technique. For a more detailed treatment we refer to the literature [9], [17], [18]. Let P be a symplectic manifold and Ω its closed nondegenerate 2-form. Let a Lie group G act on P by symplectic diffeomorphisms, i.e. G leaves Ω invariant. Suppose further that this action is Hamiltonian, that is that for every $\xi\in\mathcal{G}$, the Lie algebra of G, the corresponding vector field $\xi^{\#}$ on P satisfies

$$(3.1) \qquad \xi^{\#}\lrcorner\,\Omega=-d\phi^{\xi}$$

for some globally defined C^{∞} functions ϕ^{ξ} on P. Now define the *moment map*: $\Phi : P\to\mathcal{G}^*$ (the dual of $\mathcal{G}$) by $\langle\Phi(x),\xi\rangle=\phi^{\xi}(x)$ where $\langle\,,\rangle$ is the pairing between $\mathcal{G}$ and $\mathcal{G}^*$. Now the map Φ is G-equivariant, i.e. Φ intertwines the G-action on P with the co-adjoint action of G on $\mathcal{G}^*$. Pick a point $u\in\mathcal{G}^*$ and assume that $\Phi^{-1}(u)$ is an embedded submanifold of P. Denote by G_u the isotropy subgroup of G under the co-adjoint action. Suppose that G_u acts freely and properly on $\Phi^{-1}(u)$ so that the quotient map $\pi_u:\Phi^{-1}(u)\to\Phi^{-1}(u)/G_u$ is a submersion. Then $P_u=\Phi^{-1}(u)/G_u$ is a symplectic manifold in a natural way. It is called the *reduced phase space*. Furthermore if H is a Hamiltonian on P which is invariant under G, then the reduced Hamiltonian H_u on P_u is obtained by

$$(3.2) \qquad \pi_u^*H_u=i_u^*H$$

where $i_u:\Phi^{-1}(u)\to P$ is the inclusion map. This brief description of reduction is that of Marsden and Weinstein [17], whereas the reduction technique of [18], is more general in that one chooses a co-adjoint orbit $\mathcal{O}$ of $\mathcal{G}^*$ rather than a point u. In the case considered below the two techniques coincide since the group G is abelian.

Let us see how reduction applies in our case. The symplectic manifold in question is $P=T^*(\mathbb{C}P^n)$, the contangent bundle of $\mathbb{C}P^n$. The symplectic 2-form is $\Omega=d\theta$ where θ is the canonical 1-form on $T^*(\mathbb{C}P^n)$, $G=T_n$ the maximal torus in $SU(n+1)$ and its action on P is Hamiltonian since P is a cotangent bundle. We can characterize $T^*(\mathbb{C}P^n)$ in homogeneous coordinates $(\omega^1,\cdots,\omega^{n+1})$ by using the Hopf fibration. Identifying $T^*_{[\omega]}(\mathbb{C}P^n)$ and $T_{[\omega]}(\mathbb{C}P^n)$ canonically by using the metric on $\mathbb{C}P^n$, $T^*(\mathbb{C}P^n)$ can be identified with set of points $([\omega],p)$ satisfying

$$(3.3) \qquad |\omega|^2=1, \quad \omega\cdot p+\bar{\omega}\cdot\bar{p}=0, \quad \omega\cdot p-\bar{\omega}\cdot\bar{p}=0.$$

The moment map Φ: $T^*(\mathbb{C}P^n)\to t_n^*$ is given in homogeneous coordinates by

$$(3.4) \qquad \Phi_i([\omega],p)=\omega_i p_i-\bar{\omega}_i\bar{p}_i \quad \text{(no sum)}$$

where $t_n^*\sim i\mathbb{R}^n$ is the dual of the Lie algebra t_n of T_n. Notice that $\Sigma_i\Phi_i=0$, so $\Phi(x)$ is in $t_n^*\subset t_{n+1}^*$ and we can check that Φ is G-equivariant. Now pick a regular point ($d\Phi$ has rank n) $u=(iu_1,\cdots,iu_{n+1})\in t_n^*$, $u_i\in\mathbb{R}$, with $\Sigma_{i=1}^{n+1}u_i=0$ and look at $\Phi^{-1}(u)$. We may choose $u_i\neq 0$ for all $i=1,\cdots,n+1$ so on $\Phi^{-1}(u)$ we have $\omega_i\neq 0$. $\Phi^{-1}(u)$ is a $3n$-dimensional manifold. The isotropy subgroup G_u of $G=T_n$ is T_n itself since $G=T_n$ is a maximal torus. Moreover, since $\omega_i\neq 0$ on $\Phi^{-1}(u)$, T_n acts freely and properly there, so $P_u=\Phi^{-1}(u)/T_n$ is a manifold. A point of P_u can be represented by

$$(3.5) \qquad \begin{aligned} &\omega_i=s_i, \quad \text{real} \quad \mathbf{s}\cdot\mathbf{s}=1, \quad \mathbf{s}\sim-\mathbf{s},\\ &p_i=y_i+iu_i \qquad (u_i \text{ fixed}),\\ &\mathbf{s}\cdot\mathbf{y}=0. \end{aligned}$$

But this is just the cotangent bundle of the real projective space $\mathbb{R}P^n$ minus the $n+1$ copies of $\mathbb{R}P^{n-1}$ obtained by putting $s_i=0$.

Consider the free Hamiltonian on $T^*(\mathbb{C}P^n)$ given in local projective coordinates $z_i=\omega_i/\omega_{n+1}$, $\omega_{n+1}\neq 0$, by

$$H=4\left(1+|z|^2\right)\left[|p|^2+|z\cdot p|^2\right]. \tag{3.6}$$

Let us write the reduced Hamiltonian H_u in terms of the coordinates (s_i,y_i) on $\mathbb{R}^{2n+2}$. From (3.5) P_u is the immersed submanifold of $\mathbb{R}^{2n+2}$ given by the set $([s],y_i)$ which satisfies

$$\mathbf{s}\cdot\mathbf{s}=1,\quad \mathbf{s}\cdot\mathbf{y}=0,\quad s_i\neq 0$$

where $[s]$ denotes the equivalence class under the equivalence relation $\mathbf{s}\sim\mathbf{s}'$ if and only if $\mathbf{s}'=\pm\mathbf{s}$. The reduced Hamiltonian is easily found to be

$$H_u=\sum_{i=1}^{n+1} y_i^2+\sum_{i=1}^{n+1}\frac{u_i^2}{s_i^2},\qquad \sum_{i=1}^{n+1}u_i=0. \tag{3.7}$$

The fact that this Hamiltonian does not extend to a regular Hamiltonian on all of $T^*(\mathbb{R}P^n)$ reflects the fact that the moment map (3.4) is singular along the surface $\omega_i=p_i=0$ as well as the fact that the coordinates (2.10) break down at $s_i=0$. These, of course, are general features of reduction by angular ignorable coordinates.

A completely analogous argument can be given to find the reduction for $T^*(S^{2n+1})$ and we will again obtain the Hamiltonian (3.7) on the reduced manifold P_u but *without* the constraint $\Sigma u_i=0$ by starting with the free Hamiltonian on S^{2n+1}.

Let us now consider the reduction corresponding to the Hopf fibration. On $T^*(S^{2n+1})$ the moment map is

$$\Phi(\omega,p)=\omega\cdot p-\bar{\omega}\cdot\bar{p}. \tag{3.8}$$

So $\Phi^{-1}(0)$ is just the set of $(\omega,p)\in\mathbb{C}^{2n}$ which satisfy (3.3). But the circle group S^1 acts freely and properly on $\Phi^{-1}(0)$. Moreover, $\Phi^{-1}(0)/S^1$ is just $T^*(\mathbb{C}P^n)$ and the free Hamiltonians on S^{2n+1} and $\mathbb{C}P^n$ are related by (3.2). This completes the our discussion of the reduction technique applied to the commutative diagram (2.12).

4. Constants of the motion. We are interested in all elements of $S^2(su(n+1))$ which commute (with respect to the induced Lie bracket) with the maximal torus. For any $x,y\in S^2(su(n+1))$ put $x\sim y$ if $x\equiv y \bmod S^2(t_n)$. Let $\bar{x}$ denote the equivalence class of x. Define $\tilde{C}=\{x\in S^2(su(n+1)):[x,h]=0,\ h\in t_n\}$. Since t_n is abelian, $[x,h]=0$ implies $[y,h]=0$ if $x\sim y$. Moreover, $S^2(t_n)\subset\tilde{C}$. Set

$$\bar{C}=\tilde{C}/S^2(t_n). \tag{4.1}$$

We wish to determine $\bar{C}$. To do so consider the root space decomposition [11] of $su(n+1)$, viz. $A_n=t_n\oplus r^+\oplus r^-$. It is more convenient to work with $u(n+1)\otimes\mathbb{C}$ and then construct the corresponding real form $su(n+1)$ by considering traceless skew-Hermitian matrices. Consider the matrices E_j^i defined by putting a 1 in the ith row and jth column, and zeros elsewhere. A basis for t_n is given by iE_j^j, $j=1,\cdots,n+1$ with the one relation $E_1^1+\cdots+E_{n+1}^{n+1}=0$. There are precisely $n(n+1)/2$ A_1 subalgebras generated by $E_i^i-E_j^j$, E_j^i, E_i^j, $1\leq i<j\leq n+1$. Let A_1^{ij} denote these subalgebras. We are interested in the real forms $su(2)_{ij}$. Let c_{ij} denote the corresponding Casimir invariant [11] of $su(2)_{ij}$ and $\bar{c}_{ij}$ its class in $\bar{C}$. Denote by C the free vector space spanned by c_{ij}, $1\leq i<j\leq n+1$. Let A_1^{ij} denote these subalgebras. We are interested in the real forms $su(2)_{ij}$. Let c_{ij} denote the corresponding Casimir invariant [11] of $su(2)_{ij}$ and $\bar{c}_{ij}$ its

class in $\overline{\mathbb{C}}$. Denote by C the free vector space spanned by c_{ij}, $1 \leq i < j \leq n+1$. Let $su(k+1)$ be the compact real form corresponding to the A_k algebra generated by E^i_j, $1 \leq i < j \leq k+1$, and let $c_2(su(k+1))$ denote its Casimir invariant. Then

$$c_2(su(k+1)) \equiv \sum_{1 \leq i < j \leq k+1} c_{ij} \mod S^2(t_n). \tag{4.2}$$

LEMMA 1. *As a vector space C has dimension $n(n+1)/2$ and the set $\{\bar{c}_{ij}\}$ is a basis for C. Thus $C \sim \overline{C}$.*

Proof. Since $c_{ij} \equiv E^i_j \odot E^j_i \mod S^2(t_n)$ where $\odot$ denotes symmetric tensor product, it is enough to show that $E^i_j \odot E^j_i$ span $\overline{C}$ and that there are no relations. The last statement is clear since as vector spaces $C \subset U_2(su(n+1))$. Now let $\overline{X} \in \overline{C}$ and choose a lifting X of the form

$$X = \sum_{\substack{j \neq k \\ l \neq m}} \alpha^{km}_{jl} E^j_k \odot E^l_m + \sum_{\substack{j \neq k \\ l}} \beta^k_{jl} E^i_k \odot E^l_l.$$

X must satisfy $[E^i_i, X] = 0$, $i = 1, \cdots, n$, and this is independent of the choice of lifting. The Lie bracket relations are given by

$$\left[E^i_j, E^k_l\right] = \delta_{jk} E^i_l - \delta_{il} E^k_j. \tag{4.3}$$

One readily sees that $\beta^k_{jl} = 0$ and that the only nonvanishing α's are α^{kj}_{jk}. This gives the desired result.

Now consider the Lie algebra homomorphism $su(n+1) \to C^\infty(T^*(\mathbb{C}P^n))$ sending $\xi \in su(n+1) \to \phi^\xi \in C^\infty(T^*(\mathbb{C}P^n))$. Lie brackets in $su(n+1)$ go over to Poisson brackets on $C^\infty(T^*(\mathbb{C}P^n))$. This induces a homomorphism of $U(su(n+1)) \to C^\infty(T^*(\mathbb{C}P^n))$ with multiplication in U going over to symmetric multiplication in $C^\infty(T^*(\mathbb{C}P^n))$. In particular, we are interested in the image of $U_2(su(n+1)) \sim S^2(su(n+1))$. Let $C^\infty_2 \subset C^\infty(T^*(\mathbb{C}P^n))$ denote the subspace consisting of homogeneous polynomials in the p's of degree 2. Denote by $S_2(su(n+1))$ the image of $S^2(su(n+1))$ in C^∞_2. We can check that this map is injective.

Denote by $\hat{c}_{ij}$ the images of c_{ij} in $S_2(su(n+1))$ and by $\hat{C}$ the image of C. On real projective space $\mathbb{R}P^n$ consider the isometry group $SO(n+1)$ and its Lie algebra $so(n+1)$. Let $\{I_{ij}\}$ with $1 \leq i < j \leq n+1$, be the basis for the Lie algebra $so(n+1)$ of functions on $T^*(\mathbb{R}P^n)$ given by

$$I_{ij} = s_i y_j - s_j y_i \tag{4.4}$$

where (s_i, y_i) are given by (3.5). Using the notation of (3.2) and (3.5), we have

LEMMA 2. *For every $\hat{A} \in \hat{C}$ there is a $A_u \in C^\infty(P_u)$ such that $i^*_u \hat{A} = \pi^*_u A_u$. Furthermore,*

$$i^*{}_u c_{ij} = \pi^*_u \left[I^2_{ij} + \left(1 + \frac{s^2_j}{s^2_i}\right) u^2_i + \left(1 + \frac{s^2_i}{s^2_j}\right) u^2_j \right] \tag{4.5}$$

where $\sum_{i=1}^{n+1} u_i = 0$.

Proof. By Lemma 1 any $\hat{A} \in \hat{C}$ is invariant under the action of the maximal torus. Thus $i^*_u A(q)$ depends only on $\pi(q)$; hence there is an $A \in C^\infty(P_u)$ such that $i^*_u A = \pi^*_u A$. To verify (4.5) let $(\omega^1, \cdots, \omega^{n+1})$ be homogeneous coordinates on $\mathbb{C}^{n+1}$. The Lie algebra $u(n+1)$ of functions on $T^*(\mathbb{C}^{n+1})$ is spanned by

$$\tilde{T}_{\mu\nu} = \omega^\mu P_\nu - \omega^\nu P_\mu + c.c, \qquad \tilde{S}_{\mu\nu} = i\left(\omega^\mu P_\nu + \omega^\nu P_\mu\right) + c.c$$

where $c.c$ denotes complex conjugate and (ω^{μ},P_{ν}) are the standard coordinates on $T^*(\mathbf{C}^{n+1})$. On $\mathbf{C}P^n$ choose projective coordinates which without loss of generality we take as $z^i=\omega^i/\omega^{n+1}$, $i=1,\cdots,n$. Then for $1\leq i<j\leq n$ fixed we get functions on $T^*(\mathbf{C}P^n)$ given by

$$T_{ij}=z^iP_j-z^jP_i+c.c.,\qquad S_{ij}=i\left(z^iP_j+z^jP_i\right)+c.c$$

$$\frac{S_{ii}-S_{jj}}{2}=i\left(z^iP_i-z^jP_j\right)+c.c.$$

These generate an $su(2)$ subalgebra for each $i\neq j=1,\cdots,n$. In terms of this basis the Casimir operator is

$$\hat{c}_{ij}=T_{ij}^2+S_{ij}^2+\left(\frac{S_{ii}-S_{jj}}{2}\right)^2 \tag{4.6}$$

Writing $z^j=(s^j/s^{n+1})e^{i\alpha_j}$, $j=1,\cdots,n$ and $(s^1)^2+\cdots+(s^{n+1})^2=1$, a short computation gives

$$z^jP_i=\frac{e^{i(\alpha_j-\alpha_i)}}{2}\left(s_jy_i-i\frac{s_j}{s_i}P_{\alpha^i}\right).$$

Restricting to $\Phi^{-1}(u)$ by setting $P_{\alpha^i}=u_i$ and by performing a straightforward computation using the formulas above gives the desired result.

Notice that the free Hamiltonian on S^n is

$$H_0=\sum_{1\leq i<j\leq n+1}I_{ij}^2$$

whereas the free Hamiltonian on $\mathbf{C}P^n$ is just

$$H=c_2(su(n+1))=\sum_{1\leq i<j\leq n+1}\hat{c}_{ij}-2(n-1)\left(\sum_{i=1}^{n}P_{\alpha_i}^2+\sum_{i<j}P_{\alpha_i}P_{\alpha_j}\right)$$

Thus performing the double sum over $1\leq i<j\leq n+1$ in the formula of Lemma 2 gives precisely (3.2) with the Hamiltonian (3.7).

Assuming the previous notation we have:

LEMMA 3. *$\hat{A}$ is a constant of the motion with respect to the Hamiltonian H if and only if A_u is a constant of the motion with respect to the reduced Hamiltonian H_u. Furthermore, two such constants of the motion $\hat{A},\hat{B}$ are in involution if and only if the corresponding pair A_u, B_u are in involution.*

Proof. This follows directly from

$$i_u^*\{A,B\}=\pi_u^*\{A_u,B_u\}.$$

5. Basic theorems about separable coordinates. It is clear that an understanding of the separable coordinate systems on $\mathbf{C}P^n$ entails an understanding of the separable coordinates on $\mathbf{R}P^n$ or equivalently S^n. A study of all separable coordinate system on S^n is currently in progress and we will here state and use a theorem whose proof will appear elsewhere.

On $T^*(S^n)$ the subset of functions spanned (over $\mathbb{R}$) by the functions (4.4) form a subalgebra of $C^\infty(T^*(S^n))$ under Poisson bracket isomorphic with $o(n+1)$. By abuse of notation we will denote this subalgebra by $o(n+1)$. Since the manifold S^n is class one [8], all second order constants of the motion are in $S^2(o(n+1))$. Let $\mathfrak{D}\subset S^2(o(n+1))$ be the subspace spanned by diagonal elements I_{ij}^2, $1\leq i<j\leq n+1$. $\mathfrak{D}$ has dimension $n(n+1)/2$. The free Hamiltonian $H=\Sigma_{i<j}I_{ij}^2\in\mathfrak{D}$ defines a nondegenerate definite bilinear form on $o(n+1)$. The point is that for separation of variables on S^n, it is enough to study $\mathfrak{D}$. On S^n we can strengthen Theorem 6 of [8]:

THEOREM 1. *Necessary and sufficient conditions for the existence of an orthogonal separable coordinate system* $\{x^i\}$ *for the H. J. equation* (1.1) *on* S^n *are that there are n functions* $A_1,\cdots,A_n\in\mathfrak{D}$, *one of which, say,* A_1, *is the free Hamiltonian H, which are*

(1) *linearly independent* (*locally*);

(2) *in involution.*

Remarks. (1) The conditions (4) and (5) of [8, Thm. 6] are automatically satisfied on S^n. (2) This theorem is not valid on the complex sphere nor on real hyperboloids.

Let us now use the reduction of §2 to formulate a theorem relating the separation of variables on $\mathbb{C}P^n$ and S^{2n+1} with respect to the free Hamiltonians with the separation on S^n with respect to the reduced Hamiltonian H_u. More precisely, the separation takes place on the open set $U\subset\mathbb{R}P^n$ defined by taking $s^i>0$. Since the Lie algebras of infinitesimal isometries on $\mathbb{C}P^n$ and S^{2n+1} are the compact forms $su(n+1)$ and $o(2n+2)$, respectively, the maximal abelian subalgebras are unique up to conjugacy and have dimensions n and $n+1$, respectively. Thus we apply the reduction technique of section 2 to arrive at

THEOREM 2. *The Hamilton-Jacobi equation* (1,1) *on* $\mathbb{C}P^n(S^{2n+1})$ *with the free Hamiltonian admits a separable coordinate system* $\{x^i,\alpha^j\}$, $i=1,\cdots,n$, $j=1,\cdots,n$ (*respectively* $n+1$) *with n* (*respectively* $n+1$) *ignorable coordinates* $\{\alpha_i\}$ *if and only if the corresponding coordinates* $\{x^i\}$ *on U separate the H. J.* (1.1) *on U with the reduced Hamiltonian* (3.7) (*with the relation* $\Sigma_{i=1}^{n+1}u_i=0$ *in the case of* CP^n).

Remarks. If the separable coordinates on U are orthogonal, then the corresponding separable coordinates are orthogonal on S^{2n+1} but never on $\mathbb{C}P^n$.

We now state our main result.

THEOREM 3. *Necessary and sufficient conditions for the existence of a separable coordinate system* $\{x^i\}$ *on* $\mathbb{C}P^n$ *with n ignorable coordinates are that there are n-functions* $\hat{A}_1,\cdots,\hat{A}_n\in\hat{C}$ *one of which, say,* $\hat{A}_1$, *is the free Hamiltonian that are*

1) *linearly independent* (*locally*);

2) *in involution.*

Furthermore, there is a bijective correspondence between orthogonal separable coordinate systems on U and separable coordinate systems on $\mathbb{C}P^n$ *with n ignorable coordinates.*

Before giving the proof of this theorem we will give some geometric background.

Let x^i be an orthogonal separable coordinate system for the free Hamiltonian H_0 on S^n. Then all constants of the motion $A_1,\cdots,A_n$ are in $\mathfrak{D}$. The condition that $A_1,\cdots,A_n$ be linearly independent (locally) means that $A_1,\cdots,A_n$ span an n-plane in $n(n+1)/2$-space. Clearly, changing the A_i's by any $GL(n,\mathbb{R})$ transformation does not alter the coordinate system. We have thus determined a point of the Grassmanian $G(n,n(n+1)/2)$ of n-planes in $n(n+1)/2$ space. Using I_{ij}^2 as a basis for $\mathfrak{D}$, we write

$$A_a=\sum_{i<j}\alpha_a^{ij}I_{ij}^2 \tag{5.1}$$

where the sums run over $1\leq i<j\leq n+1$, $a=1,\cdots,n$. We will view $\{\alpha_a^{ij}\}$ as coordinates in $\mathbb{R}^{(n^2(n+1))/2}$.

Since we are dealing with the free Hamiltonian $H_0 = c_2(o(n+1))$, we fix $A_1 = H_0$. In terms of the coordinates on $R^{n^2(n+1)/2}$, this is given by the linear equations

$$\alpha_1^{ij} = 1 \quad \text{for all } 1 \leq i < j \leq n+1. \tag{5.2a}$$

In order to determine a separable coordinate system the A_i's must be in involution under Poisson bracket. Since A_1 is just the Casimir operator of $o(n+1)$, $\{A_a, A_1\} = 0$ for all $a = 2, \cdots, n$. The remaining conditions $\{A_a, A_b\} = 0$, $a, b = 2, \cdots, n$ are equivalent to a system of first order partial differential equations which upon using (5.1) can be expressed as the quadratic equations

$$\Sigma\left(\alpha_a^{ij}\alpha_b^{ik} - \alpha_b^{ij}\alpha_a^{jk}\right) = 0. \tag{5.2b}$$

where the sum is taken over the cyclic permutations on (i,j,k), and $1 \leq i < j < k \leq n+1$.

Proof of Theorem 3. Let $\{\hat{A}_1 = \hat{H}, \hat{A}_2, \cdots, \hat{A}_n\}$ be n functions in $\hat{C}$ which satisfy 1) and 2) of the theorem. Let us write

$$\hat{A}_r = \sum_{i<j} \alpha_{(r)}^{ij} c_{ij}$$

where $\alpha_{(r)}^{ij}$ has rank n. By Lemmas 2 and 3 there are n functions $A_r(u) = i_u^* \hat{A}_r$, $r = 1, \cdots, n$ which are in involution and $A_1(u) = H_u$ for all $u \in \mathbb{R}^{n+1}$ satisfying $\Sigma_{i=1}^{n+1} u_i = 0$. Moreover, since $\alpha_{(r)}^{ij}$ has rank n, they are locally linearly independent for all u, in particular at $u = 0$. But then $A_1(0) = H_0$, the free Hamiltonian on $U \subset S^n$, and $A_r(0) \in \mathfrak{D}$ by (4.5). So by Theorem 1, there corresponds an orthogonal separable coordinate system $\{x^i\}$. But we claim that $\{x^i\}$ also separates the Hamiltonian H_u. To see this we change our point of view and consider $A_r(u)$ as functions on $T^*(S^{2n+1})$ with $n+1$ ignorable coordinates α^j, and $P_{\alpha^j}^2 = u_j^2$ and drop the traceless condition on u. Again applying Theorem 1 there is a separable coordinate system $\{x^i, \alpha^j\}$, $i = 1, \cdots, n$, $j = 1, \cdots, n+1$ on S^{2n+1}. By Theorem 2 the $\{x^i\}$ then separate H_u on U. Once more by Theorem 2 there is a separable coordinate system $\{x^i, \alpha^j\}$, $i,j = 1, \cdots, n$, on $\mathbb{C}P^n$ with n ignorable coordinates.

Conversely, given a separable coordinate system $\{x^i, \alpha^j\}$ on $\mathbb{C}P^n$ with n ignorable coordinates, the corresponding coordinates $\{x^i\}$ on U separate H_u for all $u \in \mathbb{R}^n$, and in particular they separate H_0. Thus by Theorem 1 there are n linearly independent elements A_r of $\mathfrak{D}$ which are in involution. We write

$$A_r = \sum_{i<j} \alpha_r^{ij} I_{ij}^2, \qquad \operatorname{rank} \alpha_r^{ij} = n.$$

Define $A_r(u)$ by

$$A_r(u) = \sum \alpha_r^{ij}\left(I_{ij}^2 + V_{ij}\right),$$

where

$$V_{ij} = \left(1 + \frac{s_j^2}{s_i^2}\right)u_i^2 + \left(1 + \frac{s_i^2}{s_j^2}\right)u_j^2, \qquad \sum u_i = 0.$$

If we can show that the $A_r(u)$'s are in involution, then we can use Lemmas 2 and 3 to get n linearly independent elements of $\hat{C}$ which are in involution and thus prove the theorem (including the last statement). We formulate this as a lemma.

LEMMA 4. *The set* $\{A_r(u)\}$, $r = 1, \cdots, n$ *is in involution for all* $u \in \mathbb{R}^{n+1}$ *if and only if the set* $\{A_r(0)\}$ *is in involution.*

Proof. The "only if" part is trivial. As before we consider $U \subset \mathbb{R}^{n+1}$ and use cartesian coordinates $\{s_i\}$, $i=1,\cdots,n+1$ in $\mathbb{R}^{n+1}$ and $\{s_i,y_j\}$ in $T^*(\mathbb{R}^{n+1})$. We must show that

$$\sum \alpha_r^{ij}\alpha_s^{kl}\left(\left[I_{ij}^2, V_{kl}\right]+\left[V_{ij}, I_{kl}^2\right]\right)=0.$$

Now a straightforward computation gives

$$\frac{1}{4}\left[V_{ij}, I_{kl}^2\right]=t_{ijkl}\delta_{il}+t_{jikl}\delta_{jl}+t_{ijlk}\delta_{ik}+t_{jilk}\delta_{jk}$$

where

$$t_{ijkl}=\left(\frac{s_i}{s_j^2}u_j^2-\frac{s_j^2}{s_i^3}u_i^2\right)\left(s_k^2y_i-s_ks_iy_k\right).$$

Defining $u_{ijki}=t_{ijki}-t_{ikji}$, we are reduced to showing that

$$\sum\left(\alpha_r^{ij}\alpha_s^{ki}u_{ijki}+\alpha_r^{ij}\alpha_s^{jk}u_{jikj}\right)=0. \tag{5.3}$$

But an explicit computation shows that

$$u_{jikj}=\frac{s_is_k}{s_j^2}u_j^2I_{ik}-\frac{s_is_j}{s_k^2}u_k^2I_{ij}-\frac{s_js_k}{s_i^2}u_i^2I_{jk}$$

which satisfies

$$u_{ijki}+u_{jikj}=0. \tag{5.4}$$

That $\{A_r(0)\}$ are in involution implies that $\{\alpha_r^{ij}\}$ satisfies equations (5.2). Combining equations (5.4) with (5.2) implies the equality (5.3) and proves the lemma.

We end this section with two corollaries to Theorem 3.

COROLLARY. *Every separable coordinate system on* $\mathbb{C}P^n$ *with n ignorable coordinates is class one and thus globally admissible.*

COROLLARY. *For every orthogonal separable coordinate system on the sphere* S^n *the locally defined functions* $1/s_i^2$ *are Stäckel multipliers for* $i=1,\cdots,n+1$.

Remark. As mentioned in §1, by a separable coordinate system we actually mean an equivalence class of separable systems, equivalent under $P\times G$ (where G is the isometry group). However the statement of Theorem 1 and its subsequent applications require a specific choice of representative, namely one for which $A_r\in\mathcal{D}$.

6. Explicit examples of separable coordinates.

A. General *n*. In this case we discuss two examples; the most and the least degenerate coordinate systems. The most degenerate is given by spherical coordinates on S^n, viz.

$$\begin{aligned}
&s_1=\sin\phi_1\cdots\sin\phi_{n-1}\sin\phi_n,\\
&s_2=\sin\phi_1\cdots\cos\phi_{n-1}\cos\phi_n,\\
&s_3=\sin\phi_1\cdots\cos\phi_{n-1},\\
&\vdots\\
&s_n=\sin\phi_1\cos\phi_2,\\
&s_{n+1}=\cos\phi_1.
\end{aligned} \tag{6.1}$$

The separated equations on $\mathbf{C}P^n$ are

$$\begin{aligned}
&P_{\alpha_i}=u_i, \qquad 1\le i\le n\\
&P_{\phi_n}^2+\frac{u_1^2}{\sin^2\phi_n}+\frac{u_2^2}{\cos^2\phi_n}=\lambda_n,\\
&P_{\phi_{n-1}}^2+\frac{\lambda_n}{\sin^2\phi_{n-1}}+\frac{u_3^2}{\cos^2\phi_{n-1}}=\lambda_{n-1},\\
&\vdots\\
&P_{\phi_1}^2+\frac{\lambda_2}{\sin^2\phi_1}+\frac{(\Sigma u_i)^2}{\cos^2\phi_1}=\lambda_1.
\end{aligned}\tag{6.2}$$

This corresponds to the group reduction $SU(2)\subset\cdots\subset SU(n)\subset SU(n+1)$, and the corresponding Casimir invariants (4.2) give the relevant constants of the motion;

$$\hat{c}_2(su(k))=\lambda_{n-k+2}.\tag{6.3}$$

The solutions of (5.2b) are given by

$$\alpha_{n+2-k}^{ij}=\begin{cases}1, & i\le j\le k,\\ 0, & \text{otherwise.}\end{cases}$$

The least degenerate system is given by the general Jacobi elliptic coordinates on S^n, viz.

$$s_i^2=\frac{\Pi_{j=1}^n(x^j-e^i)}{\Pi_{j\ne i}(e^j-e^i)}, \qquad 1\le i\le n+1\tag{6.4}$$

where the constants e^i satisfy $e^1<x^1<e^2<\cdots<x^n<e^{n+1}$. These are known to separate variables [7] on S^n.

B. $\boldsymbol{n=2}$. It is well known [19], [20] that there are precisely two separable coordinate systems on S^2, spherical and elliptical coordinates. Thus by Theorem 3 we get two separable coordinate systems on $\mathbf{C}P^2$. We will now show that these are all the separable coordinate systems on $\mathbf{C}P^2$. In fact we will prove a more general result relevant to the study of selfdual gravitational instantons. We will make use of the classification of canonical forms for four dimensional manifolds given in [5]. Since we are dealing with a Riemannian (positive definite) metric, the only relevant types are B, C, F and H of [5].

Before stating and proving our result we give some background. Let V_4 be a four dimensional Riemannian manifold. Due to the local isomorphism between the groups $SO(4)$ and $SU(2)\times SU(2)$, we can describe local four dimensional Riemannian geometry equally well in terms of local orthogonal or local spinorial moving frames. For example, if Ω denotes the curvature two-form on V_4, then with respect to an orthonormal moving coframe $\{\theta^a\}$ we have

$$\Omega_b^a=R^a{}_{bcd}\theta^c\wedge\theta^d, \qquad a,b,c,d=1,\cdots,4\tag{6.5}$$

whereas, with respect to a local spinor coframe $\theta^{A\dot{A}}$ with $\theta^{1\dot{1}}=\theta^1+i\theta^3$, $\theta^{1\dot{2}}=\theta^2+i\theta^4$, $\theta^{2\dot{2}}=-\overline{\theta^{1\dot{1}}}$, and $\theta^{2\dot{1}}=\overline{\theta^{1\dot{2}}}$, we have

$$\Omega^A{}_B=C^A{}_{BCD}S^{CD}+\frac{R}{12}S^A{}_B+C^A{}_{B\dot{C}\dot{D}}S^{\dot{C}\dot{D}}, \tag{6.6}$$

$$\Omega^{\dot{A}}{}_{\dot{B}}=C^{\dot{A}}{}_{\dot{B}\dot{C}\dot{D}}S^{\dot{C}\dot{D}}+\frac{R}{12}S^{\dot{A}}{}_{\dot{B}}+C^{\dot{A}}{}_{\dot{B}CD}S^{CD}$$

$A,B,\dot{A},\dot{B}=1,2$; and

$$S^{AB}=\frac{1}{2}\varepsilon_{AB}\theta^{A\dot{A}}\wedge\theta^{B\dot{B}}, \tag{6.7}$$

$$S^{\dot{A}\dot{B}}=\frac{1}{2}\varepsilon_{AB}\theta^{A\dot{A}}\wedge\theta^{B\dot{B}},$$

$1=\varepsilon_{12}=-\varepsilon_{21}$, $\varepsilon_{11}=\varepsilon_{22}=0$. We mention that $S^{\dot{A}\dot{B}}$ (S^{AB}) is self-dual (anti self-dual) with respect to the Hodge star operator $*$ on exterior differential forms. The decomposition (6.6) is convenient because it realizes Ω in terms of its irreducible components with respect to the group $SO(4)$. Here $C_{\dot{A}\dot{B}\dot{C}\dot{D}}$ (C_{ABCD}) are the self-dual (anti self-dual) components of the Weyl conformal tensor, $C_{AB\dot{C}\dot{D}}$ is the traceless part of the Ricci tensor, and R is the scalar curvature. V_4 is said to be *self-dual* (*anti self-dual*) if $\Omega^A{}_B=0$ ($\Omega^{\dot{A}}{}_{\dot{B}}=0$), and *conformally self-dual* (*conformally anti self-dual*) if $C_{ABCD}=0$ ($C_{\dot{A}\dot{B}\dot{C}\dot{D}}=0$).

THEOREM 4. *Let V_4 be a conformally anti self-dual Riemannian space. Suppose further that in V_4 the Laplace–Beltrami equation* (1.2) *is separable in the local coordinate system $\{x^i\}$. Then either V_4 is conformally flat or $\{x^i\}$ is type C and nonorthogonal.*

Proof. From equations (6.5)–(6.7) we find

$$C_{ABCD}=S^{ab}{}_{(AB}S^{cd}{}_{CD)}R_{abcd},$$

$$C_{\dot{A}\dot{B}\dot{C}\dot{D}}=S^{ab}{}_{(\dot{A}\dot{B}}S^{cd}{}_{\dot{C}\dot{D})}R_{abcd},$$

where the parentheses denote symmetrization and $S^{ab}{}_{AB}(S^{ab}{}_{\dot{A}\dot{B}})$ are the components of $S_{AB}(S_{\dot{A}\dot{B}})$ with respect to $\theta^a\wedge\theta^b$. Explicitly we have

$$\begin{aligned}&C_{\dot{1}\dot{1}\dot{1}\dot{1}}-C_{1111}=4(R_{1234}-R_{2314})+4i(R_{1214}+R_{2334}),\\&C_{\dot{1}\dot{1}\dot{1}\dot{2}}-C_{1112}=2(R_{1413}+R_{2324})+2i(R_{1242}+R_{3431}),\\&C_{\dot{1}\dot{1}\dot{2}\dot{2}}-C_{1122}=-8R_{1324}+4R_{1234}+4R_{1423}\end{aligned} \tag{6.8}$$

and $C_{ABCD}=\overline{C^{ABCD}}$ and the same for dotted components. Now suppose $\{x^i\}$ is type H; then by [5, Lemma 5] the only nonvanishing components of Ω are R_{abba}. Thus from (6.8) $C_{ABCD}=0$ implies $C_{\dot{A}\dot{B}\dot{C}\dot{D}}=0$. Suppose now that $\{x^i\}$ is type F. Without loss of generality (by making an $SO(4)$ gauge transformation if necessary) we can choose x^4 as the ignorable coordinate. Then [21, eq. (37.4)] implies $R_{abb4}=0$, $a,b=1,2,3$ and the result follows as before. Similarly if $\{x^i\}$ is type B, we can use results of Petrov [22, pp. 174–175] to show that V_4 is conformally flat. Now suppose $\{x^i\}$ is type C and orthogonal, then again [21, eq. (37.4)] and (6.8) imply that V_4 is conformally flat. It follows that $\{x^i\}$ is necessarily type C and nonorthogonal.

Since in an Einstein space, Hamilton–Jacobi separability implies Laplace–Beltrami separability [1], [5], it follows that Theorem 4 holds when Laplace–Beltrami separability is replaced by Hamilton–Jacobi separability, and V_4 is an Einstein space. Furthermore, since $\mathbb{C}P^2$ with the Fubini–Study metric (2.8) is a conformally anti self-dual Einstein space, we can combine Theorems 3 and 4 with the well-known fact [19], [20] that there are precisely two separable coordinate systems on S^2 to obtain:

COROLLARY. *There are precisely two separable coordinate systems on* $\mathbb{C}P^2$. *They are the two induced by Theorem* 3 *from spherical and elliptic coordinates on* S^2.

These two coordinate systems are given by equations (6.1) and (6.4) with $n=2$. Furthermore, the constants of the motion are given by $c_2(su(3))$, P_{α^1}, P_{α^2} and for

(i) spherical coordinates by $c_{12}=c_2(su(2))$

(ii) elliptic coordinates by $c_{23}+ac_{13}$

where in (6.4) we have taken $e_1=0$, $e_2=1$, $e_3=a$, and c_{ij} is given by (4.6).

C. $n=3$ and 4. For $\mathbb{C}P^3$ there are precisely six classes of separable systems with three ignorable coordinates. These come from [23, systems (1), (3), (6), (13) and (17)] In the real case there are two inequivalent classes of type (13), see [24, Table 1]. For $\mathbb{C}P^4$ there are 14 systems on S^4 which are inequivalent under $SO(5,\mathbb{C})$. These are given by [25, classes I, V(i), VI, VII(i), VIII and X]. The inequivalent types under the real group $SO(5,\mathbb{R})$ can be worked out from these. For all of the systems mentioned above the constants of the motion on S^3 and S^4 can be read off and transformed by (4.5) to constants of the motion on $\mathbb{C}P^3$ and $\mathbb{C}P^4$, respectively. It is then a straightforward task to write down the separated equations in each case.

7. Conclusions. The main result of this paper can be formulated as an algorithm. In order to find all conjugacy classes of coordinate systems in $\mathbb{C}P^n$ having an additive separation of variables in the Hamilton–Jacobi equation (1.1) (or multiplicative separation in the Laplace–Beltrami equation (1.2)), proceed as follows:

1. Introduce n complex coordinates z_k and put

$$z_k=\frac{s_k}{s_{n+1}}e^{i\alpha_k}, \qquad 1\le k\le n \tag{7.1}$$

where

$$s_1^2+\cdots+s_{n+1}^2=1, \tag{7.2}$$

i.e., s_i $(i=1,\cdots,n+1)$ are cartesian coordinates in $\mathbb{R}^{n+1}$.

2. Find all separable coordinate systems $\{\theta_1,\cdots,\theta_n\}$ on the real sphere S^n for which the free Hamilton–Jacobi (or free Laplace–Beltrami) equation on the sphere allows a separation of variables and express s_i $(1\le i\le n+1)$ in terms of these separable coordinates. The corresponding equations on the sphere with the potential induced from $\mathbb{C}P^n$ (see (3.7)) will, as we have shown, also separate. Substitute

$$s_i=s_i(\theta_1\cdots\theta_n), \qquad 1\le i\le n+1 \tag{7.3}$$

back into (7.1). Then the sets

$$(\theta_1,\cdots,\theta_n,\alpha_1,\cdots,\alpha_n) \tag{7.4}$$

provide a complete list of representatives of all conjugacy classes of separable coordinates on $\mathbb{C}P^n$ with n ignorable coordinates.

Several comments are in order here.

1. For $n=2$, i.e. the complex projective plane $\mathbb{C}P^2$ we have proven that all separable coordinate systems have precisely 2 ignorable coordinates, i.e. the maximum possible number equal to the rank n of $su(n+1)$. Thus there exist precisely 2 separable

coordinate systems in $\mathbf{C}P^2$, induced by spherical and elliptic coordinates on S^2, respectively.

2. The $2n$ integrals of motion in involution characterizing each separable system are obtained as follows. The first n of them correspond to the ignorable coordinates α_i; they form a basis for the Cartan subalgebra of $su(n+1)$ and can be identified with the canonical momenta conjugated to α_i:

$$P_{\alpha_i}, \qquad 1 \leq i \leq n. \tag{7.5}$$

The remaining n constants of motion A_r (including the Hamiltonian (1.1)) can be interpreted as second order operators in the enveloping algebra of $su(n+1)$. Writing the infinitesimal generators of $su(n+1)$ as (again by an abuse of notation)

$$T_{ik}=E_{ik}-E_{ki}, \quad S_{ik}=i(E_{ik}+E_{ki}), \quad H_{ik}=i(E_{ii}-E_{kk}) \; 1 \leq i \leq k \leq n+1, \tag{7.6}$$

in the defining representation of $su(n+1)$, we can write the quadratic constants of motion as n independent linear combinations of the $n(n+1)/2$ Casimir invariants

$$C_{ik}=T_{ik}^2+S_{ik}^2+H_{ik}^2, \qquad 1 \leq i < k \leq n+1, \tag{7.7}$$

$(i,k$ fixed) of the $su(2)$ algebras (7.6). Thus

$$A_r = \sum_{1 \leq i < k \leq n+1} \alpha_r^{ik} C_{ik}, \qquad 1 \leq r \leq n. \tag{7.8}$$

The operators A_r can easily be restricted to $\mathbf{C}P^n$ or to S^n. Upon restriction to S^n the Casimir operators C_{ik} reduce to the form (4.5). The classification of coordinate systems on S^n then reduces to a classification of sets of n operators in involution, all of them being linear combinations of the squares of the generators of $o(n+1)$.

Several problems suggested by this paper are under active consideration:

1. The first concerns special function theory and the separation of variables on $\mathbf{C}P^n$ in spherical coordinates (see §6A of this article). If we separate variables in the Laplace–Beltrami equation, then (6.2) reduces to a system of $2n$ ordinary linear equations. The eigenfunctions of the Laplace–Beltrami equations are then expressed as products of Jacobi functions (and exponentials $e^{iu_k\alpha_k}$). The role of Jacobi polynomials as basis functions of $SU(n+1)$ representations in a basis corresponding to the subgroup reduction $SU(n+1) \supset U(n) \supset U(n-1) \supset \cdots \supset U(2) \supset U(1)$ makes it possible to obtain relations for special functions, in particular addition formulas [26]. But now more poweful methods are at our disposal. We can use the techniques of [12] by constructing a simple model of $SU(n+1)$ acting on the sections of certain holomorphic line bundles over $\mathbf{C}P^n$ and relate this action to the action on harmonic polynomials—namely the Jacobi polynomials. Furthermore, we have many more sets of bases than that given by spherical coordinates. A detailed study of tractable coordinates should give a wealth of special function identities.

2. The approach of this article has been to Hermitian hyperbolic spaces $HH(n)$. The noncompact group $SU(n,1)$ then plays the role that $SU(n+1)$ plays for $\mathbf{C}P(n)$. The results are much richer for $HH(n)$ mainly because $su(n,1)$ has $n+2$ different mutually nonconjugated maximal abelian subalgebras [27]–[28], each of them being of dimension n and leading to different types of coordinate systems with n ignorable variables [29].

3. Separation of variables on a sphere S^n is being studied for arbitrary n (the results are at present known only for $n=2,3$, and 4) [30].

REFERENCES

[1] S. Benenti and M. Francaviglia, *The theory of separability of the Hamilton–Jacobi equation and its application to general relativity*, in General Relativity and Gravitation, Vol. 1, A. Held, ed., Plenum, New York, 1980.

[2] E. G. Kalnins and W. Miller, Jr., *Killing tensors and nonorthogonal variable separation for Hamilton–Jacobi equations*, this Journal, Anal. 12 (1981), pp. 617–629.

[3] N. M. J. Woodhouse, *Killing tensors and the separation of the Hamilton–Jacobi equation*, Comm. Math. Phys., 44 (1975), pp. 9–38.

[4] E. G. Kalnins and W. Miller, Jr., *Separable coordinates for three dimensional complex Riemannian spaces*, J. Diff. Geom., 14 (1979), pp. 221–236.

[5] 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.

[6] W. Miller Jr., J. Patera and P. Winternitz, *Subgroups of Lie groups and separation of variables*, J. Math. Phys., 22 (1981), pp. 251–260.

[7] L. P. Eisenhart, *Separable systems of Stäckel*, Ann. Math., 35 (1934), pp. 284–305.

[8] E. G. Kalnins and W. Miller Jr., *Killing tensors and variable separation for Hamilton–Jacobi and Helmholtz equations*, this Journal, 11 (1980), pp. 1011–1026.

[9] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, Benjamin, Reading, MA, 1978.

[10] J. Moser, *Various aspects of integrable Hamiltonian systems*, in Dynamical Systems, Progress in Mathematics No. 8, Birkhauser, Boston, 1980.

[11] N. Jacobson, *Lie Algebras*, Interscience, New York, 1962.

[12] W. Miller Jr., *Symmetry and Separation of Variables*, Addison-Wesley, Reading, MA, 1977.

[13] P. Winternitz and I. Friš, *Invariant expansions of relativistic amplitudes and subgroups of the proper Lorentz group*, Yad. Fiz 1 (1965), pp. 889–901 [Sov. J. Nucl. Phys., 1 (1965), pp. 636–653].

[14] G. W. Gibbons and C. N. Pope, $\mathbb{C}P^2$ *as a gravitational instanton*, Comm. Math. Phys., 61 (1978), pp. 239–248.

[15] C. P. Boyer, *Gravitational Instantons*, Hadronic J. 4 (1981), pp. 2–18.

[16] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry*, vol. 2, John Wiley, New York, 1968.

[17] J. Marsden and A. Weinstein, *Reduction of symplectic manifolds with symmetry*, Rep. Math. Phys., 5 (1974), pp. 121–130.

[18] D. Kazhdan, B. Kostant and S. Sternberg, *Hamiltonian group actions and dynamical systems of Calogero type*, Comm. Pure and Appl. Math., 31 (1978), pp. 481–508.

[19] P. Olevski, *The separation of variables in the equation* $\Delta_3 u + \lambda u = 0$ *for spaces of constant curvature in two and three dimensions*, Mat. Sb, 27 (1950), pp. 379–426.

[20] J. Patera and P. Winternitz, *A new basis for the representations of the rotation group: Lamé and Heun polynomials*, J. Math. Phys., 14 (1973), pp. 1130–1139.

[21] L. P. Eisenhart, *Riemannian Geometry*, Princeton Univ. Press, Princeton, 1949.

[22] N. Petrov, *Einstein Spaces*, Pergamon, Oxford, 1969.

[23] E. G. Kalnins and W. Miller Jr., *Lie theory and the wave equation in space-time* 2. *The group* $SO(4,\mathbb{C})$, this Journal, 9 (1978), pp. 12–33.

[24] 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.

[25] E. G. Kalnins and W. Miller Jr., *The wave equation and separation of variables on the complex sphere* S_4, to appear.

[26] T. Koornwinder, *The addition formula for Jacobi polynomials and spherical harmonics*, SIAM J. Appl. Math., 25 (1973), pp. 236–246.

[27] J. Patera, P. Winternitz and H. Zassenhaus, *On the maximal abelian subgroups of the linear classical algebraic groups*, Math. Rep. Acad. Sci. Canada, 2 (1980), pp. 231–236.

[28] ———, *On the maximal abelian subgroups of the quadratic classical algebraic groups*, Math. Rep. Acad. Sci. Canada, 2 (1980), pp. 237–242.

[29] C. P. Boyer, E. G. Kalnins and P. Winternitz, *Completely integrable relativistic Hamiltonian systems and separation of variables in Hermitian hyperbolic spaces*, J. Math. Phys., 24 (1983), 2022–2034.

[30] E. G. Kalnins and W. Miller Jr. *Separation of variables on n-dimensional Riemannian manifolds.* 1. *The n-sphere* S_n *and Euclidean n-space* R_n (to appear).

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

007

GLOBAL EXISTENCE AND ASYMPTOTIC STABILITY FOR A SEMILINEAR HYPERBOLIC VOLTERRA EQUATION WITH LARGE INITIAL DATA*

WILLIAM J. HRUSA[†]

Abstract. The equation

$$(*)\qquad u_{tt}(x,t)=\phi(u_x(x,t))_x-\int_0^t m(t-\tau)\psi(u_x(x,\tau))_x\,d\tau+f(x,t),\qquad x\in\mathbb{B}\subset\mathbb{R},\quad t\geq 0,$$

provides a model for the motion of a one-dimensional viscoelastic body. Here ϕ,ψ,m, and f are given smooth functions, and $u(x,t)$ denotes the unknown displacement at time t of the particle with reference position x. All functions in $(*)$ are real-valued; subscripts x and t indicate partial differentiation.

It has been shown by several authors that under physically reasonable assumptions, various initial and initial-boundary value problems associated with $(*)$ have globally defined classical solutions provided that f and the initial data are sufficiently smooth and "small". Moreover, these solutions decay to zero as $t\to\infty$. It has also been shown (for the special case with $\psi\equiv\phi$) that if $\phi''\not\equiv 0$, then $(*)$ does not have global smooth solutions if the data are too "large".

In order to isolate the effects of nonlinearity in the memory term we here study $(*)$ with ϕ linear, but ψ (generally) nonlinear. Several global existence and asymptotic stability results which permit the data to be large are established. We analyze in detail the case where $\mathbb{B}=[0,1]$ and homogeneous Dirichlet boundary conditions are imposed. Other types of boundary conditions as well as pure initial value problems (i.e. $\mathbb{B}=\mathbb{R}$) are discussed in the last section. The analysis is based on a priori estimates of energy-type.

1. Introduction.

The equation

$$(1.1)\qquad u_{tt}(x,t)=\phi(u_x(x,t))_x-\int_0^t m(t-\tau)\psi(u_x(x,\tau))_x\,d\tau+f(x,t),\qquad 0\leq x\leq 1,\quad t\geq 0,$$

provides a model for the motion of a homogeneous one-dimensional viscoelastic body that occupies the interval $[0,1]$ in a reference configuration (which we assume to be a natural state) and has unit reference density. Here $\phi,\psi\colon\mathbb{R}\to\mathbb{R}$ are assigned smooth constitutive functions, $m\colon[0,\infty)\to\mathbb{R}$ is a given smooth relaxation function, $f\colon[0,1]\times[0,\infty)\to\mathbb{R}$ is a known forcing function, and $u(x,t)$ denotes the (unknown) displacement at time t of the particle with reference position x. Subscripts x and t indicate partial derivatives, and a prime will be used to denote the derivative of a function of a single variable. A standard problem is to determine a smooth function u which satisfies (1.1) together with prescribed initial conditions at $t=0$ and appropriate boundary conditions at $x=0,1$.

Assuming that m is integrable over $[0,\infty)$, we define the equilibrium stress function $\chi\colon\mathbb{R}\to\mathbb{R}$ by[1]

$$(1.2)\qquad \chi(\xi):=\phi(\xi)-\psi(\xi)\int_0^\infty m(s)\,ds\quad\forall\xi\in\mathbb{R}.$$

On physical grounds, it is natural to assume that

$$(1.3)\qquad \phi'(0)>0,\quad \psi'(0)>0,\quad \chi'(0)>0,$$

and that m is positive and decreasing.

*Received by the editors October 3, 1983, and in revised form December 19, 1983.

[†] Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

[1] The symbol $:=$ is used to indicate an equality in which the left-hand side is defined by the right-hand side.

If m vanishes identically, then (1.1) reduces to the undamped quasilinear wave equation

$$u_{tt}=\phi(u_x)_x+f. \tag{1.4}$$

It is well known that (1.4) does not generally have globally defined smooth solutions, no matter how smooth ϕ, f, and the initial data are. (See, for example, [7] and [9].) This situation does not improve even if f and the initial data are small.

If $m\not\equiv 0$ and the appropriate sign conditions are satisfied, the memory term in (1.1) induces dissipation which is effective (and overpowers the destabilizing effects of nonlinearity) provided that u is "small". A great deal of insight into the dissipative character of memory is provided by the work of Coleman and Gurtin [1] on growth and decay of acceleration waves in materials with memory. Roughly speaking, an acceleration wave is a solution which is smooth except along a curve Σ, across which second derivatives sustain jump discontinuities. The amplitude of such a wave is defined to be the jump in acceleration across Σ.

Coleman and Gurtin derived general and explicit expressions for the change in amplitude of an acceleration wave travelling in a nonlinear material with fading memory. In particular they showed that under physically reasonable assumptions (which exclude the case $m\equiv 0$ in (1.1)), the amplitude of an acceleration wave decays to zero as $t\to\infty$ if its initial amplitude is sufficiently small. They also showed that the amplitude of an acceleration wave may become infinite in finite time if the initial amplitude is too large.

Global existence theorems for (1.1) have been obtained by several authors. In the special case where $\psi\equiv\phi$, MacCamy [8], Dafermos and Nohel [2], and Staffans [10] have shown that various initial and initial-boundary value problems associated with (1.1) have globally defined smooth solutions provided that the initial data and forcing function are suitably smooth and small. Moreover, these solutions tend to zero as $t\to\infty$. Such results were also obtained (for initial-boundary value problems) in the general case with ψ different from ϕ by Dafermos and Nohel [3].

On the other hand, Hattori [6] has shown (for the case $\psi\equiv\phi$) that if $\phi''\not\equiv 0$, then there are smooth data for which (1.1) does not have a global smooth solution. Such data must necessarily be "large" in view of the aforementioned existence results.

We remark also that in the important special case when the relaxation function is an exponential of the form $m(s)\equiv e^{-\mu s}$, where μ is a positive constant, (1.1) (with $f\equiv 0$) is equivalent to the third order partial differential equation studied by Greenberg [4]. He derived a priori estimates which show that any sufficiently smooth and small solution must decay to zero exponentially as $t\to\infty$.

In the present paper, we study (1.1) when ϕ is linear (i.e., $\phi''\equiv 0$), but ψ is allowed to be nonlinear. In particular, we consider the initial-boundary value problem

$$u_{tt}(x,t)=cu_{xx}(x,t)-\int_0^t m(t-\tau)\psi(u_x(x,\tau))_x\,d\tau+f(x,t), \qquad 0\leq x\leq 1,\quad t\geq 0, \tag{1.5}$$

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

$$u(x,0)=u_0(x), \qquad u_t(x,0)=u_1(x), \qquad 0\leq x\leq 1, \tag{1.7}$$

where c is a positive constant and u_0, u_1 are given initial data. Our motivation for studying (1.5) is to isolate the effects of nonlinearity in the memory. We would like to determine whether or not nonlinearity of ψ leads to breakdown of smooth solutions with large initial data.

Since (1.5) is a special case of (1.1), the results of Dafermos and Nohel [3] imply that under physically reasonable assumptions on m and ψ, (1.5), (1.6), (1.7) has a unique solution $u \in C^2([0,1]\times[0,\infty))$ provided that u_0, u_1, and f are sufficiently smooth and small. Moreover, u and its partial derivatives of first and second order converge to zero uniformly on $[0,1]$ as $t\to\infty$.

It is conceivable that if m satisfies appropriate decay and sign conditions and a global analogue of (1.3) holds, then (1.5) will have global smooth solutions which tend to zero as $t\to\infty$, even for large data. A strong indication that this should be the case is contained in the paper of Coleman and Gurtin [1]. When specialized to equation (1.5), their results reveal that even large discontinuities in acceleration are damped-out.

We here establish several global existence and asymptotic stability theorems for (1.5), (1.6), (1.7) which allow the initial data to be large. Our basic approach is standard: We first construct a local solution on a maximal time interval $[0,T_0)$ with the property that a certain "energy" becomes unbounded as $t\to T_0$, if $T_0<\infty$. We then derive estimates for the local solution which imply global existence and, in certain cases, guarantee that the solution decays to zero.

The paper is divided into five sections. The main results concerning global existence and decay of solutions are stated in §2. Local solutions are studied in detail in §3, and in §4 we derive certain global estimates and provide proofs of the results stated in §2. Other types of boundary conditions are discussed in §5.

For hyperbolic type problems there is essentially an even tradeoff, locally, between space and time smoothness. In other words, taking a spatial derivative of the solution causes the same loss of smoothness as does taking a time derivative. It is therefore reasonable to expect that the time integration in (1.5) has the same effect (locally) as removing a spatial derivative from the quasilinear term $\psi(u_x)_x$. This is indeed the case; the local behavior of solutions of (1.5) is quite similar to that of solutions of the semilinear equation

$$u_{tt}=cu_{xx}+\psi(u_x)+f. \tag{1.8}$$

Roughly speaking, a solution of (1.8) exists (and retains the full smoothness of its initial data) for as long as the spatial L^2 norms of its second derivatives remain bounded. The same is true for (1.5). (Stronger bounds are required to continue a solution of (1.1). See [3].)

Travis and Webb [11] have studied local existence of solutions to the abstract initial value problem

$$(1.9)\qquad \begin{aligned} &u''(t)=Au(t)+\int_0^t g(t,s,u(s),u'(s))\,ds+f(t), \qquad t\in\mathbb{R},\\ &u(0)=u^0, \qquad u'(0)=u^1, \end{aligned}$$

where u and f take values in a Banach space X,A is a linear operator generating a strongly continuous cosine family in X, and g is a (generally) unbounded nonlinear mapping from $\mathbb{R}\times\mathbb{R}\times X\times X$ into X. A local existence result for (1.5), (1.6), (1.7) quite similar to Theorem 3.1 follows easily from Propositions 3.1 and 3.2 of [11]. The questions of global existence and higher order regularity are not considered in [11].

Although Theorem 3.1 essentially follows from the work of Travis and Webb, we provide here a direct proof based on energy estimates. These estimates are useful for establishing additional properties of local solutions. Moreover, they are helpful for securing certain global bounds.

Observe that if ϕ is linear and m is positive, then the natural global analogue of (1.3) entails boundedness of ψ'. If $c>0$ and ψ' is bounded, one can establish global existence of classical solutions to (1.5), (1.6), (1.7)—even for large data—without imposing any sign restrictions on m or ψ'. In fact, rather crude bounds (requiring only local assumptions on m) can be employed for this purpose. Of course, such solutions may become unbounded as $t \to \infty$.

The estimates required to establish decay of solutions with large data are considerably more delicate and rely crucially on positivity and decay of m, and a global version of (1.3). Our discussion of asymptotic behavior is limited to the case where m is a decreasing exponential and $f \equiv 0$. Motivated by the paper of Greenberg [4], we use exponentially weighted multipliers to form our energy identities. This leads to estimates which yield exponential decay of solutions—even for large initial data. The assumption $f \equiv 0$ is made only for the sake of simplicity; a nonzero forcing term which behaves suitably as $t \to \infty$ causes no problems.

Our asymptotic analysis can be adapted to handle "nonexponential" relaxation functions. However, this involves imposing rather complicated and implicit assumptions on the resolvent kernel associated with m'. Theorems based on these assumptions seem somewhat artificial and will not be discussed here.

It should be noted that our results depend in an essential way on the fact that we are dealing with only one spatial dimension. The methods used here can also be applied to analogues of (1.5) in more than one space dimension; however, among other things, smallness of the data would be required to establish global existence of classical solutions.

We close the Introduction with some remarks on notation. We shall frequently deal with functions from a set of the form $[0,1]\times[0,T]$ into $\mathbb{R}$. All such functions are assumed to be measurable. Subscripts x and t always indicate partial differentiation with respect to the first and second argument, respectively. (Subscripts other than x or t do not indicate differentiation.) All derivatives are to be interpreted in the sense of distributions.

For a function w: $[0,1]\times[0,T]\to\mathbb{R}$, it is useful to consider the mapping $t \mapsto w(\cdot,t)$ from $[0,T]$ into various function spaces. We use the same symbol w to denote such a mapping. Moreover, we suppress the qualification "almost everywhere" and we omit obvious remarks such as "after modification on a set of measure zero" when no danger of confusion is likely. We employ standard notation for the usual function spaces.

2. Statement of main results. We first discuss the existence of global solutions to (1.5) when no assumptions are made regarding the sign of m or ψ'. Of course, decay of solutions should not be expected in this situation; in fact, the memory term may cause solutions to grow—even if ψ is linear. Roughly speaking, Coleman and Gurtin [1] have shown that for equations such as (1.5), the amplitude of an acceleration wave is influenced by ψ', but not by ψ''. Their results suggest that boundedness of ψ' precludes the development of singularities—independently of the size of the data and the sign of the memory term. This is indeed the case.

THEOREM 2.1. *Assume that* $c>0$, $\psi\in C^2(\mathbb{R})$, $m\in W^{1,1}_{\mathrm{loc}}[0,\infty)$, *and that* ψ' *is bounded, i.e.,* $\psi'\in L^\infty(\mathbb{R})$. *Let* u_0, u_1, *and* f *be given with*

$$(2.1) \qquad u_0\in H^1_0(0,1)\cap H^2(0,1), \qquad u_1\in H^1_0(0,1),$$

$$(2.2) \qquad f\in C([0,\infty);L^2(0,1)), \qquad f_t\in L^2_{\mathrm{loc}}([0,\infty);L^2(0,1)).$$

Then the initial-boundary value problem (1.5), (1.6), (1.7) *has a unique solution* u: $[0,1]\times[0,\infty)\to\mathbb{R}$ *with*

$$u, u_t, u_x, u_{tt}, u_{tx}, u_{xx} \in C([0,\infty); L^2(0,1)). \tag{2.3}$$

The solution in Theorem 2.1 is not necessarily a classical solution. However, under slightly stronger smoothness assumptions, global classical solutions exist.

THEOREM 2.2. *Assume that* $c>0$, $\psi\in C^2(\mathbb{R})$, $m\in C^1[0,\infty)$, *and that* ψ' *is bounded. Assume further that* u_0, u_1, *and* f *satisfy*

$$u_0\in H_0^1(0,1)\cap H^3(0,1), \qquad u_1\in H_0^1(0,1)\cap H^2(0,1), \tag{2.4}$$

$$u_0''(0)=u_0''(1)=0, \tag{2.5}$$

$$f, f_t, f_x\in C([0,\infty); L^2(0,1)), \qquad f_{tt}\in L^2_{\text{loc}}([0,\infty); L^2(0,1)), \tag{2.6}$$

$$f(0,t)=f(1,t)=0 \quad \forall t\geq 0. \tag{2.7}$$

Then, the solution u *in Theorem* 2.1 *has the additional regularity*

$$u_{ttt}, u_{ttx}, u_{txx}, u_{xxx}\in C([0,\infty); L^2(0,1)). \tag{2.8}$$

Thus, by the Sobolev embedding theorem, $u\in C^2([0,1]\times[0,\infty))$.

Remark 2.1. Theorem 2.2 remains valid if (2.5) and (2.7) are replaced by the weaker compatibility assumption

$$cu_0''(0)+f(0,0)=cu_0''(1)+f(1,0)=0. \tag{2.9}$$

The proof in this situation requires careful estimation of certain boundary terms which automatically vanish if (2.5) and (2.7) hold.

Remark 2.2. As is to be expected, local analogues of Theorems 2.1 and 2.2 are valid without the assumption that ψ' is bounded. These are stated and proved in the next section. (Continuous dependence on the data is also established in §3.) Moreover, a local solution u retains the full smoothness of its data for as long as u_x remains pointwise bounded. (See Corollary 4.1.) Thus shock formation does not occur for (1.5).

We now discuss asymptotic behavior of solutions of the equation

$$u_{tt}(x,t)=cu_{xx}(x,t)-\int_0^t e^{-\mu(t-\tau)}\psi(u_x(x,\tau))_x d\tau, \qquad 0\leq x\leq 1, \quad t\geq 0, \tag{2.10}$$

where μ is a positive constant. The corresponding equilibrium stress function is given by

$$\chi(\xi):=c\xi-\mu^{-1}\psi(\xi) \quad \forall\xi\in\mathbb{R}. \tag{2.11}$$

Thus, the natural global analogue of (1.3) reads

$$c>0, \quad 0<\psi'(\xi)<\mu c \quad \forall\xi\in\mathbb{R}. \tag{2.12}$$

Under a slightly stronger assumption on ψ', we show that solutions decay to zero exponentially, even if the initial data are large.

THEOREM 2.3. *Assume that* $c>0$, $\psi\in C^2(\mathbb{R})$, *and that there are constants* α *and* β *such that*

$$0<\alpha\leq\psi'(\xi)\leq\beta<\mu c \quad \forall\xi\in\mathbb{R}. \tag{2.13}$$

Then there exist constants Γ, $\delta>0$ *such that for every* u_0 *and* u_1 *satisfying* (2.1), *the corresponding solution u of* (2.10), (1.6), (1.7) (*which exists by Theorem* 2.1) *satisfies*

(2.14)

$$\int_0^1\{u^2+u_t^2+u_x^2+u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,t)\,dx\le\Gamma e^{-\delta t}\int_0^1\left[u_0''(x)^2+u_1'(x)^2\right]dx\quad\forall t\ge0.$$

Thus, by the Sobolev embedding theorem, u also satisfies

$$\max_{x\in[0,1]}\left(u^2+u_t^2+u_x^2\right)(x,t)\le\Gamma e^{-\delta t}\int_0^1\left[u_0''(x)^2+u_1'(x)^2\right]dx\quad\forall t\ge0. \tag{2.15}$$

Remark 2.3. If a forcing term f is added to (2.10), one can still establish decay of solutions if $f(\cdot,t)$ behaves suitably as $t\to\infty$. Of course, the rate of decay of solutions depends on the rate of decay of f. An estimate quite similar to (2.14) holds if f decays exponentially in time. If it is merely assumed that f and f_t are square integrable over $[0,1]\times[0,\infty)$, it is still possible to show that $u(\cdot,t)$, $u_t(\cdot,t)$, and $u_x(\cdot,t)$ converge to zero uniformly on $[0,1]$ as $t\to\infty$. This is discussed further in §4.

Remark 2.4. Analogous results for other types of boundary conditions are discussed in §5.

3. Local solutions. This section is concerned with properties of local solutions of (1.5), (1.6), (1.7). In particular, we discuss existence, uniqueness, continuation, continuous dependence on initial data, and regularity.

THEOREM 3.1. *Assume that* $c>0$, $\psi\in C^2(\mathbb{R})$, $m\in W^{1,1}_{\mathrm{loc}}[0,\infty)$, *and let* u_0, u_1, f *be given with*

$$u_0\in H_0^1(0,1)\cap H^2(0,1),\qquad u_1\in H_0^1(0,1), \tag{3.1}$$

$$f\in C\big([0,\infty);L^2(0,1)\big),\qquad f_t\in L^2_{\mathrm{loc}}\big([0,\infty);L^2(0,1)\big). \tag{3.2}$$

Then (1.5), (1.6), (1.7) *has a unique solution u, defined on a maximal time interval* $[0,T_0)$, $T_0>0$, *with*

$$u,u_t,u_x,u_{tt},u_{tx},u_{xx}\in C\big([0,T_0);L^2(0,1)\big). \tag{3.3}$$

Moreover, if

$$\sup_{t\in[0,T_0)}\int_0^1\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,t)\,dx<\infty, \tag{3.4}$$

then $T_0=\infty$.

The procedure used to prove Theorem 3.1 is familiar: The solution of (1.5), (1.6), (1.7) is constructed as a fixed point of the "solution operator" associated with a related family of linear problems. We begin by recording some standard results concerning solutions of the linear wave equation

$$u_{tt}(x,t)=cu_{xx}(x,t)+g(x,t),\qquad 0\le x\le1,\quad t\ge0. \tag{3.5}$$

To simplify our notation, we make the following definition. For each $T>0$, we denote by $\mathcal{C}_T^2$ the set of all functions w: $[0,1]\times[0,T]\to\mathbb{R}$ with

$$w,w_t,w_x,w_{tt},w_{tx},w_{xx}\in C\big([0,T];L^2(0,1)\big). \tag{3.6}$$

PROPOSITION 3.1. *Assume that* $c>0$ *and that* (3.1) *holds. Let* $T>0$ *and* g: $[0,1]\times[0,T]\to\mathbb{R}$ *be given with* $g\in C([0,T];L^2(0,1))$, $g_t\in L^2([0,T];L^2(0,1))$. *Then*, (3.5), (1.6)

(1.7) *has a unique solution* $u\in\mathcal{C}_T^2$. *Moreover,*

$$(3.7)\qquad \int_0^1\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,t)\,dx\le\bar c\int_0^1\left[u_0''(x)^2+u_1'(x)^2+g(x,t)^2\right]dx$$
$$+\bar c\int_0^t\int_0^1\{g_t^2+u_{tt}^2\}(x,s)\,dx\,ds\quad\forall t\in[0,T],$$

where $\bar c$ *is a positive constant which depends only on* c. *Thus, by Gronwall's inequality,* u *also satisfies*

(3.8)

$$\max_{t\in[0,T]}\int_0^1\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,t)\,dx\le\bar ce^{\bar cT}\left\{\int_0^1\left[u_0''(x)^2+u_1'(x)^2\right]dx\right.$$
$$\left.+\max_{t\in[0,T]}\int_0^1 g(x,t)^2dx+\int_0^T\int_0^1 g_t(x,t)^2dx\,dt\right\}.$$

Remark 3.1. By virtue of linearity, (3.8) can be used to establish continuous dependence of the solution of (3.5), (1.6), (1.7) on u_0, u_1, and g.

For M, $T>0$, let $\mathcal{X}_{M,T}$ denote the set of all $w\in\mathcal{C}_T^2$ which satisfy

$$(3.9)\qquad w(0,t)=w(1,t)=0\quad\forall t\in[0,T],$$

and

$$(3.10)\qquad \max_{t\in[0,T]}\int_0^1\{w_{tt}^2+w_{tx}^2+w_{xx}^2\}(x,t)\,dx\le M^2,$$

equipped with the (complete) metric defined by

$$(3.11)\quad \rho(w,\hat w):=\max_{t\in[0,T]}\left(\int_0^1\{(w_{tt}-\hat w_{tt})^2+(w_{tx}-\hat w_{tx})^2+(w_{xx}-\hat w_{xx})^2\}(x,t)\,dx\right)^{1/2}.$$

Note that $\mathcal{X}_{M,T}$ is nonempty for each M, $T>0$. The Sobolev embedding theorem and (3.6) imply $w\in C^1([0,1]\times[0,T])$. Moreover, it follows from (3.9) and (3.10) that

$$(3.12)\qquad \max_{\substack{x\in[0,1]\\ t\in[0,T]}}|w_x(x,t)|\le M\quad\forall w\in\mathcal{X}_{M,T}.$$

For $w\in\mathcal{X}_{M,T}$ we consider the linear equation

(3.13)

$$u_{tt}(x,t)=cu_{xx}(x,t)-\int_0^t m(t-\tau)\psi(w_x(x,\tau))_x\,d\tau+f(x,t),\quad x\in[0,1],\quad t\in[0,T],$$

and we let S denote the map which carries w into the unique solution of (3.13), (1.6), (1.7). Our goal is to show that S has a unique fixed point in $\mathcal{X}_{M,T}$ for appropriately chosen M and T. For this purpose we employ the contraction mapping principle and Proposition 3.1. As a first step, we prove

LEMMA 3.1. *Under the assumptions of Theorem* 3.1, S *maps* $\mathcal{X}_{M,T}$ *into* $\mathcal{X}_{M,T}$ *for* M *sufficiently large and* T *sufficiently small relative to* M.

Proof. Let M, $T>0$ and $w\in\mathcal{X}_{M,T}$ be given. Define g: $[0,1]\times[0,T]\to\mathbb{R}$ by

$$(3.14)\quad g(x,t):=f(x,t)-\int_0^t m(t-\tau)\psi(w_x(x,\tau))_x\,d\tau,\qquad x\in[0,1],\quad t\in[0,T],$$

and observe that $g\in C([0,T];L^2(0,1))$, $g_t\in L^2([0,T];L^2(0,1))$. Thus, by Proposition 3.1, the solution u of (3.13), (1.6), (1.7) belongs to $\mathcal{C}_T^2$ and satisfies (3.8) (with g defined by (3.14)).

It is important to keep track of how the bound for u which we obtain from (3.8) depends on M, T, and the initial data, ect. Therefore, let us define $\bar f, \bar m, \bar\psi$: $[0,\infty)\to[0,\infty)$ by

$$\bar f(s):=\max_{t\in[0,s]}\int_0^1 f(x,t)^2dx+\int_0^s\int_0^1 f_t(x,t)^2dx\,dt\quad\forall s\geq 0, \tag{3.15}$$

$$\bar m(s):=\max_{t\in[0,s]}m(t)^2+\left(\int_0^s|m'(t)|\,dt\right)^2\quad\forall s\geq 0, \tag{3.16}$$

$$\bar\psi(\eta):=\max_{|\xi|\leq\eta}\left[\psi'(\xi)^2+\psi''(\xi)^2\right]\quad\forall\eta\geq 0, \tag{3.17}$$

and set

$$U_0:=\int_0^1\left[u_0''(x)^2+u_1'(x)^2\right]dx. \tag{3.18}$$

Observe that $\bar f, \bar m$ and $\bar\psi$ are continuous and nondecreasing on $[0,\infty)$.

It follows easily from (3.14) that

$$\int_0^1 g(x,t)^2dx\leq 2\int_0^1 f(x,t)^2dx+2\int_0^1\left(\int_0^t m(t-\tau)\psi(w_x(x,\tau))_x d\tau\right)^2dx,\qquad t\in[0,T], \tag{3.19}$$

from which we conclude that

$$\max_{t\in[0,T]}\int_0^1 g(x,t)^2dx\leq 2\bar f(T)+2T^2\bar m(T)\bar\psi(M)M^2. \tag{3.20}$$

Moreover, we have

$$g_t(x,t)=f_t(x,t)-m(0)\psi(w_x(x,t))_x-\int_0^t m'(t-\tau)\psi(w_x(x,\tau))_x d\tau, \tag{3.21}$$

$$x\in[0,1],\quad t\in[0,T],$$

which yields

$$\begin{aligned}\int_0^T\int_0^1 g_t(x,t)^2dx\,dt\leq 3\int_0^T\int_0^1 f_t(x,t)^2dx\,dt&\\ +3m(0)^2\int_0^T\int_0^1\left(\psi(w_x(x,\tau))_x\right)^2dx\,d\tau&\\ +3\int_0^T\int_0^1\left(\int_0^t m'(t-\tau)\psi(w_x(x,\tau))_x d\tau\right)^2dx\,dt,&\end{aligned} \tag{3.22}$$

and hence

$$\int_0^T\int_0^1 g_t(x,t)^2dx\,dt\leq 3\bar f(T)+6T\bar m(T)\bar\psi(M)M^2. \tag{3.23}$$

It now follows from (3.8), (3.20), and (3.23) that

$$(3.24)\qquad \max_{t\in[0,T]}\int_0^1\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,t)\,dx$$
$$\le 6\bar{c}e^{\bar{c}T}\{U_0+\bar{f}(T)+(T^2+T)\bar{m}(T)\bar{\psi}\,(M)M^2\}.$$

If M and T satisfy

$$(3.25)\qquad M^2\ge 12\bar{c}e^{\bar{c}}(U_0+\bar{f}(1)),$$
$$(3.26)\qquad T\le\min\left(1,(24\bar{c}e^{\bar{c}}\bar{m}(1)\bar{\psi}\,(M)+1)^{-1}\right),$$

then the right-hand side of (3.24) is dominated by M^2 and consequently

$$(3.27)\qquad \max_{t\in[0,T]}\int_0^1\{(Sw)_{tt}^2+(Sw)_{tx}^2+(Sw)_{xx}^2\}(x,t)\,dx\le M^2\quad \forall w\in\mathcal{X}_{M,T}.$$

This completes the proof of Lemma 3.1. □

LEMMA 3.2. *Let the assumptions of Theorem* 3.1 *hold. Then* S: $\mathcal{X}_{M,T}\to\mathcal{X}_{M,T}$ *is strictly contractive if M is sufficiently large and T is sufficiently small relative to M.*

Proof. Let M, $T>0$ and w, $\hat{w}\in\mathcal{X}_{M,T}$ be given. Set $u:=Sw$, $\hat{u}:=S\hat{w}$, $U:=u-\hat{u}$, $W:=w-\hat{w}$, and note that U satisfies

$$(3.28)\qquad U_{tt}(x,t)=U_{xx}(x,t)+G(x,t),\qquad x\in[0,1],\quad t\in[0,T],$$
$$(3.29)\qquad U(0,t)=U(1,t)=0,\qquad t\in[0,T],$$
$$(3.30)\qquad U(x,0)=U_t(x,0)=0,\qquad x\in[0,1],$$

where G is defined by

$$(3.31)\qquad G(x,t):=\int_0^t m(t-\tau)[\psi(\hat{w}_x(x,\tau))_x-\psi(w_x(x,\tau))_x]\,d\tau,$$
$$x\in[0,1],\quad t\in[0,T].$$

Observe that G has the smoothness required to apply Proposition 3.1 to (3.28), (3.29), (3.30). In particular, (3.8) can be used to estimate U in terms of W.

A simple computation shows that

$$(3.32)\qquad G(x,t)=-\int_0^t m(t-\tau)\psi'(w_x(x,\tau))W_{xx}(x,\tau)\,d\tau$$
$$-\int_0^t m(t-\tau)[\psi'(w_x(x,\tau))-\psi'(\hat{w}_x(x,\tau))]\hat{w}_{xx}(x,\tau)\,d\tau,$$
$$x\in[0,1],\quad t\in[0,T],$$

and

$$(3.33)\qquad G_t(x,t)=-m(0)\psi'(w_x(x,t))W_{xx}(x,t)$$
$$-m(0)[\psi'(w_x(x,t))-\psi'(\hat{w}_x(x,t))]\hat{w}_{xx}(x,t)$$
$$-\int_0^t m'(t-\tau)\psi'(w_x(x,\tau))W_{xx}(x,\tau)\,d\tau$$
$$-\int_0^t m'(t-\tau)[\psi'(w_x(x,\tau))-\psi'(\hat{w}_x(x,\tau))]\hat{w}_{xx}(x,\tau)\,d\tau,$$
$$x\in[0,1],\quad t\in[0,T].$$

Making use of the mean value theorem and the inequality

$$\max_{\substack{x\in[0,1]\\ t\in[0,T]}} |W_x(x,t)| \le \max_{t\in[0,T]} \int_0^1 W_{xx}^2(x,t)\,dx, \tag{3.34}$$

we deduce from (3.32) and (3.33) that

$$\max_{t\in[0,T]} \int_0^1 G(x,t)^2 dx \le 2T^2 \bar{m}(T)\bar{\psi}(M)\cdot(M^2+1) \max_{t\in[0,T]} \int_0^1 W_{xx}^2(x,t)\,dx \tag{3.35}$$

and

$$\int_0^T \int_0^1 G_t(x,t)^2 dx\,dt \le 8T\bar{m}(T)\bar{\psi}(M)\cdot(M^2+1) \max_{t\in[0,T]} \int_0^1 W_{xx}^2(x,t)\,dx, \tag{3.36}$$

where $\bar{m}$ and $\bar{\psi}$ are as in the proof of Lemma 3.1.

It follows from (3.8), (3.35), and (3.36) that

$$\max_{t\in[0,T]} \int_0^1 \{U_{tt}^2 + U_{tx}^2 + U_{xx}^2\}(x,t)\,dx \le K(M,T) \max_{t\in[0,T]} \int_0^1 W_{xx}^2(x,t)\,dx, \tag{3.37}$$

where

$$K(M,T) := 8\bar{c}e^{\bar{c}T}(T^2+T)\bar{m}(T)\bar{\psi}(M)\cdot(M^2+1). \tag{3.38}$$

If M and T satisfy

$$T \le \min\left(1, \left(64\bar{c}e^{\bar{c}}\bar{m}(1)\bar{\psi}(M)\cdot(M^2+1)+1\right)^{-1}\right), \tag{3.39}$$

then $0 \le K(M,T) \le \frac{1}{4}$. Therefore, if M and T obey (3.25), (3.26), and (3.39), then S maps $\mathcal{X}_{M,T}$ into $\mathcal{X}_{M,T}$ and

$$\rho(Sw, S\hat{w}) \le \frac{1}{2}\rho(w,\hat{w}) \quad \forall w, \hat{w} \in \mathcal{X}_{M,T}. \tag{3.40}$$

□

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. It follows from Lemmas 3.1 and 3.2 and the contraction mapping principle that, for appropriately chosen M, $T>0$, S has a unique fixed point $u\in\mathcal{X}_{M,T}$ which is obviously a solution of (1.5), (1.6), (1.7) on $[0,T]$. Moreover, it is evident that for each $T'>0$, (1.5), (1.6), (1.7) has at most one solution which belongs to $\mathcal{C}^2_{T'}$.

Let $[0,T_0)$ be the maximal interval of existence for u such that (3.3) holds. It remains only to show that if (3.4) is satisfied then $T_0=\infty$. This can be done in the usual way: If (3.4) holds and $T_0<\infty$, we can reapply the contraction mapping principle to extend the solution to an interval $[0,T_0+\delta]$, $\delta>0$, contradicting the assumption that $[0,T_0)$ is maximal. This procedure involves changing the initial time and consequently requires modification of the forcing function, to account for the history of the solution prior to the new initial time.

For each $T'\in[0,T_0)$, let us define a new forcing function $f_{T'}$: $[0,1]\times[0,\infty)\to\mathbb{R}$ by

(3.41)

$$f_{T'}(x,t) := f(x,t+T') - \int_0^{T'} m(t+T'-\tau)\psi(u_x(x,\tau))_x d\tau, \qquad x\in[0,1], \quad t\ge 0,$$

and consider the initial-boundary value problem

$$(3.42)\qquad v_{tt}(x,t)=cv_{xx}(x,t)-\int_0^t m(t-\tau)\psi(v_x(x,\tau))_x d\tau+f_{T'}(x,t), \qquad x\in[0,1],\quad t\geq 0,$$

$$(3.43)\qquad v(0,t)=v(1,t)=0,\qquad t\geq 0,$$

$$(3.44)\qquad v(x,0)=u(x,T'),\qquad v_t(x,0)=u_t(x,T'),\qquad x\in[0,1].$$

Suppose that (3.4) holds and $T_0<\infty$. Then, $u(\cdot,T')$, $u_t(\cdot,T')$, and $f_{T'}$ obey bounds uniformly in $T'\in[0,T_0)$ which allow us to choose $T^*\in(0,T_0)$ such that for each $T'\in[0,T_0)$, (3.42), (3.43), (3.44) has a unique solution in $\mathcal{C}^2_{T^*}$. (The existence of such a T^* follows from Lemmas 3.1 and 3.2 and the contraction mapping principle. In particular (3.25), (3.26), and (3.39) reveal that T^* can be selected independently of $T'\in[0,T_0)$.) Now, set $T':=T_0-(T^*/2)$, let v denote the corresponding solution of (3.42), (3.43), (3.44), and define $\tilde{u}$: $[0,1]\times[0,T_0+T^*/2]\to\mathbb{R}$ by

$$(3.45)\qquad \tilde{u}(\cdot,t):=\begin{cases} u(\cdot,t), & t\in[0,T'),\\ v(\cdot,t-T'), & t\in[T',T_0+T^*/2].\end{cases}$$

By construction, $\tilde{u}$ is a solution of (1.5), (1.6), (1.7) on $[0,T_0+T^*/2]$, and by local uniqueness, $\tilde{u}$ extends u. This violates maximality of $[0,T_0)$. Therefore if (3.4) holds, $T_0=\infty$. □

It is not difficult to prove that the solution of (1.5), (1.6), (1.7) depends continuously on the initial data. To make this precise, let H denote the space $(H_0^1(0,1)\cap H^2(0,1))\times H_0^1(0,1)$, equipped with norm $\|\cdot\|_H$ defined by

$$(3.46)\qquad \|(\varphi,\theta)\|_H:=\left(\int_0^1\left[\varphi''(x)^2+\theta'(x)^2\right]dx\right)^{1/2}\quad \forall(\varphi,\theta)\in H.$$

THEOREM 3.2. *Assume that $c>0$, $\psi\in C^2(\mathbb{R})$, $m\in W^{1,1}_{\text{loc}}[0,\infty)$, and that f satisfies* (3.2). *Suppose that $u^*\in\mathcal{C}^2_{T^*}$ is a solution of* (1.5), (1.6) *on* $[0,T^*]$ *for some $T^*>0$. Then, there is a neighborhood $\mathcal{O}$ of $(u^*(\cdot,0),u_t^*(\cdot,0))$ in H such that for each $(u_0,u_1)\in\mathcal{O}$,* (1.5), (1.6), (1.7) *has a unique solution $u\in\mathcal{C}^2_{T^*}$. Moreover, for each $t\in[0,T^*]$, the mapping $(u_0,u_1)\mapsto(u(\cdot,t),u_t(\cdot,t))$ is continuous from $\mathcal{O}$ into H, the continuity being uniform in t.*

We first establish continuous dependence on an interval $[0,\hat{T}]$, where $\hat{T}$ is (possibly) smaller than T^*. We then use a stepping argument (if necessary) to reach T^*. The stepping procedure involves the "shifted problem" (3.42), (3.43), (3.44), and thus our local result must allow for variation of f. At the expense of possibly needing more steps to reach T^*, we prove local continuous dependence on an interval whose length cannot exceed one.

Let $\mathcal{F}$ denote the set of all functions f: $[0,1]\times[0,1]\to\mathbb{R}$ such that $f\in C([0,1];\mathrm{L}^2(0,1))$ and $f_t\in L^2([0,1];L^2(0,1))$, equipped with the norm $|||\cdot|||_{\mathcal{F}}$ defined by

$$(3.47)\qquad |||f|||_{\mathcal{F}}:=\left(\max_{t\in[0,1]}\int_0^1 f(x,t)^2dx+\int_0^1\int_0^1 f_t(x,t)^2dx\,dt\right)^{1/2}\quad \forall f\in\mathcal{F}.$$

For each $r>0$, let B_r and $\mathcal{B}_r$ denote the (open) ball of radius r in H and $\mathcal{F}$, respectively.

LEMMA 3.3. *Assume that $c>0$, $\psi\in C^2(\mathbb{R})$, $m\in W^{1,1}_{\text{loc}}[0,\infty)$, and let r, $R>0$ be given. Then, there exists $\hat{T}\in(0,1]$ such that for each $(u_0,u_1)\in B_r$ and each $f\in\mathcal{B}_R$,* (1.5), (1.6), (1.7) *has a unique solution $u\in\mathcal{C}^2_{\hat{T}}$. Moreover, for each $t\in[0,\hat{T}]$, the mapping $((u_0,u_1),f)\mapsto(u(\cdot,t),u_t(\cdot,t))$ is continuous from $B_r\times\mathcal{B}_R$ into H, the continuity being uniform in t.*

Proof. Examination of the proofs of Lemmas 3.1 and 3.2 reveals that we may choose $\hat{M}>0$ and $\hat{T}\in(0,1]$ such that for each $((u_0,u_1),f)\in B_r\times\mathcal{B}_R$, S maps $\mathcal{X}_{\hat{M},\hat{T}}$ into $\mathcal{X}_{\hat{M},\hat{T}}$ and satisfies

$$\rho(Sw,S\hat{w})\le\frac{1}{2}\rho(w,\hat{w})\quad\forall w,\ \hat{w}\in\mathcal{X}_{\hat{M},\hat{T}},\tag{3.48}$$

i.e., S is a uniform contraction. Thus, for each $((u_0,u_1),f)\in B_r\times\mathcal{B}_R$, S has a unique fixed point in $\mathcal{X}_{\hat{M},\hat{T}}$.

Moreover, for each $w\in\mathcal{X}_{\hat{M},\hat{T}}$, the mapping $((u_0,u_1),f)\mapsto Sw$ is continuous (in fact, Lipschitz continuous) from $B_r\times\mathcal{B}_R$ into $\mathcal{X}_{\hat{M},\hat{T}}$. (See Remark 3.1.). Since S is a uniform contraction, this implies that the fixed point of S also depends continuously on $((u_0,u_1),f)$. (See, for example, [5, Thm. 0.3.2].) □

Proof of Theorem 3.2. Set $l:=\max_{t\in[0,T^*]}\|(u^*(\cdot,t),u_t^*(\cdot,t))\|_H$ and choose $R>0$ large enough so that for each $T'\in[0,T^*]$ and each $u\in\mathcal{C}^2_{T'}$ with $\max_{t\in[0,T']}\|(u(\cdot,t),u_t(\cdot,t))\|_H\le l+1$, we have $|||f_{T'}|||_{\mathcal{F}}\le R$, where $f_{T'}$ is given by (3.41). Now, by Lemma 3.3, we may choose $\hat{T}\in(0,1]$ such that given any $T'\in[0,T^*]$ and any $u\in\mathcal{C}^2_{T'}$ with $((u(\cdot,T'),u_t(\cdot,T')),f_{T'})\in\mathcal{B}_{l+1}\times\mathcal{B}_R$, the initial-boundary value problem (3.42), (3.43), (3.44) has a unique solution $v\in\mathcal{C}^2_{\hat{T}}$. Moreover v depends continuously on $((u(\cdot,T'),u_t(\cdot,T')),f_{T'})$ in the appropriate topologies.

If $\hat{T}\ge T^*$, then we are done. If $\hat{T}<T^*$, we make steps of length $\hat{T}$ until T^* is reached, as follows. By virtue of continuous dependence on $[0,\hat{T}]$, we choose a neighborhood $\mathcal{O}_1$ of $(u^*(\cdot,0),u_t^*(\cdot,0))$ in H such that for each $(u_0,u_1)\in\mathcal{O}_1$, the corresponding solution u of (1.5), (1.6), (1.7) satisfies $\|(u(\cdot,t)-u^*(\cdot,t),u_t(\cdot,t)-u_t^*(\cdot,t))\|_H<1$ for all $t\in[0,\hat{T}]$, and hence $(u(\cdot,t),u_t(\cdot,t))\in B_{l+1}$ for all $t\in[0,\hat{T}]$. Now, set $T':=\hat{T}$ and note that for each $(u_0,u_1)\in\mathcal{O}_1$ we have $((u(\cdot,T'),u_t(\cdot,T'),f_{T'}))\in B_{l+1}\times\mathcal{B}_R$ (where u is the corresponding solution of (1.5), (1.6), (1.7) and $f_{T'}$ is defined in terms of u by (3.41)); therefore (3.41), (3.42), (3.43) has a unique solution $v\in\mathcal{C}^2_{\hat{T}}$ which depends continuously on its data. Using (3.45) to extend solutions of (1.5), (1.6), (1.7) onto $[0,2\hat{T}]$, we conclude that for each $(u_0,u_1)\in\mathcal{O}_1$, (1.5), (1.6), (1.7) has a unique solution $u\in\mathcal{C}^2_{2\hat{T}}$. Moreover, the mapping $(u_0,u_1)\mapsto(u(\cdot,t),u_t(\cdot,t))$ is continuous from $\mathcal{O}_1$ into H, uniformly in $t\in[0,2\hat{T}]$. This argument can be repeated (if $2\hat{T}<T^*$), choosing a smaller neighborhood $\mathcal{O}_2$ of $(u^*(\cdot,0),u_t^*(\cdot,0))$ for which solutions of (1.5), (1.6), (1.7) exist on $[0,3\hat{T}]$, etc.

After a finite number of applications of the above procedure, we eventually obtain a neighborhood $\mathcal{O}$ of (u_0,u_1) in H with the desired properties. This completes the proof of Theorem 3.2 □

As u_0, u_1 and f become smoother, the corresponding solution of (1.5), (1.6), (1.7) gets smoother. Moreover, the maximal interval of existence of a smoother solution is exactly the same as the maximal interval of existence of a solution with the regularity (3.3), i.e. a bound of the form (3.4) is also sufficient to continue a smoother solution globally. More precisely, we have

THEOREM 3.3. *Assume that* $c>0$, $\psi\in C^2(\mathbb{R})$, $m\in C^1[0,\infty)$, *and let* u_0, u_1, f *be given with*

$$u_0\in H_0^1(0,1)\cap H^3(0,1),\qquad u_1\in H_0^1(0,1)\cap H^2(0,1),\tag{3.49}$$

$$u_0''(0)=u_0''(1)=0,\tag{3.50}$$

$$f,f_t,f_x\in C([0,\infty);L^2(0,1)),\quad f_{tt}\in L^2_{\mathrm{loc}}([0,\infty);L^2(0,1)),\tag{3.51}$$

$$f(0,t)=f(1,t)=0\quad\forall t\ge0.\tag{3.52}$$

Then, (1.5), (1.6), (1.7) *has a unique solution* u, *defined on a maximal time interval* $[0,T_0)$, $T_0>0$, *with*

$$u,\, u_t,\, u_x,\, u_{tt},\, u_{tx},\, u_{xx},\, u_{ttt},\, u_{ttx},\, u_{txx},\, u_{xxx} \in C([0,T_0); L^2(0,1)). \tag{3.53}$$

In addition,

$$u_{xx}(0,t)=u_{xx}(1,t)=0 \quad \forall t\in[0,T_0), \tag{3.54}$$

and if (3.4) *holds then* $T_0=\infty$.

Remark 3.2. Under the supplementary smoothness assumptions of Theorem 3.3, the existence of a local solution with the regularity (3.53) follows[2] from Dafermos and Nohel [3, Thm. 2.1]. However, the statement that "(3.4) implies $T_0=\infty$", which is important for our purposes, is not valid for the more general equation studied in [3].

Remark 3.3. It is interesting to observe that no additional smoothness of ψ is required in Theorem 3.3. However, if we want to prove that the solution depends continuously on the initial data in the natural norm associated with (3.49), the assumption $\psi\in C^2(\mathbb{R})$ should be strengthened.

It is possible to prove Theorem 3.3 by using the local result in [3] to establish the existence of a solution u which satisfies (3.53), and then make estimates to show that (3.4) implies $T_0=\infty$. However, the estimates required for this purpose, when used in conjunction with standard regularity theory for the linear wave equation, actually yield the existence of a solution of (1.5), (1.6), (1.7) satisfying (3.53). Therefore, we provide a complete (but brief) proof of Theorem 3.3. The relevant regularity result for (3.5) is recorded below. To simplify the notation, we introduce an appropriate subclass of $\mathcal{C}_T^2$. For each $T>0$, let $\mathcal{C}_T^3$ denote the set of all functions w which belong to $\mathcal{C}_T^2$ and satisfy $w_{ttt}, w_{ttx}, w_{txx}, w_{xxx}\in C([0,T]; L^2(0,1))$.

PROPOSITION 3.2. *Assume that* $c>0$ *and that* (3.49), (3.50) *hold. Let* $T>0$ *and* g: $[0,1]\times[0,T]\to\mathbb{R}$ *be given with* $g, g_t, g_x\in C([0,T]; L^2(0,1))$, $g_{tt}\in L^2([0,T]; L^2(0,1))$, *and* $g(0,t)=g(1,t)=0$ *for all* $t\in[0,T]$. *Then, the solution* u *of* (3.5), (1.6), (1.7) *belongs to* $\mathcal{C}_T^3$ *and satisfies* $u_{xx}(0,t)=u_{xx}(1,t)=0$ *for all* $t\in[0,T]$. *Moreover*

$$\begin{aligned}\max_{t\in[0,T]}\int_0^1\{u_{ttt}^2+u_{ttx}^2+u_{txx}^2+u_{xxx}^2\}(x,t)\,dx &\le \bar{C}e^{\bar{C}T}\bigg\{\int_0^1\left[u_0'''(x)^2+u_1''(x)^2\right]dx \\ &\quad+\max_{t\in[0,T]}\int_0^1\{g_t^2+g_x^2\}(x,t)\,dx+\int_0^T\int_0^1 g_{tt}(x,t)^2\,dx\,dt\bigg\}\end{aligned} \tag{3.55}$$

where $\bar{C}$ *is a constant which depends only on* c.

Proof of Theorem 3.3. For $M,N,T>0$, let $\mathcal{X}_{M,T}^N$ denote the set of all functions w which belong to $\mathcal{X}_{M,T}$ and satisfy

$$w_{ttt}, w_{ttx}, w_{txx}, w_{xxx}\in L^\infty([0,T]; L^2(0,1)), \tag{3.56}$$

$$w_{xx}(0,t)=w_{xx}(1,t)=0 \quad \forall t\in[0,T], \tag{3.57}$$

and

$$\operatorname*{ess\text{-}sup}_{t\in[0,T]}\int_0^1\{w_{ttt}^2+w_{ttx}^2+w_{txx}^2+w_{xxx}^2\}(x,t)\,dx\le N^2. \tag{3.58}$$

[2] Different boundary conditions are used in [3], but the difference is not important.

We note that (3.6) and (3.56) imply $w \in C^2([0,1]\times[0,T])$ so that (3.57) is meaningful. Moreover, it follows from (3.57) and (3.58) that

$$(3.59)\qquad \max_{\substack{x\in[0,1]\\ t\in[0,T]}} |w_{xx}(x,t)| \le N \quad \forall w \in \mathcal{X}^N_{M,T}.$$

Clearly, $\mathcal{X}^N_{M,T}$ is nonempty for each $M,N,T>0$.

As before, we let S denote the map which carries w into the solution of (3.13), (1.6), (1.7). Our first project is to show that S maps $\mathcal{X}^N_{M,T}$ into itself for appropriately chosen M,N,T. Actually, we will show that if M and N are sufficiently large, then S maps $\mathcal{X}^N_{M,T}$ into itself provided that T is small enough relative to M. The fact that T does not need to be small relative to N plays the essential role in showing that (3.4) implies $T_0=\infty$.

Let $M,N,T>0$ and $w\in\mathcal{X}^N_{M,T}$ be given. Then, the function g defined by (3.14) satisfies g, g_t, $g_x \in C([0,T];L^2(0,1))$, $g_{tt}\in L^2([0,T];L^2(0,1))$, and $g(0,t)=g(1,t)=0$ for all $t\in[0,T]$. Thus, by Proposition 3.2, the corresponding solution u of (3.13), (1.6), (1.7) belongs to $\mathcal{C}^3_T$ and satisfies $u_{xx}(0,t)=u_{xx}(1,t)=0$ for all $t\in[0,T]$.

Clearly, all of the estimates derived in the proof of Lemma 3.1 remain valid under the present circumstances. We supplement these with bounds for third derivatives of u which we obtain from (3.55). It is particularly important to keep track of how these bounds depend on M,N,T, and the initial data, etc. For this purpose we define $\bar{f}^*$, $\bar{m}^*$: $[0,\infty)\to\mathbb{R}$ by

$$(3.60)\qquad \bar{f}^*(s) := \max_{t\in[0,s]} \int_0^1 \{f_t^2+f_x^2\}(x,t)\,dx + \int_0^s\int_0^1 f_{tt}^2(x,t)\,dx\,dt \quad \forall s\ge 0,$$

$$(3.61)\qquad \bar{m}^*(s) := \max_{t\in[0,s]} \{m(s)^2+m'(s)^2\} + \left(\int_0^s |m'(t)|\,dt\right)^2 \quad \forall s\ge 0,$$

set

$$(3.62)\qquad U_0^* := \int_0^1 \left[u_0'''(x)^2+u_1''(x)^2\right]dx,$$

and let $\bar{f},\bar{m},\bar{\psi}$, and U_0 be as in the proof of Lemma 3.1. Observe that $\bar{f}^*$ and $\bar{m}^*$ are continuous and nondecreasing, and that $\bar{m}^*(s)\ge\bar{m}(s)$ for all $s\ge 0$.

We conclude from (3.14) and a routine sequence of estimations that

$$(3.63)\qquad \max_{t\in[0,T]} \int_0^1 \{g_t^2+g_x^2\}(x,t)\,dx \\ \le 3\bar{f}^*(T)+3\bar{m}^*(T)\bar{\psi}(M)M^2+3T^2\bar{m}^*(T)\bar{\psi}(M)\cdot(M^2+1)N^2,$$

and

$$(3.64)\qquad \int_0^T\int_0^1 g_{tt}(x,t)^2\,dx\,dt \le 6\bar{f}^*(T)+18T\bar{m}^*(T)\bar{\psi}(M)\cdot(M^2+1)N^2.$$

It follows from (3.55), (3.63), and (3.64) that

$$(3.65)\qquad \max_{t\in[0,T]} \int_0^1 \{u_{ttt}^2+u_{ttx}^2+u_{txx}^2+u_{xxx}^2\}(x,t)\,dx \\ \le 18\bar{C}e^{\bar{C}T}\{U_0^*+\bar{f}^*(T)+\bar{m}^*(T)\bar{\psi}(M)M^2 \\ +(T^2+T)\bar{m}^*(T)\bar{\psi}(M)\cdot(M^2+1)N^2\}.$$

If M, N, and T satisfy

$$N^2 \geq 36\bar{C}e^{\bar{C}}\left(U_0^* + \bar{f}^*(1) + \bar{m}^*(1)\bar{\psi}(M)M^2\right) \tag{3.66}$$

and

$$T \leq \min\left(1, \left(72\bar{C}e^{\bar{C}}\bar{m}^*(1)\bar{\psi}(M)\cdot(M^2+1)+1\right)^{-1}\right), \tag{3.67}$$

then the right-hand side of (3.65) is dominated by N^2. Therefore, S maps $\mathcal{X}_{M,T}^N$ into itself if $M, N, T > 0$ obey (3.25), (3.26), (3.66), and (3.67). The crucial point here is that the right-hand side of (3.67) is independent of N.

The estimates derived in the proof of Lemma 3.2 also remain valid under the present circumstances. Thus S maps $\mathcal{X}_{M,T}^N$ into itself and

$$\rho(S\tilde{w}, S\hat{w}) \leq \frac{1}{2}\rho(\tilde{w}, \hat{w}) \quad \forall \tilde{w}, \hat{w} \in \mathcal{X}_{M,T}^N \tag{3.68}$$

if M, N, and T satisfy (3.25), (3.26), (3.39), (3.66), and (3.67). Moreover, by virtue of Alouglu's theorem and sequential weak $*$ lower semicontinuity of the norm in $L^\infty([0,T]; L^2(0,1))$, $\mathcal{X}_{M,T}^N$ is complete under the metric ρ. We are now ready to synthesize the proof.

Choose $M, N, T > 0$ such that S maps $\mathcal{X}_{M,T}^N$ into $\mathcal{X}_{M,T}^N$ and (3.67) holds. By the contraction mapping principle, S has a unique fixed point $u \in \mathcal{X}_{M,T}^N$ which is a solution of (1.5), (1.6), (1.7) on $[0,T]$. Examining (3.25), (3.26), (3.39), (3.66), and (3.67), we see that the length of the time interval on which we can produce a solution by this procedure depends on U_0 and $\bar{f}$, but not on U_0^* or $\bar{f}^*$. Moreover, it is clear that u belongs to $\mathcal{C}_T^3$ and that for each $T' > 0$ there is at most one solution in $\mathcal{C}_{T'}^3$.

Let $[0, T_0)$ be the maximal interval of existence of u such that (3.53) is satisfied. Gronwall's inequality, (1.5), and (3.52) imply (3.54). For each $T' \in [0, T_0)$, let $f_{T'}$ be given by (3.41).

Suppose now that (3.4) holds and $T_0 < \infty$. Then, $u(\cdot, T')$, $u_t(\cdot, T')$, and $f_{T'}$ are sufficiently smooth, satisfy the appropriate boundary conditions, and obey bounds uniformly in $T' \in [0, T_0)$ such that we may choose $T^* > 0$ with the property that (3.42), (3.43), (3.44) has a unique solution in $\mathcal{C}_{T^*}^3$ for each $T' \in [0, T_0)$. The construction used at the end of the proof of Theorem 3.1 yields a smooth continuation of u, which violates maximality of $[0, T_0)$. □

4. Global existence and decay of solutions. We now discuss global behavior of solutions of (1.5) and (2.10). Throughout this section we use $\bar{f}, \bar{m}$ and U_0 as defined by (3.15), (3.16), and (3.18), respectively. We begin by deriving an a priori bound which can be used (in conjunction with the results of §3) to prove Theorems 2.1 and 2.2.

LEMMA 4.1. *Let the assumptions of Theorem* 3.1 *hold and let u denote the corresponding solution of* (1.5), (1.6), (1.7) *on the maximal interval* $[0, T_0)$. *If*

$$\sup_{\substack{x \in [0,1] \\ t \in [0, T_0)}} |\psi'(u_x(x,t))| < \infty, \tag{4.1}$$

then for each $T \in (0, T_0)$,

(4.2)

$$\max_{t \in [0,T]} \int_0^1 \{u_{tt}^2 + u_{tx}^2 + u_{xx}^2\}(x,t)\,dx \leq \tilde{c}\left(U_0 + \bar{f}(T)\right)\exp\left\{\tilde{c}\left[\Lambda\bar{m}(T)\cdot(T^2+T)+T\right]\right\},$$

where

$$\Lambda := \sup_{\substack{x\in[0,1]\\ t\in[0,T_0)}} \psi'(u_x(x,t))^2 \tag{4.3}$$

and $\tilde c$ is a positive constant which depends only on c.

Proof. Let $T\in(0,T_0)$ be given. Observe that u satisfies (3.5) with g defined by

$$g(x,t) := f(x,t) - \int_0^t m(t-\tau)\psi(u_x(x,\tau))_x d\tau, \qquad x\in[0,1], \quad t\in[0,T]. \tag{4.4}$$

The procedure used to derive (3.20) and (3.23) yields

$$\int_0^1 g(x,t)^2 dx \le 2\bar f(T) + 2\Lambda T\bar m(T)\int_0^t\int_0^1 u_{xx}^2(x,s)\,dx\,ds \quad \forall t\in[0,T], \tag{4.5}$$

and

$$\int_0^t\int_0^1 g_t(x,s)^2 dx\,ds \le 3\bar f(T) + 6\Lambda\bar m(T)\int_0^t\int_0^1 u_{xx}^2(x,s)\,dx\,ds \quad \forall t\in[0,T]. \tag{4.6}$$

Therefore, we deduce from (3.7) that

$$\begin{aligned}\int_0^1 \{u_{tt}^2 + u_{tx}^2 + u_{xx}^2\}(x,t)\,dx \le \bar c U_0 + 5\bar c\bar f(T) \\ + 3\bar c\Lambda\bar m(T)\cdot(T+2)\int_0^t\int_0^1 u_{xx}^2(x,s)\,dx\,ds \\ + \bar c\int_0^t\int_0^1 u_{tt}^2(x,s)\,dx\,ds \quad \forall t\in[0,T].\end{aligned} \tag{4.7}$$

The desired result follows from Gronwall's inequality and (4.7). In fact, we can use $\tilde c := 6\bar c$. □

As an immediate consequence of Lemma 4.1, we have the following corollary.

COROLLARY 4.1. *Let the assumptions of Theorem* 3.1 *or Theorem* 3.3 *hold and let u and T_0 be as in Theorem* 3.1 *or* 3.3, *respectively. If* (4.1) *holds, then $T_0=\infty$. In particular, if either*

$$\sup_{\xi\in\mathbb{R}} |\psi'(\xi)| < \infty, \tag{4.8}$$

or

$$\sup_{\substack{x\in[0,1]\\ t\in[0,T_0)}} |u_x(x,t)| < \infty, \tag{4.9}$$

then $T_0=\infty$.

Proof. Suppose that (4.1) holds and $T_0<\infty$. It then follows from Lemma 4.1 that (1.4) holds. This implies that $T_0=\infty$, which is a contradiction. Therefore, if (4.1) holds, $T_0=\infty$. □

Thus we have proved Theorems 2.1 and 2.2. As mentioned previously, the estimates required to prove Theorem 2.3 are considerably more involved. The following observation allows us to simplify some of the computations.

Remark 4.1. To prove Theorem 2.3, there is no loss in assuming that $\mu=1$. Indeed, if $\mu>0$ we can always convert (2.10), (1.6), (1.7) into an equivlaent problem of the same form which has $\mu=1$ by a linear rescaling of time. If the original equation satisfies the assumptions of Theorem 2.3, then so will the modified equation. Moreover, the new initial data will have precisely the same smoothness as the original initial data.

Therefore, we consider

$$u_{tt}(x,t)=cu_{xx}(x,t)-\int_0^t e^{-(t-\tau)}\psi(u_x(x,\tau))_x d\tau, \qquad 0\leq x\leq 1, \quad t\geq 0, \tag{4.10}$$

in place of (2.10). Formally differentiating (4.10) with respect to t and substituting for the integral term from (4.10) we obtain

$$u_{ttt}(x,t)+u_{tt}(x,t)=cu_{xxt}(x,t)+\chi(u_x(x,t))_x, \qquad 0\leq x\leq t, \quad t\geq 0, \tag{4.11}$$

where the equilibrium stress function χ is now given by

$$\chi(\xi):=c\xi-\psi(\xi) \quad \forall\xi\in\mathbb{R}. \tag{4.12}$$

We first establish an estimate of the form (2.14) for solutions of (4.10) under supplementary smoothness assumptions on the initial data. The additional smoothness permits us to use equation (4.11) to derive energy identities. The extra assumptions on the data will then be removed by using Theorem 3.2 and a density argument.

LEMMA 4.2. *Assume that $\psi\in C^2(\mathbb{R})$ and that there are constants α and β such that*

$$0<\alpha\leq\psi'(\xi)\leq\beta<c \quad \forall\xi\in\mathbb{R}. \tag{4.13}$$

Then there exist constants Γ, $\delta>0$ such that for every u_0, u_1 which satisfy

$$u_0\in H_0^1(0,1)\cap H^3(0,1), \qquad u_1\in H_0^1(0,1)\cap H^2(0,1), \tag{4.14}$$

$$u_0''(0)=u_0''(1)=0, \tag{4.15}$$

the corresponding solution u of (4.10), (1.6), (1.7) *satisfies*

$$\int_0^1\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,t)\,dx\leq\Gamma e^{-\delta t}U_0 \quad \forall t\geq 0. \tag{4.16}$$

The proof of this lemma involves the analysis of several energy identities which we obtain from (4.11). The "potential function" Φ defined by

$$\Phi(\xi):=2\int_0^\xi\chi(s)\,ds \quad \forall\xi\in\mathbb{R} \tag{4.17}$$

arises quite naturally in these identities. To establish positivity of certain combinations of terms, we make crucial use of several properties of ψ,χ, and Φ which follow from (4.13). In particular, we require the following elementary proposition.

PROPOSITION 4.1. *Assume that $\psi\in C^1(\mathbb{R})$, $\psi(0)=0$, and that* (4.13) *holds for some α and β. Then, there exist $\varepsilon\in(0,\alpha/c)$ and $\lambda>1/c$ such that*

$$(1-\varepsilon)\Phi(\xi)-\lambda\chi(\xi)^2\geq 0 \quad \forall\xi\in\mathbb{R}, \tag{4.18}$$

where χ and Φ are defined in terms of ψ by (4.12) *and* (4.17), *respectively.*

Proof. Observe that $\chi(0)=\Phi(0)=0$, $s\chi(s)\geq 0$ for all $s\in\mathbb{R}$, and $\chi'(s)\leq c-\alpha$ for all $s\in\mathbb{R}$. Thus, we have

$$2\chi(s)\chi'(s)\leq 2(c-\alpha)\chi(s) \quad \forall s\geq 0, \tag{4.19}$$

$$2\chi(s)\chi'(s)\geq 2(c-\alpha)\chi(s) \quad \forall s\leq 0. \tag{4.20}$$

Integration of these inequalities from 0 to ξ yields

$$(4.21)\qquad \chi(\xi)^2 \le (c-\alpha)\Phi(\xi) \quad \forall \xi \in \mathbb{R}.$$

Consequently, for each $\varepsilon \in (0,1)$

$$(4.22)\qquad (1-\varepsilon)\Phi(\xi) - \frac{(1-\varepsilon)}{(c-\alpha)}\chi(\xi)^2 \ge 0 \quad \forall \xi \in \mathbb{R}.$$

Now, choose $\varepsilon \in (0, \alpha/c)$ and set $\lambda := (1-\varepsilon)(c-\alpha)^{-1}$. A simple computation shows that $\lambda > 1/c$. The conclusion thus follows from (4.22). □

Proof of Lemma 4.2. We assume without loss of generality that $\psi(0)=0$. Let u_0 and u_1 satisfying (4.14), (4.15) be given and let u denote the corresponding solution of (4.10), (1.6), (1.7). It follows from Theorem 2.2 that $u \in C^2([0,1]\times[0,\infty))$ and u_{ttt}, u_{ttx}, u_{txx}, $u_{xxx} \in C([0,\infty); L^2(0,1))$. We remark that this degree of regularity is sufficient to justify the computations which follow. In particular, u satisfies (4.11).

Let δ be an arbitrary positive constant (to be specified later). We multiply (4.11) by $2e^{\delta t}u_{tt}(x,t)$ and integrate the resulting expression over space and time, using integration by parts and exploiting the boundary conditions, to arrive at

$$(4.23)\qquad \int_0^1 e^{\delta t}\{u_{tt}^2 + cu_{xt}^2 + 2\chi(u_x)u_{xt}\}(x,t)\,dx$$
$$+2\int_0^t\int_0^1 e^{\delta s}\{u_{tt}^2 - \chi'(u_x)u_{xt}^2\}(x,s)\,dx\,ds$$
$$= \int_0^1 \left[cu_0''(x)^2 + cu_1'(x)^2 + 2\chi(u_0'(x))u_1'(x)\right]dx$$
$$+\delta\int_0^t\int_0^1 e^{\delta s}\{u_{tt}^2 + cu_{xt}^2 + 2\chi(u_x)u_{xt}\}(x,s)\,dx\,ds \quad \forall t \in [0,\infty).$$

(Note that $u_{tt}(\cdot,0) = cu_0''(\cdot)$ by (4.10).) To obtain our next identity, we multiply (4.11) by $2e^{\delta t}u_t(x,t)$ and integrate as above. The outcome of this computation is

$$(4.24)\qquad \int_0^1 e^{\delta t}\{u_t^2 + \Phi(u_x) + 2u_t u_{tt}\}(x,t)\,dx$$
$$+2\int_0^t\int_0^1 e^{\delta s}\{cu_{xt}^2 - u_{tt}^2\}(x,s)\,dx\,ds$$
$$= \int_0^1 \left[u_1(x)^2 + \Phi(u_0'(x)) + 2cu_1(x)u_0''(x)\right]dx$$
$$+\delta\int_0^t\int_0^1 e^{\delta s}\{u_t^2 + \Phi(u_x) + 2u_t u_{tt}\}(x,s)\,dx\,ds \quad \forall t \in [0,\infty).$$

Making use of Proposition 4.1, we can construct a linear combination of the left-hand sides of (4.23) and (4.24) which is positive definite. For this purpose, we choose $\varepsilon \in (0,\alpha/c)$ and $\lambda > 1/c$ such that (4.18) holds. In what follows, we use Γ to denote a generic positive which can be chosen independently of δ and U_0.

We multiply (4.24) by $(1-\varepsilon)$ and add the resulting expression to (4.23). (Note that $1-\varepsilon > 0$.) After majorizing the right-hand side and rearranging certain terms, we obtain

$$\begin{aligned}(4.25)\qquad &\int_0^1 e^{\delta t}\{u_{tt}^2+2(1-\varepsilon)u_tu_{tt}+(1-\varepsilon)u_t^2\}(x,t)\,dx\\ &+\int_0^1 e^{\delta t}\{cu_{tx}^2+2\chi(u_x)u_{tx}+(1-\varepsilon)\Phi(u_x)\}(x,t)\,dx\\ &+2\int_0^t\int_0^1 e^{\delta s}\{\varepsilon u_{tt}^2+(\psi'(u_x)-c\varepsilon)u_{tx}^2\}(x,s)\,dx\,ds\\ &\le \Gamma U_0+\delta\Gamma\int_0^t\int_0^1 e^{\delta s}\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,s)\,dx\,ds\quad \forall t\in[0,\infty).\end{aligned}$$

The first and third integrals on the left-hand side of (4.25) can be bounded from below rather easily; the first by completing the square and the third since $c\varepsilon<\alpha\le\psi'(\xi)$ for all $\xi\in\mathbb{R}$. The second integral on the left-hand side of (4.25) deserves special attention. Observe that for each $\eta>0$, we have

$$(4.26)\qquad |2\chi(u_x)u_{xt}|\le\eta\chi(u_x)^2+(1/\eta)u_{xt}^2.$$

Now, using (4.18) and (4.26) with $\eta:=\lambda$, we deduce that

$$(4.27)\qquad cu_{xt}^2+2\chi(u_x)u_{xt}+(1-\varepsilon)\Phi(u_x)\ge(c-1/\lambda)u_{xt}^2.$$

Since $(c-1/\lambda)>0$, this yields a lower bound for the integral in question. We thus have an estimate of the form

$$\begin{aligned}(4.28)\qquad &\int_0^1 e^{\delta t}\{u_{tt}^2+u_{tx}^2\}(x,t)\,dx+\int_0^t\int_0^1 e^{\delta s}\{u_{tt}^2+u_{tx}^2\}(x,s)\,dx\,ds\\ &\le\Gamma U_0+\delta\Gamma\int_0^t\int_0^1 e^{\delta s}\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,s)\,dx\,ds\quad\forall t\in[0,\infty).\end{aligned}$$

Multiplying (4.11) by $2e^{\delta t}u_{xx}(x,t)$ and integrating over space and time as before produces the identity

$$\begin{aligned}(4.29)\qquad &\int_0^1 e^{\delta t}\{cu_{xx}^2-u_{tx}^2-2u_{tt}u_{xx}\}(x,t)\,dx\\ &+2\int_0^t\int_0^1 e^{\delta s}\{\chi'(u_x)u_{xx}^2-u_{tt}u_{xx}\}(x,s)\,dx\,ds\\ &=\int_0^1\left[cu_0''(x)^2-u_1'(x)^2-2cu_0''(x)^2\right]dx\\ &+\delta\int_0^t\int_0^1 e^{\delta s}\{cu_{xx}^2-u_{tx}^2-2u_{tt}u_{xx}\}(x,s)\,dx\,ds\quad\forall t\in[0,\infty).\end{aligned}$$

For each $\eta>0$, we have

$$(4.30)\qquad |2u_{tt}u_{xx}|\le\eta u_{xx}^2+(1/\eta)u_{tt}^2.$$

Recall that $\chi'(\xi)\ge c-\beta>0$ for all $\xi\in\mathbb{R}$. Therefore, if we apply (4.30) with $\eta\in(0,c-\beta]$ to (4.29) and combine the resulting inequality with (4.28), we finally arrive at an estimate of the form

$$\begin{aligned}(4.31)\qquad &\int_0^1 e^{\delta t}\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,t)\,dx+\int_0^t\int_0^1 e^{\delta s}\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,s)\,dx\,ds\\ &\le\Gamma U_0+\delta\Gamma\int_0^t\int_0^1 e^{\delta s}\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,s)\,dx\,ds\quad\forall t\in[0,\infty).\end{aligned}$$

The desired result is obtained by choosing δ sufficiently small in (4.31). □

It follows from Theorem 3.2 and a simple density argument that Lemma 4.2 remains valid for initial data $(u_0,u_1)\in H$. (Recall that $H:=(H_0^1(0,1)\cap H^2(0,1))\times H_0^1(0,1)$ equipped with norm given by (3.46).) Indeed, the set of all (u_0,u_1) which satisfy (4.14) and (4.15) is dense in H. Moreover, by Theorem 3.2 the mapping $(u_0,u_1)\mapsto(u(\cdot,t),u_t(\cdot,t))$ is continuous from H into H for each $t>0$. Since u satisfies (4.10), this implies that the mapping $(u_0,u_1)\mapsto u_{tt}(\cdot,t)$ is continuous from H into $L^2(0,1)$ for each $t>0$. Therefore, (4.16) holds under the weaker assumption $(u_0,u_1)\in H$.

By virtue of the Poincaré inequalities and the boundary conditions, (4.16) and (2.14) are equivalent. The uniform decay estimate (2.15) follows immediately from (2.14) and the Sobolev embedding theorem. In view of Remark 4.1, we have now proved Theorem 2.3.

It is not difficult to modify the preceding argument to establish decay of solutions of

(4.32)
$$u_{tt}(x,t)=cu_{xx}(x,t)-\int_0^t e^{-\mu(t-\tau)}\psi(u_x(x,\tau))_x\,d\tau+f(x,t),\qquad 0\le x\le 1,\quad t\ge 0,$$

provided that f is sufficiently smooth and $f(\cdot,t)$ decays suitably as $t\to\infty$. If the assumptions of Theorem 2.3 hold and there exists a constant $\gamma>0$ such that

$$\int_0^\infty\int_0^1 e^{\gamma s}\{f^2+f_t^2\}(x,s)\,dx\,ds<\infty, \tag{4.33}$$

then the procedure used to prove Lemma 4.2 yields an estimate of the form

$$\begin{aligned}\int_0^1\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,t)\,dx \\ \le\Gamma e^{-\delta t}\left(U_0+\int_0^\infty\int_0^1 e^{\gamma s}\{f^2+f_t^2\}(x,s)\,dx\,ds\right)\quad \forall t\ge 0\end{aligned} \tag{4.34}$$

for solutions of (4.32), (1.6), (1.7). Here Γ, $\delta>0$ are constants which are independent of u_0, u_1, and f. (Necessarily, $\delta\le\gamma$.)

If the assumptions of Theorem 2.3 hold and $f, f_t\in L^2([0,1]\times[0,\infty))$, i.e.

$$\int_0^\infty\int_0^1\{f^2+f_t^2\}(x,s)\,dx\,ds<\infty, \tag{4.35}$$

then one can show that solutions u of (4.32), (1.6), (1.7) satisfy

$$u(\cdot,t),\,u_t(\cdot,t),\,u_x(\cdot,t)\ \underset{[0,1]}{\overset{\text{unif}}{\to}}\ 0\quad \text{as } t\to\infty. \tag{4.36}$$

The basic idea of the proof is to follow the procedure used to prove Lemma 4.2, but with $\delta=0$. This leads to a bound of the form

$$\begin{aligned}\sup_{s\in[0,\infty)}\int_0^1\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,s)\,dx+\int_0^\infty\int_0^1\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,s)\,dx\,ds \\ \le\Gamma\left(U_0+\int_0^\infty\int_0^1\{f^2+f_t^2\}(x,s)\,dx\,ds\right)\end{aligned} \tag{4.37}$$

where Γ is a positive constant which is independent of u_0, u_1, and f. Standard embedding inequalities, (1.6), and (4.37) imply (4.36).

5. Other boundary conditions. With only minor revisions, the arguments used in the preceding sections remain valid for other types of boundary conditions. For easy reference, we have collected here several analogues of Theorems 2.1 through 2.3. In order to avoid repetition in the statements of these results, we record several of the common hypotheses below.

(H1): $c>0$, $\psi\in C^2(\mathbb{R})$, $\psi'\in L^\infty(\mathbb{R})$, $m\in W^{1,1}_{\mathrm{loc}}[0,\infty)$.

(H2): $c>0$, $\psi\in C^2(\mathbb{R})$, $\psi'\in L^\infty(\mathbb{R})$, $m\in C^1[0,\infty)$.

(H3): $c>0$, $\psi\in C^2(\mathbb{R})$, and there exist constants α and β such that $0<\alpha\leq\psi'(\xi)\leq\beta<\mu c$ for all $\xi\in\mathbb{R}$.

A. *Neumann conditions.* If, in place of (1.6), we impose the boundary conditions

$$u_x(0,t)=u_x(1,t)=0,\qquad t\geq 0, \tag{5.1}$$

then nontrivial rigid motions are possible. This, of course, affects the asymptotic behavior of solutions. It is convenient to first analyze (5.1) with the data normalized so as to eliminate the possibility of nontrivial rigid motions. The general problem can always be reduced to one with normalized data by superposition of a rigid motion. The appropriate conditions read

$$\int_0^1 u_0(x)\,dx=\int_0^1 u_1(x)\,dx=0, \tag{5.2}$$

$$\int_0^1 f(x,t)\,dx=0\quad\forall t\geq 0. \tag{5.3}$$

Observe that (5.2) and (5.3) imply that solutions of (1.5), (5.1), (1.7) will have zero average spatially.

THEOREM 2.1A. *Assume that* (H1) *holds and let* u_0, u_1, *and* f *be given with*

$$u_0\in H^2(0,1),\qquad u_1\in H^1(0,1), \tag{5.4}$$

$$u_0'(0)=u_0'(1)=0, \tag{5.5}$$

$$f\in C([0,\infty);L^2(0,1)),\qquad f_t\in L^2_{\mathrm{loc}}([0,\infty);L^2(0,1)). \tag{5.6}$$

Assume further that (5.2) *and* (5.3) *hold. Then, the initial-boundary value* (1.5), (5.1), (1.7) *has a unique solution* u: $[0,1]\times[0,\infty)\to\mathbb{R}$ *with*

$$u,\ u_t,\ u_x,\ u_{tt},\ u_{tx},\ u_{xx}\in C([0,\infty);L^2(0,1)). \tag{5.7}$$

Moreover, u has zero average spatially, i.e.,

$$\int_0^1 u(x,t)\,dx=0\quad\forall t\geq 0. \tag{5.8}$$

THEOREM 2.2A. *Assume that* (H2) *holds and let* u_0, u_1, *and* f *be given with*

$$u_0\in H^3(0,1),\qquad u_1\in H^2(0,1), \tag{5.9}$$

$$u_0'(0)=u_0'(1)=u_1'(0)=u_1'(0)=0, \tag{5.10}$$

$$f,f_t,f_x\in C([0,\infty);L^2(0,1)),\qquad f_{tt}\in L^2_{\mathrm{loc}}([0,\infty);L^2(0,1)). \tag{5.11}$$

Assume further that (5.2) *and* (5.3) *hold. Then, the solution u in Theorem* 2.2A *has the additional regularity*

$$u_{ttt},\ u_{ttx},\ u_{txx},\ u_{xxx}\in C([0,\infty);L^2(0,1), \tag{5.12}$$

whence $u\in C^2([0,1]\times[0,\infty))$.

THEOREM 2.3A. *Assume that* (H3) *holds. Then there are constants* Γ, $\delta>0$ *such that for every* u_0, u_1 *satisfying* (5.2), (5.4), *and* (5.5), *the corresponding solution* u *of* (2.10), (5.1), (1.7) *satisfies*

(5.13)

$$\int_0^1\{u^2+u_t^2+u_x^2+u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,t)\,dx\le\Gamma e^{-\delta t}\int_0^1\left[u_0''(x)^2+u_1'(x)^2\right]dx\quad\forall t\ge0,$$

and hence also

$$\max_{x\in[0,1]}(u^2+u_t^2+u_x^2)(x,t)\le\Gamma e^{-\delta t}\int_0^1\left[u_0''(x)^2+u_1'(x)^2\right]dx\quad\forall t\ge0.\tag{5.14}$$

The proofs of these results are virtually identical to the proofs of Theorems 2.1, 2.2, and 2.3. Therefore, we merely point out the necessary changes.

In the definition of $\mathcal{X}_{M,T}$, the condition

$$\int_0^1 w(x,t)\,dx=0\quad\forall t\in[0,T]\tag{5.15}$$

should be added and (3.9) should be replaced by

$$w_x(0,t)=w_x(1,t)=0\quad\forall t\in[0,T].\tag{5.16}$$

Moreover, (3.57) should be dropped from the definition of $\mathcal{X}_{M,T}^N$. With these modifications all of the estimates used in §§3 and 4 remain valid. In particular, the boundary contributions are once again annihilated in the integrations by parts used to derive (4.23), (4.24) and (4.29).

We now show how to reduce a problem with "un-normalized" data to one with normalized data. Given u_0, u_1, and f satisfying (5.4), (5.5), and (5.6), define $\bar{u}_0$, $\bar{u}_1\in\mathbb{R}$ and $\bar{u}$: $[0,\infty)\to\mathbb{R}$ by

$$\bar{u}_0:=\int_0^1 u_0(x)\,dx,\qquad \bar{u}_1:=\int_0^1 u_1(x)\,dx,\tag{5.17}$$

$$\bar{u}(t):=\bar{u}_0+t\bar{u}_1+\int_0^t\int_0^s\int_0^1 f(x,\tau)\,dx\,d\tau\,ds\quad\forall t\ge0.\tag{5.18}$$

Then, define $\tilde{u}_0$, $\tilde{u}_1$: $[0,1]\to\mathbb{R}$, $\tilde{f}$: $[0,1]\times[0,\infty)\to\mathbb{R}$ by

$$\tilde{u}_0(x):=u_0(x)-\bar{u}_0,\qquad \tilde{u}_1(x):=\bar{u}_1(x)-u_1,\qquad x\in[0,1],\tag{5.19}$$

$$\tilde{f}(x,t):=f(x,t)-\int_0^1 f(x,t)\,dx,\tag{5.20}$$

and consider the initial-boundary value problem

$$\tilde{u}_{tt}(x,t)=c\tilde{u}_{xx}(x,t)-\int_0^t m(t-\tau)\psi(\tilde{u}_x(x,\tau))_x\,d\tau+\tilde{f}(x,t),\qquad 0\le x\le1,\quad t\ge0,\tag{5.21}$$

$$\tilde{u}_x(0,t)=\tilde{u}_x(1,t)=0,\qquad t\ge0,\tag{5.22}$$

$$\tilde{u}(x,0)=\tilde{u}_0(x),\ \tilde{u}_t(x,0)=\tilde{u}_1(x),\qquad 0\le x\le1.\tag{5.23}$$

Clearly, $\tilde{u}_0$, $\tilde{u}_1$, and $\tilde{f}$ have zero average spatially. Thus, the above results are applicable to (5.21), (5.22), (5.23). The solution u of (1.5), (5.1), (1.7) is determined from the solution $\tilde{u}$ of (5.21), (5.22), (5.23) by the formula

$$u(x,t)=\tilde{u}(x,t)+\bar{u}(t), \qquad 0\le x\le 1,\, t\ge 0. \tag{5.24}$$

B. *Mixed conditions.* Mixed boundary conditions such as

$$u(0,t)=u_x(1,t)=0, \qquad t\ge 0, \tag{5.25}$$

can also be handled. In this case no nontrivial rigid motions are possible, so there is no need to normalize the data. Making obvious amendments in the proofs, it is straightforward to establish the following analogues of Theorem 2.1 through 2.3. (Clearly, similar results hold for the boundary conditions $u_x(0,t)=u(1,t)=0$, $t\ge 0$.)

THEOREM 2.1B. *Assume that* (H1) *holds and let* u_0, u_1, *and* f *satisfying* (5.4), (5.6), *and*

$$u_0(0)=u_1(0)=u_0'(1)=0 \tag{5.26}$$

be given. Then, the initial-boundary value problem (1.5), (5.25), (1.7) *has a unique solution* u *which satisfies* (5.7).

THEOREM 2.2B. *Assume that* (H2) *holds and let* u_0, u_1 *and* f *satisfying* (5.9), (5.11),

$$u_0(0)=u_1(0)=u_0'(1)=u_1'(1)=0, \tag{5.27}$$

and

$$u_0''(0)=0, \qquad f(0,t)=0 \quad \forall t\ge 0 \tag{5.28}$$

be given. Then, the solution u *in Theorem* 2.1B *satisfies* (5.12), *and consequently* $u\in C^2([0,1]\times[0,\infty))$.

Remark 5.1. Theorem 2.2B remains valid if (5.28) is replaced by the weaker compatibility assumption

$$cu_0''(0)+f(0,0)=0. \tag{5.29}$$

The proof in this situation requires estimation of certain boundary terms which automatically vanish if (5.28) holds.

THEOREM 2.3B. *Assume that* (H3) *holds. Then, there exist constants* Γ, $\delta>0$ *such that for every* u_0, u_1 *satisfying* (5.4) *and* (5.26), *the corresponding solution of* (2.10), (5.25), (1.7) *satisfies* (5.13) *and* (5.14).

C. *Cauchy problems.* We conclude with a few remarks regarding pure initial value problems. In place of (1.5) and (2.10), we now consider

(5.30)

$$u_{tt}(x,t)=cu_{xx}(x,t)-\int_0^t m(t-\tau)\psi(u_x(x,\tau))_x\,d\tau+f(x,t), \qquad -\infty<x<\infty, \quad t\ge 0,$$

and

$$u_{tt}(x,t)=cu_{xx}(x,t)-\int_0^t e^{-\mu(t-\tau)}\psi(u_x(x,\tau))_x\,d\tau, \qquad -\infty<x<\infty, \quad t\ge 0, \tag{5.31}$$

together with initial conditions

$$u(x,0)=u_0(x), \qquad u_t(x,0)=u_1(x), \qquad -\infty<x<\infty. \tag{5.32}$$

Due to the lack of Poincaré inequalities on all of space, additional estimates are now required for certain lower order derivatives. It is not too difficult to prove analogues of

Theorems 2.1 and 2.2; however, a great deal of information concerning the decay of solutions is lost when the spatial region is unbounded.

THEOREM 2.1C. *Assume that* (H1) *holds and let* u_0, u_1, f *be given with*

$$u_0 \in H^2(\mathbb{R}), \qquad u_1 \in H^1(\mathbb{R}), \tag{5.33}$$

$$f \in C([0,\infty); L^2(\mathbb{R})), \qquad f_t \in L^2_{\mathrm{loc}}([0,\infty); L^2(\mathbb{R})). \tag{5.34}$$

Then, the initial value problem (5.30), (5.32) *has a unique solution* u: $(-\infty,\infty)\times[0,\infty)\to\mathbb{R}$ *with*

$$u, u_t, u_x, u_{tt}, u_{tx}, u_{xx} \in C([0,\infty); L^2(\mathbb{R})). \tag{5.35}$$

THEOREM 2.2C. *Assume that* (H2) *holds and let* u_0, u_1, f *be given with*

$$u_0 \in H^3(\mathbb{R}), \qquad u_1 \in H^2(\mathbb{R}), \tag{5.36}$$

$$f, f_t, f_x \in C([0,\infty); L^2(\mathbb{R})), \qquad f_{tt} \in L^2_{\mathrm{loc}}([0,\infty); L^2(\mathbb{R})). \tag{5.37}$$

Then, the solution u in Theorem 2.1C *has the additional regularity*

$$u_{ttt}, u_{ttx}, u_{txx}, u_{xxx} \in C([0,\infty); L^2(\mathbb{R})), \tag{5.38}$$

whence $u \in C^2(\mathbb{R}\times[0,\infty))$.

The procedure used to prove Lemma 4.2 also yields a result in the same spirit for the Cauchy problem (5.31), (5.32); however the conclusion is substantially weaker. In particular, we can no longer claim exponential decay.

THEOREM 2.3C. *Assume that* (H3) *holds. Then, there exists a positive constant* Γ *such that for every* $u_0 \in H^2(\mathbb{R})$, $u_1 \in H^2(\mathbb{R})$, *the corresponding solution u of* (5.31), (5.32) *satisfies*

$$\int_{-\infty}^{\infty}\{u_t^2+u_x^2+u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,t)\,dx+\int_0^t\int_{-\infty}^{\infty}\{u_{tt}^2+u_{tx}^2+u_{xx}^2\}(x,s)\,dx\,ds \tag{5.39}$$
$$\leq \Gamma\int_{-\infty}^{\infty}\left[u_1(x)^2+u_1'(x)^2+u_0'(x)^2+u_0''(x)^2\right]dx \quad \forall t\geq 0.$$

To prove Theorem 2.3C, we assume without loss that $\mu=1$ and proceed as in the proof of Lemma 4.2, but with $\delta=0$, and replace integrations over [0, 1] with integrations over $(-\infty,\infty)$. Because of the lack of Poincaré inequalities, more care must be used to combine the energy identities. For this purpose, we employ the following straightforward refinement of Proposition 4.1.

PROPOSITION 5.1. *Under the assumptions of Proposition* 4.1, *there exist constants* $\nu>0$, $\varepsilon\in(0,\alpha/c)$, *and* $\lambda>1/c$ *such that*

$$(1-\varepsilon)\Phi(\xi)-\lambda\chi(\xi)^2\geq\nu\xi^2 \quad \forall\xi\in\mathbb{R}. \tag{5.40}$$

We choose ν, ε, and λ as in the above proposition, and consider the analogue[3] of (4.25). We bound the first integral from below by means of the inequality

$$u_{tt}^2+2(1-\varepsilon)u_tu_{tt}+(1-\varepsilon)u_t^2\geq\frac{\varepsilon}{2}u_{tt}^2+\frac{\varepsilon(1-\varepsilon)}{(2-\varepsilon)}u_t^2, \tag{5.41}$$

[3] For the Cauchy problem, U_0 should be defined in a different way.

which follows from

$$|2(1-\varepsilon)u_t u_{tt}| \leq \eta(1-\varepsilon)u_{tt}^2 + \eta^{-1}(1-\varepsilon)u_t^2 \tag{5.42}$$

with $\eta := (2-\varepsilon)/2(1-\varepsilon)$. Now, using (4.26) with $\eta := \lambda$ and (5.40), we conclude that

$$cu_{xt}^2 + 2\chi(u_x)u_{xt} + (1-\varepsilon)\Phi(u_x) \geq (c-\lambda^{-1})u_{xt}^2 + \nu u_x^2, \tag{5.43}$$

which yields an obvious lower bound for the second integral in the analogue of (4.25). The rest of the estimations are carried out as in the proof of Lemma 4.2.

Acknowledgment. The author wishes to thank Professor R. C. MacCamy for several enlightening discussions on this material.

Note added in proof. See the recent paper of Heard [12] for some related existence theorems. See also the survey article [13] for a much more complete summary of results concerning (1.1).

REFERENCES

[1] B. D. Coleman and M. E. Gurtin, *Waves in materials with memory*. II. *On the growth and decay of one-dimensional acceleration waves*, Arch. Rational Mech. Anal. 19 (1965), pp. 239–265.

[2] C. M. Dafermos and J. A. Nohel, *Energy methods for nonlinear hyperbolic Volterra integrodifferential equations*, Comm. PDE, 4 (1979), pp. 219–278.

[3] ______, *A nonlinear hyperbolic Volterra equation in viscoelasticity*, Amer. J. Math. Supplement (1981), pp. 87–116.

[4] J. M. Greenberg, *A priori estimates for flows in dissipative materials*, J. Math. Anal. Appl., 60 (1977), pp. 617–630.

[5] J. K. Hale, *Ordinary Differential Equations*, Wiley Interscience, New York, 1969.

[6] H. Hattori, *Breakdown of smooth solutions in dissipative nonlinear hyperbolic equations*, Q. Appl. Math., 40 (1982), pp. 113–127.

[7] P. D. Lax, *Development of singularities of solutions of nonlinear hyperbolic partial differential equations*, J. Math. Phys., 5 (1964), pp. 611–613.

[8] R. C. MacCamy, *A model for one-dimensional, nonlinear viscoelasticity*, Q. Appl. Math., 35 (1977), pp. 21–33.

[9] R. C. MacCamy and V. J. Mizel, *Existence and nonexistence in the large of solutions of quasilinear wave equations*, Arch. Rational Mech. Anal., 25 (1967), pp. 299–320.

[10] O. Staffans, *On a nonlinear hyperbolic Volterra equation*, this Journal, 11 (1980), pp. 793–812.

[11] C. C. Travis and G. F. Webb, *An abstract second order semilinear Volterra integrodifferential equation*, this Journal, 10 (1979), pp. 412–424.

[12] M. L. Heard, *A class of hyperbolic Volterra integrodifferential equations*, Nonlinear Anal., 8 (1984), pp. 79–93.

[13] W. J. Hrusa and J. A. Nohel, *Global existence and asymptotics in one-dimensional nonlinear viscoelasticity*, Proc. 5th Symposium on Trends in Applications of Pure Mathematics to Mechanics Lecture Notes in Physics, Springer-Verlag, Berlin, to appear.

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

ON VOLTERRA'S POPULATION EQUATION WITH DIFFUSION*

REINHARD REDLINGER[†]

Abstract. In this paper Volterra's population equation with diffusion for a single, isolated species u is considered. Generalizing a result of R. K. Miller it is shown that every nonnegative solution $u \not\equiv 0$ tends, as $t \to \infty$, to a spatially homogeneous distribution u^*, independent of the initial distribution of u. For proof, a recursively defined sequence of pairs of lower and upper solutions is used.

1. Introduction. As a simple model to describe the evolution of a single population V. Volterra [9] proposes the equation

$$u' = au - bu^2 - u\int_r^t f(t-s)u(s)\,ds, \qquad t>0, \tag{1a}$$

where $r=0$ or $-\infty$, u is the population size, a and b are positive rate constants, and the integral is an hereditary term containing the effect of the past history on the present growth rate. The initial condition for (1a) has the form

$$u(0)=u_0 \quad \text{if } r=0 \quad (\text{resp. } u(t)=g(t) \text{ for } t\le 0 \text{ if } r=-\infty). \tag{1b}$$

Concerning the asymptotic behavior of solutions u of (1), R. K. Miller [3] obtained the following result: Let

$$\begin{aligned} &a,b>0, \quad f\in C(0,\infty)\cap L_1(0,\infty), \quad f\not\equiv 0,\\ &b>\int_0^\infty |f(s)|\,ds. \end{aligned} \tag{2}$$

Then, for any positive u_0 resp. for any positive, continuous and bounded function $g(t)$ there exists a unique positive solution u of (1). This solution is defined for all $t>0$ and satisfies

$$\lim_{t\to\infty} u(t) = a\left[b+\int_0^\infty f(s)\,ds\right]^{-1}. \tag{3}$$

It is the object of this paper to extend the above result to the case that a diffusion term Δu is added to the right-hand side of (1a). We hereby suppose that the population has no interaction with the exterior. Thus, we will consider the initial boundary value problem

$$\begin{aligned} &(4a) && u_t = \Delta u + au - bu^2 - uFu && \text{in } D=(0,\infty)\times\Omega,\\ &(4b) && \partial u/\partial n = 0 && \text{on } \Gamma=(0,\infty)\times\partial\Omega,\\ &(4c) && u(0,x)=u_0(x) && \text{for } x\in\bar\Omega \end{aligned}$$

with

$$(Fu)(t,x) = \int_0^t f(t-s)u(s,x)\,ds, \qquad (t,x)\in D,$$

* Received by the editors January 18, 1983, and in revised form August 1, 1983.

[†] Mathematisches Institut I, Universität Karlsruhe, D-7500 Karlsruhe, West Germany.

where Ω is a bounded domain in $\mathbb{R}^n$ whose boundary $\partial\Omega$ is a C^2-manifold, and where $\partial/\partial n$ denotes the exterior normal derivative to $\partial\Omega$.

In the formulation of problem (4) we have assumed $r=0$. The case $r=-\infty$ will be dealt with at the end of this paper.

By a regular solution u of (4) we understand any real-valued function $u\in C(\overline{D})$ with partial derivatives u_t, u_{x_i}, $u_{x_ix_j}\in C(D)$ for i, $j=1,2,\cdots,n$ and with $\operatorname{grad} u\in C(D\cup\Gamma,\mathbb{R}^n)$ such that (4) is identically satisfied.

The following theorem will be proved:

THEOREM. *Let the coefficients of* (4a) *satisfy* (2) *and suppose* $u_0\in C^1(\overline{\Omega})$ *with* $u_0\geq 0$. *Then the initial boundary value problem* (4) *has a unique bounded, nonnegative and regular solution u. Moreover, if* $u_0\not\equiv 0$, *we have* $u(t,x)>0$ *for all* $t>0$, $x\in\overline{\Omega}$ *and*

$$\lim_{t\to\infty} u(t,x)=u^*=a\left[b+\int_0^\infty f(s)\,ds\right]^{-1},\quad \textit{uniformly for } x\in\overline{\Omega}. \tag{5}$$

Remarks. (i) The theorem remains true if the term Δu in (4a) is replaced by the linear, uniformly elliptic operator

$$Lu\equiv\sum_{i,j=1}^{n}a_{ij}(t,x)u_{x_ix_j}+\sum_{i=1}^{n}a_i(t,x)u_{x_i}$$

with coefficient functions a_{ij}, a_i that are uniformly Hölder continuous in $\overline{D}$ and satisfy $a_{ij}=a_{ji}$, $a_{ij}(0,\cdot)\in C^1(\overline{\Omega})$ for i, $j=1,2,\cdots,n$. Instead of (4b) we then require the outward conormal derivative of u to vanish on Γ. The proof is the same as that for the case $Lu\equiv\Delta u$.

(ii) For nonnegative f, the stated theorem has been proved by A. Schiaffino [6] if f is decreasing and by Y. Yamada [11] if $f\in C^1(0,\infty)$ with $tf\in L_1(0,\infty)$. A. Tesei [8] established relation (5) in case u_0 is near to u^* and $n\leq 3$. There is also a number of papers on equation (4a) with Dirichlet boundary conditions; see, e.g., A. Schiaffino and A. Tesei [7].

To prove the theorem, we will make use of the method of lower and upper solutions developed in [5] for parabolic differential equations with functionals. We will first deal with the existence part of the theorem. Relation (5) will then be established by an iterative process concerning the step-by-step improvement of the lower and upper solutions found thus far.

2. Existence. Let us introduce the function class Z consisting of all functions z: $\overline{D}\to\mathbb{R}$ such that

(α) z is continuous in $\overline{D}$,

(β) the partial derivatives z_t, z_{x_i}, $z_{x_ix_j}$ exist in D for $i,j=1,\cdots,n$,

(γ) the exterior normal derivative $\partial z/\partial n$ exists on Γ.

As is usual with parabolic problems, in (β) the existence of z_t as a one-sided derivative from below is sufficient for our purposes.

DEFINITION. By a pair of lower and upper solutions for problem (4) we understand any pair of functions v, $w\in Z$ such that

(i) $v\leq w$ in $\overline{D}$;

(ii) $v_t\leq\Delta v+av-bv^2-vF\phi$ in D, $w_t\geq\Delta w+aw-bw^2-wF\phi$ in D for all functions $\phi\in C(\overline{D})$ with $v\leq\phi\leq w$ in $\overline{D}$;

(iii) $\partial v/\partial n\leq 0$, $\partial w/\partial n\geq 0$ on Γ;

(iv) $v(0,x)\leq u_0(x)\leq w(0,x)$ for $x\in\overline{\Omega}$.

The following lemma is a special case of [5, Thm. 3.4]:

LEMMA. *Let v, w be a pair of lower and upper solutions for the initial boundary value problem* (4) *and suppose $u_0 \in C^1(\overline{\Omega})$. Then there exists a unique regular solution u of* (4) *such that $v \leq u \leq w$ in $\overline{D}$.*

We give a short outline of the proof. Set $D_T = (0, T] \times \Omega$, $\Gamma_T = (0, T] \times \partial\Omega$ with $T > 0$, and define

$$Nu = -b(Au)^2 - (Au)F(Au),$$

where

$$(Au)(t,x) = \min(w(t,x), \max(u(t,x), v(t,x))), \qquad (t,x) \in \overline{D}.$$

Consider the system

$$\text{(4a*)} \qquad u_t = \Delta u + au + Nu \quad \text{in } D_T,$$

$$\text{(4b*)} \qquad \partial u / \partial n = 0 \quad \text{on } \Gamma_T$$

with initial condition (4c) and the corresponding integral equation $u = Su$ with (Q denotes the fundamental solution of $u_t - \Delta u = 0$ in $\overline{D}_T$)

$$\begin{aligned}(Su)(t,x) = &\int_0^t \int_\Omega Q(t,x;\tau,\xi)[au(\tau,\xi) + (Nu)(\tau,\xi)]\, d\xi\, d\tau \\ &+ \int_0^t \int_{\partial\Omega} Q(t,x;\tau,\xi) h(\tau,\xi)\, do_\xi\, d\tau \\ &+ \int_\Omega Q(t,x;0,\xi) u_0(\xi)\, d\xi \quad \text{in } D\end{aligned}$$

(compare [2, §5.3]). The density h depends on the unknown function u. Since Nu as well as h are globally Lipschitz continuous with respect to u, the fixed point principle of Banach may be applied to give the existence of a unique solution $\tilde{u} \in C(\overline{D}_T)$ of $u = Su$. By using considerations of the same type as in [2, Chap. 1], the function $S\tilde{u}$ is seen to be Hölder continuous in $\overline{D}_T$. Hence, $\tilde{u}$ is a regular solution of (4a*), (4b*), (4c) by [2, §1.5].

It remains to prove that $\tilde{u}$ solves (4), i.e., that $v \leq \tilde{u} \leq w$ in $\overline{D}_T$. Assume that we have strict inequalities in (i)–(iv) above. Then $v < \tilde{u} < w$ in $\overline{D}_T$ follows by a method of proof that is well known for parabolic differential equations; see, e.g., [10, proof of Lemma 24.I]. The case of weak inequalities can be handled in the same way by making use of a family $(\rho_\varepsilon)_{\varepsilon > 0}$ of positive auxiliary functions,

$$\rho_\varepsilon(t,x) = C\varepsilon(e^{(2a+1)t} + 1) + \varepsilon h(x), \qquad (t,x) \in \overline{D}_T,$$

where $h \in C^2(\overline{\Omega})$ is a function with $\partial h / \partial n > 0$ on $\partial\Omega$ (compare [10, 31.VI (γ_2)]) and $C = C(a,h)$ is a suitably chosen positive constant. We then have $v - \rho_\varepsilon < \tilde{u} < w + \rho_\varepsilon$ for all $\varepsilon > 0$, which gives $v \leq \tilde{u} \leq w$ in $\overline{D}_T$.

Noting that $T > 0$ is arbitrary in the above reasoning, we arrive at the desired result.

With the help of the above lemma the existence part of the theorem is easily proved. It suffices to choose $v \equiv 0$ and $w \equiv K$, where K is any constant such that

$$\text{(6)} \qquad K > \max\left(\max_{x \in \overline{\Omega}} |u_0(x)|, a\left[b - \int_0^\infty |f(s)|\, ds \right]^{-1} \right).$$

Indeed, for all $(t,x)\in D$ and any function $\phi\in C(\overline{D})$ with $0\le\phi\le K$ in $\overline{D}$ we then get

$$w_t-\Delta w-aw+bw^2+w\int_0^t f(t-s)\phi(s,x)\,ds\ge -aK+bK^2-K^2\int_0^\infty |f(s)|\,ds>0,$$

as required. The other inequalities in the above Definition are trivially satisfied. It should be noted that in view of this result it is possible to replace the functional Fu in (4a) by $F(A^*u)$, where

$$(A^*u)(t,x)=\min(K,\max(0,u(t,x))),\qquad (t,x)\in\overline{D},$$

without changing the solution u of (4). In the following we will tacitly assume that such a cut-off has been made.

Let us add the remark that the Lemma still yields the existence of a unique bounded, nonnegative and regular solution u for problem (4) if instead of (2) we assume

(2′) $$a\le 0,\quad b>0,\quad f\in C(0,\infty)\cap L_1(0,\infty),\quad f\ge 0,$$

or

(2″) $$a>0,\quad 0<b\le\int_0^\infty f(s)\,ds<\infty,\quad f\in C(0,\infty),\quad f\ge 0.$$

The first case is mentioned by R. K. Miller [3], whereas the second one has been considered by J. M. Cushing in [1]. In both cases we can use the lower solution $v\equiv 0$. As upper solution w we choose

(2′) the solution of the initial value problem

$$w'=aw-bw^2,\qquad w(0)=\max_{x\in\overline{\Omega}}|u_0(x)|$$

or respectively

(2″) the function $w\equiv K_1$, where K_1 is any constant such that

$$K_1\ge\max\Big(\max_{x\in\overline{\Omega}}|u_0(x)|,a/b\Big).$$

Note that $\lim_{t\to\infty}w(t)=0$ if (2′) holds. Hence, in this case all nonnegative solutions tend to zero as $t\to\infty$.

The asymptotic behavior for (2″) is by far more complicated and will not be considered here. As regards problem (1), a discussion of the possible behavior of the solutions can be found in [1, Chaps. 3, 5]. For (4), we refer to the above-mentioned papers [8] and [11].

3. Asymptotic behavior. We will now prove the second part of the Theorem, i.e., the validity of (5). Suppose (2) to hold and let $u_0\not\equiv 0$. By u we denote the unique regular solution of (4) whose existence has been shown in §2. Write $f=f^+-f^-$, where

$$f^+(s)=\max(0,f(s)),\quad f^-(s)=\max(0,-f(s))\quad\text{for } s\ge 0,$$

and set

$$c^+=\int_0^\infty f^+(s)\,ds,\qquad c^-=\int_0^\infty f^-(s)\,ds.$$

Equation (4a) then takes the form

$$(4a') \qquad u_t = \Delta u + au - bu^2 - u\int_0^t f^+(t-s)u(s,x)\,ds + u\int_0^t f^-(t-s)u(s,x)\,ds.$$

Further define

$$\underline{u}(t) = \min_{x\in\bar\Omega} u(t,x), \qquad \bar u(t) = \max_{x\in\bar\Omega} u(t,x)$$

and set

$$I = \left[\liminf_{t\to\infty} \underline{u}(t),\ \limsup_{t\to\infty} \bar u(t)\right].$$

We have to show that the interval I consists of the single point $a[b+\int_0^\infty f(s)\,ds]^{-1}$.

Let us first prove that

$$(7) \qquad I \subset [0, a/(b-c^-)].$$

To see (7), we consider the function $p_1(t)$ defined by

$$p_1' = ap_1 - bp_1^2 + p_1 Kc^- \quad \text{for } t>0, \qquad p_1(0) = \bar u(0)$$

with K given by (6). Using the results of §2, it is easily seen that $0, p_1$ is a pair of lower and upper solutions for problem (4). By the lemma there exists a unique regular solution $\tilde u$ of (4) such that $0 \le \tilde u(t,x) \le p_1(t)$ in $\bar D$. Since $p_1(t) \le K$ for all $t \le 0$, we have $\tilde u = u$. It follows that $I \subset [0,\gamma_1]$, where

$$\gamma_1 = \lim_{t\to\infty} p_1(t) = (a + Kc^-)/b.$$

Remark. The estimate $0 \le u(t,x) \le K$ made it possible to replace Fu in (4a) by $F(A^*u)$ without affecting the solution u of (4), and this led us to the improved upper solution p_1. In a similar way, use can be made of the inclusion $I \subset [\alpha,\beta]$, where $0 \le \alpha \le \beta$. It suffices to treat the case $I \subset [0,\gamma_1]$ in detail. The general case can be handled in the same manner.

Choose $\varepsilon > 0$. Then there exists a $t_0 > 1$ such that

$$\bar u(t) \le \gamma_1 + \varepsilon \quad \text{for all } t \ge t_0 - 1$$

and a $t_1 > t_0$ such that

$$\int_{t-t_0}^t f^-(s)\,ds < \varepsilon \quad \text{for all } t \ge t_1.$$

Set

$$z(t) = \gamma_1 + \varepsilon + (K - \gamma_1 - \varepsilon)\max(0, \min(1, t_0 - t)) \quad \text{for } t \ge 0.$$

Without changing the solution u of (4) the functional $F(A^*u)$ in (4a) may then be replaced by $F(A'u)$ where

$$(A'u)(t,x) = \max(0, \min(u(t,x), z(t))), \qquad (t,x) \in \bar D.$$

Hence, the functions $v = 0$ and $w = p_2$, where $p_2(t)$ is defined by

$$p_2' = ap_2 - bp_2^2 + \varepsilon Kp_2 + (\gamma_1 + \varepsilon)c^- p_2 \quad \text{for } t > t_1,$$
$$p_2(t) = K \quad \text{for } 0 \le t \le t_1,$$

are a pair of lower and upper solutions for (4). We have

$$\lim_{t\to\infty} p_2(t) = \gamma_2 + \varepsilon(K + c^-)/b \quad \text{with } \gamma_2 = (a + \gamma_1 c^-)/b.$$

Since ε is arbitrarily small, it follows that

$$I \subset [0, \gamma_2].$$

Replacing γ_1 by γ_2 in the above reasoning, we get

$$I \subset [0, \gamma_3], \quad \text{where } \gamma_3 = (a + \gamma_2 c^-)/b,$$

and so on. The sequence (γ_n) is decreasing and nonnegative. Hence there exists $\gamma = \lim_{n\to\infty} \gamma_n$. A little computation shows $\gamma = a/(b - c^-)$, and this proves (7).

Let us now improve the lower solution $v = 0$. We first observe that by the strong maximum principle for linear parabolic equations (see [4, §3.3]) we have

$$\underline{u}(t) > 0 \quad \text{for } t > 0.$$

Choose $\varepsilon > 0$. Then there exists a $t_2 > 1$ such that

$$\bar{u}(t) \leq \gamma + \varepsilon \quad \text{for all } t \geq t_2 - 1$$

and a $t_3 > t_2$ such that

$$\int_{t-t_2}^{t} f^+(s)\,ds < \varepsilon \quad \text{for all } t > t_3.$$

Set

$$\delta_1 = \frac{1}{2}\min\left\{\underline{u}(t) \colon \frac{1}{2}t_3 \leq t \leq t_3\right\}$$

and define a function $v_1(t)$ by

$$\begin{aligned} v_1' &= av_1 - bv_1^2 - \varepsilon K v_1 - c^+(\gamma + \varepsilon)v_1 && \text{for } t > t_3, \\ v_1(t) &= \max(0, \delta_1(2t - t_3/)t_3) && \text{for } 0 \leq t \leq t_3. \end{aligned}$$

Then $v = v_1$ and, say, $w = K$ is a pair of lower and upper solutions for (4). In view of

$$\lim_{t\to\infty} v_1(t) = \gamma(b - c^+ - c^-)/b - \varepsilon(K + c^+)/b$$

(for all sufficiently small ε) we conclude that

$$I \subset [\mu_1, \nu_1] = [\gamma(b - c^+ - c^-)/b, \gamma], \quad \text{where } \gamma = a/(b - c^-). \tag{8}$$

Remark. Note that the inequality $\underline{u}(t) > v_1(t)$ is obviously satisfied for $0 \leq t \leq t_3$. Therefore, the proof of the Lemma still goes through (see the outline of proof in §2), even though hypothesis (ii) in the above definition of a lower solution is violated for some $t \in (0, t_3]$.

To prove (5) we now use an iterative argument starting with (8). Choose $\varepsilon \in (0, \mu_1/2)$. Then there exists a $t_4 > t_2$ such that

$$\mu_1 - \varepsilon \leq \underline{u}(t) \leq \bar{u}(t) \leq \nu_1 + \varepsilon \quad \text{for all } t \geq t_4 - 1$$

and a $t_5 > t_4$ such that

$$\int_0^{t-t_4} f^+(s)\,ds > c^+ - \varepsilon, \qquad \int_{t-t_4}^t f^+(s)\,ds < \varepsilon \quad \text{for all } t > t_5$$

and

$$\int_0^{t-t_4} f^-(s)\,ds > c^- - \varepsilon, \qquad \int_{t-t_4}^t f^-(s)\,ds < \varepsilon \quad \text{for all } t > t_5.$$

Set

$$\delta_2 = \frac{1}{2}\min\left\{ \underline{u}(t)\colon \frac{1}{2}t_5 \le t \le t_5 \right\}.$$

Define the functions $v_2(t)$ and $w_2(t)$ respectively by

$$\begin{aligned} v_2' &= av_2 - bv_2^2 - \varepsilon K v_2 - c^+(\nu_1+\varepsilon)v_2 + (c^- - \varepsilon)(\mu_1 - \varepsilon)v_2 && \text{for } t > t_5, \\ v_2(t) &= \max(0, \delta_2(2t - t_5)/t_5) && \text{for } 0 \le t \le t_5, \end{aligned}$$

and by

$$\begin{aligned} w_2' &= aw_2 - bw_2^2 - (c^+ - \varepsilon)(\mu_1 - \varepsilon)w_2 + \varepsilon K w_2 + c^-(\nu_1 + \varepsilon)w_2 && \text{for } t > t_5, \\ w_2(t) &= K && \text{for } 0 \le t \le t_5. \end{aligned}$$

Then v_2, w_2 is a pair of lower and upper solutions for (4). We have

$$\lim_{t\to\infty} v_2(t) = (a + c^-\mu_1 - c^+\nu_1)/b - \varepsilon(K + c^+ + c^- + \mu_1 - \varepsilon)/b$$

and

$$\lim_{t\to\infty} w_2(t) = (a - c^+\mu_1 + c^-\nu_1)/b + \varepsilon(K + c^+ + c^- + \mu_1 - \varepsilon)/b.$$

Since $\varepsilon > 0$ is arbitrarily small, we find

$$I \subset [\mu_2, \nu_2] = [(a + c^-\mu_1 - c^+\nu_1)/b, (a - c^+\mu_1 + c^-\nu_1)/b].$$

Repeating the above argument, we get two sequences (μ_n), (ν_n) defined by

(9a) $\quad \mu_{n+1} = (a + c^-\mu_n - c^+\nu_n)/b \quad \text{for } n \in \mathbb{N}, \qquad \mu_1 = \gamma(b - c^+ - c^-)/b,$

and by

(9b) $\quad \nu_{n+1} = (a - c^+\mu_n + c^-\nu_n)/b \quad \text{for } n \in \mathbb{N}, \qquad \nu_1 = \gamma,$

with $\gamma = a/(b - c^-)$. It is an easy matter to show by induction that

$$\mu_1 \le \mu_2 \le \cdots \le \mu_n \le \nu_n \le \cdots \nu_2 \le \nu_1 \quad \text{for all } n \in \mathbb{N}.$$

Hence there exist

$$\mu = \lim_{n\to\infty} \mu_n \quad \text{and} \quad \nu = \lim_{n\to\infty} \nu_n,$$

and from (9) it follows that

$$\mu = \nu = a/(b + c^+ - c^-),$$

as desired. The proof of the Theorem is complete.

Remark. The case $r=-\infty$ presents no new difficulties. Instead of (4) we now have

$$u_t=\Delta u+au-bu^2-u\int_{-\infty}^{t} f(t-s)u(s,x)\,ds \quad \text{in } D, \tag{10a}$$

$$\frac{\partial u}{\partial n}=0 \quad \text{on } \Gamma, \tag{10b}$$

$$u(t,x)=g(t,x) \quad \text{in } \Gamma_0=(-\infty,0]\times\bar{\Omega}. \tag{10c}$$

In the definitions of §2 replace $\bar{D}$ by $\bar{D}\cup\Gamma_0$ and (iv) by the assumption $v\le g\le w$ in Γ_0. Then, under suitable regularity conditions on g, the lemma is still valid.

For example, suppose that $g\in C(\Gamma_0)$ is a bounded nonnegative function which is uniformly Hölder continuous in Γ_0. Assume further that the function $g_0(x)=g(0,x)$ satisfies the same hypotheses as u_0 above. Let (2) be fulfilled. Then (10) has a unique bounded, nonnegative and regular solution u that satisfies (5) in case $g_0\not\equiv 0$. Apart from some obvious modifications the proof of this assertion is the same as that of the theorem.

We finally remark that the above assertions concerning the cases (2′) and (2″) are also valid for problem (10).

REFERENCES

[1] J. M. Cushing, *Integrodifferential Equations and Delay Models in Population Dynamics*, Lecture Notes in Biomathematics 20, Springer, Berlin, 1977.

[2] A. Friedman, *Partial Differential Equations of Parabolic type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[3] R. K. Miller, *On Volterra's population equation*, SIAM J. Appl. Math., 14 (1966), pp. 446–452.

[4] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[5] R. Redlinger, *Existence theorems for semilinear parabolic systems with functionals*, Nonlinear Anal., to appear.

[6] A. Schiaffino, *On a diffusion Volterra equation*, Nonlinear Anal., 3 (1979), pp. 595–600.

[7] A. Schiaffino and A. Tesei, *Monotone methods and attractivity results for Volterra integro-partial differential equations*, Proc. Roy. Soc. Edinburgh, 89A (1981), pp. 135–142.

[8] A. Tesei, *Stability properties for partial Volterra integrodifferential equations*, Ann. Mat. Pura Appl., (IV) 126 (1980), pp. 103–115.

[9] V. Volterra, *Leçons sur la théorie mathématique de la lutte pour la vie*, Gauthier-Villars, Paris, 1931.

[10] W. Walter, *Differential and integral inequalities*, Ergebnisse der Mathematik und ihrer Grenzgebiete 55, Springer, Berlin, 1970.

[11] Y. Yamada, *On a certain class of semilinear Volterra diffusion equations*, J. Math. Anal. Appl., 88 (1982), pp. 433–451.

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

009

SPEED OF PROPAGATION OF SOLUTIONS OF A LINEAR INTEGRODIFFERENTIAL EQUATION WITH NONCONSTANT COEFFICIENTS*

GUY CANADAS†

Abstract. Using an energetic method we prove a property of propagation with a finite speed for the solutions of the system of linear viscoelasticity in nonhomogeneous media. Our results generalize those obtained by P. L. Davis [SIAM J. Math. Anal., 19 (1979), pp. 570–576] for constant coefficients.

1. Introduction. The notion of hyperbolicity was extended to linear integrodifferential equations of Volterra type by P. L. Davis [7] in the case of equations with constant coefficients. Such an equation is said to be hyperbolic if the solutions of an appropriate problem propagate with finite speed.

In [8], [9], [10] the same author studies equations arising in electromagnetic theory, in heat condution for materials with memory and an equation of linear viscoelasticity, always with constant coefficients. It is precisely this last equation we are interested in here, in the case of non constant coefficients, i.e. for an inhomogeneous viscoelastic medium.

The tools used by P. L. Davis to study the evolution of the solution support are the Fourier transform and the Paley–Wiener theorem. This method can obviously not be used when the medium is nonhomogeneous. To extend the results to this case we use an energetic method in a domain limited by the characteristics of the principal part (in Hörmander's sense) of the operator.

2. The linear viscoelastic problem. The linear viscoelastic system is, using the summation convention for repeated subscripts:

$$(\mathrm{I})\quad \begin{aligned} &(1)\quad \rho(x)\frac{\partial v_i}{\partial t}=\partial_j\sigma_{ij}(x,t)+f_i(x,t),\\ &(2)\quad \frac{\partial}{\partial t}\varepsilon_{ij}(x,t)=\frac{1}{2}\left(\partial_j v_i(x,t)+\partial_i v_j(x,t)\right),\\ &(3)\quad \sigma_{ij}(x,t)=\mu_{ijkl}(x)\left\{\varepsilon_{kl}(x,t)-\int_0^t b_{klpq}(x,t-\tau)\varepsilon_{pq}(x,\tau)\,d\tau\right\}, \end{aligned}$$

where ρ is the mass per unit volume, $f=(f_1,f_2,f_3)$ a force per unit volume, $v=(v_1,v_2,v_3)$ the velocity, σ the Cauchy stress tensor, ε the linearized strain tensor, μ the instantaneous elastic tensor. The tensor b describes the relaxation behaviour of the material ($b\equiv 0$ when the material is purely elastic).

We denote by E the linear space of 2nd order symmetric tensors on $\mathbb{R}^3$, endowed with the usual scalar product:

$$\forall\alpha,\beta\in E\quad \langle\alpha,\beta\rangle=\alpha_{ij}\beta_{ij},$$

and usual norm: $|\alpha|=(\alpha_{ij},\alpha_{ij})^{1/2}$. The same notation $|\cdot|$ serves also for the Euclidean norm in $\mathbb{R}^3$. The linear space of linear symmetric operators of E is denoted by $\mathcal{L}_s(E)$, and its natural norm by $\|\cdot\|$:

$$\forall\mu\in\mathcal{L}_s(E)\quad \|\mu\|=\mathrm{Max}\{|\mu\cdot\alpha|,\alpha\in E,|\alpha|=1\}.$$

*Received by the editors December 27, 1982, and in revised form, July 27, 1983.

† Département de Mathématiques, Université de Pau, 64000 Pau, France.

Definition. We say that (v,ε,σ) is a solution of system (I) on $\Omega\times]0,T[$, where Ω is an open subset of $\mathbb{R}^3$, if $v\in H^1_{\mathrm{loc}}(\Omega\times[0,T[,\mathbb{R}^3)$, $\mu\varepsilon$, $\sigma\in H^1_{\mathrm{loc}}(\Omega\times[0,T[,E)$ and the equations (1), (2), (3) are satisfied almost everywhere on $\Omega\times]0,T[$.

In [3], [4], [5], [6], [11], one can find results of existence, uniqueness and stability for certain limit problems associated with the system (I).

These authors employ various hypotheses on the coefficients, and in [1] one can find a discussion on natural hypotheses in the problems of solid visco-elasticity. In this work we will use only:

$$\begin{array}{ll} \mathrm{H1} & \begin{array}{l}\mu,\mu^{-1}\in L^\infty_{\mathrm{loc}}(\mathbb{R}^3,\mathcal{L}_s(E)),\\ b\in L^\infty_{\mathrm{loc}}(\mathbb{R}^3\times[0,+\infty[,\mathcal{L}_s(E)).\end{array}\\ \mathrm{H2} & \dfrac{\partial b}{\partial t}\in L^\infty_{\mathrm{loc}}(\mathbb{R}^3\times[0,+\infty[,\mathcal{L}_s(E)).\\ \mathrm{H3} & \begin{array}{l}\rho>0 \text{ and } \rho,\rho^{-1}\in L^\infty_{\mathrm{loc}}(\mathbb{R}^3),\\ \mu(x) \text{ is positive definite a.e. } x\in\mathbb{R}^3.\end{array}\end{array}$$

3. Inversion of the constitutive law. The constitutive equation (3) may be inverted to express ε as a function of σ. This result, which generalizes an analogous result of R. C. MacCamy [2] is given in:

Theorem 1. *Under the hypotheses* H1 *the constitutive law* (3) *may be inverted, and for* $\varepsilon,\sigma\in L^2_{\mathrm{loc}}(\mathbb{R}^3\times[0,+\infty[,E)$ *we have*:

$$\varepsilon(x,t)=\mu^{-1}(x)\left\{\sigma(x,t)+\int_0^t k(x,t-\tau)\sigma(x,\tau)\,d\tau\right\},\tag{4}$$

where $k\in L^\infty_{\mathrm{loc}}(\mathbb{R}^3\times[0,+\infty[,\mathcal{L}_s(E))$ *depends only on* b.

If moreover H2 *is satisfied, then* $\partial k/\partial t\in L^\infty_{\mathrm{loc}}(\mathbb{R}^3\times[0,+\infty[,\mathcal{L}_s(E))$.

Proof. Let K be a compact subset of $\mathbb{R}^3$, $T>0$ and $Q=K\times[0,T]$. For $\varphi\in L^2(Q,E)$ we define

$$(T_b\varphi)(x,t)=\int_0^t b(x,t-\tau)\varphi(x,\tau)\,d\tau \quad \text{a.e. } (x,t)\in Q.$$

T_b is a continuous linear operator in $L^2(Q,E)$, with norm smaller than $(T/\sqrt{2})\|b\|_{L^\infty(Q,\mathcal{L}_s(E))}$.

We show first that the operator $I-T_b$, where I is the identity on $L^2(Q,E)$, is invertible. For a given $f\in L^2(Q,E)$ let us prove that the equation $\varphi-T_b\varphi=f$ has a unique solution $\varphi\in L^2(Q,E)$.

Let N be an integer verifying $0<\alpha=T/N<\sqrt{2}\,\|b\|^{-1}_{L^\infty(Q,\mathcal{L}_x(E))}$ and suppose that φ is known on $K\times[0,n\alpha]$ for a value $n\in\{0,1,\cdots,N\}$. For a.e. $(x,t)\in K\times[n\alpha,(n+1)\alpha]$ $\varphi(x,t)$ verifies:

$$\varphi(x,t)-\int_{n\alpha}^t b(x,t-\tau)\varphi(x,\tau)\,d\tau=f(x,t)+\int_0^{n\alpha} b(x,t-\tau)\varphi(x,\tau)\,d\tau.$$

The right-hand member is known, and the norm of the operator $T_b^{(n)}$ on $L^2(K\times[n\alpha,(n+1)\alpha],E)$ defined by $(T_b^{(n)}\psi)(x,t)=\int_{n\alpha}^t b(x,t-\tau)\psi(x,\tau)\,d\tau$ is smaller than $\alpha/\sqrt{2}\,\|b\|_{L^\infty(Q,E)}<1$. A classical fixed point theorem shows that φ is then uniquely determined on $K\times[n\alpha,(n+1)\alpha]$ and by induction that it is uniquely determined on Q.

In a similar manner one can prove that if $f \in L^\infty(Q,E)$, then $\varphi \in L^\infty(Q,E)$.

We show now that $(I-T_b)^{-1}=I+T_k$ where $k \in L^\infty(Q,E)$ is the solution of $k-T_bk=b$. Fubini's theorem shows that $T_bT_k=T_{T_bk}$, and the relation $T_bk=k-b$ gives $T_bT_k=T_{k-b}=T_k-T_b$ and therefore $(I-T_b)(I-T_k)=I$.

For the second part of the theorem we consider $k^{\boldsymbol{\cdot}} \in L^\infty_{\text{loc}}(\mathbb{R}^3 \times \mathbb{R}_+, \mathfrak{L}_s(E))$ defined by

$$k^{\boldsymbol{\cdot}}(x,t)=\frac{\partial b}{\partial t}(x,t)+b(x,0)k(x,t)+\int_0^t \frac{\partial b}{\partial t}(x,t-\tau)k(x,\tau)\,d\tau,$$

where $b(\cdot,0) \in L^\infty_{\text{loc}}(\mathbb{R}^3, \mathfrak{L}_s(E))$ is the trace of b in $t=0$. It is easy to see that $k^{\boldsymbol{\cdot}}=\partial k/\partial t$.

4. Propagation with a finite speed. We define $c \in L^\infty_{\text{loc}}(\mathbb{R}^3)$ by

(5) $$c(x)=\big(|\mu(x)|/\rho(x)\big)^{1/2}$$

(with the hypothesis of symmetry, $|\mu(x)|$ is equal to the greatest eigenvalue of μ). The quantity c, homogeneous to a speed, appears with regard to Theorem 2 as the local maximal speed of propagation. More precisely this theorem says that in a domain of $\mathbb{R}^3$, if c is a.e. bounded by a constant c_0, then the local speed of propagation of solutions is bounded above by c_0 in this domain.

Notice that the maximal speed of propagation does not depend on b, and is the same as in the elastic case when $b=0$.

THEOREM 2. *We assume* H1, H2 *and* H3. *Fix* $r>0$, *define* B *as* $\{x \in \mathbb{R}^3; |x-x_0|<r\}$ *and choose* c_0 *a positive constant so that* $c(x) \leq c_0$ *a.e.* $x \in B$. *If a solution of* (I) *vanishes a.e. on* $B \times [0,t_0[$, *then this solution still vanishes on*

$$Q=\big\{(x,t) \in \mathbb{R}^3 \times \mathbb{R}_+; t \geq t_0, c_0(t-t_0)+|x-x_0|<r\big\}$$

provided that f also vanishes on Q.

Proof. Multiplying (1) by $v_i(x,t)$ and integrating on $Q_s=\{(x,t) \in Q; t_0 \leq t \leq s\}$ with $s \in [t_0, t_0+r/c_0]$, we get:

$$\int_{Q_s} \frac{\partial}{\partial t}\left(\frac{1}{2}\rho v^2\right) dx\,dt=\int_{Q_s} \partial_j \sigma_{ij} v_i\,dx\,dt=\int_{Q_s} \left\{\partial_j(\sigma_{ij}v_i)-\left\langle \sigma, \frac{\partial \varepsilon}{\partial t}\right\rangle\right\} dx\,dt.$$

In view of Theorem 1 we have

$$\frac{\partial \varepsilon}{\partial t}(x,t)=\mu^{-1}(x)\left\{\frac{\partial \sigma}{\partial t}(x,t)+k(x,0)\sigma(x,t)+\int_0^t \frac{\partial k}{\partial t}(x,t-\tau)\sigma(x,\tau)\,d\tau\right\},$$

and with the symmetry of μ:

(6) $$\int_{Q_s}\left\{\frac{\partial}{\partial t}\left[\frac{1}{2}\rho v^2+\frac{1}{2}\left\langle \sigma, \mu^{-1}\sigma\right\rangle\right]-\partial j(\sigma_{ij}v_i)\right\} dx\,dt$$

$$=-\int_{Q_s}\left\langle \mu^{-1}(x)\sigma(x,t), k(x,0)\sigma(x,t)+\int_0^t \frac{\partial k}{\partial t}(x,t-\tau)\sigma(x,\tau)\,d\tau\right\rangle dx\,dt.$$

The right-hand member of (6) can easily be bounded above by $A\int_{Q_s}|\sigma|^2\,dx\,d\tau$ where A is a constant depending only on $\|\mu^{-1}k(\cdot,0)\|_{L^\infty(B,\mathcal{L}_s(E))}$ and $\|\mu^{-1}\frac{\partial k}{\partial t}\|_{L^\infty(Q,\mathcal{L}_s(E))}$. To transform the left-hand member, let us establish the Green formula

$$(7)\quad \int_{Q_s}\left\{\frac{\partial}{\partial t}\left[\frac{1}{2}\rho v^2+\frac{1}{2}\langle\sigma,\mu^{-1}\sigma\rangle\right]-\partial_j(\sigma_{ij}v_i)\right\}dx\,dt$$

$$=\int_{\partial Q_s}\left\{\left[\frac{1}{2}\rho^2+\frac{1}{2}\langle\sigma,\mu^{-1}\sigma\rangle\right]n_t-\sigma_{ij}v_in_j\right\}d\gamma,$$

where $n=(n_1,n_2,n_3,n_t)=(n_x,n_t)$ is the outward normal to ∂Q_s. In fact this formula is true with ρ and μ regular enough, for example when they are continuously differentiable. We consider then two sequences (ρ_n) and (μ_n) which verify:

ρ_n and μ_n^{-1} are uniformly bounded for $x\in B$ and $n\in\mathbb{N}$,

$\rho_n\to\rho$ in $L^1(B)$ when $n\to+\infty$,

$\mu_n^{-1}\to\mu^{-1}$ in $L^1(B,\mathcal{L}_s(E))$ when $n\to+\infty$.

We obtain (6) as a limit of the same formula written with ρ_n and μ_n instead of ρ and μ. For the left-hand member there are no difficulties, because ρ_n and μ_n do not depend on t. For the right-hand member, the same argument shows that $\rho_nv^2+\langle\sigma,\mu_n^{-1}\sigma\rangle$ converges a.e. on ∂Q_s to $\rho v^2+\langle\sigma,\mu^{-1}\sigma\rangle$ with $\|\rho_nv^2+\langle\sigma,\mu_n^{-1}\sigma\rangle\|_{L^1(\partial Q_s)}$ bounded. Lebesgue's dominated convergence theorem then applies and gives the result. The left-hand member of (5) can be written

$$\int_{\Gamma_s}\left\{\frac{1}{2}\rho v^2+\frac{1}{2}\langle\sigma,\mu^{-1}\sigma\rangle\right\}dx+\int_{\Gamma_s'}\left\{\left[\frac{1}{2}\rho v^2+\frac{1}{2}\langle\sigma,\mu^{-1}\sigma\rangle\right]n_t-\sigma_{ij}v_in_j\right\}d\gamma,$$

where $\Gamma_s=\{(x,t)\in\partial Q_s, t=s\}$ and $\Gamma_s'=\{(x,t)\in\partial Q_s; c_0(t-t_0)+|x-x_0|+r\}$. We have a.e. on Γ_s':

$$\left[\frac{1}{2}\rho v^2+\frac{1}{2}\langle\sigma,\mu^{-1}\sigma\rangle\right]n_t-\sigma_{ij}v_in_j\geq c_0|n_x|\left\{\frac{1}{2}\rho v^2+\frac{1}{2}\langle\sigma,\mu^{-1}\sigma\rangle-\frac{|\sigma|\cdot|v|}{c_0}\right\}\geq 0.$$

The left-hand member of (6) is then bounded below by the integral term on Γ_s, and therefore:

$$F(s)=\int_{\Gamma_s}\left\{\frac{1}{2}\rho v^2+\frac{1}{2}\langle\sigma,\mu^{-1}\sigma\rangle\right\}dx\leq A\int_{Q_s}|\sigma|^2dx\,dt.$$

This inequality gives:

$$F(s)\leq\frac{A}{2\|\mu\|_{L^\infty(B,\mathcal{L}(E))}}\int_{t_0}^{s}F(t)\,dt.$$

Gronwall's lemma shows that F vanishes a.e. on $[t_0,t_0+r/c_0]$ and therefore that the solution of (I) vanishes on Q.

5. The one-dimensional case. The simplicity of the geometry permits in that case the integration in a domain Q limited by the characteristic curves of the principal part of the operator, instead of a cone contained in this domain, as previously. We obtain therefore a more precise result expressed in Theorem 3. The hypotheses H1, H2 and H3 remain the same if we change $\mathbb{R}^3$, E and $\mathcal{L}_s(E)$ into $\mathbb{R}$. $\mu(x)$ is therefore a positive number, and $c(x)=(\mu(x)/\rho(x))^{1/2}$.

THEOREM 3. *We assume* H1, H2 *and* H3. *Let* $t_0 \ge 0$, $X_+ < X_-$ *two real numbers. Let* $B =]X_+, X_-[$ *and* $Q=\{(x,t)\in\mathbb{R}^2;\ X_+ \le x \le X_-,\ 0 \le t \le \mathrm{Inf}(t_+(x), t_-(x))\}$, *where* $t=t_+(x)$ (*respectively* $t=t_-(x)$) *verifies* $dt_+/dx = 1/c(x)$ *and* $t_+(X_+)=t_0$ (*respectively* $dt_-/dx=-1/c(x)$ *and* $t_-(X_-)=t_0$). *If a solution of* (I) *vanishes on* $B\times[0,t_0]$ *and* f *vanishes on* Q, *then this solution vanishes on* Q.

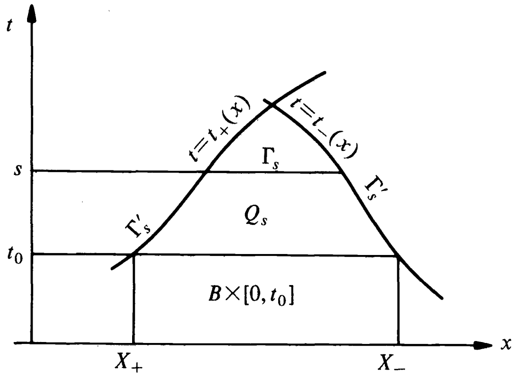


FIG 1.

The proof of this theorem is identical to the previous one if we change the definitions of B and Q as indicated. The hypotheses on ρ and μ ensure that ∂Q_s is Lipschitz continuous and the same arguments hold.

Remark. The hypothesis H2 is not essential. Theorem 3 remains valid if we suppose, instead of hypothesis H2, that for a.e. $x\in\mathbb{R}$, $t\mapsto k(x,t)$ is a positive decreasing function. In this case one has to consider, for fixed x, a sequence k_n vérifying: $k_n(0)=k(x,0)$, k_n positive decreasing on $\mathbb{R}_+$, and $k_n \to k(x,\cdot)$ in $L^1_{\mathrm{loc}}(\mathbb{R}_+)$.

REFERENCES

[1] F. BLOOM, *Ill-Posed Problems for Integrodifferential Equations in Mechanics and Electromagnetic Theory*, SIAM Studies in Applied Mathematics 3, Society for Industrial and Applied Mathematics, Philadelphia.

[2] R. C. MACCAMY, *Stability theorems for a class of functional differential equations*, SIAM J. Appl. Math., 30 (1976), pp. 557–576.

[3] G. CANADAS, *Sur une équation hyperbolique du troisime ordre intervenant en sismique*, thèse de 3ème cycle, Université de Pau, 1979.

[4] B. D. COLEMAN AND W. NOLL, *Foundations of linear viscoelasticity*, Rev. Modern Physics, 33 (1961), pp. 239–249.

[5] B. D. COLEMAN AND V. MIZEL, *On the stability of solutions of functional differential equations*, Arch. Rat. Mech. Anal., 30 (1968), pp. 173–196.

[6] C. A. DAFERMOS, *An abstract Volterra equation with applications to linear viscoelasticity*, J. Differential Equations, 7 (1970), pp. 554–569.

[7] P. L. DAVIS, *Hyperbolic integrodifferential equations*, Proc. Amer. Math. Soc., 47 (1975), pp. 155–160.

[8] ———, *Hyperbolic integrodifferential equations arising in the electromagnetic theory of dielectrics*, J. Differential Equations, 18 (1975), pp. 170–178.

[9] ———, *On the hyperbolicity of the equations of the linear theory of heat conduction for materials with memory*, SIAM J. Appl. Math., 30 (1976), pp. 75–80.

[10] ———, *On the speed of propagation of solutions of integrodifferential equations*, SIAM J. Math. Anal., 10 (1979), pp. 570–576.

[11] G. DAVAUT AND J. L. LIONS, *Let inéquations en mécanique et en physique*, Dunod, Paris, 1972.

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

010

THE LEAST CONSTANT IN FRIEDRICHS' INEQUALITY IN ONE DIMENSION*

J. T. MARTI†

Abstract. Friedrichs' inequality for the Sobolev space $H^1(0,s)$, $s>0$, is $\|f\|_1 \leq c(f(0)^2+f(s)^2+\|f'\|^2)^{1/2}$, $f \in H^1(0,s)$, where f' is the first derivative of f and $\|\cdot\|_1$ and $\|\cdot\|$ are the norms of $H^1(0,s)$ and $L_2(0,s)$ respectively. The usual proofs for the existence of such a constant c are based on nonconstructive functional analytic ideas. It is shown that the least constants c can be evaluated by variational techniques and are given by the unique solution of a simple transcendental equation. In the special case $s=1$ one has, for example, $c=1.07869\cdots$.

1. Introduction. Let $C^k(0,s)$ be the vector space of bounded continuous real functions f on an open interval $(0,s)$ of $\mathbb{R}$ with bounded continuous derivatives $f',\cdots,f^{(k)}$, and let $\|\cdot\|$ be the norm of $L_2(0,s)$. The Sobolev space $H^1(0,s)$ [1, 3.1 and 3.16] is obtained by completing $C^1(0,s)$ with respect to the Sobolev norm given by

$$\|f\|_1 := \left(\|f\|^2+\|f'\|^2\right)^{1/2}, \qquad f \in C^1(0,s).$$

It is well known [1, 5.4] that $H^1(0,s)$ may be considered as a subset of $C^0(0,s)$, imbedded such that for some constant $d>0$ depending on s the Sobolev inequality $\sup\{|f(x)|: x \in (0,s)\} \leq d\|f\|_1$ holds for all f in $H^1(0,s)$. Therefore, the so-called traces Tf in $\mathbb{R}^2$ of all f in $H^1(0,s)$ are well defined by

$$Tf=(f(0+),f(s-))^T=\lim_{\varepsilon \searrow 0}(f(\varepsilon),f(s-\varepsilon))^T.$$

Of course, the trace operator $T: H^1(0,s) \to \mathbb{R}^2$ given this way is a linear operator with norm $\|T\| := \{|Tf|: f \in H^1(0,s), \|f\|_1 \leq 1\} \leq 2^{1/2}d$, where $\mathbb{R}^2$ has the usual Euclidean norm $|\cdot|$. For the special case of the Sobolev space $H^1(0,s)$, *Friedrichs' inequality* [2, Théorème 1.9] is given by

$$(1) \qquad \|f\|_1 \leq c\left(|Tf|^2+\|f'\|^2\right)^{1/2}, \qquad f \in H^1(0,s),$$

where $c>0$ is a constant depending on s. The inequality of Friedrichs is an important tool in the theory of elliptic operators, where it can be used to show the V-ellipticity of certain differential operators which is needed in the existence theory for the corresponding boundary value problems.

The existence of the constants c in (1) is usually shown by an argument using the classical inequality of Poincaré and Banach's open mapping theorem based on a verification that the right-hand side of (1) defines a norm on $H^1(0,s)$ for which $H^1(0,s)$ is complete again. The defect of such a (nonconstructive) proof is the fact that one has no idea how large c is. However, the following concrete bounds are given by Rektorys [3, 18.5, 19]: $c=(1+4s^2/\pi^2)^{1/2}$ which gives 1.18545 for $s=1$.

The aim of this paper is to use variational techniques in order to compute the exact value of the best (i.e. least) constant c in (1). A simple example shows that c is greater than or equal to $(1+s^2/12)^{1/2}>1$: Let f be the function given by $f(x)=x$, $0<x\leq s/2$; $=s/2-x$, $s/2<x<s$; then by (1) one has $c^2 \geq (s^3/12+s)/s$.

* Received by the editors May 10, 1983.

† Seminar für Angewandte Mathematik ETH CH-8092 Zürich, Switzerland.

2. Evaluation of the exact value of the least constant in Friedrichs' inequality. By the following argument it is sufficient to determine the least constant c in (1) on the set $C^1(0,s)$. Let the functional F on $H^1(0,s)$ related to (1) be given by

$$F(f):=\|f\|_1/\left(|Tf|^2+\|f'\|^2\right)^{1/2},\qquad f\in H^1(0,s). \tag{2}$$

Then since

$$c=\sup\{F(f):0\neq f\in H^1(0,s)\},$$

we obviously have

$$c_0:=\sup\{F(f):0\neq f\in C^1(0,s)\}\leq c.$$

Assuming now that $c_0<c$, one would have an f in $H^1(0,s)$ such that $F(f)>c_0$. However, by the continuity of $\|\cdot\|_1$, T and $\|\cdot\|$, this would imply the existence of a g in $C^1(0,s)$ such that $F(g)>c_0$, which would contradict the definition of c_0.

Next, if there is a nonzero f in $C^2(0,s)$ for which F assumes a maximum value, then for each g in $C^1(0,s)$ such that $Tg=0$ one necessarily has $0\leq F(f)^2-F(f+g)^2$. Using (2) and the fact that $T(f+g)=Tf$, one obtains

$$\begin{aligned}0&\leq\left(|Tf|^2\|f\|_1^2+\|f'+g'\|^2\|f\|_1^2-|Tf|^2\|f+g\|_1^2-\|f'\|^2\|f+g\|_1^2\right)/2\\&=|Tf|^2(f,g)_1+\|f\|^2(f',g')-\|f'\|^2(f,g)+o\left(\|g\|_1^2\right),\end{aligned}$$

where $(\cdot,\cdot)$ and $(\cdot,\cdot)_1$ are inner products of $L_2(0,s)$ and $H^1(0,s)$ respectively. Since a partial integration yields $(f',g')=-(f'',g)$ one has

$$0\leq\left(-\left[|Tf|^2+\|f'\|^2\right]f+\left[|Tf|^2-\|f\|^2\right]f'',g\right)+o\left(\|g\|_1^2\right).$$

If $|Tf|=\|f\|$, then the density of $C_0^1(0,s)$ (the subset of functions in $C^1(0,s)$ with compact support in $(0,s)$) in $L_2(0,s)$ and the above inequality imply that $(|Tf|^2+\|f'\|^2)g=0$ and thus $f=0$, a contradiction to $f\in C^2(0,s)\setminus\{0\}$. Therefore, one may assume that $|Tf|\neq\|f\|$ and define

$$\lambda:=-\left(|Tf|^2+\|f'\|^2\right)/\left(|Tf|^2-\|f\|^2\right). \tag{3}$$

Applying the above density argument again, one then obtains Euler's differential equation for the described variational problem:

$$f''+\lambda f=0.$$

It is clear that for the solutions f of this differential equation, $F(f)$ attains a maximum. Since $f\neq0$ implies $\lambda\neq0$ one has to consider the two cases $\lambda=\pm a^2$, $a>0$ with corresponding general solutions

$$f(x)=\cos a\left(x-\frac{s}{2}\right)+b\sin a\left(x-\frac{s}{2}\right),\qquad x\in(0,s),$$

$b\in\mathbb{R}$, where the positive homogeneity of the nominator and denominator of F defined in (2) allows to set the multiplicative constant of the cosine function equal to one and where in the case $\lambda=-a^2$ the notation sin and cos for the corresponding hyperbolic functions sinh and cosh is kept for simplicity of notation.

Using the symmetries with respect to the midpoint $s/2$ of the interval $(0,s)$ one gets

$$(4)\qquad |Tf|^2 = 1+\cos as \pm b^2(1-\cos as),$$

$$(5)\qquad 2\|f\|^2 = s + a^{-1}\sin as \pm b^2(s-a^{-1}\sin as),$$

$$(6)\qquad 2a^{-2}\|f'\|^2 = \pm(s-a^{-1}\sin as)+b^2(s+a^{-1}\sin as),$$

hence by (3) and $\lambda = \pm a^2$

$$a^2 = \pm\frac{2+2\cos as \pm 2b^2(1-\cos as)\pm a^2 s \mp a\sin as + a^2b^2(s+a^{-1}\sin as)}{2+2\cos as \pm 2b^2(1-\cos as) - s - a^{-1}\sin as \mp b^2(s+a^{-1}\sin as)}.$$

Solving for b^2 yields

$$(7)\qquad b^2 = \pm\frac{a\sin as - (a^2\pm 1)(\cos as + 1)}{a\sin as-(a^2\pm 1)(\cos as - 1)}$$

and thus by (2), (4), (5), (6) and a lengthy but elementary calculation

$$(8)\qquad F(f)^2 = 1\pm a^{-2}.$$

From (7) it follows immediately that

$$(9)\qquad a(a^2\pm 1)^{-1}\sin as - \cos as = (1\pm b^2)/(1\mp b^2), \qquad b\in\mathbb{R},$$

where the right-hand side of (9) lies in $\mathbb{R}\setminus[-1,1)$ or $(-1,1]$ for the upper and lower sign respectively. Therefore, for the lower case sign (hyperbolic functions), the possible solutions a of (9) are to be found in the interval $(1,\infty)$. In this case, in view of (8) and the example of §1 which shows that $c>1$, the functions $F(f)$ assume values in $(0,1)$ so that the hyperbolic functions fall out of competition.

On the other hand, for the upper case sign (where $\lambda = a^2$ and the functions in (9) are circular functions) one obtains from the considerations at the beginning of this section

$$(10)\qquad c = (1+a_0^{-2})^{1/2},$$

where a_0 is the smallest number a in $(0,\infty)$ satisfying (9) for some b in $\mathbb{R}$. It is now clear that a_0 is the unique solution a in $(0,\pi/s)$ of

$$(11)\qquad \cos as - a(a^2+1)^{-1}\sin as = -1.$$

From the computational point of view, the root a_0 of (11) in $(0,\pi/s)$ can easily be determined e.g. by Newton's method. For the special case $s=1$ we obtain (up to 5 decimal places) $a_0 = 2.47254$, and thus by (10) the value 1.07869 for c. Finally, it is of interest how a_0 and c behave as s goes to 0 or ∞: It is easy to see that in both cases a_0 is close to π/s asymptotically, hence that c is close to $(1+s^2/\pi^2)^{1/2}$ which shows that c tends to 1 or ∞ respectively.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[3] K. Rektorys, *Variational Methods in Mathematics, Science and Engineering*, Reidel, Dordrecht, 1980.

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

011

STUDY OF THE CORRECTOR OF THE EIGENVALUE OF A TRANSPORT OPERATOR*

RÉMI SENTIS†

Abstract. We consider the transport operator

$$A^\varepsilon = \frac{-1}{\varepsilon}\sum_i v_i \frac{\partial}{\partial x_i} + \frac{1}{\varepsilon^2} Q$$

on the space $L^2(\Omega \times V)$ where Ω is a bounded open set of R^N, V a compact set of R^N and Q a Markovian generator. We show that its largest eigenvalue ω_ε converges (when $\varepsilon \to 0$) to the largest eigenvalue ω of a diffusion operator A on $L^2(\Omega)$ and we calculate the limit of $(1/\varepsilon)(\omega_\varepsilon - \omega)$.

Introduction. Let A^ε be the transport operator defined on $L^2(\Omega \times V)$ by

$$f = f(x,v) \to A^\varepsilon f(w,v) = \frac{-1}{\varepsilon}\sum_i v_i \frac{\partial f}{\partial x_i} + \frac{1}{\varepsilon^2} Qf \qquad (x \in \Omega, \quad v \in V)$$

with zero boundary conditions on $\partial\Omega$, where Ω is a bounded open set of R^N with smooth boundary $\partial\Omega$; V is a compact set of R^N symmetrical with respect to 0, provided with a probability measure (this measure is denoted as a Lebesgue measure for simplicity); and where Q is the operator (depending on the parameter $x \in \Omega$) on $L^2(V)$ defined by

$$g = g(v) \to Q_x g(v) = \int_v \sigma_1(x,v,w) g(w)\, dw - \sigma(x,v) g(v)$$

with σ_1 and σ strictly positive (we emphasize the various assumptions on V, σ_1, σ in §1). ε is a positive real number (assumed to be small). The operator $(A^\varepsilon + \alpha)$ [with α being a real number] is said to be critical (with respect to the domain Ω) if there exists a positive function u in $L^2(\Omega \times V)$ such that:

$$A^\varepsilon u + \alpha u = 0.$$

We recall in §2 that if ω_ε is the maximal real eigenvalue of A^ε, then $(A^\varepsilon + \omega_\varepsilon)$ is critical (that is to say that the corresponding eigenfunction is positive). The goal of this paper is to give an asymptotic expansion of ω_ε with respect to ε.

The main assumptions are the following (let Q'_x be the adjoint of Q_x):

$$(*) \quad \begin{array}{ll} \text{(i)} & Q_x 1 = 0 \quad \forall x \in \Omega, \\ \text{(ii)} & \exists \pi \in L^2(V) |\, Q'_x \pi = 0 \quad \int_V \pi(v) v_i\, dv = 0 \quad (i = 1, 2, \cdots, N). \end{array}$$

Now let us show how a general "collision operator" $\bar{Q}$ may be reduced to one of the above-mentioned type. Let V be a union of spheres centered in 0 (that is to say $V = G \times S^2$ with G being a compact set of R^+) and let $\bar{Q}_x$ be defined by:

$$\bar{Q}_x g(v) = \int \bar{\sigma}_1\left(x, |v|, |w|, \frac{v \cdot w}{|v||w|}\right) g(w)\, dw - \bar{\sigma}(x, |v|) g(v),$$

*Received by the editors December 28, 1981.

†C. E. A. Limeil, B.P. 27, 94190 Villeneuve Saint Georges, France.

where $\bar{\sigma}_1$ is a strictly positive function defined on $\Omega\times G\times G\times[-1,+1]$ and $\bar{\sigma}$ is a strictly positive function defined on $\Omega\times G$. (This corresponds to a "multi-group" model in neutron physics, where the cross sections depend only on the position and the energy of the incident neutron. Modeling with a small parameter ε means that the dimension of the domain Ω is large with respect to the mean free path of the neutrons.)

According to the Krein–Rutman theorem, we know that the maximal eigenvalue of $\bar{Q}_x$ is real and simple. We assume that this eigenvalue is 0. (This restriction is not very strong because the nature of an eigenvalue problem is not changed if one adds a constant.) Thus there exist two functions ρ_x and ρ'_x of $L^2(V)$ such that:

$$\begin{aligned}&\rho_x(v)>0, \quad &&\rho'_x(v)>0, \quad \forall x\in\Omega, \quad \forall v\in V,\\ &\bar{Q}_x\rho_x=0, \quad &&\bar{Q}'_x\rho'_x=0.\end{aligned}$$

Since ρ_x is strictly positive, we may let

$$Q_x g=\frac{1}{\rho_x}\bar{Q}_x(\rho_x g), \qquad \pi(v)=\rho_x(v)\rho'_x(v).$$

Then we have:

$$Q'_x g=\rho_x\bar{Q}'_x\left(\frac{1}{\rho_x}g\right).$$

Assume that π does not depend on x. (This is true in "monogroup" models, where $V=S^2$ and $\rho_x=\rho'_x=1$. This is also true, obviously, in the case where Q_x does not depend on x.) Then Q satisfies (*). Indeed, due to the spherical symmetry, we have:

$$\int_v \rho_x(v)\rho'_x(v)v_i\,dv=0.$$

An outline of the paper follows: In §1, we give preliminary results. Particularly we recall that there exists an unbounded diffusion operator A on $L^2(\Omega)$:

$$A=\sum_{ij}\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial\cdot}{\partial x_j}\right)$$

with zero boundary Dirichlet conditions, which is the limit of A^ε in the following sense:

$$(A^\varepsilon-\alpha)^{-1}f \xrightarrow[L^2(\Omega\times V)]{} (A-\alpha)^{-1}\left[\int_V f(\cdot,v)\pi(v)\,dv\right] \quad \forall f\in L^2(\Omega\times V), \quad \forall\alpha>0.$$

In §2, we show that the types ω_ε and ω of the semigroups $e^{A^\varepsilon t}$ and e^{At} are also simple eigenvalues of A^ε and A. In §3, we use a characterisation of the type of a semigroup in order to find a lower bound for ω_ε. And in §4 we prove that an eigenfunction of A^ε corresponding to ω_ε converges in $L^2(\Omega\times V)$ to an eigenfunction of A corresponding to ω, and that:

$$\omega_\varepsilon\to\omega \quad \text{when } \varepsilon\to 0.$$

Finally the calculation of the limit of $(\omega_\varepsilon-\omega)/_\varepsilon$ is developed in §5.

1. The problem. Let Ω be an open bounded subset of R^N, with a smooth boundary $\partial\Omega$. Assume that Ω is connected.

Let V be a compact subset of R^N symmetrical with respect to 0, provided with a probability measure symmetrical with respect to 0, whose support is not contained in any hyperplane of R^N, that is to say there exists a positive constant c_0, such that:

$$(1) \qquad \sum_{i,j}\left(\int_V v_i v_j dv\right)\xi_i\xi_j \geq c_0|\xi|^2 \quad \forall\xi\in R^N.$$

We assume that

$$(2) \qquad 0\notin V.$$

In the Hilbert spaces $L^2(\Omega\times V)$, $L^2(V)$, $L^2(\Omega)$ we denote the scalar products by:

$$\langle\cdot,\cdot\rangle, (\cdot,\cdot)_V, (\cdot,\cdot)_\Omega \qquad (\text{or } (\cdot,\cdot))$$

and the norms by:

$$\|\cdot\|, |\cdot|_{L^2(V)}, |\cdot|_{L^2(\Omega)} \qquad \left(\text{or } |\cdot|_{L^2} \text{ if there is no ambiguity}\right),$$

respectively.

Let us denote also by $C(\Omega\times V)$, $C(V)$, $C(\Omega)$ the spaces of continuous functions from $(\Omega\times V)$, V or Ω into R.

Let us define the operator Q on $L^2(V)$ (depending on the parameter x) by

$$Q_x f = K_x f - \sigma f,$$

with

$$K_x f = \int_V \sigma_1(x,v,w)f(w)\,dw, \qquad \sigma(x,v) = K_x 1(v),$$

where $\sigma_1\in C(\Omega\times V\times V)$. σ_1 is smooth in x and satisfies:

$$(3) \qquad \exists\sigma_l\in\mathbb{R}^+, \quad 0<\sigma_l\leq\sigma_1(x,v,w).$$

We may consider Q as an operator defined on $C(V)$ (and also on $L^2(\Omega\times V)$ or $C(\Omega\times V)$). From (3) we know (see for example Blankenship and Papanicolaou [4]) that 0 is a simple eigenvalue of Q and that there exists a unique eigenfunction π of Q' which is a probability measure density (Q' is the adjoint of Q). From (3) we also know that

$$\exists\pi_l\in\mathbb{R}^+, \quad 0<\pi_l\leq\pi(v).$$

Assume that π does not depend on x and that

$$(4) \qquad (\pi,v_i)_V = 0 \quad \forall i.$$

If $\sigma_1(x,v,w)=\sigma_1(x,w,v)$ then we have $\pi=1$ and (4) is satisfied. Moreover, if $f\in L^2(\Omega)$, it is necessary and sufficient for the existence of a solution u of

$$Qu+f=0,$$

that $(\pi,f)=0$.

Thus let $\zeta_i=\zeta_i^x$ be the function of $C(V)$ satisfying:

$$Q_x\zeta_i^x - v_i = 0, \qquad (\zeta_i^x,\pi)=0.$$

For any x on $\partial\Omega$, let us denote by n_x the unit outward normal to $\partial\Omega$,

$$\Gamma_x^- = \{v\in V \text{ such that } n_x\cdot v<0\}, \qquad \Gamma^- = \left\{(x,v)\in\partial\Omega\times V | v\in\Gamma_x^-\right\}.$$

For any positive ε, let us define the operators Λ and A^ε on $L^2(\Omega\times V)$ by:

$$\Lambda f=-v\frac{\partial f}{\partial x}\qquad \left(v\frac{\partial}{\partial x}=\sum_i v_i\frac{\partial}{\partial x_i}\right),$$

$$A^\varepsilon f=\frac{1}{\varepsilon}\Lambda f+\frac{1}{\varepsilon^2}Qf,$$

$$D(A^\varepsilon)=D(\Lambda)=\left\{f\in L^2(\Omega\times V)\text{ such that }v\frac{\partial f}{\partial x}\in L^2(\Omega\times V),\ f|_{\Gamma^-}=0\right\}.$$

The following proposition will be useful.

PROPOSITION 1. i) *Assume that* (3) *holds. We have for any f in* $L^2(V)$:

$$(Qf,\pi f)_V\leq-\frac{1}{2}\sigma_l\pi_l\int_V\int_V[f(v)-f(w)]^2dv\,dw.$$

ii) A^ε *is the infinitesimal generator of a semigroup* T_t^ε *of class* C^0 *on* $L^2(\Omega\times V)$, *which is bounded uniformly with respect to t and* ε.

Proof. i) Here we do not write the parameter x. Since π satisfies

$$\int_V\sigma_1(w,v)\pi(w)\,dw=\sigma(v)\pi(v),$$

we have

$$(\sigma f,\pi f)_V=\int_V\int_V\sigma_1(v,w)\pi(v)f(v)^2dw\,dv=\int_V\int_V\sigma_1(w,v)\pi(w)f^2(v)\,dw\,dv$$
$$=\frac{1}{2}\int_V\int_V\sigma_1(v,w)\pi(v)\left[f(v)^2+f(w)^2\right]dv\,dw.$$

This yields the result, indeed we have:

$$(Qf,\pi f)_V=-\frac{1}{2}\int_V\int_V\sigma_1(v,w)\pi(v)\left[-2f(v)f(w)+f(v)^2+f(w)^2\right]dv\,dw.$$

ii) In the Hilbert space $L^2(V)$ provided by the scalar product $\{f;\ g\}\mapsto(f,\pi g)_V$, the operator Q is dissipative. Since we have

$$\int_V\int_V v\frac{\partial f}{\partial x}(x,v)f(x,v)\pi\,dx\,dv\leq 0\quad\forall f\in L^2(\Omega\times V),$$

the operator A^ε is also dissipative in the Hilbert space $L^2(\Omega\times V)$ provided by the scalar product $\{f;\ g\}\to\langle f,\pi g\rangle$. Therefore, we have

$$\|\pi^{1/2}T_t^\varepsilon f\|\leq\|\pi^{1/2}f\|\qquad\forall f\in L^2(\Omega\times V),$$
$$\pi_l^{1/2}\|T_t^\varepsilon f\|\leq|\pi|_0^{1/2}\|f\|\quad\forall f.$$

Q.E.D.

Let us define the unbounded operator A on $L^2(\Omega)$ by

$$Af=\sum_{i,j}\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial f}{\partial x_j}\right),\quad\text{where } a_{ij}(x)=\left(\pi,-v_i\zeta_j^x\right)_V,$$

$$D(A)=H^2(\Omega)\cap H_0^1(\Omega)\text{ (using the usual Sobolev notation).}$$

We know that the coefficients a_{ij} are smooth with respect to x. Let us show now that the matrix a_{ij} is strictly positive definite. Indeed, let ξ belong to R^N and $g(v)=\Sigma_i v_i \xi_i$; then we have

$$\sum_{i,j} a_{ij}\xi_i\xi_j = -(\pi g, Qg)_V \geq \frac{1}{2}\sigma_l \pi_l \int_V \int_V [g(v)-g(w)]^2 dv\,dw \geq 0.$$

And if we had $\Sigma a_{ij}\xi_i\xi_j=0$, then g would be constant, that is to say, $g(v)=0$ almost everywhere (because $(g,1)_V=0$). Therefore $\xi=0$ (due to (1)). So A is the generator of a contraction semigroup T_t, of class C^0, defined on $L^2(\Omega)$. (We may also consider A and T_t as operators on $C(\Omega)$.)

The following result is classical (see Blankenship and Papanicolaou [4], Bensoussan, Lions and Papanicolaou [3] for the case where $\Omega=R^N$, or Williams [16], or Sentis [13], [14, §3]). When ε goes to 0, for any $t>0$, we have:

$$\begin{aligned} &T_t^\varepsilon f \to T_t f && \text{in } C(\Omega\times V) && \forall f\in C^\infty(\Omega) \text{ s.t. } f|_{\partial\Omega}=0,\\ &T_t^\varepsilon f \to T_t \Pi f && \text{in } C(\Omega\times V) && \forall f\in C^0(\Omega\times V) \text{ s.t. } f|_{\partial\Omega}=0 \text{ and } f(\cdot,v)\in C^\infty(\Omega), \end{aligned}$$

where Π is the projection defined by

$$(\Pi f)(x)=(\pi,f(x,\cdot))_V \quad \forall f\in L^2(\Omega\times V).$$

Since T_t^ε is uniformly bounded with respect to ε, on $L^2(\Omega\times V)$ and $C(\Omega\times V)$, using the previous convergence, we can see that, for any $t>0$, when $\varepsilon\to 0$.

$$T_t^\varepsilon f \to T_t f \quad \text{in } C(\Omega\times V) \quad \forall f\in C(\Omega) \text{ s.t. } f|_{\partial\Omega}=0, \tag{5}$$

$$T_t^\varepsilon f \to T_t \Pi f \quad \text{in } L^2(\Omega\times V) \quad \forall f\in L^2(\Omega\times V). \tag{6}$$

Hence, we see that, for any positive real number α:

$$(A^\varepsilon-\alpha)^{-1}\to(A-\alpha)^{-1}\Pi$$

by the strong convergence of the operators on $L^2(\Omega\times V)$, when $\varepsilon\to 0$.

These results (which may be interpreted as singular perturbation results) mean that the transport process associated with the Markovian generator A^ε may be approximated well by the diffusion process associated with the diffusion operator A. (More precisely the "spatial part" of the transport process converges weakly, as $\varepsilon\to 0$, to the diffusion process.) This kind of result has been well known for a long time to the specialists in neutronics (e.g., Larsen and Keller [10]) and in probabilities (e.g., Khasminskii [7]).

Let ω_ε and ω be the types of the semigroups T_t^ε and T_t (considered as operators on $L^2(\Omega\times V)$ and $L^2(\Omega)$). Recall that the type ω of a semigroup T_t is defined by:

$$\omega=\lim_{t\to\infty}\frac{1}{t}\log\|T_t\|.$$

These two types are nonpositive. Now we shall show that they are also eigenvalues of A^ε and A.

2. Spectral properties of A and A^ε. It is well known that, for the operator A, the eigenvalue with the largest real part is real (see, for example, Protter and Weinberger

[12]). But we shall prove here, following Amann [1], that:

PROPOSITION 2. i) *The type* ω *of* T_t *is an eigenvalue of* A *and there exists a unique function* ϕ *such that*[1]

$$\phi \in D(A)\cap L^2_+(\Omega),\quad A\phi=\omega\phi,\quad |\phi|_{L^2}=1. \tag{7}$$

ii) *Furthermore* ϕ *is smooth, and for any compact set* Ω' *with* $\Omega'\subset\Omega$

$$\inf\left(\phi(x)|x\in\Omega'\right)>0. \tag{8}$$

iii) *If we let* $E=\{f\in C(\Omega)$ *such that* $\exists\alpha>0, \exists\beta>0\ -\beta\phi\le f\le\alpha\phi\}$, *then there exists a function* $\hat{\phi}$ *of* $C(\Omega)$ *having a lower bound like* (8) *and*:

$$\forall f\in E\quad e^{-\omega t}T_t f\to\phi(f,\hat{\phi})\quad \text{in } C(\Omega) \text{ when } t\to\infty. \tag{9}$$

Proof. We can write A in the following form:

$$Af=\alpha_{ij}\frac{\partial^2 f}{\partial x_i\partial x_j}+b_i\frac{\partial f}{\partial x_i}\qquad \text{with } \alpha_{ij}=\frac{1}{2}(a_{ij}+a_{ji}),$$

$$D(A)=H^2(\Omega)\cap H^1_0(\Omega),\quad b_i=\frac{\partial}{\partial x_j}a_{ji}.$$

Let us denote by $\tilde{A}$ the operator A considered as an unbounded operator on $C(\Omega)$. Then according to Amann [1, Thm. 1.16] applied to the Dirichlet boundary value problem, we know that there exists a real eigenvalue ω_0 of $\tilde{A}$ (with finite multiplicity) such that $\mathrm{Re}\lambda<\omega_0$ for any eigenvalue λ ($\lambda\neq\omega_0$) of $\tilde{A}$, such that the corresponding eigenfunction ϕ is unique up to a multiplicative constant and such that

$$\tilde{A}\phi=\omega_0\phi,\qquad \phi\in C_+(\Omega).$$

We know also that the semigroup generated by $\tilde{A}$ is compact on $C(\Omega)$ thus $e^{\omega_0 t}$ is equal to its spectral radius which is $e^{\omega t}$. Then: $\omega_0=\omega$.

Now, any eigenfunction f of A is smooth (indeed we have $f\in D(A^n)\subset H^{2n}$ for any n in $\mathbb{N}$) and the eigenfunctions of A and $\tilde{A}$ are the same. Thus there exists a unique function $\hat{\phi}$ satisfying (7).

Since ϕ may reach its minimum only on $\partial\Omega$, we have (8). If we denote by A' the adjoint of A, we have the same result for A'. That is to say, there exists a unique function $\hat{\phi}$ satisfying

$$\hat{\phi}\in D(A')\cap L^2_+(\Omega),\quad A'\hat{\phi}=\omega\hat{\phi},\quad (\phi,\hat{\phi})=1.$$

Moreover inf $(\hat{\phi}(x)|x\in\Omega')>0$, for any compact subset Ω' with $\Omega'\subset\Omega$. Finally we can see that E provided with the norm:

$$\|f\|_E=\sup_{\Omega}|f/\phi|$$

is a Banach space. Since $e^{\tilde{A}t}\phi=e^{\omega t}\phi$, the semigroup $e^{\tilde{A}t}$ can be restricted to E and is of class C^0 on E and its infinitesimal generator is $\tilde{A}$ with the domain

$$D(\tilde{A})=\left\{f\in E \text{ s.t. } \alpha_{ij}\frac{\partial^2 f}{\partial x_i\partial x_j}\in E\right\}.$$

[1] Let $L^2_+(\Omega)$ be the cone of nonnegative functions in $L^2(\Omega)$ and $C_+(\Omega)$ the cone of nonnegative functions in $C(\Omega)$.

Since ω is an eigenvalue of A with finite multiplicity and since we can show easily that:

$$\left\|e^{-\omega t}e^{\tilde{A}t}f\right\|_E \leq \|f\|_E \quad \forall f\in E, \quad \forall t>0.$$

We know that the eigenvalue ω is semisimple. (That is to say: $(\lambda-\omega)(\lambda-\tilde{A})^{-1}$ converges strongly on E, when $\lambda \to \omega^+$. This limit is the spectral projector P related to ω. See Kato [8, §§III 6.5 and I 5.4].) Thus, we have

$$\forall f\in E, \quad e^{-\omega t}e^{\tilde{A}t}f \to Pf \quad \text{in } E, \quad \text{when } t\to\infty.$$

Thus the range of P is the eigenspace related to ω. On the other hand the dimension of this eigenspace is 1, thus there exists m in the dual of E satisfying:

$$Pf=\phi m(f), \qquad m(\phi)=1.$$

Since we have:

$$e^{-\omega t}\left(e^{\tilde{A}t}f,\hat{\phi}\right)_{L^2}=(f,\hat{\phi})_{L^2} \quad \forall f\in E,$$

we see that $m(f)=(f,\hat{\phi})_{L^2}$. Then (9) follows. Q.E.D.

Now, let us give a general result which we shall apply to the operator A^ε:

PROPOSITION 3. *Let B be a Banach space and B_+ a convex closed cone such that B is the closed hull of (B_+-B_-). Let L be the infinitesimal generator of a semigroup T_t (strongly continuous on B) such that for t large enough*:

i) *T_t is compact*,

ii) $T_t(B_+)\subset B_+$.

Then the type ω of this semigroup is (if $\omega\neq-\infty$) an eigenvalue with finite multiplicity of L and

$$\exists\phi\in B_+\cap D(L), \qquad L\phi=\omega\phi, \quad \phi\neq 0.$$

Proof. Let us fix t such that i) and ii) are satisfied. The type ω of the semigroup satisfies:

$$\omega t=\lim_{n\to\infty}\frac{1}{n}\log\|T_{nt}\|=\log\lim_{n\to\infty}\left(\|T_t{}^n\|\right)^{1/n}.$$

Thus $e^{\omega t}$ is the spectral radius of T_t if $\omega\neq-\infty$. For t large enough, T_t is compact; therefore (since $T_t(B_+)\subset B_+$) we can apply the weak Krein–Rutman theorem (Krein and Rutman [6, Thm. 6.17]) and we see that the spectral radius of T_t is an eigenvalue associated with an eigenfunction ϕ_t of B_+:

$$\phi_t\in B_+, \quad T_t\phi_t=e^{\omega t}\phi_t, \quad \phi_t\neq 0.$$

On the other hand we know that the eigenspace corresponding to $e^{\omega t}$ is independent of t. Hence, ϕ_t does not depend on t and for any $t>0$ we have

$$T_t\phi=e^{\omega t}\phi.$$

Since

$$\frac{T_t\phi-\phi}{t}\to L\phi \quad \text{when } t\to 0,$$

we have the desired result. Q.E.D.

According to Jörgens [5] and (2), for t large enough T_t^ε is compact from $L^2(\Omega\times V)$ to $L^2(\Omega\times V)$. Since T_t^ε leaves invariant the cone $L_+^2(\Omega\times V)$ of positive functions of $L^2(\Omega\times V)$, we can apply the foregoing and if $\omega_\varepsilon\neq-\infty$, there exists an eigenfunction ϕ_ε satisfying:

$$\phi_\varepsilon\in D(A^\varepsilon)\cap L_+^2(\Omega\times V),\quad A^\varepsilon\phi_\varepsilon=\omega_\varepsilon\phi_\varepsilon,\qquad \|\phi_\varepsilon\|=1.$$

We call ϕ_ε a principal eigenfunction of A^ε. The uniqueness of ϕ_ε is now proved.

Remark 1. For ε small enough, the dimension of the eigenspace of A^ε related to ω_ε is 1.

Sketch of the proof (*due to Bardos* [2]). Let us choose ε such that $(\sigma+\varepsilon^2\omega_\varepsilon)$ is strictly positive and let $q=\varepsilon^{-1}(\sigma+\varepsilon^2\omega_\varepsilon)$ and $H=\varepsilon^{-1}K$. If there exists an eigenfunction X linearly independent of ϕ_ε, there also exists a real number α such that $\psi=\phi_\varepsilon+\alpha X$ satisfies

$$(10)\qquad \begin{array}{lll} \text{i)} & (\Lambda+H-q)\psi=0, & \psi\in L^2(\Omega\times V),\\ \text{ii)} & \psi^+\neq0, & \psi^-\neq0,\end{array}$$

where ψ^+ and ψ^- are the positive and negative parts of ψ ($\psi=\psi^+-\psi^-$). Let $B=(q-\Lambda)^{-1}$ which is a bounded operator on $L^2(\Omega\times V)$. Thanks to (10i), we have

$$\psi=BH\psi=(BH)^2\psi.$$

Let us denote by Λ', B' and H' the adjoint operators of Λ, B and H, respectively. We know that there exists an eigenfunction $f\in L_+^2(\Omega\times V)$ of $(\Lambda'+H'-q)$ and thus

$$f=B'H'f,\qquad f\neq0.$$

If $u=H'f$, we have $u=H'B'u$. Therefore, we have

$$(11)\qquad \langle|\psi|,u\rangle=\langle|\psi|,(H'B')^2u\rangle=\langle(BH)^2|\psi|,u\rangle=\langle|(BH)^2\psi|,u\rangle.$$

On the other hand we can show easily that there exists a constant $\beta>0$ such that

$$(HBH)g(x,v)\geq\beta\langle g,1\rangle\quad\forall(x,v),\quad\forall g\in L_+^2(\Omega\times V).$$

Thus let us define the function γ from Ω to $\mathbb{R}$, by

$$\gamma(x)=\frac{\beta}{\sup q}\left[1-\exp\left(-(\inf q)\cdot\frac{d(x,\Omega)}{\sup_V|v|}\right)\right]$$

Then we have $\gamma(x)>0$ for any x in Ω and

$$(BH)^2g(x,v)\geq\gamma(x)\langle g,1\rangle\quad\forall(x,v)\quad\forall g\in L_+^2(\Omega\times V).$$

Similarly, we can show that $u(x)>0$ for any x in Ω. Thus, we have

$$\begin{aligned}(BH)^2\psi(x,v)&\leq(BH)^2|\psi|(x,v)-2\gamma(x)\langle\psi_-,1\rangle,\\ -(BH)^2\psi(x,v)&\leq(BH)^2|\psi|(x,v)-2\gamma(x)\langle\psi_+,1\rangle.\end{aligned}$$

Thanks to (10ii), this yields:

$$\langle|(BH)^2\psi|,u\rangle<\langle(BH)^2|\psi|,u\rangle$$

which contradicts (11). Q.E.D.

3. Lower bound for the type ω_ε. Let us denote by $\underline{\lim}$ and $\overline{\lim}$ the lower limit and upper limit.

PROPOSITION 4. *When ε goes to* 0, *we have*

$$\underline{\lim}_{\varepsilon}\, \omega_\varepsilon \geq \omega.$$

For the proof we use an idea due to S. Varadhan communicated by G. C. Papanicolaou [11], introducing the quantities $\gamma^\varepsilon(t)$ and $\gamma(t)$.

We need only the positivity of T_t^ε and T_t and the properties (5), (8), (9), to prove this proposition.

Proof. The usual norm of the operators on $L^2(\Omega\times V)$ is denoted by $\|\cdot\|$. Let Ω_0 and Ω_1 be two open sets such that

$$\varnothing \neq \Omega_1 \subset \overline{\Omega}_1 \subset \Omega_0 \subset \overline{\Omega}_0 \subset \Omega.$$

Let f be a continuous function on Ω such that

$$0 \leq f(x) \leq 1 \quad \text{where} \quad f(x)=\begin{cases} 0 & \text{if } x \notin \Omega_0, \\ 1 & \text{if } x \in \Omega_1. \end{cases}$$

From (8) and (9) we know there exist two constants ρ_0 and ρ_1 such that for t large enough,

$$0<\rho_0 \leq e^{-\omega t}(T_t f)(x) \leq \rho_1 \quad \forall x \in \Omega_0. \tag{12}$$

On the other hand, let us define:

$$\gamma^\varepsilon(t)=\frac{1}{t}\log\left[\inf_{x\in\Omega_0}(T_t^\varepsilon f)(x)\right],$$

$$\gamma(t)=\frac{1}{t}\log\left[\inf_{x\in\Omega_0}(T_t f)(x)\right].$$

On account of (12), $\gamma(t)$ is finite for t large enough and when $t\to\infty$, we have

$$0=\lim_{t\to\infty}\left(\frac{1}{t}\log\rho_0\right) \leq \underline{\lim}_{t\to\infty}\,(\gamma(t)-\omega) \leq \overline{\lim}_{t\to\infty}\left(\frac{1}{t}\log\rho_1\right)=0.$$

Thus,

$$\lim_{t\to\infty}\gamma(t)=\omega. \tag{13}$$

On the other hand, by (5), when ε goes to 0

$$\inf_{x\in\Omega_0} T_t^\varepsilon f(x) \to \inf_{x\in\Omega_0} T_t f(x).$$

Hence, we have, when ε goes to 0

$$\gamma^\varepsilon(t)\to\gamma(t). \tag{14}$$

In particular, we see that $\gamma^\varepsilon(t)$ is finite for ε small enough. Let us assume for the moment the following lemma.

LEMMA 1. *For fixed* ε, *we have*:

$$\lambda^\varepsilon = \sup_{t\in\mathbb{R}_+}\gamma^\varepsilon(t)=\overline{\lim}_{t\to\infty}\gamma^\varepsilon(t).$$

Thanks to (14) we have, for any t, when ε goes to 0,

$$\underline{\lim}_{\varepsilon}\lambda^{\varepsilon}\geq\underline{\lim}_{\varepsilon}\gamma^{\varepsilon}(t)=\gamma(t).$$

And due to (13), we conclude that:

$$\underline{\lim}_{\varepsilon}\lambda^{\varepsilon}\geq\omega.$$

To complete the proof it suffices to show that $\lambda^{\varepsilon}\leq\omega^{\varepsilon}$ for any ε. Now, since $\|f\|^2_{L^2(\omega\times V)}\leq c|\Omega_0|$ (where $|\Omega_0|$ denotes the measure of Ω_0), we have

$$\|T_t^{\varepsilon}\|^2\geq\frac{1}{\|f\|^2}\|T_t^{\varepsilon}f\|^2\geq\frac{|\Omega_0|}{\|f\|^2}\left|\inf_{\Omega_0}T_t^{\varepsilon}f\right|^2\geq\left(e^{t\gamma^{\varepsilon}(t)}\right)^2c\quad\forall t.$$

Therefore the type ω_{ε} of T_t^{ε} is finite. For ω' larger than ω_{ε}, there exists a constant $M_{\omega'}$ such that

$$\|T_t^{\varepsilon}\|\leq M_{\omega'}e^{\omega' t}\quad\forall t.$$

Hence,

$$c^{1/2}e^{t\gamma^{\varepsilon}(t)}\leq M_{\omega'}e^{\omega' t}\quad\forall t$$

and

$$\lambda^{\varepsilon}=\overline{\lim}_{t\to\infty}\gamma^{\varepsilon}(t)\leq\overline{\lim}_{t\to\infty}\left(\omega'+\frac{1}{t}\log M_{\omega'}\right)=\omega'.$$

Therefore

$$\lambda^{\varepsilon}\leq\omega^{\varepsilon}.$$

Proof of Lemma 1. Let $\lambda^{\varepsilon}=\sup_t\gamma^{\varepsilon}(t)$.

Let χ_0 be the characteristic function of the open set Ω_0 and let

$$\delta_t=\inf_{x\in\Omega_0}(T_t^{\varepsilon}f)(x),\quad\left(\gamma^{\varepsilon}(t)=\frac{1}{t}\log\delta_t\right).$$

Since $T_s^{\varepsilon}f\geq\chi_0T_s^{\varepsilon}f$ and $T_t^{\varepsilon}\chi_0(x)\geq T_t^{\varepsilon}f(x)$, $\forall x\in\Omega$ we have:

$$\delta_{t+s}\geq\inf_{\Omega_0}T_t^{\varepsilon}(\chi_0T_s^{\varepsilon}f)\geq\inf_{\Omega_0}T_t^{\varepsilon}\left(\chi_0\left(\inf_{\Omega_0}T_s^{\varepsilon}f\right)\right)\geq\delta_t\delta_s. \tag{15}$$

Due to the definition of λ_{ε}, for any small ρ ($\rho>0$), there exists a positive number a such that:

$$\lambda^{\varepsilon}-\rho\leq\frac{1}{a}\log\delta_a\leq\lambda_{\varepsilon}.$$

By (15), it follows that

$$\frac{1}{na}\log\delta_{na}\geq\frac{1}{na}\log(\delta_a)^n=\frac{1}{a}\log\delta_a\geq\lambda^{\varepsilon}-\rho\quad\forall n\in N.$$

We conclude that for any small ρ ($\rho>0$), we have

$$\overline{\lim}_{t\to\infty}\left(\frac{1}{t}\log\delta_t\right)\geq\lambda^{\varepsilon}-\rho.$$

4. Convergence of ω_ε and of the principal eigenfunction.

PROPOSITION 5. *Let ϕ_ε be a principal eigenvalue of A^ε. Then there exists a function $\psi \in L^2(\Omega)$ such that*

$$\phi_\varepsilon \to \psi, \quad \text{in } L^2(\Omega \times V), \text{ when } \varepsilon \to 0.$$

Moreover we have

$$|\psi|_{L^2(\Omega)} = 1, \qquad \psi \geq 0.$$

First let us give a lemma in which the parameter x does not appear.

LEMMA 2. *Let f_n be a sequence of elements of $L^2(V)$ such that:*

$$(16) \qquad \begin{array}{ll} \text{i)} & \overline{\lim_{n\to\infty}} |f_n|_{L^2(V)} < +\infty, \\ \text{ii)} & (Qf_n, \pi f_n)_V \to 0. \end{array}$$

Then the sequence $|f_n|_{L^2(V)}$ converges and

$$f_n \to \pm \lim |f_n|_{L^2} \quad \text{in } L^2(V).$$

Proof of Lemma 2. First let us show the following inequality (due to Tartar [15])

$$(17) \qquad |f-(f,1)_V|^2_{L^2} \leq \frac{-1}{\sigma_l \pi_l}(Qf, \pi f)_V \quad \forall f \in L^2(V).$$

Indeed, we have, thanks to Proposition 1,

$$(Qf, \pi f)_V \leq -\frac{1}{2}\sigma_l \pi_l \int_V \int_V |f(w)-f(v)|^2 dv\, dw.$$

Now we have (provided that $\int_V dv = 1$)

$$\int_V \int_V |f(w)-f(v)|^2 dv\, dw = 2\int_V |f(v)|^2 dv - 2(f,1)^2_V$$

$$= 2\int_V \left[|f(v)|^2 - (f,1)^2_V\right] dv = 2\int_V [f-(f,1)_V]^2 dv.$$

Therefore (17) holds, and by (16ii) we have that when n goes to ∞

$$f_n - (f_n, 1)_V \to 0 \quad \text{in } L^2(V).$$

On the other hand, by (16i), there exists a subsequence of f_n, still denoted by f_n, which converges to a function f of $L^2(V)$ in $L^2(V)$ weakly. Thus we have

$$(f_n, 1)_V \to (f, 1)_V \quad \text{when } n \to \infty,$$

and it follows that

$$f_n \to (f,1)_V \quad \text{in } L^2(V).$$

Thus the whole sequence $|f_n|_{L^2}$ converges, and we have

$$(f,1)_V = \pm \lim |f_n|_{L^2}. \qquad \text{Q.E.D.}$$

Proof of Proposition 5. Since $\phi_\varepsilon \in D(A^\varepsilon)$ we know that:

$$\left\langle -v\frac{\partial\phi_\varepsilon}{\partial x}, \pi\phi_\varepsilon \right\rangle \leq 0.$$

Then, because of Proposition 2, we have

$$-\frac{1}{\varepsilon^2}\langle Q\phi_\varepsilon, \phi_\varepsilon\pi\rangle = \frac{1}{\varepsilon}\left\langle -v\frac{\partial\phi_\varepsilon}{\partial x}, \pi\phi_\varepsilon\right\rangle - \omega_\varepsilon\langle\phi_\varepsilon^2, \pi\rangle \leq |\omega_\varepsilon|\cdot C' \leq C.$$

Hence it follows:

$$\langle Q\phi_\varepsilon, \phi_\varepsilon\pi\rangle \to 0.$$

Since $(Q_x\phi_\varepsilon(x,\cdot), \pi\phi_\varepsilon(x,\cdot))_V$ is nonpositive, there exists a measurable set Ω_0 (such that $\Omega\setminus\Omega_0$ is a negligible set) satisfying, $\forall x \in \Omega_0$.

i) $\overline{\lim_\varepsilon} |\phi_\varepsilon(x,\cdot)|_{L^2} < +\infty$,

ii) $(Q_x\phi_\varepsilon(x,\cdot), \pi\phi_\varepsilon(x,\cdot))_V \to 0$.

Using the preceding lemma, we can see that for any x in Ω_0 there exists a number ψ_x in R^+ such that, when $\varepsilon \to 0$

$$\phi_\varepsilon(x,\cdot) \to \psi_x \quad \text{in } L^2(V), \qquad \psi_x = \lim_\varepsilon |\phi_\varepsilon(x,\cdot)|_{L^2}. \tag{18}$$

We see that the function $x \to \psi_x$ is measurable, and if we denote by $\psi \in L^2(\Omega)$ equal to ψ_x on Ω_0 and zero elsewhere, we have

$$\int_\Omega |\psi_x|^2 dx = \int_\Omega \lim_\varepsilon |\phi_\varepsilon(x,\cdot)|^2_{L^2(V)} dx = \lim_\varepsilon \|\phi_\varepsilon\|^2 = 1,$$

and

$$\phi_\varepsilon \to \psi \quad \text{in } L^2(\Omega \times V). \qquad \text{Q.E.D.}$$

PROPOSITION 6. *When ε goes to* 0, *we have*

$$\omega_\varepsilon \to \omega$$

and the limit ψ of ϕ_ε is equal to the eigenfunction ϕ of A satisfying

$$\phi \in D(A), \quad A\phi = \omega\phi, \quad \phi \geq 0, \quad |\phi|_{L^2(\Omega)} = 1. \tag{19}$$

Before proving this proposition, let us state a lemma.

LEMMA 3.[2] *For any f in $\mathcal{D}(\Omega)$, there exists a family f_ε in $C(\Omega \times V)$ such that, when ε goes to* 0,

i) $f_\varepsilon \to \pi f$ *in* $C(\Omega \times V)$,

ii) $A^{\varepsilon *} f_\varepsilon \to A^* f$ *in* $C(\Omega \times V)$.

Proof. (Here and in the following we use summation convention.) It may be seen that the solvability condition for the equation $Q'u + g = 0$ is:

$$(1, g)_V = 0.$$

[2] The adjoint of an operator (with respect to the L^2 duality) is denoted by *, and the set of smooth functions defined on Ω, with compact support in Ω, is denoted by $\mathcal{D}(\Omega)$.

Therefore, let $\bar{\zeta}_i^x$ be the solution of

$$Q'_x\bar{\zeta}_i^x+v_i\pi=0, \qquad (\bar{\zeta}_i^x,1)=0.$$

Let us define:

$$f_\varepsilon=\pi f+\varepsilon\bar{\zeta}_i^x\frac{\partial f}{\partial x_i}+\varepsilon^2 f_2,$$

where $f_2(x,\cdot)$ is a solution of

$$Q'_x f_2(x,\cdot)+v_i\frac{\partial}{\partial x_i}\left(\bar{\zeta}_j^x\frac{\partial f}{\partial x_j}\right)-\frac{\partial}{\partial x_i}\left(\left(1,v_i\bar{\zeta}_j^x\right)\frac{\partial f}{\partial x_j}\right)=0.$$

We see that f_2 belongs to $C(\Omega\times V)$ and i) follows. We have also $(\partial f_2/\partial x_i)\in C(\Omega\times V)$. On the other hand, we have

$$A^{\varepsilon *}f_\varepsilon=\frac{1}{\varepsilon}\left(Q'\bar{\zeta}_i\frac{\partial f}{\partial x_i}+v_i\pi\frac{\partial f}{\partial x_i}\right)+\left[Q'f_2+v_i\frac{\partial}{\partial x_i}\left(\bar{\zeta}_j\frac{\partial f}{\partial x_j}\right)\right]+\varepsilon v_i\frac{\partial f_2}{\partial x_i}$$

$$=\frac{\partial}{\partial x_i}\left(a_{ji}\frac{\partial f}{\partial x_j}\right)+\varepsilon v_i\frac{\partial f_2}{\partial x_i}\to A^*f \quad \text{in } C(\Omega\times V).$$

Indeed we have $a_{ji}(x)=(-\pi v_j,\zeta_i^x)_V=(Q'\bar{\zeta}_j^x,\zeta_i^x)_V=(\bar{\zeta}_j^x,Q\zeta_i^x)_V=(v_i,\bar{\zeta}_j^x)_V$. Q.E.D.

Proof of Proposition 6. Using the preceding lemma for any f in $\mathfrak{D}(\Omega)$ we have:

$$\langle A^\varepsilon\phi_\varepsilon,f_\varepsilon\rangle=\omega_\varepsilon\langle\phi_\varepsilon,f_\varepsilon\rangle=\langle\phi_\varepsilon,A^{\varepsilon *}f_\varepsilon\rangle,$$
$$f_\varepsilon\to\pi f, \quad A^{\varepsilon *}f_\varepsilon\to A^*f \quad \text{in } L^2(\Omega\times V).$$

Let $\omega'=\overline{\lim}_\varepsilon\,\omega_\varepsilon$. There exists a sequence ε_n converging to 0 such that

$$\omega_{\varepsilon_n}\to\omega'.$$

Then we have, due to Proposition 5,

$$\omega_{\varepsilon_n}\langle\phi_{\varepsilon_n},f_{\varepsilon_n}\rangle=\langle\phi_{\varepsilon_n},A^{\varepsilon_n^*}f_{\varepsilon_n}\rangle\to\omega'(\psi,f)_\Omega=(\psi,A^*f)_\Omega.$$

Since $\mathfrak{D}(\Omega)$ is dense in $L^2(\Omega)$ we conclude that:

$$\psi\in D(A), \quad A\psi=\omega'\psi \quad (\psi\neq 0).$$

Hence ω' is an eigenvalue of A. But, thanks to Proposition 4, $\omega'\geq\omega$. Since ω is the maximal real eigenvalue, it follows that

$$\omega'=\omega.$$

Since ψ satisfies $|\psi|_{L^2}=1$, $\psi\geq 0$, we also have $\phi=\psi$. Q.E.D.

5. The corrector of ω_ε. Before calculating the corrector of ω_ε (i.e., $\lim_{\varepsilon\to 0}(1/\varepsilon)(\omega_\varepsilon-\omega)$), we give a result which one can easily prove with the techniques used in Bensoussan, Lions and Papanicolaou [3, §3.5] (a similar calculation is performed in Sentis [14, §4.4.1]).

PROPOSITION 7. *Let f be in $C^\infty(\Omega)$ and let α be a positive real number ($\alpha\geq 0$). There exists a unique solution u_ε of the equation*

$$-\alpha u_\varepsilon+A^\varepsilon u_\varepsilon=f, \qquad u_\varepsilon\in D(A^\varepsilon).$$

Let u_0 be the solution of

$$-\alpha u_0+Au_0=f,\qquad u_0\in D(A).$$

Then we have

$$\left|u_\varepsilon(x,v)-u_0(x)-\varepsilon\left[\zeta_i^x(v)\frac{\partial u_0}{\partial x_i}+w(x)+\tilde{u}_1^\varepsilon(x,v)\right]\right|\leq\varepsilon^2C^t\quad\forall x,v,$$

where the function $\tilde{u}_1^\varepsilon$ (which is called a boundary layer) satisfies:

$$|\tilde{u}_1^\varepsilon(x,v)|\leq C\exp\left(-\frac{\gamma}{\varepsilon}d(x,\partial\Omega)\right),$$

where $d(\cdot,\cdot)$ denotes distance and C and γ are positive constant, w is the solution of

$$\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial w}{\partial x_j}\right)-\alpha w+\frac{\partial}{\partial x_i}\left(\beta_{ijk}\frac{\partial^2u_0}{\partial x_j\partial x_k}\right)+\frac{\partial}{\partial x_i}\left(\mu_{ij}\frac{\partial u_0}{\partial x_j}\right)=0,$$
$$w(x)=\left(p_x,\zeta_i^x|_{\Gamma_x^-}\right)_{\Gamma_x^-}\frac{\partial u_0}{\partial x_i}\quad\forall x\in\partial\Omega,$$

μ_{ij} and β_{ijk} being smooth coefficients satisfying

$$\beta_{ijk}(x)=(\pi,v_i\eta_{jk}^x),\qquad Q_x\eta_{jk}^x-v_j\zeta_k^x-a_{jk}(x)=0,$$
$$\mu_{ij}(x)=(\pi,-v_iq_j^x),\qquad Q_xq_j^x+\frac{\partial}{\partial x_i}(-v_i\zeta_j^x-a_{ij}(x))=0,$$

and p_x is a probability measure on Γ_x^- (which depends only on Q_x), for any x in $\partial\Omega$.

Let ϕ satisfy (19). Thanks to Proposition 6 there exists a solution z_ε of the equation

(20) $$A^\varepsilon z_\varepsilon=\omega\phi,\qquad z_\varepsilon\in D(A^\varepsilon).$$

We know that $z=\phi$ is the unique solution of

$$Az=\omega\phi,\qquad z\in D(A).$$

Let w be the solution of

$$\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial w}{\partial x_j}\right)+\frac{\partial}{\partial x_i}\left(\beta_{ijk}\frac{\partial^2\phi}{\partial x_j\partial x_k}\right)+\frac{\partial}{\partial x_i}\left(\mu_{ij}\frac{\partial\phi}{\partial x_j}\right)=0,$$
$$w(x)=\left(p_x,\zeta_i^x|_{\Gamma_x^-}\right)\frac{\partial\phi}{\partial x_i}\quad\forall x\in\partial\Omega.$$

Then, thanks to Proposition 7, we have

(21) $$z_\varepsilon-\phi-\varepsilon\left(\zeta_i\frac{\partial\phi}{\partial x_i}+w\right)=\varepsilon\rho_\varepsilon\quad\text{with }\|\rho_\varepsilon\|\leq\varepsilon C^t.$$

(Indeed $\|\tilde{u}_1^\varepsilon\|$ is less than ε times a constant.)

PROPOSITION 8. *When ε goes to 0, we have (if we let: $\partial/\partial n=a_{jk}n_j\partial/\partial x_k$)*

$$\frac{1}{\varepsilon}(\omega_\varepsilon-\omega)\to\omega_1=\left(\phi,\frac{\partial}{\partial x_i}\left(\beta_{ijk}\frac{\partial^2\phi}{\partial x_j\partial x_k}+\mu_{ij}\frac{\partial\phi}{\partial x_j}\right)\right)_\Omega+\int_{\partial\Omega}(p_x,\zeta_i^x)\frac{\partial\phi}{\partial x_i}\frac{\partial\phi}{\partial n}.$$

Proof. Let ψ_ε be an eigenfunction of $A^{\varepsilon *}$ satisfying

$$\psi_\varepsilon \in D(A^{\varepsilon *}), \quad A^{\varepsilon *}\psi_\varepsilon = \omega_\varepsilon \psi_\varepsilon, \quad \|\psi_\varepsilon\| = 1.$$

Using the definition of z_ε given in (20) we have

$$\omega\langle \phi, \psi_\varepsilon\rangle = \langle A^\varepsilon z_\varepsilon, \psi_\varepsilon\rangle = \langle z_\varepsilon, A^{\varepsilon *}\psi_\varepsilon\rangle = \omega_\varepsilon \langle z_\varepsilon, \psi_\varepsilon\rangle. \tag{22}$$

Due to (21), it follows that

$$\begin{aligned} \frac{1}{\varepsilon}\frac{\omega_\varepsilon - \omega}{\omega_\varepsilon}\langle \phi, \psi_\varepsilon\rangle &= \frac{1}{\varepsilon\omega_\varepsilon}[\omega_\varepsilon\langle \phi, \psi_\varepsilon\rangle - \omega_\varepsilon \langle z_\varepsilon, \psi_\varepsilon\rangle] \\ &= \left\langle \frac{\phi - z_\varepsilon}{\varepsilon}, \psi_\varepsilon \right\rangle = -\left\langle \zeta_i \frac{\partial \phi}{\partial x_i} + w + \rho_\varepsilon, \psi_\varepsilon \right\rangle. \end{aligned} \tag{23}$$

Let ε go to 0; we can show, in exactly the same way as for ϕ_ε, that

$$\psi_\varepsilon \to \pi\phi, \quad \text{in } L^2(\Omega \times V).$$

Hence, we have

$$\frac{1}{\omega}\lim_{\varepsilon}\frac{1}{\varepsilon}(\omega_\varepsilon - \omega) = -\left\langle \zeta_i \frac{\partial \phi}{\partial x_i}, \pi\phi \right\rangle - (w, \phi)_\Omega.$$

Since $(\zeta_i, \pi)_V = 0$ for any i, we have

$$\begin{aligned} \omega_1 = \lim_{\varepsilon \to 0} \frac{\omega_\varepsilon - \omega}{\varepsilon} &= -\omega(w, \phi)_\Omega = (w, A\phi)_\Omega \\ &= \left(-\frac{\partial}{\partial x_i}\left(a_{ji}\frac{\partial w}{\partial x_j}\right), \phi\right)_\Omega + \int_{\partial\Omega} w \frac{\partial \phi}{\partial n} \\ &= \left(\phi, \frac{\partial}{\partial x_i}\beta_{ijk}\frac{\partial^2 \phi}{\partial x_j \partial x_k} + \frac{\partial}{\partial x_i}\mu_{ij}\frac{\partial \phi}{\partial x_j}\right)_\Omega + \int_{\partial\Omega}(p_x, \xi_i^x)\frac{\partial \phi}{\partial x_i}\frac{\partial \phi}{\partial n}. \end{aligned}$$

From the definition of w, this yields the result. Q.E.D.

Remark 2. The calculation of the corrector of the eigenfunction ϕ_ε is very difficult. Even when the second order operator is homogeneous there are no good results (see Kesavan [9]), we have only the corrector for the corresponding z_ε.

Remark 3. There exists a constant C such that:

$$|\omega_\varepsilon - \omega - \varepsilon\omega_1| \le \varepsilon^2 C.$$

Indeed, if we put $\delta_\varepsilon = \varepsilon^{-1}(\omega_\varepsilon - \omega)$, by (23), we have

$$\begin{aligned} \delta_\varepsilon &= -(\omega + \varepsilon\delta_\varepsilon)\left[\left\langle \zeta_i \frac{\partial \phi}{\partial x_i}, \phi \right\rangle + (w, \phi)_\Omega + \langle \rho_\varepsilon, \psi_\varepsilon\rangle\right] \\ &= -\omega(w, \phi)_\Omega - \omega_\varepsilon\langle \rho_\varepsilon, \psi_\varepsilon\rangle - \varepsilon\delta_\varepsilon(w, \phi)_\Omega = \omega_1 - \omega_\varepsilon\langle \rho_\varepsilon, \psi_\varepsilon\rangle - \varepsilon\delta_\varepsilon(w, \phi)_\Omega. \end{aligned} \tag{24}$$

Since ω_ε, δ_ε and the function ψ_ε are bounded, and since $\|\rho_\varepsilon\| \le \varepsilon C^t$, we conclude that $\varepsilon^{-1}(\delta_\varepsilon - \omega_1)$ is bounded.

Remark 4. If we assume that σ_1 does not depend on x and that V satisfies the property of spherical symmetry. Then there exist $\rho\in C(R)$ and $\sigma_2\in C(R^+,R^+)$ such that

$$\sigma_2(a,b)=\sigma_2(b,a)\geq\sigma_l>0,\quad \rho\geq 1,\qquad \sigma_1(v,v')=\sigma_2(|v|,|v'|)\rho\left(\frac{v,v'}{|v||v'|}\right)\quad \forall v,v'\in V.$$

Then it may be easily seen that

$$\eta_{ijk}=\mu_{ij}=0\quad \forall i,j,k.$$

Remark 5. All the foregoing remains true if Ω is an open set of the torus $[0,, 1[^N$, with a smooth boundary $\partial\Omega$ and such that $\bar{\Omega}\neq[0,1[^N$. A nuclear reactor where there are many periodically spaced control rods may be modeled by a wide domain with a large number of holes which are periodically spaced with period 1, and in this framework the preceding calculation of ω_1 (with Ω an open set of the torus) may be useful.

Acknowledgments. I thank G. Papanicolaou and C. Bardos for useful discussions on this problem.

REFERENCES

[1] H. AMANN, *Nonlinear operators in ordered Banach spaces and some applications to nonlinear boundary value problems*, Nonlinear Operators, and the Calculus of Variations, Proc. Bruxelles, 1975, Lecture Notes in Mathematics 543, Springer, Berlin, 1976, pp. 1–55.

[2] C. BARDOS, *Cours de 3ème cycle*, Université Paris VI, 1981.

[3] A. BENSOUSSAN, J. L. LIONS AND G. C. PAPANICOLAOU, *Boundary layers and homogenization of transport processes*, J. Publ. RIMS, Kyoto University, 15 (1979), pp. 53–157.

[4] G. BLANKENSHIP AND G. C. PAPANICOLAOU, *Stability and control of stochastic systems with wide-band noise disturbance*, I, SIAM J. Appl. Math., 34 (1978), pp. 437–476.

[5] K. JORGENS, *An asymptotic expansion in the theory of neutron transport*, Comm. Pure and Appl. Math., 11 (1958), pp. 219–242.

[6] M. G. KREIN AND M. A. RUTMAN, *Linear operators leaving invariant a cone in a Banach space*, Transl. AMS, 26, 1950.

[7] R. Z. KHASMINSKII, *On diffusion process with a small parameter*, J. Akad. Nauk. SSSR Ser. Math., 27 (1963), pp. 1280–1300.

[8] T. KATO, *Perturbation Theory for Linear Operator*, Springer, Berlin, 1976.

[9] S. KESAVAN, *Homogenization of eigenvalue problems*, Appl. Math. Optim., 5 (1979), pp. 153–168.,

[10] E. W. LARSEN AND J. B. KELLER, *Asymptotic solution of neutron transport problems for small mean free paths*, J. Math. Phys., 15 (1974), pp. 75–81.

[11] G. C. PAPANICOLAOU, Private communication.

[12] M. PROTTER AND H. WEINBERGER, *On the spectrum of general second order operators*, Bull. AMS, 72 (1966), pp. 251–255.

[13] R. SENTIS, *Approximation and homogenization of a transport process*, SIAM J. Appl. Math., 39 (1980), pp. 134–141.

[14] ———, *Analyse asymptotique d'équation de transport avec structure périodique*, Thèse, Partie A. Université Paris IX, 1981.

[15] L. TARTAR, Private communication.

[16] M. WILLIAMS, *Homogenization of linear transport problems*, Ph.D. dissertation, New York Univ., 1976.

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

THE NORM OF CERTAIN CONVOLUTION TRANSFORMS ON L_p SPACES OF ENTIRE FUNCTIONS OF EXPONENTIAL TYPE*

B. F. LOGAN†

Abstract. Denote by $B_p(\Omega)$ the subspace of $L_p(-\infty,\infty)$ whose elements are restrictions to the real line of entire functions of order 1, type at most Ω. It is shown that the norm of certain convolution transforms on $B_p(\Omega)$ is independent of p, $1\leq p\leq\infty$. It follows that a number of classical inequalities sharp for functions in $B_\infty(\Omega)$ are also sharp for functions in $B_p(\Omega)$ with $1\leq p<\infty$.

1. Introduction.

Consider the convolution transform

$$Tf(x)=g(x)=\int_{-\infty}^{\infty} f(x-t)\,dG(t) \tag{1}$$

where G is a function of bounded variation on $(-\infty,\infty)$ and f belongs to the space $L_p=L_p(-\infty,\infty)$ for some p satisfying $1\leq p\leq\infty$, with norm

$$\|f\|_p=\left\{\int_{-\infty}^{\infty}|f(t)|^p dt\right\}^{1/p}<\infty, \qquad 1\leq p<\infty \tag{2}$$

$$=\operatorname*{essup}_{-\infty<t<\infty}|f(t)|<\infty, \qquad p=\infty.$$

Here essup is the smallest number M such that

$$|f(t)|\leq M$$

is satisfied for almost all t in $(-\infty,\infty)$. If $f(t)$ is continuous, then

$$\operatorname*{essup}_{-\infty<t<\infty}|f(t)|=\sup_t|f(t)|.$$

Minkowski's inequality, which asserts that

$$\left\|\sum a_k f_k\right\|_p\leq\sum|a_k|\cdot\|f_k\|_p \tag{3}$$

where a_k are scalars, $f_k\in L_p$, generalizes to

$$\left\|\int_{-\infty}^{\infty}dG(t)f(x;t)\right\|_p\leq\int_{-\infty}^{\infty}|dG(t)|\,\|f(x;t)\|_p \tag{4}$$

where the norm is understood to be taken over the variable x.

In particular, if $f(x;t)=f(x-t)$, then

$$\|f(x;t)\|_p=\|f(t)\|_p,$$

i.e., the L_p norm is translation invariant.

Then we have for $g(x)=Tf(x)$ in (1),

$$\|g\|_p\leq\|f\|_p\int_{-\infty}^{\infty}|dG(t)|. \tag{5}$$

*Received by the editors October 23, 1982. This paper was presented at AMS Meeting #808, Special Session on Operator Theory, held in Evanston, Illinois, November 11-12, 1983.

† Bell Laboratories, Murray Hill, New Jersey, 07974.

The norm of the transform on L_p, which depends on p and G, is defined by

$$M_p(G) = \sup_{\substack{f\in L_p\\ \|f\|_p=1}} \|g\|_p. \tag{6}$$

Here we are interested in the norm of the transform (1) on certain proper subspaces of L_p. The subspaces we consider are the spaces $B_p(\Omega)$ which consist of *band-limited functions* in L_p, i.e. the subspace of L_p whose elements are restrictions to the real line of entire functions of order 1, type at most Ω. (In the terminology of Boas [1] entire functions of order 1, type at most Ω are called entire functions of "exponential type Ω".) Now functions belonging to L_p for some p greater than 2 do not necessarily have Fourier transforms. However, functions in $B_p(\Omega)$ can be essentially described as those continuous functions of L_p whose "Fourier transforms vanish outside $[-\Omega,\Omega]$". In fact, Boas [1, Thm. 6.8.14] shows that for f in $B_\infty(\Omega)$ there exists a sequence F_n of functions, each of bounded variation, such that

$$f(x) = \lim_{n\to\infty} \int_{-\Omega}^{\Omega} e^{ixt}\,dF_n(t). \tag{7}$$

Also (see [1, Thms. 6.7.17 and 6.7.18])

$$B_p(\Omega) \subset B_\infty(\Omega), \qquad 1 \le p < \infty. \tag{8}$$

Since a bounded function in L_p, i.e., a function in $L_p \cap L_\infty$, also belongs to $L_{p'}$, whenever $p < p' < \infty$, we have

$$B_p(\Omega) \subset B_{p'}(\Omega) \subset B_\infty(\Omega). \tag{9}$$

One can argue from (7) (also cf. [2]) that all functions f in $B_\infty(\Omega)$ (and hence all functions in $B_p(\Omega)$ by (8)) are orthogonal to all functions in L_1 (and also all finite measures) whose Fourier transforms vanish over $(-\Omega,\Omega)$, and that

$$f(x) = \int_{-\infty}^{\infty} r_\Omega(x-t) f(t)\,dt, \qquad f \in B_p(\Omega) \tag{10}$$

where r_Ω is any function of L_1 (or a finite measure) whose Fourier transform is 1 over $[-\Omega,\Omega]$.

Consequently, if f in (1) belongs to $B_p(\Omega)$ we may replace G by G_Ω, where G_Ω is any function of bounded variation satisfying

$$\int_{-\infty}^{\infty} e^{-i\omega t}\{dG(t) - dG_\Omega(t)\} = 0, \qquad -\Omega \le \omega \le \Omega. \tag{11}$$

The norm of the transform (1) on $B_p(\Omega)$ is defined by

$$M_p(G,\Omega) = \sup_{\substack{f\in B_p(\Omega)\\ \|f\|_p=1}} \|Tf\|_p. \tag{12}$$

Using (5) and (11) we see $M_p(G,\Omega)$ is bounded by

$$M_p(G,\Omega) \le \inf \int_{-\infty}^{\infty} |dG_\Omega(t)|, \tag{13}$$

where the inf is over functions G_Ω satisfying (11).

Now the right-hand side of (13) is independent of p. Our purpose here is to point out that there are, in fact, a number of interesting transforms for which the norm is independent of p.

We prove the following result.

THEOREM A. *Suppose there exists a function G_Ω satisfying*

$$\int_{-\infty}^{\infty} e^{-i\omega t}\{dG(t)-dG_\Omega(t)\}=0, \qquad -\Omega\leq\omega\leq\Omega$$

and

$$\max_{-\Omega\leq\omega\leq\Omega}\left|\int_{-\infty}^{\infty} e^{-i\omega t}dG(t)\right|=\int_{-\infty}^{\infty}|dG_\Omega(t)|.$$

Then the norm of the transform (1) *on $B_p(\Omega)$ is given by*

$$M_p(G;\Omega)=\int_{-\infty}^{\infty}|dG_\Omega(t)|, \qquad 1\leq p\leq\infty.$$

In particular, we use this theorem to show that a number of "classical" inequalities known to be sharp for $B_\infty(\Omega)$ are also sharp for $B_p(\Omega)$, $1\leq p\leq\infty$. This is done below.

2. Applications of the theorem. One of the simpler applications is to transforms of the special form

$$g(x)=\int_{-\infty}^{\infty} K(x-t)f(t)\,dt, \tag{14}$$

where

$$K(t)=e^{i\lambda t}P(t), \qquad P(t)\geq0 \qquad (-\infty<t<\infty) \tag{14a}$$

and λ satisfies $-\Omega\leq\lambda\leq\Omega$. Then

$$\int_{-\infty}^{\infty}|K(t)|dt=\int_{-\infty}^{\infty}P(t)\,dt=\int_{-\infty}^{\infty}e^{-i\lambda t}K(t)\,dt;$$

i.e.,

$$\max_{-\Omega\leq\omega\leq\Omega}\left|\int_{-\infty}^{\infty}K(t)e^{-i\omega t}dt\right|=\int_{-\infty}^{\infty}|K(t)|dt.$$

An example of this type is the inequality relating the norm of the analytic continuation of f on a line parallel to the real axis to the norm of f on the real line.

Application 1. If $f\in B_p(\Omega)$ then

$$\|f(x+iy)\|_p\leq e^{\Omega|y|}\|f(y)\|_p. \tag{15}$$

This inequality is sharp in the sense that $e^{\Omega|y|}$ cannot be replaced with any smaller number, for $1\leq p\leq\infty$.

The inequality (15) is due to Plancherel and Polya (see [1, Thm. 6.7 : 1]) and generalized by Boas [1, Thm. 6.7.4].

Proof of Application 1. We view $f(x+iy)=T_yf(x)$, for fixed real y, as a convolution transform.

We first choose a reproducing kernel r_Ω in (10) which has an analytic continuation belonging to L_1 on lines parallel to the real axis; e.g.,

$$(16) \qquad r_\Omega(x)=\frac{\sin(\Omega+\varepsilon)x}{\pi x}\cdot\frac{\sin\varepsilon x}{\varepsilon x},$$

where ε is a fixed positive number. It is readily verified that the Fourier transform of r_Ω is 1 over $[-\Omega,\Omega]$, decreasing linearly to zero at $\pm(\Omega+2\varepsilon)$. Then the analytic continuation of f in $B_p(\Omega)$ is given by

$$(17) \qquad f(x+iy)=\int_{-\infty}^{\infty} r_\Omega(x+iy-t)f(t)\,dt=\int_{-\infty}^{\infty} r_\Omega(t+iy)f(x-t)\,dt.$$

Now

$$r_\Omega(t)=\frac{1}{2\pi}\int_{-A}^{A}\hat{r}_\Omega(\omega)e^{i\omega t}\,d\omega$$

and hence

$$r_\Omega(t+iy)=\frac{1}{2\pi}\int_{-A}^{A}\hat{r}_\Omega(\omega)e^{-\omega y}e^{i\omega t}\,d\omega.$$

Therefore

$$(18) \qquad \int_{-\infty}^{\infty} e^{-i\omega t}r_\Omega(t+iy)\,dt=e^{-\omega y}, \qquad -\Omega\leq\omega\leq\Omega.$$

Thus in (17) we may replace r_Ω by any function of L_1 (or any bounded measure) whose Fourier transform is $e^{-\omega y}$ over $[-\Omega,\Omega]$. An appropriate choice is

$$(19) \qquad k_\Omega(t;y)=\begin{cases} e^{-i\Omega t}\cdot\dfrac{ye^{\Omega y}}{\pi(t^2+y^2)}, & y>0,\\[2ex] e^{i\Omega t}\cdot\dfrac{|y|e^{-\Omega y}}{\pi(t^2+y^2)}, & y<0.\end{cases}$$

We have

$$\hat{k}_\Omega(\omega;y)=\int_{-\infty}^{\infty} e^{-\omega t}k_\Omega(t;y)\,dt=\begin{cases} e^{-|\Omega+\omega|y}, & y>0,\\ e^{|\Omega-\omega|y-\Omega y}, & y<0.\end{cases}$$

So regardless of the sign of y we have

$$(20) \qquad \hat{k}_\Omega(\omega;y)=e^{-\omega y}, \qquad -\Omega\leq y\leq\Omega.$$

Thus we obtain (15) by replacing $r_\Omega(t+iy)$ in (17) in $k_\Omega(t;y)$.

It then follows from

$$(21) \qquad \max_{-\Omega\leq\omega\leq\Omega}\left|\int_{-\infty}^{\infty} k_\Omega(t;y)e^{-i\omega t}\,dt\right|=\int_{-\infty}^{\infty}|k_\Omega(t;y)|\,dt=e^{\Omega|y|},$$

and Theorem A, that $e^{\Omega|y|}$ in the inequality (15) cannot be replaced by a smaller number, for any p in $[1,\infty]$. □

Another interesting series of applications is to transforms (or operators) that admit the representation

$$(22)\qquad Tf(x)=g(x)=\sum_{k=-\infty}^{\infty} a_k(\theta;\Omega)f\left(x+\theta+\frac{k\pi}{\Omega}\right),\qquad f\in B_\infty(\Omega),$$

where (for some choice of θ)

$$(22a)\qquad (-1)^k a_k(\theta;\Omega)\geq 0,\qquad \sum_{-\infty}^{\infty}|a_k(\theta;\Omega)|<\infty.$$

(See Boas [1, Chapter 11, "Operators and Their Extremal Properties"].) This transformation may be regarded as a convolution transform where $G(t)=G_\Omega(t)$ in (1) is a step function of bounded variation. In this case, owing to the alternating sign of a_k, and the "(π/Ω)-translates" of f, the norm of the transform (22) on $B_\infty(\Omega)$ is readily seen to be

$$(23)\qquad M_\infty(G;\Omega)=\sum_{-\infty}^{\infty}|a_k(\theta;\Omega)|,$$

with $\|g\|_\infty=M_\infty(G;\Omega)$ attained for

$$(23a)\qquad f(x)=A\left[\alpha e^{i\Omega x}+(1-\alpha)e^{-i\Omega x}\right],\qquad 0\leq\alpha\leq 1,$$

where $|A|=1$.

Here again we have

$$(24)\qquad \max_{-\Omega\leq\omega\leq\Omega}\left|\int_{-\infty}^{\infty}dG(t)e^{-i\omega t}dt\right|=\int_{-\infty}^{\infty}|dG_\Omega(t)|=\sum_{-\infty}^{\infty}|a_k(\theta;\Omega)|$$

where the max is attained for $\omega=\pm\Omega$. Hence, according to Theorem A, the norm of the transform (22) on $B_p(\Omega)$ is

$$(25)\qquad M_p(G;\Omega)=\sum_{-\infty}^{\infty}|a_k(\theta;\Omega)|,\qquad 1\leq p\leq\infty.$$

An important example of this kind is

$$(26)\qquad Tf(x)=g_\tau(x)=f(x+\tau)-f(x-\tau),\qquad f\in B_p(\Omega),$$

where $\tau>0$. Here we regard $g_\tau(x)$ as obtained by the convolution transform

$$(27)\qquad g_\tau(x)=\int_{-\infty}^{\infty}f(x-t)\,dG(t;\tau),$$

where

$$G(t;\tau)=\begin{cases}1, & -\tau\leq t\leq\tau,\\ 0 & \text{otherwise.}\end{cases}$$

Application 2. Let $f\in B_p(\Omega)$ and for real τ set $g_\tau(x)=f(x+\tau)-f(x-\tau)$. Then

$$(28)\qquad \|g_\tau\|_p\leq\begin{cases}2\sin|\Omega\tau|\,\|f\|_p & \text{if } -\pi<2\Omega\tau<\pi,\\ 2\|f\|_p & \text{if } |2\Omega\tau|>\pi.\end{cases}$$

These inequalities are sharp in the sense that $2\sin|\Omega\tau|$ and 2 cannot be replaced by any smaller numbers, for $1\leq p\leq\infty$. Equality can be attained only in the case $p=\infty$ and in the case $p=1$ only for $|2\Omega\tau|\geq\pi$.This is an inequality due to Bernstein (see [1, Thm. 11.4.1]) for the case $p=\infty$.

The limiting case obtained by dividing both sides of (28) by 2τ and letting $\tau\rightarrow 0$ is the well-known "Bernstein's inequality" for the derivative:

$$\|f'\|_\infty\leq\Omega\|f\|_\infty, \qquad f\in B_\infty(\Omega). \tag{29}$$

We obtain the following generalization.

Application 3. Let $f\in B_p(\Omega)$. Then

$$\|f'\|_p\leq\Omega\|f\|_p. \tag{30}$$

This inequality is sharp in the sense that Ω cannot be replaced by any smaller number, for $1\leq p\leq\infty$. Equality can occur only for $p=\infty$.

We derive Applications 2 and 3 below. We prove inequalities (28) and (30) directly, and use Theorem A only to show they are sharp. The discussion of the cases of equality in Applications 2 and 3 seems to be new.

Proof of Applications 2 *and* 3. We first derive (28) in the case $\tau>\pi/2\Omega$. We have

$$\int_{-\infty}^{\infty} dG(t;\ \tau)e^{-i\omega t}dt=2i\sin\omega\tau$$

and

$$\int_{-\infty}^{\infty}|dG(t;\ \tau)|dt=2.$$

So for $\tau\geq\pi/2\Omega$ we have

$$\max_{\Omega\leq\omega\leq\Omega}|2i\sin\omega\tau|=\int_{-\infty}^{\infty}|dG(t;\ \tau)|.$$

Hence, according to Theorem A, with $G(t)=G(t;\ \tau)$ defined in (27), the norm of the transform is

$$M_p(G;\ \Omega)=2, \quad \tau\geq\frac{\pi}{2\Omega}, \quad 1\leq p\leq\infty. \tag{31}$$

For $p=\infty$ and $\tau\geq 2\pi/\Omega$, the norm of the transform is clearly attained for

$$f(x)=f_0(x)=|A|\{\alpha e^{i\lambda x}+(1-\alpha)e^{-i\lambda x}\},$$

where $\lambda=\pi/2\tau<\Omega$, $0\leq\alpha\leq 1$ and $|A|=1$. For $0<\tau<\pi/2\Omega$ we get a representation of the form (22), with $\theta=\pi/2\Omega$, by considering

$$I_n=\int_{|z|=n\pi/\Omega}\frac{f(z+t)\,dz}{(z^2-\tau^2)\cos\Omega z},$$

where $f\in B_p(\Omega)$ and $0<\tau<\pi/2\Omega$.

From the growth estimate (15) we find that

$$\lim_{n\rightarrow\infty} I_n=0;$$

i.e., by Cauchy's integral theorem the sum of the residues is 0. That is,

$$\frac{f(t+\tau)}{2\tau\cos\Omega\tau}-\frac{f(t-\tau)}{2\tau\cos\Omega\tau}-\sum_{-\infty}^{\infty}\frac{(-1)^k f(t+t_k)}{\Omega(t_k^2-\tau^2)}=0$$

or

$$f(x+\tau)-f(x-\tau)=\sum_{-\infty}^{\infty} a_k f(x+t_k), \qquad f\in B_\infty(\Omega), \tag{32}$$

where $0<\tau<\pi/2\Omega$ and

$$t_k=\left(k+\frac{1}{2}\right)\frac{\pi}{\Omega}, \qquad a_k=(-1)^k\frac{2\tau\cos\Omega\tau}{\Omega\left(t_k^2-\tau^2\right)}.$$

(The formula (32) is actually valid for all τ, since a_k is an entire function of τ.) Here $\Sigma|a_k|=2\sin\Omega\tau$ is evaluated by taking $f(t)=f_0(t)=\sin\Omega x$, or more generally,

$$f_0(x)=\alpha\exp\{i\Omega(x-t_0)\}+(1-\alpha)\exp\{-i\Omega(x-t_0)\}$$

for $0\leq\alpha<1$.

Thus we obtain an inequality due to Bernstein (see [1, Thm. 11.4.1]),

$$|f(x+\tau)-f(x-\tau)|\leq\begin{cases} 2\sin\Omega\tau\cdot\|f\|_\infty, & 0<\tau<\dfrac{\pi}{2\Omega}, \\ 2\|f\|_\infty, & \tau\geq\dfrac{\pi}{2\Omega}, \end{cases} \tag{33}$$

valid for all f in $B_\infty(\Omega)$. Moreover, we conclude from the representation (32), and the obvious upper bound $2\|f\|_p$, that (28) holds. (Clearly the norm in question is an even function of τ, vanishing for $\tau=0$; so we may extend the inequality for $\tau>0$ to all real τ.)

We now prove (30). The representation of the derivative operator is obtained from (32) as

$$f'(x)=\frac{4\Omega}{\pi^2}\sum_{k-\infty}^{\infty}\frac{(-1)^k f(x+t_k)}{(2k+1)^2}, \qquad f\in B_\infty(\Omega), \tag{34}$$

where $t_k=(k+\frac{1}{2})\pi/\Omega$. From this (30) follows.

Note that we did not use Theorem A to establish the inequalities (28) and (30). However it follows from Theorem A and the representation of the transforms (operators) that the inequalities are sharp for $1\leq p\leq\infty$; i.e.,

$$\sup_{\substack{f\in B_p(\Omega)\\ \|f\|_p=1}}\|(fx+\tau)-f(x-\tau)\|_p=\begin{cases} 2\sin|\Omega\tau|\cdot\|f\|_p, & -\pi<2\Omega\tau<\pi, \\ 2\|f\|_p, & |2\Omega\tau|\geq\pi \end{cases} \tag{35}$$

and

$$\sup_{\substack{f\in B_p(\Omega)\\ \|f\|_p=1}}\|f'\|_p=\Omega\|f\|_p, \qquad f\in B_p(\Omega), \tag{36}$$

for $1\leq p\leq\infty$.

We now treat the cases of equality in (28) and (30). We have seen that the "sups" are attained here in case $p=\infty$. For other p they are not, except in (35) for $p=1$ and $|2\Omega\tau|>\pi$. The reason for this is that the L_p norm is *strictly convex* for $1<p<\infty$, meaning that equality can hold in

$$\left\|\sum f_k\right\|_p\leq\sum\|f_k\|_p, \qquad 1<p<\infty \tag{37}$$

only if, assuming (say) $\|f_0\|_p \neq 0$,

$$f_k(x)=\lambda_k f_0(x) \quad \text{(a.e.)} \tag{37a}$$

where $\lambda_k \geq 0$. (This goes back to Hölder's inequality.)

Now suppose $f(x)$ in $B_p(\Omega)$ with $\|f\|_p=1$ and set

$$g(x)=\sum a_k f(x-t_k), \tag{38}$$

where every $a_k \neq 0$.

Then in order for equality to hold in

$$\|g\|_p \leq \|f\|_p \sum_k |a_k| = \sum_k |a_k| \qquad (1<p<\infty) \tag{39}$$

we must have for all x, according to (37a),

$$a_k f(x-t_k)=\lambda_k a_j f(x-t_j), \qquad \lambda_k>0,$$

or

$$f(x+t_j-t_k)=\frac{\lambda_k a_j}{a_k} f(x), \qquad -\infty<x<\infty, \tag{40}$$

for all j and k appearing in the sum.

But

$$\|f(x+t_j-t_k)\|_p=\|f(x)\|_p=1.$$

So in (40) we must have

$$\left|\frac{\lambda_k a_j}{a_k}\right|=1,$$

or, since $\lambda_k>0$,

$$\lambda_k=\left|\frac{a_k}{a_j}\right|. \tag{41}$$

If the a_k are real, then $\lambda_k a_j/a_k$ takes the value either $+1$ or -1, and hence for all x we have

$$f\{x+2(t_j-t_k)\}=f(x).$$

Thus, if $t_j \neq t_k$, i.e., if there are at least two distinct t_k with associated real nonzero a_k in the sum (39), then f must be periodic in order to achieve equality in (39); so if $\|f\|_p<\infty$ we must have $\|f\|_p=0$, in contradiction to the assumption $\|f\|_p=1$. If the a_k are not real but nonzero and there are at least two distinct t_k then $|\lambda_k a_j/a_k|=1$ and hence

$$|f(x+t_j-t_k)|=|f(x)|.$$

So $|f(x)|$ is periodic and the same conclusion results. Thus the "sups" in (35) and (36) cannot be attained for $1<p<\infty$.

Now for $p=1$ and ∞ the L_p norm is not strictly convex. It is clear that the proportionality (37a) is not required for equality in (37) for $p=1$, ∞.

In case $p=1$ and

$$g(x)=\sum_k a_k f_k(x), \qquad f_k \in L_1(-\infty,\infty) \tag{42}$$

we have

$$\begin{aligned}\|g\|_1 &= \int_{-\infty}^{\infty} |g(x)|dx = \int_{-\infty}^{\infty} \{\operatorname{sgn} g(x)\} g(x)\,dx \\ &= \sum_k \int_{-\infty}^{\infty} \{\operatorname{sgn} g(x)\} f_k(x)\,dx \le \sum_k \|f_k\|.\end{aligned} \tag{43}$$

In order for $\|g\|_1 = \Sigma_k \|f_k\|_1$ with $\|f_k\|_1 \neq 0$, the f_k must have a common signum function,

$$\operatorname{sgn} f_k(x) = \operatorname{sgn} g(x), \quad \text{(a.e.)} \tag{43a}$$

where for a complex number z, we define

$$\operatorname{sgn} z = \begin{cases} \dfrac{|z|}{z}, & z \neq 0, \\ 0, & z = 0. \end{cases}$$

Now suppose f is in $B_1(\Omega)$, $\|f\|_1 = 1$, and set

$$g(x) = \sum_k a_k f(x-t_k), \tag{44}$$

with every $a_k \neq 0$. Then equality in

$$\|g\|_1 \le \|f\|_1 \cdot \sum_k |a_k| = \sum_k |a_k| \tag{45}$$

requires (since the zeros of f are isolated) equality everywhere in

$$\operatorname{sgn}\{a_k f(x-t_k)\} = \operatorname{sgn}\{a_j f(x-t_j)\}, \qquad -\infty < x < \infty \tag{46}$$

for all k and j appearing in the sum.

Since the signum of a product is the product of the signums, we may write (46) as (recall a_j, $a_k \neq 0$)

$$\operatorname{sgn} f(x+t_j-t_k) = \left\{ \frac{\operatorname{sgn} a_j}{\operatorname{sgn} a_k} \right\} \operatorname{sgn} f(x) \tag{47}$$

which must hold for all j and k appearing in the sum. Now we suppose that there are at least two distinct t_k, say t_1 and t_2, appearing in the sum (44) with

$$\operatorname{sgn} a_1 = -\operatorname{sgn} a_2 \quad (\neq 0) \tag{48}$$

and

$$0 < t_2 - t_1 = T \le \frac{\pi}{\Omega}. \tag{49}$$

Then we must have

$$\operatorname{sgn} f(x+T) = -\operatorname{sgn} f(x), \qquad -\infty < x < \infty, \tag{50}$$

in order to achieve equality in (45). We may suppose (by translating f and multiplying by a scalar of unit modulus) that $f(0)>0$. Then (50) gives

$$\operatorname{sgn} f(nT)=(-1)^n \operatorname{sgn} f(0)=(-1)^n \tag{51}$$

where $n=0, \pm 1, \cdots$, i.e.

$$(-1)^n f(nT)=|f(nT)|, \qquad f \text{ in } B_1(\Omega). \tag{52}$$

Now for continuous functions f in L_1 which are also of bounded variation, Poisson's summation formula asserts that

$$T\sum_{-\infty}^{\infty}(-1)^n f(nT)=\sum_{-\infty}^{\infty}\hat{f}\left((2n-1)\frac{\pi}{T}\right), \tag{53}$$

where $\hat{f}$ is the Fourier transform of f. If f belongs to $B_1(\Omega)$, f is continuous (analytic) and also of bounded variation. In fact

$$\operatorname{var} f=\int_{-\infty}^{\infty}|f'(t)|dt<\infty,$$

since by Bernstein's inequality for $B_1(\Omega)$,

$$\|f'\|_1 \leq \Omega \|f\|_1.$$

Also the Fourier transform of f is continuous and vanishes outside $[-\Omega,\Omega]$ and hence at the endpoints. Therefore we have from (52) and (53)

$$T\sum_{-\infty}^{\infty}(-1)^n f(nT)=T\sum_{-\infty}^{\infty}|f(nT)|=0, \tag{54}$$

from which we conclude that

$$f(nT)=0, \quad n=0,\pm 1,\cdots, \qquad f\in B_1(\Omega). \tag{55}$$

But if T satisfies $0<T\leq \pi/\Omega$, (55) implies $f\equiv 0$. (see [1, 9.4.2]). This contradicts the assumption that $\|f\|_1=1$. Hence if in (44) there are two distinct t_k, say t_1 and t_2, and associated coefficients a_1 and a_2 such that

$$\operatorname{sgn} a_1=-\operatorname{sgn} a_2 \quad \text{and} \quad 0<t_2-t_1\leq \frac{\pi}{\Omega},$$

then equality cannot be attained in (45). Thus the "sup" in (36) cannot be attained for $p=1$, nor in (35) for $|2\tau|\leq \pi/\Omega$. However if $|2\tau|>\pi/\Omega$, the "sup" in (35) may be attained for $p=1$ and

$$f(x)=e^{\pm i\lambda x}f_\varepsilon(x), \tag{56}$$

where $\lambda=\pi/|2\tau|<\Omega$, f_ε is in $B_1(\varepsilon)$, $\varepsilon=\Omega-\lambda$, and

$$f_\varepsilon(x)\geq 0, \quad -\infty<x<\infty, \qquad \int_{-\infty}^{\infty} f_\varepsilon(x)\,dx=1.$$

The type of f is at most $\lambda+\varepsilon=\Omega$, so f belongs to $B_1(\Omega)$, and

$$f(x+\tau)-f(x-\tau)=e^{\pm \lambda(x+\tau)}\{f_\varepsilon(x+\tau)+f_\varepsilon(x-\tau)\}.$$

Hence, since f_ε is nonnegative

$$\int_{-\infty}^{\infty} |f(x+\tau)-f(x+\tau)|dx = 2\int_{-\infty}^{\infty} f_\varepsilon(x)\,dx = 2\int_{-\infty}^{\infty} |f(x)|dx. \qquad \square$$

3. **Proof of the theorem.** The theorem is a consequence of the following lemma.

LEMMA. *Define for G of bounded variation the convolution transform on L_p*

$$Tf(x) = g(x) = \int_{-\infty}^{\infty} f(x-t)\,dG(t).$$

Then the norm of this transform on $B_p(\Omega)$ defined by

$$M_p(G;\Omega) = \sup_{\substack{f\in B_p(\Omega)\\ \|f\|_p=1}} \|g\|_p$$

satisfies

$$M_p(G;\Omega) \geq \max_{-\Omega\leq\omega\leq\Omega} \left| \int_{-\infty}^{\infty} e^{-i\omega t}\,dG(t) \right|$$

for $1 \leq p \leq \infty$.

Now according to the hypotheses of Theorem A, there exists G_Ω satisfying

$$\int_{-\infty}^{\infty} e^{-i\omega t}\,dG(t) = \int_{-\infty}^{\infty} e^{-i\omega t}\,dG_\Omega(t), \qquad -\Omega\leq\omega\leq\Omega \tag{57}$$

and

$$\max_{-\Omega\leq\omega\leq\Omega} \left| \int_{-\infty}^{\infty} e^{-i\omega t}\,dG(t) \right| = \int_{-\infty}^{\infty} |dG_\Omega(t)|. \tag{58}$$

From (13) we have

$$M_p(G;\Omega) \leq \int_{-\infty}^{\infty} |dG_\Omega(t)|,$$

which with the lemma establishes the theorem. □

Proof of the lemma. First the case $p=\infty$ is obvious. We suppose that

$$\max_{-\Omega\leq\omega\leq\Omega} \left| \int_{-\infty}^{\infty} e^{-i\omega t}\,dG(t) \right| = \left| \int_{-\infty}^{\infty} e^{-i\lambda_0 t}\,dG(t) \right|, \tag{59}$$

where λ_0 satisfies $-\Omega\leq\lambda_0\leq\Omega$. Then we take $f=e^{-i\lambda_0 t}$ to obtain the lower bound for $M_\infty(G,\Omega)$. Now we will establish the lemma by exhibiting a function f in $B_p(\Omega)$ of norm 1, given an arbitrary p in $[1,\infty)$, such that the corresponding norm of $g=Tf$ is arbitrarily close to the maximum in (59). Now a slight difficulty arises here in case $\lambda_0 = \pm\Omega$. However, since

$$\int_{-\infty}^{\infty} e^{-i\omega t}\,dG(t)$$

is a continuous function of ω for G of bounded variation (by a standard argument) we may find λ in the open interval $(-\Omega,\Omega)$ such that given any positive δ,

$$\left| \int_{-\infty}^{\infty} (e^{-i\lambda_0 t} - e^{-i\lambda t})\,dG(t) \right| < \delta. \tag{60}$$

So we assume that $\lambda \pm \varepsilon$, for some positive ε, is in the closed interval $[-\Omega, \Omega]$. Then we let

$$f(x) = e^{i\lambda x} f_\varepsilon(x) \tag{61}$$

where f_ε belongs to $B_p(\varepsilon)$ and $\|f_\varepsilon\|_p = 1$. Consequently f belongs to $B_p(\Omega)$ with $\|f\|_p = 1$. (The type of $f(z)$ is at most $|\lambda \pm \varepsilon| \leq \Omega$.) We may take, for example,

$$f_\varepsilon(t) = (\varepsilon)^{1/p} \phi(\varepsilon t)$$

where ϕ is a fixed function of norm 1 in $B_p(1)$.

Now we wish to show for such f that, by choosing ε sufficiently small and positive, the function

$$g(x; \lambda, \varepsilon) = \int_{-\infty}^{\infty} f(x-t)\, dG(t) \tag{62}$$

can be made arbitrarily close in the L_p-norm to

$$g^*(x; \lambda) = e^{i\lambda x} f_\varepsilon(x) \int_{-\infty}^{\infty} e^{-i\lambda t} dG(t) = f(x) \int_{-\infty}^{\infty} e^{-i\lambda t} dG(t). \tag{63}$$

(This is reasonable, since the "Fourier transform" of f is supported on a small interval centered on λ, and if the Fourier–Stieltjes transform of G were constant (say $=1$) over this interval, then the convolution with f would reproduce f exactly.)

We have

$$h(x; \lambda, \varepsilon) = g^*(x; \lambda) - g(x; \lambda, \varepsilon) = \int_{-\infty}^{\infty} e^{i\lambda(x-t)} \{ f_\varepsilon(x) - f_\varepsilon(x-t) \}\, dG(t). \tag{64}$$

Then using (4) we have

$$\|h\|_p = \|g^* - g\|_p \leq \int_{-\infty}^{\infty} |dG(t)| \cdot \|f_\varepsilon(x) - f_\varepsilon(x-t)\|_p \tag{65}$$

where on the right the norm is understood to be taken over the variable x.

Now using the inequality (28), which we obtained independently of Theorem A, we have

$$\|h\|_p \leq 2 \int_{-\pi/\varepsilon}^{\pi/\varepsilon} \sin \frac{\varepsilon |t|}{2} |dG(t)| + 2 \int_{|t| \geq \pi/\varepsilon} |dG(t)|. \tag{66}$$

This will be small for sufficiently small ε, for then

$$2 \int_{-\pi/\varepsilon}^{\pi/\varepsilon} \sin \frac{\varepsilon |t|}{2} |dG(t)| \leq 2 \sin \int_{|t| \leq 2/\sqrt{\varepsilon}} |dG(t)| + 2 \int_{(2/\sqrt{\varepsilon}) < |t| < \pi/\varepsilon} |dG(t)|,$$

and hence for sufficiently small ε

$$\|h\|_p \leq 2 \sin\sqrt{\varepsilon} \int_{-\infty}^{\infty} |dG(t)| + 2 \int_{|t| > 2/\sqrt{\varepsilon}} |dG(t)|. \tag{67}$$

The second integral (the tails) necessarily tends to zero as $\varepsilon \to 0$, since the first integral converges.

So for any positive δ we have

$$\|h\|_p < \delta \tag{68}$$

for ε sufficiently small but positive. Then, since

$$g^*(x;\lambda)=g(x;\lambda,\varepsilon)+h(x;\lambda,\varepsilon)$$

we have

$$\|g^*\|_p\le\|g\|_p+\|h\|_p,$$

i.e.

$$\|g\|_p\ge\|g^*\|_p-\|h\|_p\ge\|g^*\|_p-\delta. \tag{69}$$

Now

$$\|g^*\|_p=\left|\int_{-\infty}^{\infty}e^{-i\lambda t}dG(t)\right|$$

and $|\lambda-\lambda_0|\le\varepsilon$, so that for ε sufficiently small we will have

$$\|g^*\|_p\ge\left|\int_{-\infty}^{\infty}e^{-i\lambda_0 t}dG(t)\right|-\delta, \tag{70}$$

$$\|g\|_p\ge\left|\int_{-\infty}^{\infty}e^{-i\lambda_0 t}dG(t)\right|-2\delta, \tag{71}$$

which establishes the lemma. □

Acknowledgment. The author would like to thank J. C. Lagarias for help in the preparation of this manuscript.

REFERENCES

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

[2] B. F. LOGAN, *Theory of analytic modulation systems*, Bell System Tech. J., 57 (1978), pp. 491–576.

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

ON SOME GAMMA FUNCTION INEQUALITIES*

JACQUES DUTKA†

Abstract. D. K. Kazarinoff [Edinburgh Math. Notes, 40 (1956), pp. 19–21] and G. N. Watson [Proc. Edinburgh Math. Soc., 2, 11 (1959), pp. 7–9] gave extensions and proofs of gamma function inequalities closely associated with Wallis' product. One of Watson's proofs generated considerable interest, particularly among statisticians. Here some of the extensive previous background is outlined and further results are developed.

1. Introduction. One form of Wallis' product formula is

$$\pi=\lim_{n\to\infty}\left[\frac{(2n)!!}{(2n-1)!!}\right]^2\cdot\frac{2}{2n+1}. \tag{1.1}$$

A frequently employed method of obtaining it is to integrate over $[0,\pi/2]$ the inequalities $\sin^{2n+1}t<\sin^{2n}t<\sin^{2n-1}t$ where $0<t<\pi/2$ and n is a positive integer. Then one gets

$$\left[\frac{(2n)!!}{(2n-1)!!}\right]^2\cdot\frac{2}{2n+1}<\pi<\left[\frac{(2n)!!}{(2n-1)!!}\right]^2\cdot\frac{2}{2n}. \tag{1.2}$$

Since the ratio of the bounds on π tends to unity as $n\to\infty$, (1.1) follows. The substance of the foregoing is generally given in textbooks. Much less frequently treated, however, is a closely related inequality, (1.4) below. Since from (1.2) $n<[(2n)!!/(2n-1)!!]^2/\pi<(n+\frac{1}{2})$, it is convenient to define $\theta(n)$ for n a positive integer by

$$n+\theta(n)=\left[\frac{(2n)!!}{(2n-1)!!}\right]^2\frac{1}{\pi}=\left[\frac{2^{2n}}{\binom{2n}{n}}\right]^2\frac{1}{\pi}, \tag{1.3'}$$

and more generally by

$$x+\theta(x)=\begin{cases}[\Gamma(x+1)/\Gamma(x+\frac{1}{2})]^2 & \text{for } x>-\frac{1}{2},\\ 0 & \text{for } x=-\frac{1}{2}.\end{cases} \tag{1.3}$$

From (1.3') and the foregoing, it follows that

$$0<\theta(n)<\tfrac{1}{2}\qquad(n=1,2,3,\cdots). \tag{1.4}$$

In 1956, D. K. Kazarinoff [1] obtained an improved inequality

$$\tfrac{1}{4}<\theta(n)<\tfrac{1}{2}\qquad(n=1,2,3,\cdots). \tag{1.5}$$

In 1959, G. N. Watson [2], to whom the inequality represented by (1.4) was new, gave two additional proofs of (1.5) as well as additional results. The first proof involved the application of Gauss' formula for a hypergeometric series of the form $F(a,b;c;1)$ to

* Received by the editors April 13, 1982, and in revised form June 21, 1983.

† Audits & Surveys, Inc., One Park Avenue, New York, New York 10016.

(1.3) so that

$$\theta(x)=-x+xF(-\tfrac{1}{2},-\tfrac{1}{2};x;1) \tag{1.6}$$

$$=\left\{\frac{(-\frac{1}{2})^2}{1!}+\frac{(-\frac{1}{2}\cdot\frac{1}{2})^2}{2!\,(x+1)}+\frac{(-\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{3}{2})^2}{3!\,(x+1)(x+2)}+\cdots\right\},$$

where the series on the right certainly converges for $x>-\frac{1}{2}$. The terms of the series are positive, and, after the initial constant term, each is a decreasing function of x. Hence $\theta(x)$ is a monotonic decreasing function of x which $\to\frac{1}{4}$ as $x\to\infty$. In particular, (1.5) follows from this and (1.3).

Watson's proof attracted considerable interest because of its simplicity and its applicability to related problems. Statisticians, in particular, have made use of this method in connection with the derivation of inequalities for ratios of gamma functions.

The purpose of this article is to call attention to some of the previous background associated with the foregoing and to develop further results.

2. Some consequence of results of Wallis, Stirling and Binet. From a consideration of (1.2) and (1.3′), it is convenient to regard $1/[n+\theta(n)]$ as a convergence factor in the Wallis infinite product representation of π. In 1656, when John Wallis obtained the equivalent of (1.2) as the culminating result of his greatest work, [3, Prop. 191], he gave, without proof, essentially better bounds than (1.2). From this improved bounds for $\theta(n)$ in (1.4) and better bounds in (1.5) follow.

On applying the Cauchy–Schwarz inequality to

$$\int_0^{\pi/2}(\sin t)^{2x-1}dt=\tfrac{1}{2}B(x,\tfrac{1}{2}),\qquad x>0, \tag{2.1}$$

one gets

$$[B(x,\tfrac{1}{2})]^2<B(x-\tfrac{1}{2},\tfrac{1}{2})\cdot B(x+\tfrac{1}{2},\tfrac{1}{2}),$$

whence, on setting $x=n$ and $x=n+\frac{1}{2}$ where n is a positive integer, one obtains Wallis' result

$$\left[\frac{(2n-1)!!}{(2n-2)!!}\right]^2\frac{2}{2n}\sqrt{\frac{2n+1}{2n}}<\frac{4}{\pi}<\left[\frac{(2n-1)!!}{(2n-2)!!}\right]^2\frac{2}{2n}\sqrt{\frac{2n}{2n-1}}\,. \tag{2.2}$$

(Actually, Wallis set $n=7$, but indicated that the method of procedure could be continued indefinitely.) It is readily shown, on applying (1.3′) to (2.2) that

$$n\cdot\left[\left(1+\frac{1}{2n}\right)^{1/2}-1\right]<\theta(n)<n\cdot\left[\left(1-\frac{1}{2n}\right)^{-1/2}-1\right]\qquad(n=1,2,3,\cdots). \tag{2.3}$$

The lower (upper) bound is a positive increasing (decreasing) function of n. As $n\to\infty$, the bounds, and thus $\theta(n)$ tend to $\frac{1}{4}$. Thus, as n increases, (2.3) yields pairs of narrowing bounds, uniformly better than (1.4), and providing better upper bounds than (1.5).

Series for $[2^{2n}/\binom{2n}{n}^2]$, (which by (1.3′) equals $\pi\cdot[n+\theta(n)]$) and for its reciprocal were obtained by James Stirling [4, pp. 119–124] in his book of 1730, and corrected (in part) by J. Binet [5, pp. 319–320]. In modern notation, the series are

$$\left[2^{2n}\Big/\binom{2n}{n}\right]^2=\pi n\cdot F(\tfrac{1}{2},\tfrac{1}{2};n+1;1)=\pi(n+\tfrac{1}{2})F(-\tfrac{1}{2},\tfrac{1}{2};n+1:1), \tag{2.4}$$

$$(2.5)\quad \left[\binom{2n}{n}\Big/2^{2n}\right]^2=\frac{2}{\pi(2n+1)}\cdot F(\tfrac{1}{2},\tfrac{1}{2};n+\tfrac{3}{2};1)=\frac{1}{\pi n}\cdot F(-\tfrac{1}{2},\tfrac{1}{2};n+\tfrac{1}{2};1),$$

which yield lower and upper bounds for $n+\theta(n)$ and its reciprocal. On comparing with (1.6) and applying Gauss' formula, one gets

$$(2.6)\quad x\cdot F(-\tfrac{1}{2},-\tfrac{1}{2};x;1)=x\cdot F(\tfrac{1}{2},\tfrac{1}{2};x+1;1)=(x+\tfrac{1}{2})\cdot F(-\tfrac{1}{2},\tfrac{1}{2};x+1;1).$$

Alternatively, if the values of $[2^{2n}/\binom{2n}{n}]^2$ for even moderate values of n are known, close bounds for π can be obtained. Binet [5, pp. 161–162] also obtained the series expansions

$$(2.7)\quad \frac{B(t+r,t-r)}{B(t,t)}=F(r,r;t+r;1)$$

$$=F(-r,-r;t-r;1),\qquad r\text{ real},\quad t>0,\quad t>|r|.$$

In particular, let $r=b$, $t=c-b$. Then, since all the terms in the series in (7) are nonnegative, for an arbitrary positive integer m, one gets firstly

$$(2.8)\quad \frac{\Gamma(c)\Gamma(c-2b)}{\Gamma^2(c-b)}\geq\sum_{k=0}^{m-1}\frac{[(b)_k]^2}{k!(c)_k},\qquad b\text{ real},\quad c>0,\quad c-2b>0,$$

where equality holds if and only if $b=0,-1,\cdots,-(\mathrm{m}-1)$. This result was obtained by H. Ruben [6] in 1967 as an extension of an earlier result of J. Gurland for the case $m=2$. Secondly, one gets

$$(2.9)\quad \frac{\Gamma(c)\Gamma(c-2b)}{\Gamma^2(c-b)}\geq\sum_{k=0}^{m-1}\frac{[(-b)_k]^2}{k!(c-2b)_k},\qquad b\text{ real},\quad c>0,\quad c-2b>0,$$

where equality holds if and only if $b=0,1,2,\cdots,(m-1)$. (For example, for $m=2$, (2.9) is stronger (weaker) than (2.8) if b is positive (negative).)

Inequalities related to Gurland's have been investigated by a number of statisticians. See for example, A. W. Kemp [7] and the bibliography given there.

3. Bounds for $n+\theta(n)$ and some applications. From (2.6) and their equivalent representations of $x+\theta(x)$ for $x>0$, numerous representations in terms of integrals or series can be obtained. For example, it is readily verified that

$$(3.1)\quad x+\theta(x)=x+\tfrac{1}{4}\int_0^1 F(\tfrac{1}{2},\tfrac{1}{2};x+1;t)\,dt=x+\tfrac{1}{4}\cdot{}_3F_2\left[\begin{matrix}\tfrac{1}{2},\ \tfrac{1}{2},\ 1\ ;\\ 2,\ x+1\end{matrix}\right],$$

where ${}_3F_2[\cdots]$ denotes the generalized hypergeometric function with unit argument, or

$$(3.2)\quad x+\theta(x)=x+\frac{x}{4}\int_0^1(1-t)^{x-1}\cdot F(\tfrac{1}{2},\tfrac{1}{2};2;t)\,dt$$

$$=\frac{4x^2}{\pi}\cdot\int_0^1 k(1-k^2)^{x-1}K(k)\,dk,$$

where $K(k)$ is the complete elliptic integral of the first kind. But, in many applications, it is important to obtain close bounds for $x+\theta(x)$, particularly when x is a (large) positive integer n. From (1.1),

$$\pi=\left[\frac{(2n)!!}{(2n-1)!!}\right]^2\frac{2}{2n+1}\cdot\frac{2n+2}{2n+1}\cdot\frac{2n+2}{2n+3}\cdot\frac{2n+4}{2n+3}\cdots,$$

and, by a product relation for the quotient of beta functions stemming from Euler,

$$\pi=\left[\frac{(2n)!!}{(2n-1)!!}\right]^2\frac{1}{n}\frac{B\left(n+\frac{1}{2},\frac{1}{2}\right)}{B\left(n,\frac{1}{2}\right)}. \tag{3.3}$$

(See, e.g., Nielsen [8, p. 132].) But Euler also obtained some continued fraction expansions for quotients of beta functions which will prove useful subsequently.

Euler [9, p. 380] obtained the continued fraction for $x>0$

$$\frac{B\left(x,\frac{1}{2}\right)}{B\left(x+\frac{1}{2},\frac{1}{2}\right)}=1+\frac{2}{8x-1+}\,\frac{1\cdot 3}{8x}+\frac{3\cdot 5}{8x}+\frac{5\cdot 7}{8x}+\cdots, \tag{3.4}$$

where the initial convergents on the right are

$$\frac{1}{1},\frac{8x+1}{8x-1},\frac{64x^2+8x+3}{64x-8x+3},\frac{512x^3+64x^2+144x+15}{512x^3-64x^2+144x-15},\cdots. \tag{3.5}$$

(The partial numerators in the continued fraction expansion in (3.4) are quadratic functions of the index, while the partial denominators, after the initial two, are the same. It is known that continued fractions of this type converge slowly. But H. Rutishauser [10] has shown how the convergence of such fractions can be accelerated.)

Let $x=n$ and $x=n+\frac{1}{2}$ in (3.4) where n is a positive integer. Then from (3.3) and (1.3′), one gets

$$n<n+\frac{4n+3}{2(8n+5)}<\cdots<n+\theta(n)<\cdots<n+\frac{2n}{8n-1}<n+\frac{1}{2}. \tag{3.6}$$

Another continued fraction expansion for the quotient of beta functions obtained by Euler [11, pp. 301–303, 323] follows. For $x>0$

$$(x+1)\frac{B\left(\frac{x+3}{4},\frac{1}{2}\right)}{B\left(\frac{x+1}{4},\frac{1}{2}\right)}=x+\frac{1^2}{2x+}\,\frac{3^2}{2x+}\,\frac{5^2}{2x+}\cdots, \tag{3.7}$$

where the initial convergents on the right are

$$\frac{x}{1},\frac{2x^2+1}{2x},\frac{4x^3+11x}{4x^2+9},\frac{8x^4+72x^2+25}{8x^3+68x},\cdots. \tag{3.8}$$

(Compare the remarks following (3.5).) Let $x=4n+1$ and $x=4n+3$ in (7) where n is a nonnegative integer. Then one gets

(3.9)

$$n+\frac{1}{4}<n+\frac{8n^2+13n+6}{32n^2+48n+19}<\cdots<n+\theta(n)<\cdots<n+\frac{8n+3}{8(4n+1)}<n+\frac{n+1}{4n+3}\cdots.$$

Evidently the inequalties (1.4) and (1.5) are included in (3.6) and (3.9), but numerous additional results can be obtained in a unified manner. For example, from the bounds in (9), Gurland's inequalities [12] follow

$$\frac{4n+3}{(2n+1)^2}\left[\frac{(2n)!!}{(2n-1)!!}\right]^2<\pi<\frac{4}{4n+1}\left[\frac{2n!!}{(2n-1)!!}\right]^2, \tag{3.10}$$

as well as improvements on pairs of bounds for π obtained by Ruben [6, e.g. (17)–(19), etc.].

Moreover a result of Chu [13] may be obtained:

If $c\geq -1$ *is independent of* n, *then*

$$\frac{1}{\pi}\frac{n+c}{n+\theta}\begin{cases} <\dfrac{1}{\sqrt{\pi}} & \mathit{if}\; c\leq\dfrac{1}{4}, \\[2ex] >\dfrac{1}{\sqrt{\pi}} & \mathit{if}\; c\geq\dfrac{n+1}{4n+1} \end{cases}.$$

4. Some additional results. In 1954, M. E. Wise [15] developed some interesting expansions for the ratio of two factorials and for the logarithm of this ratio. In particular, he obtained the equivalent of

$$\left[\frac{2^{2n}}{\binom{2n}{n}}\right]^2=\pi\left(n+\frac{1}{4}\right)\cdot\left[w\left(n+\frac{1}{4}\right)\right]^2 \qquad (n=1,2,3,\cdots), \tag{4.1}$$

where

$$w(t)=1+\frac{1}{2^6t^2}-\frac{19}{2^{13}t^4}+\frac{631}{2^{19}t^6}-\frac{11\cdot 13\cdot 1219}{2^{27}t^8}+\frac{13\cdot 17\cdot 1093}{2^{29}t^{10}}-\cdots. \tag{4.2}$$

The series converges rapidly for even moderate values of n but the coefficients of the powers of t in (2) become increasingly complicated.

An expression for $x+\theta(x)$, essentially as the quotient of two hypergeometric functions will now be obtained. In particular, this will yield an expression for $w[(n+\frac{1}{4})]^2$ in (4.1). By Erdélyi et al. [16, Vol. I, p. 104 (51)]

$$F(a,1-a;b+1;\tfrac{1}{2})=\Gamma\begin{bmatrix}\dfrac{b}{2} & \dfrac{b+1}{2}\\[2ex] \dfrac{b+a}{2} & \dfrac{b-a+1}{2}\end{bmatrix}, \qquad b\neq 0,-1,-2,\cdots, \tag{4.3}$$

so that for positive values of the arguments of the gamma functions,

$$\frac{F(a,1-a;b;\frac{1}{2})}{F(a,1-a;b+1;\frac{1}{2})}=\frac{2}{b}\frac{\Gamma((b+a)/2+\frac{1}{2})\cdot\Gamma((b-a)/2+1)}{\Gamma((b+a)/2)\cdot\Gamma((b-a)/2+\frac{1}{2})}. \tag{4.4}$$

Let $a=\frac{1}{2}$ and $b=(4x+1)/2$. Then

$$x+\theta(x)=(x+\tfrac{1}{4})\frac{F(\frac{1}{2},\frac{1}{2};(4x+1)/2;\frac{1}{2})}{(\frac{1}{2},\frac{1}{2};(4x+3)/2;\frac{1}{2})}=\left[\frac{\Gamma(x+1)}{\Gamma(x+\frac{1}{2})}\right]^2; \tag{4.5}$$

cf. (3.7). In particular, if x is a positive integer, from (4.1)

$$[w(n+\tfrac{1}{4})]^2 = \frac{F(\frac{1}{2},\frac{1}{2};(4n+1)/2;\frac{1}{2})}{F(\frac{1}{2},\frac{1}{2};(4n+3)/2;\frac{1}{2})}, \tag{4.6}$$

so that $[w(n+\frac{1}{4})]^2$ may be expressed in terms of the quotient of two factorial series, which converge rapidly even for moderate n.

Acknowledgment. The writer wishes to express his appreciation to Dr. S. Wrigge of Stockholm for helpful comments.

REFERENCES

[1] D. K. KAZARINOFF, *On Wallis' formula*, The Edinburgh Mathematical Notes, Edinburgh Mathematical Society, No. 40, Dec. 1956, pp. 19–21.

[2] G. N. WATSON, *A note on gamma functions*, Proc. Edinburgh Mathematical Society, Ser. 2, Vol. 11 (1959), Mathematical Notes, pp. 7–9.

[3] JOHN WALLIS, *Arithmetica Infinitorum*, Oxford, 1656.

[4] JAMES STIRLING, *Methodus Differentialis*..., London, 1730.

[5] J. BINET, *Mémoire sur les intégrales euleriennes*..., J. de l'École Polytechnique, Cahier 27 (1839), pp. 123–343.

[6] H. RUBEN, *Variance bounds and orthogonal expansions in Hilbert space with an application to inequalities for gamma functions and* π. J. Reine und Angew. Math., 225 (1967), pp. 147–153.

[7] A. W. KEMP, *On gamma function inequalities*, Skand. Akt., 1973, pp. 65–69.

[8] N. NIELSEN, *Handbuch der Gammafunktion*, 1906, reprinted by Chelsea, New York, 1965.

[9] L. EULER, *De seriebus in quibus producta ex binis terminis contiguis*..., Opera Omnia, Ser. 1, Vol. 15, pp. 338–382.

[10] H. RUTISHAUSER, *Beschleunigung der Konvergenz einer gewissen Klasse von Kettenbruchen*, Z. Angew. Math. Mech., 38 (1958), pp. 187–190.

[11] L. EULER, *De fractionibus continuis observationes*, Opera Omnia, Ser. 1, Vol. 14, pp. 291–349.

[12] J. GURLAND, *On Wallis' formula*, Amer. Math. Monthly, 63 (1956), pp. 643–645.

[13] J. T. CHU, *A modified Wallis product and some applications*, Amer. Math. Monthly, 69 (1962), pp. 402–405.

[14] M. E. WISE, *The ratio of two factorials and some fundamental probabilities*, Proc. Sect. of Sciences, Akad. van Wetenschappen. Amsterdam, Ser. A, 57 (1954), pp. 513–521.

[15] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, F. TRICOMI, ET AL, *Higher Transcendental Functions*, I, McGraw-Hill, New York, 1953.

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

014

PROJECTION FORMULAS, A REPRODUCING KERNEL AND A GENERATING FUNCTION FOR q-WILSON POLYNOMIALS*

B. NASSRALLAH AND MIZAN RAHMAN†

Abstract. A projection formula for the q-Wilson polynomials $p_n(x;\ a,b,c,d)$ is obtained which is then used to construct a reproducing kernel. Using Askey and Wilson's q-analogue of the beta integral an integral representation is obtained for a very well-poised ${}_8\phi_7$ as a q-analogue of Euler's integral formula for a ${}_2F_1$. As an application of these results a generating function is obtained for the continuous q-Jacobi polynomials introduced by Askey and Wilson.

1. Introduction. The importance of projection formulas of the types

$$q_n(x)=\int p_n(y)\,d\nu_x(y),\qquad d\nu_x(y)\geq 0 \tag{1.1}$$

and

$$q_n(x)=\sum_k \alpha_{n,k}p_k(x),\qquad \alpha_{n,k}\geq 0 \tag{1.2}$$

in the theory of orthogonal polynomials was pointed out by many authors during the 1970's (see, for example, [1]–[5], [12], [13] and [18]). Explicit formulas for $d\nu_x(y)$ have been known for some time when $q_n(x)$ and $p_n(y)$ are both Jacobi polynomials with different sets of related parameters (see [1], [2] and [18]). Formulas for $\alpha_{n,k}$ have been found by Askey and Gasper in [5], again for Jacobi polynomials, and relationships between the parameters have been obtained for which these coefficients are nonnegative.

In view of the discovery of various q-analogues of Jacobi polynomials and, more recently, of a very general set of basic orthogonal polynomials called q-Wilson polynomials [9], interest in finding explicit formulas for $d\nu_x(y)$ and $\alpha_{n,k}$ has naturally shifted to the q-world. In this paper we shall be concerned with projection formulas of the type (1.1) and postpone treatment of (1.2) to a later report.

Askey and Wilson [9] have evaluated the following integral as a q-analogue of the familiar beta integral:

$$\begin{aligned}\int_{-1}^{1} dx\,w(x)&=\frac{2\pi(abcd)}{(ab)_\infty(ac)_\infty(ad)_\infty(bc)_\infty(bd)_\infty(cd)_\infty(q)_\infty}\\&=\kappa,\quad\text{say,}\end{aligned} \tag{1.3}$$

where

$$w(x)\equiv w(x;\ a,b,c,d)=(1-x^2)^{-1/2}\frac{h(x,1)h(x,-1)h\left(x,\sqrt{q}\right)h\left(x,-\sqrt{q}\right)}{h(x,a)h(x,b)h(x,c)h(x,d)} \tag{1.4}$$

with

$$h(x,a)=\prod_{k=0}^{\infty}\left(1-2axq^k+a^2q^{2k}\right), \tag{1.5}$$

*Received by the editors July 8, 1982, and in revised form June 9, 1983. This work was supported by the Natural Sciences and Engineering Research Council of Canada under grant A6197.

† Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

provided

$$\max(|a|, |b|, |c|, |d|, |q|)<1. \tag{1.6}$$

The basic shifted factorial with base q is normally denoted by $(a;\, q)_n$. However, as we shall be using the same base q throughout the paper we shall adopt the abbreviation $(a)_n$ for the sake of printing economy. Thus

$$(a)_n\equiv(a;\, q)_n=\begin{cases}1, & n=0,\\ (1-a)(1-aq)\cdots(1-aq^{n-1}), & n=1,2,\cdots.\end{cases} \tag{1.7}$$

The symbols used on the right-hand side of (1.3) then have the meaning

$$(a)_\infty=(a;\, q)_\infty=\prod_{k=0}^{\infty}(1-aq^k), \tag{1.8}$$

whenever the infinite product converges.

Askey and Wilson [9] have shown that the q-Wilson polynomials defined by

$$p_n(x;\, a,b,c,d)={}_4\phi_3\left[\begin{matrix}q^{-n}, abcdq^{n-1}, ae^{i\theta}, ae^{-i\theta} \\ ab, ac, ad\end{matrix};q\right], \tag{1.9}$$

$x=\cos\theta$, $0\leq\theta\leq\pi$, are orthogonal with respect to the weight function $w(x;\, a,b,c,d)$. The ${}_4\phi_3$ function on the right is a special case of the basic hypergeometric series defined by

$${}_{r+1}\phi_r\left[\begin{matrix}a_1,a_2,\cdots,a_{r+1} \\ b_1,\cdots,b_r\end{matrix};z\right]=\sum_{k=0}^{\infty}\frac{(a_1)_k(a_2)_k\cdots(a_{r+1})_k}{(q)_k(b_1)_k\cdots(b_r)_k}z^k. \tag{1.10}$$

The continuous q-Jacobi polynomials $P_n^{(\alpha,\beta)}(x|q)$, as introduced by Askey–Wilson [9], are a special case of the q-Wilson polynomials:

(1.11)

$$P_n^{(\alpha,\beta)}(x|q)=\frac{(q^{\alpha+1})_n}{(q)_n}p_n(x;\, q^{(\alpha+1/2)/2}, q^{(\alpha+3/2)/2}, -q^{(\beta+1/2)/2}, -q^{(\beta+3/2)/2}).$$

Two of the most widely used projection formulas for the classical Jacobi polynomials are [1], [18]

$$(1-x)^{\alpha+\mu}\frac{P_n^{(\alpha+\mu,\beta-\mu)}(x)}{P_n^{(\alpha+\mu,\beta-\mu)}(1)}=\frac{\Gamma(\alpha+\mu+1)}{\Gamma(\alpha+1)\Gamma(\mu)}\int_x^1(1-y)^\alpha(y-x)^{\mu-1}\frac{P_n^{(\alpha,\beta)}(y)}{P_n^{(\alpha,\beta)}(1)}dy, \tag{1.12}$$

and

(1.13)

$$(1+x)^{\beta+\mu}\frac{P_n^{(\alpha-\mu,\beta+\mu)}(x)}{P_n^{(\beta+\mu,\alpha-\mu)}(1)}=\frac{\Gamma(\beta+\mu+1)}{\Gamma(\beta+1)\Gamma(\mu)}\int_{-1}^x(1+y)^\beta(x-y)^{\mu-1}\frac{P_n^{(\alpha,\beta)}(y)}{P_n^{(\beta,\alpha)}(1)}dy,$$

where $-1<\operatorname{Re}\alpha$, $\operatorname{Re}\beta$ and $0<\operatorname{Re}\mu$.

We shall obtain q-analogues of these formulas for the q-Wilson polynomials (1.9) generally, and for $P_n^{(\alpha,\beta)}(x|q)$ in particular, and will show in what sense they are q-analogues of (1.12) and (1.13). We shall also obtain a reproducing kernel for $p_n(x;\, a,b,c,d)$ by using the projection formulas. This will be done in the following section.

In §3 we shall find a q-analogue of Euler's integral representation of Gauss' hypergeometric function ${}_2F_1$ that turns out to be a very well-poised ${}_8\phi_7$. In §4 we shall apply our results to obtain a generating function for $P_n^{(\alpha,\beta)}(x|q)$.

2. Projection formulas and a reproducing kernel. By (1.9) and (1.10) we have, for arbitrary parameters a, b,. c, d, c', d', subject to the constraint (1.6),

$$\int_{-1}^{1} dy\, w(y;\, a,b,c,d)p_n(y;\, a,b,c',d') = \sum_{k=0}^{n} \frac{(q^{-n})_k(abc'd'q^{n-1})_k q^k}{(q)_k(ab)_k(ac')_k(ad')_k} \int_{-1}^{1} dy\, w(y;\, a,b,c,d)(ae^{i\phi})_k(ae^{-i\phi})_k, \tag{2.1}$$

$y=\cos\phi$. However, from (1.3) to (1.5) it is clear that

$$\begin{aligned}\int_{-1}^{1} dy\, w(y;\, a,b,c,d)(ae^{i\phi})_k(ae^{-i\phi})_k &= \int_{-1}^{1} dy\, w(y;\, aq^k,b,c,d)\\ &= \frac{2\pi(abcdq^k)_\infty}{(abq^k)_\infty(acq^k)_\infty(adq^k)_\infty(bc)_\infty(bd)_\infty(cd)_\infty(q)_\infty}\\ &= \frac{\kappa(ab)_k(ac)_k(ad)_k}{(abcd)_k}.\end{aligned} \tag{2.2}$$

Hence we get

$$\int_{-1}^{1} dy\, w(y;\, a,b,c,d)p_n(y;\, a,b,c',d') = \kappa\, {}_4\phi_3\left[\begin{matrix} q^{-n}, abc'd'q^{n-1}, ac, ad \\ ac', ad', abcd \end{matrix}; q, q\right]. \tag{2.3}$$

Setting $c=\mu e^{i\theta}$, $d=\mu e^{-i\theta}$, $x=\cos\theta$, dropping the primes in c', d' and using (1.3), we get the projection formula

$$\int_{-1}^{1} dy\, w(y;\, a,b,\mu e^{i\theta},\mu e^{-i\theta})p_n(y;\, a,b,c,d) = \frac{2\pi(ab\mu^2)_\infty}{(q)_\infty(ab)_\infty(\mu^2)_\infty|(a\mu e^{i\theta})_\infty(b\mu e^{i\theta})_\infty|^2}\, p_n(x;\, a\mu,b\mu,c\mu^{-1},d\mu^{-1}), \tag{2.4}$$

which is valid if max $(|a|,|b|,|\mu|)<.1$

Since the ${}_4\phi_3$ series defining $p_n(x;\, a,b,c,d)$ is balanced, we can transform it to various different forms by Sears' formula [17]

$${}_4\phi_3\left[\begin{matrix} q^{-n}, a, b, c \\ d, e, f \end{matrix}; q\right] = \frac{(de/bc)_n(df/bc)_n}{(e)_n(f)_n}\left(\frac{bc}{d}\right)^n {}_4\phi_3\left[\begin{matrix} q^{-n}, a, d/b, d/c \\ d, de/bc, df/bc \end{matrix}; q\right], \tag{2.5}$$

where $abcq^{1-n}=def$. Thus

$$p_n(y;\, a,b,c,d) = \frac{(bc)_n(cd)_n}{(ad)_n(ab)_n}\left(\frac{a}{c}\right)^n p_n(y;\, c,d,a,b),$$

$$p_n(x;\, a\mu,b\mu,c\mu^{-1},d\mu^{-1}) = \frac{(bc)_n(cd\mu^{-2})_n}{(ad)_n(ab\mu^2)_n}\left(\frac{a\mu^2}{c}\right)^n p_n(x;\, c\mu^{-1},d\mu^{-1},a\mu,b\mu). \tag{2.6}$$

Using (2.6) in (2.4) and interchanging $c \leftrightarrow a$, $d \leftrightarrow b$, we get a second projection formula

$$(2.7)\quad \int_{-1}^{1} dy\, w(y;\, \mu e^{i\theta}, \mu e^{-i\theta}, c, d)\, p_n(y;\, a,b,c,d)$$
$$= \frac{2\pi (cd\mu^2)_\infty}{(q)_\infty (cd)_\infty (\mu^2)_\infty \left|(c\mu e^{i\theta})_\infty (d\mu e^{i\theta})_\infty\right|^2}$$
$$\cdot \frac{(cd)_n (ab\mu^{-2})_n}{(ab)_n (cd\mu^2)_n} \mu^{2n} p_n(x;\, a\mu^{-1}, b\mu^{-1}, c\mu, d\mu)$$

with max $(|c|, |d|, |\mu|) < 1$.

We now replace x by z and θ by ψ in (2.4), multiply both sides by $w(z;\, c\mu^{-1}, d\mu^{-1}, \mu e^{i\theta}, \mu e^{-i\theta})|(a\mu e^{i\psi})_\infty (b\mu e^{i\psi})_\infty|^2$ and integrate over z to get

$$(2.8)\quad \int_{-1}^{1} dz\, w(z;\, c\mu^{-1}, d\mu^{-1}, \mu e^{i\theta}, \mu e^{-i\theta}) \left|(a\mu e^{i\psi})_\infty (b\mu e^{i\psi})_\infty\right|^2$$
$$\cdot \int_{-1}^{1} dy\, w(y;\, a, b, \mu e^{i\psi}, \mu e^{-i\psi})\, p_n(y;\, a,b,c,d)$$
$$= \frac{(ab\mu^2)_\infty (cd)_\infty}{(ab)_\infty (cd\mu^{-2})_\infty} \left\{ \frac{2\pi}{(q)_\infty (\mu^2)_\infty \left|(ce^{i\theta})_\infty (de^{i\theta})_\infty\right|} \right\}^2$$
$$\cdot \frac{(bc)_n (cd\mu^{-2})_n}{(ad)_n (ab\mu^2)_n} \left(\frac{a\mu^2}{c}\right)^n \cdot p_n(x;\, c,d,a,b),$$

which, by (2.6), leads to the integral equation

$$(2.9)\quad \int_{-1}^{1} K_\mu(x,y;\, q)\, p_n(y;\, a,b,c,d)\, dy = \lambda_n p_n(x;\, a,b,c,d)$$

where the reproducing kernel is given by

(2.10)

$$K_\mu(x,y;\, q) = \frac{(ab)_\infty (cd\mu^{-2})_\infty}{(cd)_\infty (ab\mu^2)_\infty} \left| \frac{(q)_\infty (\mu^2)_\infty (ce^{i\theta})_\infty (de^{i\theta})_\infty}{2\pi} \right|^2$$
$$\cdot \int_{-1}^{1} w(z;\, c\mu^{-1}, d\mu^{-1}, \mu e^{i\theta}, \mu e^{-i\theta})\, w(y;\, a, b, \mu e^{i\psi}, \mu e^{-i\psi})$$
$$\cdot \left|(a\mu e^{i\psi})_\infty (b\mu e^{i\psi})_\infty\right|^2 dz,$$

$\cos\psi = z$, provided max $(|a|, |b|, |c\mu^{-1}|, |d\mu^{-1}|, |\mu|, |q|) < 1$. The eigenvalue λ_n is given by

$$(2.11)\quad \lambda_n = \frac{(ab)_n (cd\mu^{-2})_n}{(cd)_n (ab\mu^2)_n} \mu^{2n}.$$

Let us now set $a = q^{(\alpha+1/2)/2}$, $b = q^{(\alpha+3/2)/2}$, $c = -q^{(\beta+1/2)/2}$, $d = -q^{(\beta+3/2)/2}$ in (2.4) and replace μ by $q^{\mu/2}$. Use of the q-gamma function [6], [19]

$$(2.12)\quad \Gamma_q(x) = \frac{(q)_\infty}{(q^x)_\infty} (1-q)^{1-x}$$

and the notation

$$(a)_\lambda=\frac{(a)_\infty}{(aq^\lambda)_\infty} \tag{2.13}$$

enables us to write (2.4) in the form

(2.14)

$$\int_{-1}^{1} G(x,y;\,q)\,{}_4\phi_3\left[\begin{matrix} q^{-n},\, q^{n+\alpha+\beta+1},\, q^{(\alpha+1/2)/2}e^{i\phi},\, q^{(\alpha+1/2)/2}e^{-i\phi} \\ q^{\alpha+1},\, -q^{(\alpha+\beta+1)/2},\, -q^{(\alpha+\beta+2)/2} \end{matrix};q\right]dy$$

$$=\frac{\Gamma_q(\alpha+1)\Gamma_q(\mu)}{\Gamma_q(\alpha+\mu+1)}\left|\left(\sqrt{q}\,e^{i\theta}\right)_{(\alpha+\mu-1/2)/2}\left(\sqrt{q}\,e^{i\theta}\right)_{(\alpha+\mu+1/2)/2}\right|^2 2^{-\alpha-\mu}$$

$$\cdot\,{}_4\phi_3\left[\begin{matrix} q^{-n},\, q^{n+\alpha+\beta+1},\, q^{(\alpha+\mu+1/2)/2}e^{i\theta},\, q^{(\alpha+\mu+1/2)/2}e^{-i\theta} \\ q^{\alpha+\mu+1},\, -q^{(\alpha+\beta+1)/2},\, -q^{(\alpha+\beta+2)/2} \end{matrix};q\right],$$

where

$$G(x,y;\,q)=\frac{(1-q)(q)_\infty^2\left|\left(\sqrt{q}\,e^{i\theta}\right)_\infty\right|^4}{2^{\alpha+\mu+1}\pi}\left(1-y^2\right)^{-1/2} \tag{2.15}$$

$$\cdot\frac{h(y,1)h(y,-1)h\left(y,\sqrt{q}\right)h\left(y,-\sqrt{q}\right)}{h\left(y,q^{(\alpha+1/2)/2}\right)h\left(y,q^{(\alpha+3/2)/2}\right)h\left(y,q^{\mu/2}e^{i\theta}\right)h\left(y,q^{\mu/2}e^{-i\theta}\right)}.$$

Since $\lim_{q\to 1}\Gamma_q(x)=\Gamma(x)$ and $\lim_{q\to 1}|(\sqrt{q}\,e^{i\theta})_\lambda|^2=2^\lambda(1-\cos\theta)^\lambda$, we obtain, from (2.14),

$$\lim_{q\to 1}\int_{-1}^{1} G(x,y;\,q)\frac{P_n^{(\alpha,\beta)}(y)}{P_n^{(\alpha,\beta)}(1)}dy=\frac{\Gamma(\alpha+1)\Gamma(\mu)}{\Gamma(\alpha+\mu+1)}(1-x)^{\alpha+\mu}\frac{P_n^{(\alpha+\mu,\beta-\mu)}(x)}{P_n^{(\alpha+\mu,\beta-\mu)}(1)}. \tag{2.16}$$

We shall now show that

$$\lim_{q\to 1}G(x,y;\,q)=(1-y)^\alpha(y-x)^{\mu-1}H(y-x), \tag{2.17}$$

where $H(t)$ is the unit step function. First, we rewrite $G(x,y;\,q)$ in the form

(2.18)

$$G(x,y;\,q)=\left[\frac{\Gamma_q(1/2)}{\sqrt{\pi}}\right]^2 2^{-(\alpha+1/2)}(1-y)^{-1/2}\left|\left(e^{i\phi}\right)_{(\alpha+1/2)/2}\left(\sqrt{q}\,e^{i\phi}\right)_{(\alpha+1/2)/2}\right|^2$$

$$\cdot 2^{-1/2}(1+y)^{-1/2}\left|\left(-e^{i\phi}\right)_{1/2}\right|^2 2^{1-\mu}\left|\left(\sqrt{q}\,e^{i\theta+i\phi}\right)_{((\mu-1)/2)}\left(\sqrt{q}\,e^{i\theta-i\phi}\right)_{((\mu-1)/2)}\right|^2$$

$$\cdot L(x,y;\,q)$$

where

$$L(x,y;\,q)=\frac{1}{2}\frac{\left(\sqrt{q}\right)_\infty^2\left|\left(\sqrt{q}\,e^{i\theta}\right)_\infty\left(-\sqrt{q}\,e^{i\phi}\right)_\infty\right|^4}{\left|\left(\sqrt{q}\,e^{i\theta+i\phi}\right)_\infty\left(\sqrt{q}\,e^{i\theta-i\phi}\right)_\infty\right|^2}. \tag{2.19}$$

It is clear that

$$\lim_{q\to 1} G(x,y;\, q)=(1-y)^{\alpha}|y-x|^{\mu-1}\lim_{q\to 1} L(x,y;\, q). \tag{2.20}$$

Using Jacobi's triple product formula:

$$(t;\, q)_{\infty}(qt^{-1};\, q)_{\infty}(q;\, q)_{\infty}=\sum_{n=-\infty}^{\infty}(-1)^{n}q^{n(n-1)/2}t^{n}, \tag{2.21}$$

and changing the base to q^2, we obtain
(2.22)

$$L(x,y;\, q^2)=\frac{1}{2}\left[\frac{(q;\, q^2)_{\infty}}{(q^2;\, q^2)_{\infty}}\right]^2\frac{\left(\sum_{n=-\infty}^{\infty}(-1)^{n}q^{n^2}e^{in\theta}\right)^2\left(\sum_{n=-\infty}^{\infty}q^{n^2}e^{in\phi}\right)^2}{\left(\sum_{n=-\infty}^{\infty}(-1)^{n}q^{n^2}e^{in(\theta+\phi)}\right)\left(\sum_{n=-\infty}^{\infty}(-1)^{n}q^{n^2}e^{in(\theta-\phi)}\right)}.$$

Using the theta functions [20, Ch. 21] this can be expressed in the form

$$L(x,y;\, q^2)=\frac{1}{2}\left[\frac{(q;\, q^2)_{\infty}\vartheta_3}{(q^2;\, q^2)_{\infty}}\right]^2\frac{\vartheta_4^2(\theta/2,q)\vartheta_3^2(\phi/2,q)}{\vartheta_4((\theta+\phi)/2,q)\vartheta_4((\theta-\phi)/2,q)\vartheta_3^2}, \tag{2.23}$$

where

$$\vartheta_3=(-q;\, q^2)_{\infty}(-q;\, q^2)_{\infty}(q^2;\, q^2)_{\infty}.$$

Using the identity $(q;\, q^2)_{\infty}(-q;\, q^2)_{\infty}(-q^2;\, q^2)_{\infty}=1$ and the addition formula [20, Ex. 2, p. 488]

$$\vartheta_4\left(\frac{\theta+\phi}{2},q\right)\vartheta_4\left(\frac{\theta-\phi}{2},q\right)\vartheta_3^2=\vartheta_4^2\left(\frac{\theta}{2},q\right)\vartheta_3^2\left(\frac{\phi}{2},q\right)+\vartheta_2^2\left(\frac{\theta}{2},q\right)\vartheta_1^2\left(\frac{\phi}{2},q\right), \tag{2.24}$$

we can write

$$L(x,y;\, q^2)=\frac{1}{2}\left[(-q;\, q^2)_{1/2}\right]^2\Big/\left(1+\frac{\vartheta_2^2(\theta/2,q)\vartheta_1^2(\phi/2,q)}{\vartheta_4^2(\theta/2,q)\vartheta_3^2(\phi/2,q)}\right). \tag{2.25}$$

Since

$$\lim_{q\to 1}\frac{1}{2}\left[(-q;\, q^2)_{1/2}\right]^2=1, \tag{2.26}$$

we need to consider the limit of

$$B(x,y;\, q^2)=\frac{\vartheta_2(\theta/2,q)\vartheta_1(\phi/2,q)}{\vartheta_4(\theta/2,q)\vartheta_3(\phi/2,q)}\quad \text{as } q\to 1. \tag{2.27}$$

Using Poisson's transformation [20, p. 476]

$$\sum_{n=-\infty}^{\infty}e^{-n^2\pi t+2niz}=\frac{e^{-z^2/\pi t}}{\sqrt{t}}\sum_{n=-\infty}^{\infty}e^{-n^2\pi/t+2nz/t},\qquad \operatorname{Re} t>0 \tag{2.28}$$

and the definitions of the theta functions we find that, with $q=e^{-\pi t}$,

$$(2.29)\qquad \vartheta_1\left(\frac{\phi}{2},q\right)=\frac{e^{-(\pi-\phi)^2/4\pi t}}{\sqrt{t}}\sum_{-\infty}^{\infty}(-1)^n\exp\left(-\frac{\pi n^2}{t}-\frac{n(\pi-\phi)}{t}\right),$$

$$\vartheta_2\left(\frac{\theta}{2},q\right)=\frac{e^{-\theta^2/4\pi t}}{\sqrt{t}}\sum_{-\infty}^{\infty}(-1)^n\exp\left(-\frac{\pi n(n-\theta/\pi)}{t}\right),$$

$$\vartheta_3\left(\frac{\phi}{2},q\right)=\frac{e^{-\phi^2/4\pi t}}{\sqrt{t}}\sum_{-\infty}^{\infty}\exp\left(-\frac{\pi n(n-\phi/\pi)}{t}\right),$$

$$\vartheta_4\left(\frac{\theta}{2},q\right)=\frac{e^{-(\pi-\theta)^2/4\pi t}}{\sqrt{t}}\sum_{-\infty}^{\infty}\exp\left(-\frac{\pi n^2}{t}-\frac{n(\pi-\theta)}{t}\right)$$

As $q\to 1-$, we have

$$(2.30)\quad B(x,y;\,q^2)\simeq\exp\left[\frac{\theta^2}{4\pi t}-\frac{(\pi-\phi)^2}{4\pi t}+\frac{\phi^2}{4\pi t}+\frac{(\pi-\theta)^2}{4\pi t}\right]=\exp\left(\frac{\phi-\theta}{2t}\right).$$

For $0\leq\theta,\ \phi\leq\pi$, it is clear that

$$(2.31)\qquad \lim_{q\to 1-}B(x,y;\,q^2)=\lim_{t\to 0+}\exp\left(\frac{\phi-\theta}{2t}\right)=\begin{cases}0 & \text{if } \theta>\phi,\\ \infty & \text{if } \theta<\phi.\end{cases}$$

Hence,

$$(2.32)\qquad \lim_{q\to 1-}L(x,y;\,q^2)=\begin{cases}0 & \text{if } x>y,\\ 1 & \text{if } x<y,\end{cases}$$

which completes the proof of (2.17).

3. A q-analogue of Euler's integral formula.

We start with Sears' identity [16]:

$$(3.1)\qquad \frac{(e)_\infty(f)_\infty}{(a)_\infty(b)_\infty(c)_\infty}\,{}_3\phi_2\left[\begin{matrix}a,b,c\\ e,f\end{matrix};q\right]$$
$$-\frac{q}{e}\,\frac{(q^2/e)_\infty(qf/e)_\infty}{(aq/e)_\infty(bq/e)_\infty(cq/e)_\infty}\,{}_3\phi_2\left[\begin{matrix}aq/e,\,bq/e,\,cq/e\\ q^2/e,\,fq/e\end{matrix};q\right]$$
$$=\frac{(e)_\infty(q/e)_\infty(f/a)_\infty(f/b)_\infty(f/c)_\infty}{(a)_\infty(b)_\infty(c)_\infty(aq/e)_\infty(bq/e)_\infty(cq/e)_\infty},$$

where both ${}_3\phi_2$'s on the left-hand side are balanced, that is

$$(3.2)\qquad ef=abcq.$$

Let $a \to A$, $b \to ae^{i\theta}$, $c \to ae^{-i\theta}$, $e \to E$. Then (3.1) together with (3.2) gives

$$(3.3)\quad {}_3\phi_2\left[\begin{matrix} A, ae^{i\theta}, ae^{-i\theta} \\ E, Aqa^2/E \end{matrix}; q\right] \frac{-q}{E} \frac{(A)_\infty (q^2/E)_\infty (Aq^2a^2/E^2)_\infty (ae^{i\theta})_\infty (ae^{-i\theta})_\infty}{(E)_\infty (qAa^2/E)_\infty (Aq/E)_\infty (aq^{i\theta}/E)_\infty (aqe^{-i\theta}/E)_\infty}$$

$$\cdot\, {}_3\phi_2\left[\begin{matrix} Aq/E, aqe^{i\theta}/E, aqe^{-i\theta}/E \\ q^2/E, Aq^2a^2/E^2 \end{matrix}; q\right]$$

$$= \frac{(q/E)_\infty (qa^2/E)_\infty (aAqe^{i\theta}/E)_\infty (aAqe^{-i\theta}/E)_\infty}{(qAa^2/E)_\infty (Aq/E)_\infty (aqe^{i\theta}/E)_\infty (aqe^{-i\theta}/E)_\infty}.$$

Assuming that $|aq/E| < 1$ and that (1.6) holds, we now multiply both sides of (3.3) by $w(x;\, a, b, c, d)$ and integrate over x to get

(3.4)

$$\frac{(q/E)_\infty (qa^2/E)_\infty}{(Aq/E)_\infty (Aqa^2/E)_\infty} \int_{-1}^{1} dx\, w(x;\, a, b, c, d) \left| \frac{(qaAe^{i\theta}/E)_\infty}{(aqe^{i\theta}/E)_\infty} \right|^2$$

$$= \kappa\, {}_4\phi_3\left[\begin{matrix} A, ab, ac, ad \\ E, qAa^2/E, abcd \end{matrix}; q\right]$$

$$- \frac{2\pi (abcdq/E)_\infty (A)_\infty (q^2/E)_\infty (q^2a^2A/E^2)_\infty qE^{-1}}{(abq/E)_\infty (acq/E)_\infty (bc)_\infty (bd)_\infty (cd)_\infty (E)_\infty (qAa^2/E)_\infty (qA/E)_\infty (q)_\infty}$$

$$\cdot\, {}_4\phi_3\left[\begin{matrix} Aq/E, abq/E, acq/E, adq/E \\ q^2/E, q^2a^2A/E^2, abcdq/E \end{matrix}; q\right].$$

Noting that $-qE^{-1}(q^2/E)_\infty/(E)_\infty = (q/E)_\infty/(E/q)_\infty$ we now divide both sides of (3.4) by κ and simplify to get

(3.5)

$$\frac{(Aa^2cdq/E)_\infty (acq/E)_\infty (adq/E)_\infty (Aq/E)_\infty}{(a^2cdq/E)_\infty (Aacq/E)_\infty (Aadq/E)_\infty (q/E)_\infty}\, {}_4\phi_3\left[\begin{matrix} A, ab, ac, ad \\ E, qAa^2/E, abcd \end{matrix}; q\right]$$

$$+ \frac{(Aa^2cdq/E)_\infty (ab)_\infty (ac)_\infty (ad)_\infty (A)_\infty (q^2a^2A/E^2)_\infty (abcdq/E)_\infty}{(a^2cdq/E)_\infty (Aadq/E)_\infty (Aacq/E)_\infty (abcd)_\infty (E/q)_\infty (abq/E)_\infty (qAa^2/E)_\infty}$$

$$\cdot\, {}_4\phi_3\left[\begin{matrix} Aq/E, abq/E, acq/E, adq/E \\ q^2/E, q^2a^2A/E^2, abcdq/E \end{matrix}; q\right]$$

$$= \frac{(Aa^2cdq/E)_\infty (acq/E)_\infty (adq/E)_\infty (ab)_\infty (ac)_\infty}{(a^2cdq/E)_\infty (Aacq/E)_\infty (Aadq/E)_\infty (abcd)_\infty (qa^2A/E)_\infty 2\pi}$$

$$\cdot\, (ad)_\infty (q)_\infty (bc)_\infty (bd)_\infty (cd)_\infty (qa^2/E)_\infty$$

$$\cdot \int_{-1}^{1} dx\, w(x;\, a, b, c, d) \left| \frac{(qaAe^{i\theta}/E)_\infty}{(qae^{i\theta}/E)_\infty} \right|^2.$$

By [10, 8.5 (3)] the left-hand side equals a very well-poised ${}_8\phi_7$. Replacing aq/E by t for the sake of notational simplicity we have thus found the integral representation:

(3.6)

$$
{}_8\Phi_7\left[\begin{matrix} Aacdtq^{-1}, q\sqrt{Aacdtq^{-1}}, -q\sqrt{Aacdtq^{-1}}, At/b, A, ac, ad, dc \\ \sqrt{Aacdtq^{-1}}, -\sqrt{Aacdtq^{-1}}, abcd, acdt, Adt, Act, Aat \end{matrix}; bt\right]
$$

$$
=\frac{(Aacdt)_\infty(ct)_\infty(dt)_\infty(at)_\infty}{(acdt)_\infty(Act)_\infty(Adt)_\infty(Aat)_\infty}\cdot\frac{(q)_\infty(ab)_\infty(ac)_\infty(ad)_\infty(bc)_\infty(bd)_\infty(cd)_\infty}{2\pi(abcd)_\infty}
$$

$$
\cdot\int_{-1}^{1} dx\, w(x;\, a,b,c,d)\left|\frac{(Ate^{i\theta})_\infty}{(te^{i\theta})_\infty}\right|^2 .
$$

It appears that as far as Askey–Wilson's q-extension (1.3) of the beta-integral is concerned, the appropriate q-analogue of a ${}_2F_1$ is a very well-poised ${}_8\phi_7$. It is not hard to see that (3.6) does approach Euler's well-known integral formula [11, 2.1.3(10)] in the limit $q\to 1$ after one replaces the parameters A, a, b, c, d by powers of q.

4. A generating function for $P_n^{(\alpha,\beta)}(x|q)$. Rogers' q-ultraspherical polynomials [7], [8], [15] are defined by the generating function

$$
(4.1)\qquad \sum_{n=0}^{\infty} C_n(x;\, A\,|\,q)t^n=\left|\frac{(Ate^{i\theta})_\infty}{(te^{i\theta})_\infty}\right|^2,
$$

where $x=\cos\theta$, $0\le\theta\le\pi$, $|t|<1$ and $C_n(x;\, A\,|\,q)$ has the series representation

$$
(4.2)\qquad C_n(x;\, A\,|\,q)=\sum_{i=0}^{n}\frac{(A)_k(A)_{n-k}}{(q)_k(q)_{n-k}}\cos(n-2k)\theta.
$$

However, $C_n(x;\, A\,|\,q)$ can also be represented [8] by a balanced ${}_4\phi_3$:

$$
(4.3)\qquad C_n(x;\, A\,|\,q)=\frac{(A^2)_n}{(q)_n}A^{-n/2}{}_4\Phi_3\left[\begin{matrix} q^{-n}, A^2q^n, \sqrt{A}\,e^{i\theta}, \sqrt{A}\,e^{-i\theta} \\ A\sqrt{q}, -A\sqrt{q}, -A \end{matrix}; q\right]
$$

$$
=\frac{(A^2)_n}{(q)_n}A^{-n/2}p_n\left(x;\, \sqrt{A}, \sqrt{Aq}, -\sqrt{A}, -\sqrt{Aq}\right).
$$

Setting $a=\sqrt{A}$, $b=\sqrt{Aq}$, $c=-\sqrt{A}$, $d=-\sqrt{Ae}$ in (2.4), we have the projection formula

$$
(4.4)\quad \int_{-1}^{1} dy\, w\left(y;\, \sqrt{A}, \sqrt{Aq}, \mu e^{i\theta}, \mu e^{-i\theta}\right)C_n(y;\, A\,|\,q)
$$

$$
=\frac{2\pi\left(A\sqrt{q}\,\mu^2\right)_\infty}{(q)_\infty\left(A\sqrt{q}\right)_\infty(\mu^2)_\infty\left|\left(\sqrt{A}\,\mu e^{i\theta}\right)_\infty\left(\sqrt{Aq}\,\mu e^{i\theta}\right)_\infty\right|^2}
$$

$$
\cdot\frac{(A^2)_n}{(q)_n}A^{-n/2}p_n\left(x;\, \mu\sqrt{A}, \mu\sqrt{Aq}, -\mu^{-1}\sqrt{A}, -\mu^{-1}\sqrt{Aq}\right).
$$

Multiply both sides by t^n, sum over n from 0 and ∞ and assume that the order of summation and integration can be interchanged. We get, using (4.1),

$$(4.5)\quad \int_{-1}^{1} dy\, w\left(y;\sqrt{A},\sqrt{Aq},\mu e^{i\theta},\mu e^{-i\theta}\right)\left|\frac{(Ate^{i\phi})_\infty}{(te^{i\phi})_\infty}\right|^2$$

$$=\frac{2\pi\left(A\sqrt{q}\,\mu^2\right)_\infty}{(q)_\infty(\mu^2)_\infty\left(A\sqrt{q}\right)_\infty\left|\left(\sqrt{A}\,\mu e^{i\theta}\right)_\infty\left(\sqrt{Aq}\,\mu e^{i\theta}\right)_\infty\right|^2}$$

$$\cdot\sum_{n=0}^{\infty}\frac{(A^2)_n\left(t/\sqrt{A}\right)^n}{(q)_n}P_n\left(x;\mu\sqrt{A},\mu\sqrt{Aq},-\mu^{-1}\sqrt{A},-\mu^{-1}\sqrt{Aq}\right).$$

However, setting $a=\sqrt{A}$, $b=\sqrt{Aq}$, $c=\mu e^{i\theta}$, $d=\mu e^{-i\theta}$ in (3.6) we have
(4.6)

$$\int_{-1}^{1} dy\, w\left(y;\sqrt{A},\sqrt{Aq},\mu e^{i\theta},\mu e^{-i\theta}\right)\left|\frac{(Ate^{i\phi})_\infty}{(te^{i\phi})_\infty}\right|^2$$

$$=\frac{2\pi\left(A\sqrt{q}\,\mu^2\right)_\infty}{(q)_\infty\left(A\sqrt{q}\right)_\infty(\mu^2)_\infty\left|\left(\sqrt{A}\,\mu e^{i\theta}\right)_\infty\left(\sqrt{Aq}\,\mu e^{i\theta}\right)_\infty\right|^2}$$

$$\cdot\frac{\left(\sqrt{A}\,\mu^2t\right)_\infty(A^{3/2}t)_\infty(\mu Ae^{i\theta}t)_\infty(\mu Ae^{-i\theta}t)_\infty}{(A^{3/2}\mu^2t)_\infty\left(\sqrt{A}\,t\right)_\infty(\mu e^{i\theta}t)_\infty(\mu e^{-i\theta}t)_\infty}$$

$$\cdot\,{}_8\phi_7\left[\begin{matrix}A^{3/2}\mu^2tq^{-1},\,q\sqrt{A^{3/2}\mu^2tq^{-1}},\,-q\sqrt{A^{3/2}\mu^2tq^{-1}},\,t\sqrt{A/q},\,A,\\ \sqrt{A^{3/2}\mu^2tq^{-1}},\,-\sqrt{A^{3/2}\mu^2tq^{-1}},\,A\mu^2\sqrt{q},\,\sqrt{A}\,\mu^2t,\,\mu Ate^{-i\theta},\end{matrix}\right.$$

$$\left.\begin{matrix}\mu\sqrt{A}\,e^{i\theta},\,\mu\sqrt{A}\,e^{-i\theta},\,\mu^2\\ \mu Ate^{i\theta},\,A^{3/2}t\end{matrix}\;;t\sqrt{Aq}\right].$$

Comparing (4.5) and (4.6) we get
(4.7)

$$\sum_{n=0}^{\infty}\frac{(A^2)_n}{(q)_n}\left(\frac{t}{\sqrt{A}}\right)^n p_n\left(x;\mu\sqrt{A},\mu\sqrt{Aq},-\mu^{-1}\sqrt{A},-\mu^{-1}\sqrt{Aq}\right)$$

$$=\frac{\left(\sqrt{A}\,\mu^2t\right)_\infty(A^{3/2}t)_\infty}{\left(\sqrt{A}\,t\right)_\infty(A^{3/2}\mu^2t)_\infty}\left|\frac{(\mu Ate^{i\theta})_\infty}{(\mu te^{i\theta})_\infty}\right|^2$$

$$\cdot\,{}_8\phi_7\left[\begin{matrix}A^{3/2}\mu^2tq^{-1},\,q\sqrt{A^{3/2}\mu^2tq^{-1}},\,-q\sqrt{A^{3/2}\mu^2tq^{-1}},\\ \sqrt{A^{3/2}\mu^2tq^{-1}},\,-\sqrt{A^{3/2}\mu^2tq^{-1}},\,\mu Ate^{-i\theta},\,\mu Ate^{i\theta},\end{matrix}\right.$$

$$\left.\begin{matrix}\mu\sqrt{A}\,e^{i\theta},\,\mu\sqrt{A}\,e^{-i\theta},\,\mu^2,\,A,\,t\sqrt{A/q}\\ A^{3/2}t,\,\sqrt{A}\,\mu^2t,\,A\mu^2\sqrt{q}\end{matrix}\;;\,t\sqrt{Aq}\right].$$

For convergence of the series on the right we require $|t\sqrt{Aq}|<1$. This formula was recently obtained by Gasper and Rahman [14] without the benefit of the integral representation (3.6), and, consequently they required a good deal more computation.

For the continuous q-Jacobi polynomials defined in (1.11) we may set $A=q^{(\alpha+\beta+1)/2}$, $\mu=q^{(\alpha-\beta)/2}$ in (4.7) and obtain

(4.8)

$$\sum_{n=0}^{\infty}\frac{\left(q^{\alpha+\beta+1}\right)_n}{\left(q^{\alpha+1}\right)_n}\left(tq^{-(\alpha+\beta+1)/4}\right)^n P_n^{(\alpha,\beta)}(x|q)$$

$$=\frac{\left(tq^{(3\alpha-\beta+1)/4}\right)_\infty\left(tq^{3(\alpha+\beta+1)/4}\right)_\infty}{\left(tq^{(\alpha+\beta+1)/4}\right)_\infty\left(tq^{(5\alpha+\beta+3)/4}\right)_\infty}\left|\frac{\left(q^{(3\alpha+\beta+2)/4}te^{i\theta}\right)_\infty}{\left(q^{(\alpha-\beta)/4}te^{i\theta}\right)_\infty}\right|^2$$

$$\cdot {}_8\phi_7\left[\begin{matrix}tq^{(5\alpha+\beta-1)/4},\ q\sqrt{tq^{(5\alpha+\beta-1)/4}},\ -q\sqrt{tq^{(5\alpha+\beta-1)/4}},\ q^{(\alpha+1/2)/2}e^{i\theta},\ q^{(\alpha+1/2)/2}e^{-i\theta},\\ \sqrt{tq^{(5\alpha+\beta-1)/4}},\ -\sqrt{tq^{(5\alpha+\beta-1)/4}},\ tq^{(3\alpha+\beta+2)/4}e^{-i\theta},\ tq^{(3\alpha+\beta+2)/4}e^{i\theta},\end{matrix}\right.$$

$$\left.\begin{matrix}q^{(\alpha-\beta)/2},\ q^{(\alpha+\beta+1)/2},\ tq^{(\alpha+\beta-1)/4}\\ tq^{3(\alpha+\beta+1)/4},\ tq^{(3\alpha-\beta+1)/4},\ q^{\alpha+1}\end{matrix};tq^{(\alpha+\beta+3)/4}\right].$$

Note that the right-hand side is positive if $0<t<1$, $\alpha+\beta+1\geq 0$ and $\alpha\geq\beta$.

Acknowledgment. We would like to thank Professor R. Askey for suggesting the proof of (2.17).

REFERENCES

[1] R. Askey, *Orthogonal polynomials and positivity*, in Special Functions and Wave Propagation, D. Ludwig and F. W. J. Olver, Eds., Studies in Applied Mathematics, 6, Society for Industrial and Applied Mathematics, Philadelphia, 1970.

[2] ———, *Orthogonal Polynomials and Special Functions*, CBMS Regional Conference Series in Applied Mathematics, 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[3] ———, *Summability of Jacobi series*, Trans. Amer. Math. Soc., 179 (1973), pp. 71–84.

[4] R. Askey and J. Fitch, *Integral representations for Jacobi polynomials and some applications*, J. Math. Anal. Appl., 26 (1969), pp. 411–437.

[5] R. Askey and G. Gasper, *Jacobi polynomial expansions of Jacobi polynomials with nonnegative coefficients*, Proc. Camb. Phil. Soc., 70 (1970), pp. 243–255.

[6] R. Askey, *The q-gamma and q-beta functions*, Appl. Anal., 8 (1978), pp. 125–141.

[7] R. Askey and M. Ismail, *The Rogers' q-ultraspherical polynomials*, Approximation Theory, III, E. W. Cheney, ed., Academic Press, New York, 1980, pp. 175–182.

[8] ———, *A generalization of ultraspherical polynomials*, Studies in Pure Mathematics, R. Erdös, ed., Birkhäuser, 1983, pp. 55–78.

[9] R. Askey and J. Wilson, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., to appear.

[10] W. N. Bailey, *Generalized Hypergeometric Series*, Stechert-Hafner Service Agency, New York and London, 1964.

[11] A. Erdélyi et al., *Higher Transcendental Functions*, Vol. I, McGraw-Hill, New York, 1953.

[12] G. Gasper, *Positivity and special functions*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 375–433.

[13] ______, *Projection formulas for orthogonal polynomials of a discrete variable*, J. Math. Anal. Appl., 45 (1974), pp. 176–198.

[14] G. Gasper and Mizan Rahman, *A non-terminating extension of the Sears-Carlitz formula and some applications*, in preparation.

[15] L. J. Rogers, *Third memoir on the expansion of certain infinite products*, Proc. Lond. Math. Soc., 26 (1895), pp. 15–32.

[16] D. B. Sears, *Transformation of basic hypergeometric functions of special type*, Proc. Lond. Math. Soc., 52 (1951), pp. 467–483.

[17] ______, *On the transformation theory of basic hypergeometric functions*, Proc. Lond. Math. Soc., 53 (1951), pp. 158–180.

[18] G. Szegö, *Orthogonal Polynomials*, AMS Colloquium Publications 23, fourth edition, American Mathematical Society, Providence, RI, 1975.

[19] J. Thomae, *Beiträge zur Theorie der durch die Heinesche Reihe*: $1+((1-q^{\alpha})(1-q^{\beta})/(1-q)(1-q))x + \dots.$ *darstellbaren Functionen*, J. Reine Angew, Math., 70 (1869), pp. 258–291.

[20] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th Ed., Cambridge Univ. Press, Cambridge, 1952.

SIAM J. MATH. ANAL.
Vol. 16, No. 1, January 1985

ON CONVERGENCE AND DEGENERACY IN RATIONAL PADÉ AND CHEBYSHEV APPROXIMATION*

LLOYD N. TREFETHEN[†] AND MARTIN H. GUTKNECHT[‡]

Abstract. We study two questions associated with rational approximation of a function $f(z)$ near the origin $z=0$: continuity of the Padé approximation operator, and convergence of Chebyshev to Padé approximants as the domain of approximation shrinks to a point. Both become delicate in the case of degenerate approximations, i.e. approximations whose numerator and denominator are deficient in degree. In this situation various distinct definitions of convergence of sequences of rational functions make sense, and we give a unified treatment that explains their interrelationships. Our results show that the answers to the above questions are generally affirmative only in the nondegenerate case.

AMS-MOS subject classification (1980). Primary 41A20, Secondary 30E10, 41A21, 41A50

Introduction. This paper is concerned with two problems connected with approximation by rational functions:

(1) continuity of the Padé approximation operator;

(2) convergence of Chebyshev to Padé approximants as the domain of approximation shrinks to the origin.

The first question has been investigated previously in [4], [8], [14], [15], and the second in [2], [3], [6], [10], [11], [12]. Our purpose is to unify, correct, and extend some of the results of these papers.

Both problems turn upon questions of the convergence of sequences of functions within a fixed space R_{mn}, the set of rational functions having at most m zeros and at most n poles. Such convergence can be defined naturally in many different ways, and it is not obvious a priori which of these is most appropriate. Since each of the papers cited above considers only one or two of these definitions, the scope of the existing results, and the connections between them, have been unclear. We hope to improve this situation.

In particular we will investigate approximations involving a *degenerate* rational function $r\in R_{mn}$—that is, one with $\mu<m$ zeros and $\nu<n$ poles, hence with a *defect* $d=\min\{m-\mu, n-\nu\}$ that is positive. It is in the degenerate situation where the various definitions of convergence become distinct, and also where the answers to (1) and (2) are least obvious. The explanation for this is that in the degenerate case, r can be multiplied by one or more pole-zero pairs $(z-\zeta)/(z-\zeta')$ and the result will still belong to R_{mn}; if ζ and ζ' are nearly equal, the effect of such a perturbation will be large near these points but can be made arbitrarily small elsewhere. It is natural that this possibility should render convergence results somewhat complicated.

If $r\in R_{mn}$ has defect d, and $\tilde{r}\in R_{mn}$ is arbitrary, then $r-\tilde{r}$ belongs to $R_{m+n-d,2n-d}$ and so can have at most $m+n-d$ zeros. As a consequence the degree of agreement of r with a function $\tilde{r}$ is in some sense determined—in the absence of troublesome pole-zero pairs—by how closely they agree at any $m+n+1-d$ points. In both problems (1) and (2) above the origin is a distinguished point, and so we are led to the following notion of "H" convergence: "$r_\varepsilon \to_H r$" denotes convergence as $\varepsilon\to 0$ of the Taylor coefficients

* Received by the editors July 5, 1983.

[†] Courant Institute of Mathematical Sciences, New York University, New York, New York 10012. The research of this author was supported by a National Science Foundation Mathematical Sciences Postdoctoral Fellowship and by the U.S. Department of Energy under contract DE-AC02-76-ER03077-V.

[‡] Seminar für Angewandte Mathematik, ETH-Zurich, Switzerland.

of degrees 0 through $m+n-d$ of $r_\varepsilon \in R_{mn}$ to the corresponding coefficients of r. This is one of a sequence of convergence definitions we consider (precise statements in §1):

cw: coefficientwise,
au: almost uniform away from poles of r,
Δ: uniform on some disk Δ about the origin,
Tay: all Taylor coefficients,
H: Taylor coefficients of degree $\leq m+n-d$,
μ: measure.

In addition four other convergence definitions will be mentioned, mainly at the end of §1:

I: uniform on some interval I about the origin,
χ: chordal metric on all of $\mathbb{C}$,
χ_K: chordal metric on compact subsets of $\mathbb{C}$,
cap: capacity.

Definition I is of interest because it has been used in the papers of Werner and Wuytack [14], [15] and Chui et al. [2], [3]. Definition χ is stronger than all of the others, and becomes relevant for rational functions deficient in neither numerator nor denominator degree. Definition χ_K is equivalent to cw, and cap to μ.

Our main results can be abbreviated as follows, where cw is short for $r_\varepsilon \to_{\text{cw}} r$, and so on.

THEOREM 1. (a) *For arbitrary r one has*

$$\text{cw} \Rightarrow \text{au} \Rightarrow \Delta \Rightarrow \text{Tay} \Rightarrow H \Rightarrow \mu .$$

If r is nondegenerate, all six definitions are equivalent. (b) *If r is degenerate, they are all distinct.*

THEOREM 2. (a) *The Padé approximation operator is always H-continuous, regardless of degeneracy, hence also always μ-continuous.* (b) *It is continuous in other senses only when this follows from Theorem* 1, *i.e. only at a function f whose Padé approximant is nondegenerate.*

THEOREM 3. (a) *Chebyshev approximations on a small domain εK containing a neighborhood of the origin always converge in H as $\varepsilon \to 0$ to the Padé approximant r^P, regardless of degeneracy, hence also in μ.* (b) *If r^P is nondegenerate, K can be an arbitrary set with at least $m+n+1$ points* (*e.g.* $[-1, 1]$ *or* $[0, 1]$) *and they will still converge, in all senses.* (c) *If r^P is degenerate they do not in general converge in any sense stronger than H.*

Theorems 2-3 show that the solutions to problems (1) and (2) are closely related: desired properties typically hold in the relatively weak H sense, but hold in stronger senses only when this follows from general considerations involving sequences of rational functions.

In addition we discuss at the end a variant of the Chebyshev vs. Padé question: not whether Chebyshev approximants converge to Padé as the domain shrinks, but whether the magnitude of the error in Chebyshev approximation converges to that for Padé. One sees easily that in general it need not, even when r^P is nondegenerate.

Before beginning, it remains to make some specific remarks on how our results relate to those obtained previously.

(1) *Continuity of the Padé operator*. The basic theorem in this area is due to Werner and Wuytack [15]: in approximation of a real function f, the Padé operator is I-continuous at f if and only if $r^P(f)$ is nondegenerate. (The "if" half of this result was known earlier.) Our Theorem 2 shows that the same statement extends to continuity with respect to cw, au, Δ, and Tay, and that there is no need to restrict attention to real functions.

(2) *Convergence of Chebyshev to Padé.* In 1964 Walsh showed that $r_\varepsilon^* \to_{\mathrm{au}} r^p$ must hold as $\varepsilon \to 0$ for complex approximation on small disks $|z| \le \varepsilon$, if r^p is nondegenerate [11]. Our Theorem 3a is a generalization of this result. In 1974 he extended the convergence statement to real approximation on $[0,\varepsilon]$ [12], but the proof he gave is erroneous. Theorem 3b here gives a correct proof of this theorem, as well as generalizing it with regard to domain and definition of convergence. On the other hand in 1974 Chui, Shisha, and Smith claimed to show $r_\varepsilon^* \to_I r^p$ for real approximation on $[0,\varepsilon]$, regardless of degeneracy [2], [3]. However, our Theorem 3c shows that this conclusion is false. The upshot of our results is that it appears there is very little difference regarding the $r_\varepsilon^* \to r^p$ problem between real and complex approximation, or between approximation on $|z| \le \varepsilon$, $[-\varepsilon,\varepsilon]$, and $[0,\varepsilon]$. In all cases convergence in cw, au, Δ, or Tay is assured only if r^p is nondegenerate.

1. Convergence of sequences of rational functions. Let $\mathbb{C}$ denote the complex plane topologized by the absolute value metric $d(w,z)=|w-z|$. Let S denote the extended plane $\mathbb{C} \cup \{\infty\}$ topologized by the *chordal* or *spherical metric* χ defined by

$$\chi(w,z)=\frac{|w-z|}{\left(1+|w|^2\right)^{1/2}\left(1+|z|^2\right)^{1/2}}$$

for w, $z \in \mathbb{C}$, and by continuity for $w=\infty$ or $z=\infty$ [1], [7]. Under this definition S is a compact 2-manifold, and $\chi(w,z)$ can be interpreted as the Euclidean distance in $\mathbb{R}^3$ between the points w and z on the Riemann sphere of diameter 1. For any two functions f,g: $S \to S$, and any set $K \subseteq S$, define the uniform-norm distance between f and g on K (possibly infinite) by

$$\|f-g\|_K=\sup_{z\in K}|f(z)-g(z)|,$$

and the chordal-metric distance $\chi_K(f,g)$ on K (at most 1) by

$$\chi_K(f,g)=\sup_{z\in K}\chi(f(z),g(z)).$$

Let χ_S be abbreviated by χ.

Let $m,n \ge 0$ be fixed integers, and let R_{mn} be the space of complex rational functions r with at most m zeros in $\mathbb{C}$ (unless $r \equiv 0$) and at most n poles in $\mathbb{C}$, counted with multiplicity, and satisfying the additional condition $r(0) \ne \infty$. A function r: $S \to S$ belongs to R_{mn} if and only if it can be written as a fraction

$$(1.1) \qquad r(z)=\frac{p(z)}{q(z)}=\frac{a_0+\cdots+a_m z^m}{b_0+\cdots+b_n z^n}, \qquad b_0=1$$

for some coefficients a_k, $b_k \in \mathbb{C}$. We assume that all common factors have been removed from p and q, which makes this representation unique, and we refer to $\{a_k\}=\{a_k(r)\}$ and $\{b_k\}=\{b_k(r)\}$ as "the coefficients of r as a rational function." Let $\mu \le m$ and $\nu \le n$ denote the exact degrees of p and q, so that if $r \not\equiv 0$, then r has exactly μ zeros and ν poles in $\mathbb{C}$. If $r \equiv 0$, then $\mu=-\infty$ and $\nu=0$.

DEFINITION. The *defect* of r is the nonnegative integer

$$d=d(r)=\min\{m-\mu,n-\nu\}.$$

r is nondegenerate if $\mu=m$ or $\nu=n$ (i.e. $d=0$); otherwise it is *degenerate* (i.e. $d>0$).

Any $r \in R_{mn}$ has a Taylor series

$$r(z)=\sum_{k=0}^{\infty} c_k z^k \tag{1.2}$$

converging in a neighborhood of $z=0$. We refer to $\{c_k\}=\{c_k(r)\}$ as "the Taylor coefficients of r", and for convenience we define also $c_k=0$ for $k<0$. The coefficients $\{a_k\}$, $\{b_k\}$ and $\{c_k\}$ are related linearly. To obtain this relation, equate (1.1) and (1.2) and then multiply through by $q(z)$. The result is the following infinite system of equations:

$$\begin{bmatrix} c_{-n} & c_{-n+1} & \cdots & c_{-1} \\ c_{-n+1} & c_{-n+2} & \cdots & c_0 \\ \vdots & \vdots & \cdots & \vdots \\ c_{m-n} & c_{m-n+1} & \cdots & c_{m-1} \\ c_{m-n+1} & c_{m-n+2} & \cdots & c_m \\ c_{m-n+2} & c_{m-n+3} & \cdots & c_{m+1} \\ \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} b_n \\ b_{n-1} \\ \vdots \\ b_1 \end{bmatrix} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \\ 0 \\ 0 \\ \vdots \end{bmatrix} - \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_m \\ c_{m+1} \\ c_{m+2} \\ \vdots \end{bmatrix}. \tag{1.3}$$

Of particular interest is the $n \times n$ subsystem

$$\begin{bmatrix} c_{m-n+1} & \cdots & c_m \\ \vdots & & \vdots \\ c_m & \cdots & c_{m+n-1} \end{bmatrix} \begin{bmatrix} b_n \\ \vdots \\ b_1 \end{bmatrix} = - \begin{bmatrix} c_{m+1} \\ \vdots \\ c_{m+n} \end{bmatrix}. \tag{1.4}$$

Let H denote the matrix in (1.4). Since H has the form $h_{ij}=h_{i+j}$, it is a *Hankel matrix*. If H is nonsingular, i.e. $\det H \neq 0$, then the coefficients $\{b_k\}$ are uniquely determined by $(c_{m-n+1},\cdots,c_{m+n})$ as the solution to (1.4). Once these are known, the coefficients $\{a_k\}$ are uniquely determined by $(c_0,\cdots,c_m)$ from the first $m+1$ rows of (1.3). All together, $\{a_k\}$ and $\{b_k\}$ depend upon $c_0,\cdots,c_{m+n}$ in this case, but not on the remaining coefficients c_k.

The following result is well known [1], [5].

PROPOSITION. *r is degenerate if and only if H is singular.*

Proof. The solutions $\{a_k\}$, $\{b_k\}$ to (1.3), (1.4) are unique if and only if H is nonsingular. Such solutions correspond to all possible representations of r as a fraction of the form (1.1), including those that are not in lowest terms. On the other hand a lowest-terms quotient p/q is the unique representation for r if and only if it can be multiplied by no fraction $(z-a)/(z-a)$ and still remain a quotient of type (m,n), which is to say, if and only if r is nondegenerate. □

Thus when H is singular, the defect d is positive. It can be seen that in general $\{a_k\}$ and $\{b_k\}$ are determined by $c_0,\cdots,c_{m+n-d}$ (but not by $c_0,\cdots,c_{m+n-d-1}$).

Suppose $f(z)=\sum_{k=0}^{\infty} c_k z^k$, $c_k \in \mathbf{C}$, is a formal power series. The *Padé approximant* $r^p \in R_{mn}$ to f is defined to be that rational function in R_{mn} whose Taylor series agrees with f to as high an order as possible. It can be shown that if the matrix H formed from the coefficients $\{c_k\}$ is nonsingular, then the coefficients of r^p are the unique solution of (1.3) and (1.4), and $(f-r^p)(z)=O(z^{m+n+1})$. In general H may be singular, but r^p is always uniquely defined and satisfies $(f-r^p)(z)=O(z^{m+n+1-d})$. (Neither of these estimates need be sharp.)

Now let $\{r_\varepsilon\}_{\varepsilon>0}$ be a family of functions in R_{mn}, and let r belong to R_{mn} also. We will make use of the following six definitions of convergence of r_ε to r as $\varepsilon\to 0$.

cw ("coefficientwise") $r_\varepsilon \to_{\text{cw}} r$ if $\lim_{\varepsilon\to 0} a_k(r_\varepsilon)=a_k(r)$ for $0\le k\le m$ and $\lim_{\varepsilon\to 0} b_k(r_\varepsilon)=b_k(r)$ for $1\le k\le n$.

au ("almost uniform") $r_\varepsilon\to_{\text{au}} r$ if $\lim_{\varepsilon\to 0}\|r_\varepsilon - r\|_K=0$ for any compact $K\subseteq\mathbb{C}$ that contains no poles of r.

Δ ("wrt disk Δ") $r_\varepsilon\to_\Delta r$ if $\lim_{\varepsilon\to 0}\|r_\varepsilon - r\|_\Delta=0$ for some disk $\Delta=\{z\in\mathbb{C}:|z|\le\delta\}$, $\delta>0$.

Tay ("Taylor") $r_\varepsilon\to_{\text{Tay}} r$ if $\lim_{\varepsilon\to 0} c_k(r_\varepsilon)=c_k(r)$ for all $k\ge 0$.

H ("Hankel") $r_\varepsilon\to_H r$ if $\lim_{\varepsilon\to 0} c_k(r_\varepsilon)=c_k(r)$ for $0\le k\le m+n-d$.

μ ("measure") $r_\varepsilon\to_\mu r$ if for any $\delta>0$ and any compact $K\subseteq\mathbb{C}$, $\lim_{\varepsilon\to 0}\mu\{z\in K: |r_\varepsilon(z)-r(z)|>\delta\}=0$, where μ is the Lebesgue measure on $\mathbb{C}$.

The following theorem describes the relationships between these definitions of convergence. In the statement "cw" is an abbreviation for $r_\varepsilon\to_{\text{cw}} r$, and so on.

THEOREM 1. (a) *If r is nondegenerate, then*

$$\text{cw}\Leftrightarrow\text{au}\Leftrightarrow\Delta\Leftrightarrow\text{Tay}\Leftrightarrow H\Leftrightarrow\mu. \tag{1.5}$$

(b) *If r is degenerate, then*

$$\text{cw}\underset{\not\Leftarrow}{\Rightarrow}\text{au}\underset{\not\Leftarrow}{\Rightarrow}\Delta\underset{\not\Leftarrow}{\Rightarrow}\text{Tay}\underset{\not\Leftarrow}{\Rightarrow}H\underset{\not\Leftarrow}{\Rightarrow}\mu, \tag{1.6}$$

except that au $\Rightarrow$ cw *holds if r has no poles in* $\mathbb{C}$.

Proofs—arbitrary r. First we prove those implications asserted to hold regardless of whether r is nondegenerate, namely the five rightward implications in (1.5)–(1.6).

(a) cw $\Rightarrow$ au. If $r_\varepsilon\to_{\text{cw}} r$, then the denominator polynomials q_ε converge coefficientwise to q, which implies that the zeros of q_ε converge to zeros of q or to ∞. If $K\subseteq\mathbb{C}$ is compact and contains no poles of r, it follows that for all sufficiently small ε, the poles of r_ε are uniformly bounded away from K. Therefore for small enough ε, the values $r_\varepsilon(z)$ $(z\in K)$ depend continuously on the coefficients of r_ε, hence on ε, in a manner uniform in z for $z\in K$. This implies $\lim_{\varepsilon\to 0}\|r_\varepsilon - r\|_K=0$.

(b) au $\Rightarrow\Delta$. Trivial.

(c) $\Delta\Rightarrow$Tay. If $r_\varepsilon\to_\Delta r$, there is a disk Δ on which r is analytic with $\lim_{\varepsilon\to 0}\|r_\varepsilon - r\|_\Delta =0$. If $\|r_\varepsilon - r\|_\Delta<\infty$, then r_ε is analytic on Δ too. Therefore the Taylor coefficients for both r and r_ε can be computed by Cauchy integrals around $|z|=\delta$, and the uniform convergence on that circle implies that these integrals converge.

(d) Tay $\Rightarrow H$. Trivial.

(e) $H\Rightarrow\mu$. If $r_\varepsilon\to_H r$, then $c_k(\Delta r_\varepsilon)\to 0$ for $0\le k\le m+n-d$, where $\Delta r_\varepsilon = r_\varepsilon - r\in R_{m+n-d,2n-d}$. Setting $M=m+n-d$ and $N=2n-d$, we see that it is enough to show that if $r_\varepsilon\in R_{MN}$ satisfies $c_k(r_\varepsilon)\to 0$ for $0\le k\le M$, then $r_\varepsilon\to_\mu 0$.

For each ε, let $r_\varepsilon(z)$ be written as a quotient $p_\varepsilon(z)/q_\varepsilon(z)$ with the normalization $\|q_\varepsilon\|_\Delta=1$, where Δ is the unit disk. (This is a different normalization from that of (1.1). Further specification regarding common factors and a constant of modulus 1 is unnecessary.) The condition $\|q_\varepsilon\|_\Delta=1$ implies $|b_k|\le 1$ for each coefficient of q_ε, and since $p_\varepsilon(z)=q_\varepsilon(z)\Sigma_{k=0}^\infty c_k z^k$, the conclusion $\lim_{\varepsilon\to 0}\|p_\varepsilon\|_\Delta=0$ then follows by the $c_k\to 0$ hypothesis.

Now let $K\subseteq\mathbb{C}$ be compact, and let $\delta>0$ be arbitrary. Clearly $\|p_\varepsilon\|_K\to 0$ also as $\varepsilon\to 0$. On the other hand we have $\{z\in K:|r_\varepsilon(z)|>\delta\}\subseteq\{z\in K:|q_\varepsilon(z)|<\|p_\varepsilon\|_K/\delta\}$, and it is readily seen that the latter set has measure bounded by const$(\|p_\varepsilon\|_K/\delta)^{2/N}$. Therefore the measure of this set goes to 0 with ε, which is just what is required to establish $r_\varepsilon\to_\mu 0$ (see [1, vol. 1, §6.6]).

Proofs—degenerate r. Next we prove those relationships asserted to hold if and only if r is degenerate, namely the leftward nonimplications in (1.6). If $r \in R_{mn}$ is degenerate, then $\mu < m$ and $\nu < n$ hold. Therefore for any $a, z_0 \in \mathbb{C}$ with $z_0 \neq 0$, the function

$$r_\varepsilon(z) = r(z)\left(1 + \frac{a}{1 - z/z_0}\right)$$

belongs to R_{mn} too. By choosing a and z_0 judiciously, we construct a sequence of counterexamples that establish the required results. Detailed verifications are left to the reader. In the case $r(z) \equiv 0$, each construction should be modified by setting simply $r_\varepsilon(z) = a/(1 - z/z_0)$.

(f) au $\nRightarrow$ cw. Assuming r has a finite pole at z_0, take $r_\varepsilon(z) = r(z)(1 + \varepsilon/(z - z_0))$.

(g) $\Delta \nRightarrow$ au. Take $r_\varepsilon(z)$ as in (f), but with z_0 equal to any nonzero complex number that is not a pole of r.

(h) Tay $\nRightarrow \Delta$. Take $r_\varepsilon(z) = r(z)(1 + \varepsilon^{-1/\varepsilon}/(1 - z/\varepsilon))$.

(i) $H \nRightarrow$ Tay. Take $r_\varepsilon(z) = r(z)(1 + \varepsilon^{m+n+1-d}/(1 - z/\varepsilon))$.

(j) $\mu \nRightarrow H$. Take $r_\varepsilon(z) = r(z)(1 + 1/(1 - z/\varepsilon))$.

Proof—nondegenerate r. Finally, assume that $r \in R_{mn}$ is nondegenerate. To complete the proof of Theorem 1, it is enough to show $\mu \Rightarrow$cw:

(k) $\mu \Rightarrow$cw. If r is nondegenerate, assume it has n finite poles $z_1, \cdots, z_n$; the case of m zeros is analogous. It is clear that if $r_\varepsilon \to_\mu r$, then for all sufficiently small ε, r_ε must have n poles $z_k^{(\varepsilon)}$ satisfying $z_k^{(\varepsilon)} \to z_k$ as $\varepsilon \to 0$. This implies $q_\varepsilon \to_{\text{cw}} q$, hence $q_\varepsilon \to_\mu q$. From this and $r_\varepsilon \to_\mu r$, one can conclude $p_\varepsilon \to_\mu p$, hence $p_\varepsilon \to_{\text{cw}} p$, hence $r_\varepsilon \to_{\text{cw}} r$. □

We now make some remarks on the additional notions of convergence mentioned in the Introduction. They are defined as follows:

I ("wrt interval I") $r_\varepsilon \to_I r$ if $\lim_{\varepsilon \to 0} \|r_\varepsilon - r\|_I = 0$ for some interval $I = [-\delta, \delta]$, $\delta > 0$.

χ ("chordal") $r_\varepsilon \to_\chi r$ if $\lim_{\varepsilon \to 0} \chi(r_\varepsilon, r) = 0$.

χ_K ("almost chordal") $r_\varepsilon \to_{\chi_K} r$ if $\lim_{\varepsilon \to 0} \chi_K(r_\varepsilon, r) = 0$ for any compact $K \subseteq \mathbb{C}$.

cap ("capacity") $r_\varepsilon \to_{\text{cap}} r$ if for any $\delta > 0$ and any compact $K \subseteq \mathbb{C}$, $\lim_{\varepsilon \to 0} \text{cap}\{z \in K : |r_\varepsilon(z) - r(z)| > \delta\} = 0$, where cap is the logarithmic capacity [7].

We state without proof some basic facts relating these definitions to the others.

THEOREM 1c.

(i) *If r is nondegenerate, then $\Delta \Leftrightarrow I$. Otherwise $\Delta \Rightarrow I$ but $I \nRightarrow \Delta$.*

(ii) *If both $\mu = m$ and $\nu = n$ hold, then $\chi \Leftrightarrow$*cw. *Otherwise $\chi \Rightarrow$*cw *but* cw $\nRightarrow \chi$.

(iii) $\chi_K \Leftrightarrow$cw.

(iv) cap $\Leftrightarrow \mu$.

Result (iv) is, of course, quite different from the more familiar situation cap $\Rightarrow \mu$ $\nRightarrow$ cap that holds for approximation by arbitrary rational functions rather than rational functions of fixed type (m, n) [1].

2. Continuity of the Padé approximation operator. Let $m, n \geq 0$ be fixed and let $f(z) = \sum c_k z^k$ be a formal power series. Then f has a unique Padé approximant $r^P \in R_{mn}$, and we let P denote the operator

$$P: f \mapsto r^P.$$

In fact r^P depends only on the coefficients $c_0, \cdots, c_{m+n}$, and if the defect is $d > 0$, it depends only on $c_0, \cdots, c_{m+n-d}$. (To be precise, $\tilde{f} - f = O(z^{m+n+1-d})$ implies $P(\tilde{f}) = P(f)$, but $\tilde{f} - f = O(z^{m+n-d})$ does not.) Therefore the most reasonable way to define convergence of f_ε to f in the Padé approximation context is:

$$f_\varepsilon \to f \text{ if } \lim_{\varepsilon \to 0} c_k(f_\varepsilon) = c_k(f) \quad \text{for } 0 \leq k \leq m+n.$$

Since only finitely many terms of f have any influence, we can be careless as to whether f is a full power series or just a set of numbers $c_0,\cdots,c_{m+n}$. On the other hand for defining convergence of $P(f_\varepsilon)$ to $P(f)$ all of the choices discussed in §1 are reasonable candidates. To each definition of convergence corresponds a different definition of continuity of the Padé approximation operator. We say that P is *H-continuous at f* if $f_\varepsilon \to f$ implies $P(f_\varepsilon) \to_H P(f)$, and so on.

If $r^p = P(f)$ is nondegenerate, then for most senses of continuity it is an easy matter of linear algebra to show directly that P is continuous at f. Essentially the required argument is given in [8, Thm. 3.17] (for rational interpolation), [4, Thm. 8] (for Newton–Padé approximation), in §II of [14] (under the stricter assumption that r^p is normal), and probably elsewhere too. An explicit statement that P is continuous in the nondegenerate case appears perhaps first as [15, Thm. 4.1], where I-continuity is established. The case where r^p is normal was treated earlier in [14]. Our approach here is to show that P is H-continuous regardless of degeneracy, from which continuity in other sense follows as a corollary of Theorem 1a, if r^p is nondegenerate.

THEOREM 2a. *Let f be arbitrary. The Padé approximation operator P is H-continuous at f.*

COROLLARY (*by Theorem* 1). *If r^p is nondegenerate, then P is also* cw-, au-, Δ-, *and* Tay-*continuous at f. Whether or not r^p is nondegenerate, P is μ-continuous at f.*

Werner and Wuytack have established μ-continuity previously in [15, Thm. 6].

Proof. In fact one has local Lipschitz H-continuity with a constant of exactly 1. For we have already mentioned that $r^p - f = O(z^{m+n+1-d})$, and we claim that the analogous identity holds for sufficiently nearby perturbations $\tilde f$ of f. To see this, observe that for either $(\mu,\nu)=(m-d,n)$ or $(\mu,\nu)=(m,n-d)$ (with the obvious modification if $r^p \equiv 0$), r^p is also the Padé approximant to f in $R_{\mu\nu}$, and is nondegenerate with respect to that class. By the Proposition of §1 the same nondegeneracy holds for nearby $\tilde f$, since small perturbations of a nonsingular matrix H are nonsingular. Therefore for all $\tilde f$ sufficiently near to f one has $P_{\mu\nu}(\tilde f) - f = O(z^{\mu+\nu+1}) = O(z^{m+n+1-d})$, hence a fortiori $P_{mn}(\tilde f) - \tilde f = O(z^{m+n+1-d})$, as claimed. □

The main result of [15] is the following converse to Theorem 2a: if r^p is degenerate, then P is I-discontinuous at f. The proof involves multiplications of r^p by cleverly chosen pole-zero pairs. Our proof below generalizes this result to Tay-discontinuity, hence also discontinuity in cw, au, and Δ. Also, in [15] Werner and Wuytack present their argument only for the case in which r^p lies in a 2×2 square block in the Padé table, and they suggest that the proof for the general case will require the introduction of several pole-zero pairs rather than one. However the following proof, which has no block size restriction, shows that one is enough.

THEOREM 2b. *If r^p is degenerate, then P is* Tay-*discontinuous at f.*

COROLLARY (*by Theorem* 1b). *If r^p is degenerate, then P is also* cw-, au-, *and* Δ-*discontinuous at f.*

Proof. Let f,m,n be given, let f have the form $f(z)=c_l z^l + c_{l+1}z^{l+1} + \cdots$ with $c_l \neq 0$, and let $P(f)=r^p \in R_{mn}$ have defect $d>0$. Then

$$f(z) = r^p(z) + az^{m+n+1-d} + O(z^{m+n+2-d}) \tag{2.1}$$

for some $a \in \mathbf{C}$, possibly zero. To begin with, assume $r^p \not\equiv 0$, which implies $l \le m+n-d$.

If $a \neq 0$, then for each $\varepsilon > 0$, define

$$r_\varepsilon(z) = r^p(z)\left(1 + \frac{a\varepsilon^{m+n+1-d-l}/c_l}{1-z/\varepsilon}\right). \tag{2.2}$$

Since r^p is degenerate, r_ε belongs to R_{mn} for each ε and has defect at least $d-1$. This implies that for $r_\varepsilon = P(f_\varepsilon)$ to hold for some f_ε, it is enough that f_ε satisfy

$$(2.3) \qquad f_\varepsilon(z) = r_\varepsilon(z) + O(z^{m+n+2-d}).$$

To achieve this, let f_ε be that function which has the Taylor coefficients of f for degree $\geq m+n+2-d$ and those of r_ε for degree $\leq m+n+1-d$. Now from (2.1) and (2.2) it follows that the coefficients of f and r_ε agree up to $O(\varepsilon)$ for degrees $\leq m+n+1-d$. Therefore $f_\varepsilon \to f$ as $\varepsilon \to 0$, while from (2.2), $r_\varepsilon \not\to_{\mathrm{Tay}} r^p$. This establishes discontinuity as claimed.

If $a=0$, replace $a\varepsilon^{m+n+1-d-l}$ by $\varepsilon^{m+n+2-d-l}$ in (2.2).

It remains to treat the case $r^p \equiv 0$, which will occur whenever $f(z) = O(z^{m+1})$. If $m \geq n-1$, we set

$$(2.4) \qquad r_\varepsilon(z) = \frac{a\varepsilon^{m+n+1-d}}{1-z/\varepsilon},$$

or $r_\varepsilon(z) = \varepsilon^{m+n+2-d}/(1-z/\varepsilon)$ if $a=0$, and then the proof is again valid. Therefore assume $m \leq n-2$. In this event r^p has defect $d=n$, while (2.4) has defect $m \leq n-2$, and so that proof breaks down at (2.3). If $f(z) = O(z^{n+1})$, let f_ε have the Taylor coefficients of f for $k \geq n+1$ and those of $\varepsilon^{n+1}/(1-z/\varepsilon)$ for $k \leq n$. Then $r_\varepsilon = P(f_\varepsilon) = \varepsilon^{n+1}/(1-z/\varepsilon) \not\to_{\mathrm{Tay}} 0$ as $\varepsilon \to 0$, but $f_\varepsilon \to f$, and discontinuity is established.

On the other hand if $f(z) = az^K + O(z^{K+1})$ with $a \neq 0$ and $m+1 \leq K \leq n$, set $f_\varepsilon(z) = \varepsilon + f(z)$ and $r_\varepsilon = P(f_\varepsilon)$. Then for any $\varepsilon > 0$, r_ε will have Kth coefficient $a \neq 0$, while r^p has Kth coefficient 0. Thus again one has $r_\varepsilon \not\to_{\mathrm{Tay}} r^p$. □

In summary, the Werner–Wuytack result that P is continuous at f if and only if $r^p(f)$ is nondegenerate holds not only for I-continuity, which seems after all a somewhat unnatural definition of continuity for a problem with no intrinsic restriction to the real axis, but also for continuity with respect to definitions cw, au, Δ, and Tay.

3. Best approximation on small domains. Suppose $K \subseteq \mathbf{C}$ is a compact set, f is a fixed function, and for each $\varepsilon > 0$, $r^*_{\varepsilon K}$ is a best approximation to f in R_{mn} on εK. In 1934 Walsh posed the question [10]: as $\varepsilon \to 0$, must $r^*_{\varepsilon K}$ approach r^p? We are especially interested in three choices of K:

$$\Delta = \{z : |z| \leq 1\}, \quad I = [-1,1], \quad J = [0,1].$$

In his original paper Walsh settled the question in the affirmative for polynomial approximation on these regions, showing [10, pp. 175–176]

$$(3.1) \qquad r^*_{\varepsilon\Delta}, r^*_{\varepsilon I}, r^*_{\varepsilon J} \to_{\mathrm{au}} r^p \quad \text{if } n=0$$

provided f is analytic (case Δ) or sufficiently differentiable (cases I, J) at the origin.

To obtain analogous theorems for rational approximation, it is convenient to make use of the linear system (1.3), and therefore natural to assume that r^p is nondegenerate. In 1964 Walsh extended (3.1) to

$$(3.2) \qquad r^*_{\varepsilon\Delta} \to_{\mathrm{au}} r^p \quad \text{if } r^p \text{ is nondegenerate}$$

for f analytic in a neighborhood of the origin [11]. In Theorem 3a below we generalize this result as follows: if K is any region containing a ball around the origin, then $r^*_{\varepsilon K} \to_H r^p$ as $\varepsilon \to 0$, regardless of degeneracy. Thus H- and μ-convergence always occur, by Theorem 1a, and if r^p is nondegenerate, one also has convergence with respect to cw, au, Δ, and Tay.

A decade later Walsh published the analogous result for the half-interval:

$$(3.3)\qquad r^*_{\varepsilon J}\to_{\mathrm{au}} r^p \quad \text{if } r^p \text{ is nondegenerate}$$

for $f\in C^{m+n+1}[0,\delta]$, some $\delta>0$ [12]. This theorem is correct, but Walsh's proof has an error in it: his equation (13) does not follow from his equation (12), and it appears that no simple modification can get around this problem. In Theorem 3b below we give an alternate proof that avoids this error, and in the process generalize the domain: we show that if K is any bounded set with at least $m+n+1$ points, then $r^*_{\varepsilon K}\to_{\mathrm{cw}} r^p$ if r^p is nondegenerate. In particular K can be disconnected or discrete, and it need not contain the origin.

Walsh did not speculate as to whether the nondegeneracy condition is necessary for (3.2) and (3.3) to hold. This question was taken up by Chui, Shisha, and Smith, also in 1974. For the problem of approximation of a real function $f\in C^{m+n+1}[0,\delta]$ by rational functions with real coefficients, they claimed [2], [3]

$$(3.4)\qquad r^*_{\varepsilon J}\to_I r^p \quad \text{regardless of degeneracy.}$$

However, *this assertion is false*. We will demonstrate this in Theorem 3c by exhibiting a counterexample that is a modification of some related examples derived in [6]. The error in the proof of [2] comes in the last sentence of the paper, where the authors appeal to the fact cw$\Rightarrow I$ (in the notation of our Theorem 1), without having imposed the normalization $b_0=1$ (eq. (1.1)) in the definition of cw that is needed for this implication to hold.

In general it appears that if r^p is degenerate, then nothing can be said about convergence in senses stronger than H, regardless of what domain K is under consideration, and in fact Theorem 3c will give examples with $r^*_{\varepsilon K}\not\to_{\mathrm{Tay}} r^p$ for $K=\Delta$, I, and J, for both real and complex approximation.

THEOREM 3a. *Let f be analytic in a neighborhood of the origin and have the (m,n) Padé approximant r^p with defect d. Let $K\subset\mathbb{C}$ be a bounded set that contains a disk about the origin, and for each ε, let $r^*_{\varepsilon K}$ be a best approximation in R_{mn} to f on εK. Then*

$$(3.5)\qquad r^*_{\varepsilon K}\to_H r^p \quad \text{as } \varepsilon\to 0.$$

COROLLARY (*by Theorem* 1a). *Under the same hypotheses one has $r^*_{\varepsilon K}\to_\mu r^p$, and if in addition $d=0$, one has convergence also with respect to* cw, au, Δ, *and* Tay.

Remark. The assumptions that f is analytic and that $r^*_{\varepsilon K}$ is the best approximation are unnecessarily strict. All that is needed for the proof is $\|r^*_{\varepsilon K}-r^p\|_{\varepsilon K}=o(\varepsilon^{m+n-d})$.

Proof. By definition r^p is analytic at the origin and its Taylor coefficients agree with those of f through degree $m+n-d$. Since K is bounded, this implies

$$\|f-r^p\|_{\varepsilon K}=o(\varepsilon^{m+n-d})$$

and therefore also

$$\|f-r^*_{\varepsilon K}\|_{\varepsilon K}=o(\varepsilon^{m+n-d}).$$

Subtracting these estimates yields

$$(3.6)\qquad \|r^p-r^*_{\varepsilon K}\|_{\varepsilon K}=o(\varepsilon^{m+n-d}).$$

Now without loss of generality assume K contains the disk Δ. Then (3.6) will hold in particular on the boundary of $\varepsilon\Delta$, and by a Cauchy integral this implies the estimate

$$c_k(r^p-r^*_{\varepsilon K})=o(\varepsilon^{m+n-d-k})$$

for the kth Taylor coefficient of $r^p - r^*_{\varepsilon K}$. For $k \le m+n-d$ one therefore has $c_k(r^*_{\varepsilon K}) \to c_k(r^p)$ as $\varepsilon \to 0$, and this is the definition of $r^*_{\varepsilon K} \to_H r^p$. □

THEOREM 3b. *Let f be analytic in a neighborhood of the origin and have the nondegenerate (m,n) Padé approximant r^p. Let $K \subseteq \mathbb{C}$ be an arbitrary bounded set containing at least $m+n+1$ points, which need not include the origin. Then*

$$r^*_{\varepsilon K} \to_{\text{cw}} r^p \quad \text{as } \varepsilon \to 0. \tag{3.7}$$

COROLLARY (*by Theorem* 1a). *Under the same hypotheses one also has convergence with respect to* au, Δ, Tay, H, *and* μ.

Remark. The remark following Theorem 3a applies again here, and now it is more important. To guarantee $\|r^*_{\varepsilon K} - r^p\| = o(\varepsilon^{m+n})$, it will be enough for f to have $m+n+1$ derivatives at the origin with respect to the set $\bigcup \varepsilon K$, which may consist of a union of rays through the origin (such as J or I) rather than a complex neighborhood. Also, if $f(\bar{z}) = \overline{f(z)}$, then the conclusion holds for real best approximations $r^*_{\varepsilon K}$ as well as complex ones.

Proof. Let r^p and $r^*_{\varepsilon K}$ be represented as

$$r^p(z) = \frac{P(z)}{Q(z)}, \qquad r^*_{\varepsilon K}(z) = \frac{p_\varepsilon(z)}{q_\varepsilon(z)}$$

normalized by $Q(0) = 1$ and $\|q_\varepsilon\|_{\varepsilon K} = 1$. As in the previous proof, one obtains the estimate (3.6) with $d = 0$,

$$\left\| \frac{p_\varepsilon}{q_\varepsilon} - \frac{P}{Q} \right\|_{\varepsilon K} = o(\varepsilon^{m+n}).$$

By the normalization of Q and q_ε we can multiply through to obtain

$$\|p_\varepsilon Q - P q_\varepsilon\|_{\varepsilon K} = o(\varepsilon^{m+n}).$$

Since the function inside the norm is a polynomial of degree at most $m+n$, and since K contains at least $m+n+1$ points, this estimate can only hold if in fact

$$\|p_\varepsilon Q - P q_\varepsilon\|_{\varepsilon \Delta} = o(\varepsilon^{m+n}).$$

By a Cauchy integral over $|z| = \varepsilon$, this implies that the polynomials $p_\varepsilon Q$ and $P q_\varepsilon$ have approximately equal coefficients,

$$c_k(p_\varepsilon Q) = c_k(P q_\varepsilon) + o(1), \qquad k = 0, \cdots, m+n. \tag{3.8}$$

Now if $n = 0$, then $Q \equiv q_\varepsilon \equiv 1$, and (3.8) is the conclusion (3.7) we are looking for. Therefore assume $n > 0$. In this event the nondegeneracy assumption implies $P \not\equiv 0$, and since P is independent of ε and $\|q_\varepsilon\|_{\varepsilon K} = 1$, it follows that for each sufficiently small ε, $P q_\varepsilon$ has a coefficient bounded below in modulus by a fixed constant. Together with (3.8) this implies that the sets of zeros of $p_\varepsilon Q$ and $P q_\varepsilon$ must converge to each other as $\varepsilon \to 0$ in the following sense: if $z_1, \cdots, z_{m+n}$ and $\zeta_1, \cdots, \zeta_{m+n}$ are the zeros of these polynomials, counted with multiplicity, and padded with numbers $z_k = \infty$ or $\zeta_k = \infty$ when the degree is less than $m+n$, then for some ordering of the subscripts one has $\chi(z_k, \zeta_k) \to 0$ for each k as $\varepsilon \to 0$. Now the zeros of P and Q are independent of ε, and if r^p is nondegenerate, either there are m of the former or there are n of the latter, or both. Suppose P has m zeros; the other case is analogous. Then the convergence of the zero sets of $P q_\varepsilon$ and $p_\varepsilon Q$ implies that for all sufficiently small $\varepsilon, p_\varepsilon$ has degree exactly m, with its zeros converging as $\varepsilon\to 0$ to those of P. It follows then that the zeros of q_ε also converge to those of Q, with possibly some additional zeros converging to ∞. From here (3.7) is a ready consequence. □

Remark. Comparing Theorem 3b to Theorem 3a, one sees that in permitting a general region K, we have lost the ability to conclude $r^*_{\varepsilon K}\to_H r^p$ in the absence of nondegeneracy. An example shows that this cannot be helped: take $f(x)=x$, $(m,n)=(0,1)$, $K=\{1,-1\}$. For each ε, the best approximation is then

$$r^*_{\varepsilon K}(z)=\frac{\varepsilon^2}{z},$$

which has a pole at the origin, so $r^*_{\varepsilon K}\to_H r^p$ certainly does not hold. On the other hand it is still conceivable that some conditions on f and K weaker than the assumption $\Delta\subseteq K$ would be enough to ensure $r^*_{\varepsilon K}\to_H r^p$.

Having established $r^*_{\varepsilon K}\to r^p$ under appropriate hypotheses, we come now to the task of giving examples to show that if r^p is degenerate, then convergence in stronger senses than H will not in general take place. (Of course, degeneracy will not always cause nonconvergence; for example, the best real approximation in R_{0n} to $f(x)=x$ on εI is 0 for all ε, which converges to the degenerate Padé approximant $r^p\equiv 0$ in every sense.)

THEOREM 3c. *There exist examples of integers m,n and entire functions f with the following properties*:

(i) $r^*_{\varepsilon\Delta}\not\to_{\mathrm{Tay}} r^p$,
(ii) $r^*_{\varepsilon I}\not\to_{\mathrm{Tay}} r^p$,
(iii) $r^*_{\varepsilon I}\not\to_{\mathrm{Tay}} r^p$.

Analogous examples also exist if each problem is restricted to approximation of real functions by real rational functions, in which case one also has $r^*_{\varepsilon K}\not\to_I r^p$. (By a "real" function on Δ, we mean a function f with $f(\bar z)=\overline{f(z)}$.)

COROLLARY (*by Theorem* 1b). *The same nonconvergence results hold with respect to* cw, au, *and* Δ.

Proof. There are six statements to prove, which we label Δ-$\mathbb{C}$, I-$\mathbb{C}$, J-$\mathbb{C}$, Δ-$\mathbb{R}$, I-$\mathbb{R}$, J-$\mathbb{R}$. Probably examples exist in each category for arbitrary $m\geq 0$, $n\geq 1$, but we will not worry about achieving this generality.

Δ-$\mathbb{C}$. Take $(m,n)=(0,1)$ and $f(z)=z^2-z^5$, hence $r^p\equiv 0$. In the proof of Theorem 4 in [6] it was shown that for all $\varepsilon, r^*_{\varepsilon\Delta}$ has a pole in the region $|z|\leq\rho\varepsilon$, for some fixed constant ρ. It follows also from the arguments there that one has $\|f-0\|_{\varepsilon\Delta}-\|f-r^*_{\varepsilon\Delta}\|_{\varepsilon\Delta}\geq\text{const}\,\varepsilon^5$ as $\varepsilon\to 0$. This implies $\|r^*_{\varepsilon\Delta}(\varepsilon\omega)\|\geq\text{const}\,\varepsilon^5$ for each of the three roots of $\omega^3=-1$, and therefore $r^*_{\varepsilon\Delta}$ must have the form

$$r^*_{\varepsilon\Delta}(z)=\frac{a(\varepsilon)\varepsilon^6}{z-\varepsilon b(\varepsilon)} \tag{3.9}$$

with $a(\varepsilon)$ bounded below and $b(\varepsilon)$ bounded both above and below by constants. It follows that the Taylor coefficients $c_k(r^*_{\varepsilon\Delta})$ for $k\geq 6$ diverge to ∞ as $\varepsilon\to 0$, so in particular $r^*_{\varepsilon\Delta}\not\to_{\mathrm{Tay}} r^p$

Δ-$\mathbb{R}$. The example and argument above are suitable here too. Since $b(\varepsilon)$ is real, obviously $r^*_{\varepsilon\Delta}\not\to_I r^p$ also.

I-$\mathbb{C}$. Consider $f(x)=x$, $(m,n)=(0,1)$. It is shown in [6] that $r^*_{\varepsilon I}$ is not a constant, but has a pole somewhere in $\mathbb{C}$. A scale invariance argument shows that $r^*_{\varepsilon I}$ must therefore have the form

$$r^*_{\varepsilon I}(z)=\frac{a\varepsilon^2}{z-\varepsilon b} \tag{3.10}$$

for some constants a and b, independent of ε. Obviously $r^*_{\varepsilon I}\not\rightarrow_{\text{Tay}} r^p\equiv 0$.

I-$\mathbb{R}$. Consider $f(x)=x^2$, $(m,n)=(0,2)$. By the equioscillation theorem, $r^*_{\varepsilon I}$ cannot be a constant, for a constant yields at best three equioscillation points. So it must have a pair of real poles, symmetric with respect to the origin, and a scaling argument gives

$$r^*_{\varepsilon I}(x)=\frac{a\varepsilon^4}{x^2-\varepsilon^2 b^2},$$

which again implies $r^*_{\varepsilon I}\not\rightarrow_{\text{Tay}} r^p$, and also $r^*_{\varepsilon I}\not\rightarrow_I r^p$.

J-$\mathbb{C}$ *and* J-$\mathbb{R}$. For both of these situations the example and argument of case I-$\mathbb{C}$ apply. The equioscillation theorem shows that the best approximation to $f(x)=x$ on εJ is not a constant, either in real or complex approximation; therefore in each case it must have the form (3.10), which implies $r^*_{\varepsilon J}\not\rightarrow_{\text{Tay}} r^p$. For real approximation one also has $r^*_{\varepsilon J}\not\rightarrow_I r^p$, and in fact in this case the coefficients of the solution have been calculated explicitly by Maehly and Witzgall [8]: they are $a=-\frac{1}{4}$, $b=(1+\sqrt{2})/2$. □

All of the above examples have $r^p\equiv 0$, and it may seem that this might make them exceptional. However, examples with $r^p\not\equiv 0$ can also be invented. For example, consider type (2,1) approximation to $f(z)=z+z^7+z^{15}$ on $\varepsilon\Delta$. Now $r^p(z)=z$, but the arguments of [6] show that $r^*_{\varepsilon\Delta}\not\rightarrow_{\text{Tay}} r^p$ holds regardless.

* * *

Throughout this section we have investigated whether $r^*_{\varepsilon K}$ and r^p approach each other as $\varepsilon\to 0$. However, for some purposes it may be more interesting to know whether $\|f-r^*_{\varepsilon K}\|_{\varepsilon K}$ and $\|f-r^p\|_{\varepsilon K}$ approach each other. Let

$$\sigma_{\varepsilon K}=\frac{\|f-r^p\|_{\varepsilon K}}{\|f-r^*_{\varepsilon K}\|_{\varepsilon K}}\geq 1$$

be a measure of the agreement between these two. For Padé approximation to be asymptotically best, one should have $\sigma_{\varepsilon K}\to 1$ as $\varepsilon\to 0$. The following examples reveal that in general this need not occur, even if r^p is nondegenerate, but that on the other hand it may occur even if $r^*_{\varepsilon K}\not\rightarrow_{\text{Tay}} r^p$.

First, suppose K is the disk Δ and f is analytic at 0, as in the proofs Δ-$\mathbb{C}$ and Δ-$\mathbb{R}$ above. Then whether or not r^p is degenerate, for small enough ε the function $f-r^p$ maps $|z|=\varepsilon$ onto a curve of winding number at least $m+n+1-d$ whose modulus is constant up to a factor $1+O(\varepsilon)$. By Rouche's theorem one concludes $\sigma_{\varepsilon K}=1+O(\varepsilon)$ [9]. Thus $\sigma_{\varepsilon K}\to 1$ does not imply $r^*_{\varepsilon K}\to_{\text{Tay}} r^p$.

Second, let K be any bounded region that contains a disk about the origin. An extension of the above argument shows $\sigma_{\varepsilon K}\leq$const as $\varepsilon\to 0$, but it is easy to devise situations in which $\sigma_{\varepsilon K}$ is bounded away from 1, even when r^p is nondegenerate. (For example: let K be the eccentric disk $|z-\frac{1}{2}|\leq 1$, and take $f(z)=z$, $(m,n)=(0,0)$. Then $\|f-r^p\|_{\varepsilon K}=3\varepsilon/2$, but $\|f-r^*_{\varepsilon K}\|_{\varepsilon K}=\varepsilon$, so $\sigma_{\varepsilon K}=\frac{3}{2}$ for all ε.) Thus $r^*_{\varepsilon K}\to_{\text{Tay}} r^p$ does not imply $\sigma_{\varepsilon K}\to 1$.

Third, consider any of the examples in the proofs I-$\mathbb{C}$, I-$\mathbb{R}$, J-$\mathbb{C}$, J-$\mathbb{R}$ above. Here one has $\sigma_{\varepsilon K}\equiv$const$>1$ as $\varepsilon\to 0$, independent of ε, and so neither $\sigma_{\varepsilon K}\to 1$ nor $r^*_{\varepsilon K}\to_{\text{Tay}} r^p$ holds.

Finally, take $K=[\frac{1}{2},1]$ and let f be a C^{∞} function on $[0,\infty)$ whose Taylor series at the origin is that of some fixed $r_{\infty}\in R_{mn}$, degenerate or nondegenerate, but which equals some slightly different $r_k\in R_{mn}$ on each interval $4^{-k}K$, $k\geq 0$. Then $\|f-r^*_{\varepsilon K}\|_{\varepsilon K}=0$ but $\|f-r^p\|_{\varepsilon K}\neq 0$ for each $\varepsilon=4^{-k}$, so in this case the ratios $\sigma_{\varepsilon K}$ are not even bounded as $\varepsilon\to 0$.

Acknowledgments. We are grateful to Edward Saff for several suggestions that contributed significantly to the development of this paper, and to Lothar Reichel for calling our attention to the question discussed at the end of §3.

REFERENCES

[1] G. Baker and P. Graves-Morris, *Padé Approximants* (2 vols.), Encyclopedia of Mathematics, vols. 13 and 14, Addison-Wesley, Reading, MA, 1981.

[2] C. K. Chui, O. Shisha and P. W. Smith, *Padé approximants as limits of best rational approximants*, J. Approx. Theory, 12 (1974), pp. 201–204.

[3] C. K. Chui, *Recent results on Padé approximants and related problems*, *in* Approximation Theory III, Lorentz, Chui, and Schumaker, eds., Academic Press, New York, 1976.

[4] M. A. Gallucci and W. B. Jones, *Rational approximations corresponding to Newton series* (*Newton-Padé approximants*), J. Approx. Theory, 17 (1976), pp. 366–392.

[5] W. B. Gragg, *The Padé table and its relation to certain algorithms of numerical analysis*, SIAM Rev., 14 (1972), pp. 1–62.

[6] M. H. Gutknecht and L. N. Trefethen, *Nonuniqueness of best rational Chebyshev approximations on the unit disk*, J. Approx. Theory, 39 (1983), pp. 275–288.

[7] E. Hille, *Analytic Function Theory*, Vol. II, Chelsea, New York, 1973.

[8] C. Maehly and C. Witzgall, *Tschebyscheff-Approximationen in kleinen Intervallen* II, Numer. Math., 2 (1960), pp. 293–307.

[9] L. N. Trefethen, *Near-circularity of the error curve in complex Chebyshev approximation*, J. Approx. Theory, 31 (1981), pp. 344–367.

[10] J. L. Walsh, *On approximation to an analytic function by rational functions of best approximation*, Math. Z., 38 (1934), pp. 163–176.

[11] ———, *Padé approximants as limits of rational functions of best approximation*, J. Math. Mech., 13 (1964), pp. 305–312.

[12] ———, *Padé approximants as limits of rational functions of best approximation, real domain*, J. Approx. Theory, 11 (1974), pp. 225–230.

[13] H. Werner, *On the rational Tschebyscheff operator*, Math. Z., 86 (1964), pp. 317–326.

[14] L. Wuytack, *On the conditioning of the Padé approximation problem*, *in* Padé Approximation and its Applications, de Bruin and van Rossum, eds., Springer, New York, 1981.

[15] H. Werner and L. Wuytack, *On the continuity of the Padé operator*, SIAM J. Numer. Anal., 20 (1983), pp. 1273–1280.

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

FORCED OSCILLATIONS OF EXTENSIBLE BEAMS*

NORIO YOSHIDA†

Abstract. Forced oscillations of extensible beams are studied and sufficient conditions are given that all classical solutions of boundary value problems for extensible beam equations are oscillatory in a cylindrical domain. Various end conditions are considered.

1. Introduction. We are concerned with the oscillatory behavior of solutions of the extensible beam equation

$$(*)\qquad \frac{\partial^2 u}{\partial t^2}+\alpha\frac{\partial^4 u}{\partial x^4}-\left(\beta+\gamma\int_0^L\left(\frac{\partial u(\xi,t)}{\partial\xi}\right)^2 d\xi\right)\frac{\partial^2 u}{\partial x^2}+c(x,t,u)=f(x,t),\qquad (x,t)\in I\times R_+,$$

where $I=(0,L)$, $R_+=(0,\infty)$, α is a positive constant and β and γ are constants. Equation $(*)$ with $c(x,t,u)\equiv f(x,t)\equiv 0$ was proposed by Woinowsky-Krieger [17] as a model for the transverse deflection of an extensible beam of natural length L. The existence of solutions to initial-boundary value problems for $(*)$ was discussed by numerous authors; see, for example, [1], [3], [4], [6], [7], [12], [17], [18]. The purpose of this paper is to obtain sufficient conditions for all solutions of boundary value problems for $(*)$ to be oscillatory in $I\times R_+$. Our method is an adaptation of that used in studying the oscillatory behavior of solutions of hyperbolic equations (cf. [2], [5], [8], [15], [19]).

In §2 we consider the case of hinged ends and in §3 we consider the case of sliding ends. Various end conditions are studied in §4.

We assume that the following conditions are satisfied throughout this paper:

(A-I) $c(x,t,\eta)$ is a real-valued continuous function in $I\times R_+\times R^1$;

(A-II) $\eta c(x,t,\eta)\geqq 0$ for all $(x,t,\eta)\in I\times R_+\times R^1$;

(A-III) $c(x,t,-\eta)=-c(x,t,\eta)$ for all $(x,t,\eta)\in I\times R_+\times R_+$;

(A-IV) $f(x,t)$ is a real-valued continuous function in $I\times R_+$.

DEFINITION. A function u: $I\times R_+\to R^1$ is said to be *oscillatory* in $I\times R_+$ if it has a zero in $I\times(t,\infty)$ for any $t>0$.

2. Hinged ends. In this section we deal with the case of hinged ends for which

$$(\mathrm{HE})\qquad u(0,t)=u(L,t)=\frac{\partial^2 u}{\partial x^2}(0,t)=\frac{\partial^2 u}{\partial x^2}(L,t)=0.$$

THEOREM 2.1. *Assume that* $\gamma\geqq 0$, *and that there exists a positive function* $\phi(x)\in C^4(I)$ *such that*

$$(1)\qquad \alpha\phi^{(4)}(x)-\beta\phi''(x)\geqq k\phi(x)\quad \textit{in } I \textit{ for some constant } k\geqq 0,$$

$$(2)\qquad \phi''(x)\leqq 0\quad \textit{in } I,$$

$$(3)\qquad \phi(0)=\phi(L)=\phi''(0)=\phi''(L)=0.$$

* Received by the editors March 31, 1983, and in revised form August 16, 1983.

† Department of Mathematics, Faculty of Engineering, Iwate University, Morioka, Japan.

Every classical solution u of $(*)$ *satisfying the boundary condition* (HE) *is oscillatory in* $I \times R_+$ *if the ordinary differential inequalities*

$$y''(t) + ky(t) \leqq \int_0^L f(x,t)\phi(x)\,dx, \tag{4}$$

$$y''(t) + ky(t) \leqq -\int_0^L f(x,t)\phi(x)\,dx \tag{5}$$

are oscillatory at $t = \infty$ in the sense that neither (4) *nor* (5) *has a solution which is positive on $[t\ , \infty)$ for any $t > 0$.*

Proof. Suppose to the contrary that there exists a solution u which has no zero in $I \times [t_0\ , \infty)$ for some $t_0 > 0$. Fitst we assume $u > 0$ in $I \times [t_0\ , \infty)$. By assumption (A-II) we obtain $c(x,t,u) \geqq 0$, and therefore

$$\frac{\partial^2 u}{\partial t^2} + \alpha \frac{\partial^4 u}{\partial x^4} - \left(\beta + \gamma \int_0^L \left(\frac{\partial u(\xi,t)}{\partial \xi} \right)^2 d\xi \right) \frac{\partial^2 u}{\partial x^2} \leqq f(x,t). \tag{6}$$

Multiplying (6) by $\phi(x)$ and then integrating over I yields

$$\int_0^L \frac{\partial^2 u}{\partial t^2}\phi(x)\,dx + \alpha \int_0^L \frac{\partial^4 u}{\partial x^4}\phi(x)\,dx - \left(\beta + \gamma \int_0^L \left(\frac{\partial u(\xi,t)}{\partial \xi} \right)^2 d\xi \right) \int_0^L \frac{\partial^2 u}{\partial x^2}\phi(x)\,dx \leqq \int_0^L f(x,t)\phi(x)\,dx. \tag{7}$$

Integrating by parts and using (HE), (3), we have

$$\int_0^L \frac{\partial^2 u}{\partial x^2}\phi(x)\,dx = \int_0^L u\phi''(x)\,dx, \tag{8}$$

$$\int_0^L \frac{\partial^4 u}{\partial x^4}\phi(x)\,dx = \int_0^L u\phi^{(4)}(x)\,dx. \tag{9}$$

Combining (7)–(9) yields

$$\frac{d^2}{dt^2}\int_0^L u\phi(x)\,dx + \int_0^L u(\alpha\phi^{(4)}(x) - \beta\phi''(x))dx$$
$$-\gamma \int_0^L \left(\frac{\partial u(\xi,t)}{\partial \xi} \right)^2 d\xi \int_0^L u\phi''(x)\,dx \leqq \int_0^L f(x,t)\phi(x)\,dx, \qquad t \geqq t_0.$$

We observe, using (1) and (2), that

$$\frac{d^2}{dt^2}\int_0^L u\phi(x)\,dx + k\int_0^L u\phi(x)\,dx \leqq \int_0^L f(x,t)\phi(x)\,dx, \qquad t \geqq t_0.$$

Hence, $M[u](t) \equiv \int_0^L u\phi(x)dx$ is a positive solution of (4) in $[t_0\ , \infty)$. Next we consider the case where $u < 0$ in $I \times [t_0\ , \infty)$ for some $t_0 > 0$. Letting $U \equiv -u$, we see that

$$\frac{\partial^2 U}{\partial t^2} + \alpha \frac{\partial^4 U}{\partial x^4} - \left(\beta + \gamma \int_0^L \left(\frac{\partial U(\xi,t)}{\partial \xi} \right)^2 d\xi \right) \frac{\partial^2 U}{\partial x^2} + c(x,t,U) = -f(x,t).$$

Proceeding as in the case where $u > 0$, we conclude that $M[U](t)$ is a positive solution of (5) in $[t_0\ , \infty)$. This contradicts the hypothesis and completes the proof.

COROLLARY 2.1. *Assume that* $\gamma \geqq 0$ *and* $\alpha(\pi/L)^4+\beta(\pi/L)^2 \geqq 0$. *Every classical solution u of* (*) *satisfying* (HE) *is oscillatory in* $I\times R_+$ *if*

$$\liminf_{t\to\infty}\int_T^t\left(1-\frac{\theta}{t}\right)\left(\int_0^L f(x,\theta)\sin\frac{\pi}{L}x\,dx\right)d\theta=-\infty,$$

$$\limsup_{t\to\infty}\int_T^t\left(1-\frac{\theta}{t}\right)\left(\int_0^L f(x,\theta)\sin\frac{\pi}{L}x\,dx\right)d\theta=\infty$$

for all large T.

Proof. It is easy to see that $\phi(x)\equiv\sin(\pi/L)x$ satisfies conditions (1)–(3) with $k=\alpha(\pi/L)^4+\beta(\pi/L)^2$. Using a result of Kusano and Naito [9, Thm. 2], we find that

$$(10)\qquad y''(t)+\left(\alpha(\pi/L)^4+\beta(\pi/L)^2\right)y(t)\leqq\pm\int_0^L f(x,t)\sin\frac{\pi}{L}x\,dx$$

are oscillatory at $t=\infty$. Hence, the conclusion follows from Theorem 2.1.

COROLLARY 2.2. *Assume that* $\gamma\geqq 0$ *and* $\alpha(\pi/L)^4+\beta(\pi/L)^2>0$. *Every classical solution u of* (*) *satisfying* (HE) *is oscillatory in* $I\times R_+$ *if there exists a* C^2 *function* ρ: $[1,\infty)\to R^1$ *with the following properties*:

(i) $\rho(t)$ *is oscillatory*;

(ii) $\rho''(t)=\int_0^L f(x,t)\sin\frac{\pi}{L}x\,dx, \qquad t\geqq 1$;

(iii) $\lim_{t\to\infty}\rho(t)=0.$

Proof. Conditions (1)–(3) are satisfied with $\phi(x)=\sin(\pi/L)x$. We easily see that the ordinary differential inequality

$$(11)\qquad y''(t)+\left(\alpha(\pi/L)^4+\beta(\pi/L)^2\right)y(t)\leqq 0$$

has no eventually positive solution (cf. Kahane [5, p. 185]). It follows from a result of Kusano and Naito [9, Thm. 3] that (4) and (5) are oscillatory at $t=\infty$. Hence, the conclusion follows from Theorem 2.1.

COROLLARY 2.3. *Assume that* $\gamma\geqq 0$ *and* $\alpha(\pi/L)^4+\beta(\pi/L)^2>0$. *Every classical solution u of* (*) *satisfying* (HE) *is oscillatory in* $I\times R_+$ *if the function*

$$\int_t^{t+\pi/\omega}\left(\int_0^L f(x,s)\sin\frac{\pi}{L}x\,dx\right)\sin\omega(s-t)\,ds$$

is oscillatory in (t_0,∞) *for some* $t_0>0$, *where* $\omega=(\alpha(\pi/L)^4+\beta(\pi/L)^2)^{1/2}$.

Proof. It is sufficient to show that (10) are oscillatory at $t=\infty$. Suppose to the contrary that there exist eventually positive solutions of (10). Multiplying (10) by $g(t,s)\equiv\sin\omega(t-s)$ and integrating over $(s,s+\pi/\omega)$ with respect to t yield

$$(12)\qquad \int_s^{s+\pi/\omega}y''g(t,s)\,dt+\left(\alpha(\pi/L)^4+\beta(\pi/L)^2\right)\int_s^{s+\pi/\omega}yg(t,s)\,dt$$

$$\leqq\pm\int_s^{s+\pi/\omega}\left(\int_0^L f(x,t)\sin\frac{\pi}{L}x\,dx\right)g(t,s)\,dt.$$

Integration by parts gives

$$(13)\qquad \int_s^{s+\pi/\omega} y''g\,dt=[y'g]_{t=s}^{t=s+\pi/\omega}-[yg_t]_{t=s}^{t=s+\pi/\omega}+\int_s^{s+\pi/\omega} yg_{tt}dt$$

$$=\omega(y(s+\pi/\omega)+y(s))+\int_s^{s+\pi/\omega} yg_{tt}dt,$$

where the subscript t denotes the partial differentiation with respect to t. Combining (12) with (13) yields

$$\omega(y(s+\pi/\omega)+y(s))+\int_s^{s+\pi/\omega} y\Big(g_{tt}+\big(\alpha(\pi/L)^4+\beta(\pi/L)^2\big)g\Big)dt$$

$$\leqq \pm\int_s^{s+\pi/\omega}\left(\int_0^L f(x,t)\sin\frac{\pi}{L}x\,dx\right)\sin\omega(t-s)\,dt,$$

and therefore

$$(14)\qquad \omega(y(t+\pi/\omega)+y(t))\leqq \pm\int_t^{t+\pi/\omega}\left(\int_0^L f(x,s)\sin\frac{\pi}{L}x\,dx\right)\sin\omega(s-t)\,ds.$$

The left-hand side of (14) is positive, but the right-hand side of (14) oscillates. Thus we have a contradiction and (10) are oscillatory at $t=\infty$. The conclusion follows from Theorem 2.1.

COROLLARY 2.4. *Assume that* $\gamma\geqq 0$ *and* $\alpha(\pi/L)^4+\beta(\pi/L)^2>0$. *Every classical solution u of the extensible beam equation*

$$\frac{\partial^2 u}{\partial t^2}+\alpha\frac{\partial^4 u}{\partial x^4}-\left(\beta+\gamma\int_0^L\left(\frac{\partial u(\xi,t)}{\partial \xi}\right)^2 d\xi\right)\frac{\partial^2 u}{\partial x^2}+c(x,t,u)=0$$

is oscillatory in $I\times R_+$.

Proof. We choose $\phi(x)=\sin(\pi/L)x$. Since $f(x,t)\equiv 0$, (4) and (5) reduce to (11), which has no eventually positive solution. Hence, the conclusion follows from Theorem 2.1.

Example 1. We consider the extensible beam equation

$$(15)\qquad \frac{\partial^2 u}{\partial t^2}+\alpha\frac{\partial^4 u}{\partial x^4}-\left(\beta+\gamma\int_0^L\left(\frac{\partial u(\xi,t)}{\partial \xi}\right)^2 d\xi\right)\frac{\partial^2 u}{\partial x^2}+c(x,t,u)=\left(\sin\frac{\pi}{L}x\right)e^{-t}\cos t,$$

where $\gamma\geqq 0$ and $\alpha(\pi/L)^4+\beta(\pi/L)^2>0$. Since

$$\left|\int_T^t\left(1-\frac{\theta}{t}\right)\left(\int_0^L\left(\sin\frac{\pi}{L}x\right)^2 e^{-\theta}\cos\theta\,dx\right)d\theta\right|$$

$$\leqq\frac{L}{2}\int_T^t\left|\left(1-\frac{\theta}{t}\right)e^{-\theta}\cos\theta\right|d\theta\leqq\frac{L}{2}\int_0^\infty e^{-\theta}d\theta=\frac{L}{2}<\infty,$$

Corollary 2.1 is not applicable to (15). We define the function $\rho(t)$ by $\rho(t)=-(L/4)e^{-t}\sin t$. It can be shown that $\rho(t)$ satisfies conditions (i)–(iii) of Corollary 2.2. Hence, every classical solution u of (15) satisfying (HE) is oscillatory in $I\times R_+$.

Example 2. We consider the extensible beam equation

$$(16)\qquad \frac{\partial^2 u}{\partial t^2}+\alpha\frac{\partial^4 u}{\partial x^4}-\left(\beta+\gamma\int_0^L\left(\frac{\partial u(\xi,t)}{\partial \xi}\right)^2 d\xi\right)\frac{\partial^2 u}{\partial x^2}=0,$$

where $\gamma \geqq 0$ and $\alpha(\pi/L)^4+\beta(\pi/L)^2>0$. Corollary 2.4 implies that every classical solution u of (16) satisfying (HE) is oscillatory in $I\times R_+$. In fact, there exists an oscillatory solution $u=(\sin(\pi/L)x)T(t)$, where $T(t)$ is an oscillatory solution of the Duffing's equation

$$T''+\left(\alpha(\pi/L)^4+\beta(\pi/L)^2\right)T+\gamma(\pi/L)^4(L/2)T^3=0$$

(cf. Woinowsky-Krieger [17, p. 36]).

Remark 1. Our result in this section can be generalized to the equation

$$(17)\qquad \frac{\partial^2 u}{\partial t^2}+\Delta^2 u-Q\left(\|\operatorname{grad} u(\cdot,t)\|_{L^2}^2\right)\Delta u+c(x,t,u)=f(x,t), \qquad (x,t)\in\Omega\times R_+,$$

where Δ denotes the Laplacian in R^n, $\Omega=(0,L_1)\times\cdots\times(0,L_n)$ and $Q(s)\geqq 0$ for $s\geqq 0$. The existence of solutions of (17) was studied by Medeiros [10] and Menzala [11].

3. Sliding ends. This section is devoted to the case of sliding ends for which

$$(\text{SE})\qquad \frac{\partial u}{\partial x}(0,t)=\frac{\partial u}{\partial x}(L,t)=\frac{\partial^3 u}{\partial x^3}(0,t)=\frac{\partial^3 u}{\partial x^3}(L,t)=0.$$

THEOREM 3.1. *Every classical solution u of* $(*)$ *satisfying* (SE) *is oscillatory in $I\times R_+$ if the ordinary differential inequalities*

$$(18)\qquad y''(t)\leqq\int_0^L f(x,t)\,dx,$$

$$(19)\qquad y''(t)\leqq-\int_0^L f(x,t)\,dx$$

are oscillatory at $t=\infty$.

Proof. Suppose to the contrary that there exists a solution u which has no zero in $I\times[t_0,\infty)$ for some $t_0>0$. First we assume $u>0$ in $I\times[t_0,\infty)$. As in the proof of Theorem 2.1, we obtain the inequality (6). Integrating (6) over I, we obtain

$$\frac{d^2}{dt^2}\int_0^L u\,dx+\alpha\int_0^L\frac{\partial^4 u}{\partial x^4}dx-\left(\beta+\gamma\int_0^L\left(\frac{\partial u(\xi,t)}{\partial\xi}\right)^2 d\xi\right)\int_0^L\frac{\partial^2 u}{\partial x^2}dx\leqq\int_0^L f(x,t)\,dx.$$

Since

$$\int_0^L\frac{\partial^4 u}{\partial x^4}dx=\frac{\partial^3 u}{\partial x^3}(L,t)-\frac{\partial^3 u}{\partial x^3}(0,t)=0,$$

$$\int_0^L\frac{\partial^2 u}{\partial x^2}dx=\frac{\partial u}{\partial x}(L,t)-\frac{\partial u}{\partial x}(0,t)=0,$$

we see that

$$\frac{d^2}{dt^2}\int_0^L u\,dx\leqq\int_0^L f(x,t)\,dx.$$

Hence, $\int_0^L u\,dx$ is a positive solution of (18) in $[t_0,\infty)$. This contradicts the hypothesis. In the case where $u<0$ in $I\times[t_0,\infty)$, $U\equiv -u$ satisfies

$$\frac{\partial^2 U}{\partial t^2}+\alpha\frac{\partial^4 U}{\partial x^4}-\left(\beta+\gamma\int_0^L\left(\frac{\partial U(\xi,t)}{\partial\xi}\right)^2 d\xi\right)\frac{\partial^2 U}{\partial x^2}+c(x,t,U)=-f(x,t).$$

Using the same arguments as in the case where $u>0$, we are led to a contradiction. The proof is complete.

THEOREM 3.2. *Assume that the following condition holds*:

(A-V) $c(x,t,\eta)\geqq p(t)H(\eta)$ *for all* $(x,t,\eta)\in I\times R_+\times R_+$, *where* p *is continuous, positive in* R_+ *and* H *is continuous, nonnegative and convex in* R_+.

Every classical solution u *of* $(*)$ *satisfying* (SE) *is oscillatory in* $I\times R_+$ *if the ordinary differential inequalities*

$$y''(t)+p(t)H(y(t))\leqq\frac{1}{L}\int_0^L f(x,t)\,dx, \tag{20}$$

$$y''(t)+p(t)H(y(t))\leqq-\frac{1}{L}\int_0^L f(x,t)\,dx \tag{21}$$

are oscillatory at $t=\infty$.

Proof. Suppose to the contrary that there exists a solution u which is positive in $I\times[t_0,\infty)$ for some $t_0>0$. As in the proof of Theorem 3.1, we obtain

$$\frac{d^2}{dt^2}\int_0^L u\,dx+\int_0^L c(x,t,u)\,dx=\int_0^L f(x,t)\,dx. \tag{22}$$

An application of Jensen's inequality [13, p. 160] shows that

$$\int_0^L c(x,t,u)\,dx\geqq p(t)\int_0^L H(u)\,dx\geqq p(t)L\cdot H\left(\frac{1}{L}\int_0^L u\,dx\right). \tag{23}$$

Combining (22) with (23) yields

$$\frac{d^2}{dt^2}\int_0^L u\,dx+Lp(t)H\left(\frac{1}{L}\int_0^L u\,dx\right)\leqq\int_0^L f(x,t)\,dx.$$

Hence, $(1/L)\int_0^L u\,dx$ is a positive solution of (20) in $[t_0,\infty)$. In the case where $u<0$ in $I\times[t_0,\infty)$, we see that $(1/L)\int_0^L(-u)\,dx$ is a positive solution of (21) in $[t_0,\infty)$. This contradicts the hypothesis and completes the proof.

COROLLARY 3.1. *Every classical solution* u *of* $(*)$ *satisfying* (SE) *is oscillatory in* $I\times R_+$ *if*

$$\liminf_{t\to\infty}\int_T^t\left(1-\frac{\theta}{t}\right)\left(\int_0^L f(x,\theta)dx\right)d\theta=-\infty, \tag{24}$$

$$\limsup_{t\to\infty}\int_T^t\left(1-\frac{\theta}{t}\right)\left(\int_0^L f(x,\theta)\,dx\right)d\theta=\infty \tag{25}$$

for all large T.

Proof. Conditions (24) and (25) imply that (18) and (19) are oscillatory at $t=\infty$ (see Kusano and Naito [9, Theorem 2]). The conclusion follows from Theorem 3.1.

COROLLARY 3.2. *Assume that* (A-V) *holds, and that* $H(\eta)=\eta^\sigma$, *where* $\sigma\geqq1$ *is the quotient of odd integers. Every classical solution* u *of* $(*)$ *satisfying* (SE) *is oscillatory in* $I\times R_+$ *if*

$$\int^\infty t^{1-\varepsilon}p(t)\,dt=\infty \quad \textit{for some } \varepsilon>0 \quad (\sigma=1), \tag{26}$$

$$\int^\infty tp(t)\,dt=\infty \qquad (\sigma>1), \tag{27}$$

and there exists a C^2 function ρ: $[1, \infty) \to R^1$ with the following properties:

(i) $\rho(t)$ *is oscillatory*;

(ii) $$\rho''(t) = \frac{1}{L}\int_0^L f(x,t)\,dx, \qquad t \geqq 1;$$

(iii) $$\lim_{t\to\infty} \rho(t) = 0.$$

Proof. It follows from (26) and (27) that

$$y''(t) + p(t)(y(t))^\sigma \leqq 0$$

has no eventually positive solution (see, e.g., [16, p. 633]). Using a result of Kusano and Naito [9, Thm. 3], we find that

$$y''(t) + p(t)(y(t))^\sigma \leqq \pm \frac{1}{L}\int_0^L f(x,t)\,dx$$

are oscillatory at $t = \infty$. Hence, the conclusion follows from Theorem 3.2.

COROLLARY 3.3. *Assume that $c(x,t,\eta) = p_0\eta$ (p_0 is a positive constant), i.e.*

$$\frac{\partial^2 u}{\partial t^2} + \alpha\frac{\partial^4 u}{\partial x^4} - \left(\beta + \gamma\int_0^L\left(\frac{\partial u(\xi,t)}{\partial \xi}\right)^2 d\xi\right)\frac{\partial^2 u}{\partial x^2} + p_0 u = f(x,t). \tag{28}$$

Every classical solution u of (28) *satisfying* (SE) *is oscillatory in $I \times R_+$ if the function*

$$\int_t^{t+\pi/\tilde{p}}\left(\int_0^L f(x,s)\,dx\right)\sin\tilde{p}(s-t)\,ds$$

is oscillatory in (t_0, ∞) for some $t_0 > 0$, where $\tilde{p} = (p_0)^{1/2}$.

Proof. By the same arguments as were used in Corollary 2.3, we conclude that

$$y''(t) + p_0 y(t) \leqq \pm\frac{1}{L}\int_0^L f(x,t)\,dx$$

are oscillatory at $t = \infty$. Hence, the conclusion follows from Theorem 3.2.

Example 3. We consider the extensible beam equation

$$\frac{\partial^2 u}{\partial t^2} + \alpha\frac{\partial^4 u}{\partial x^4} - \left(\beta + \gamma\int_0^L\left(\frac{\partial u(\xi,t)}{\partial \xi}\right)^2 d\xi\right)\frac{\partial^2 u}{\partial x^2} + p(x,t)u^\sigma = f_1(x)t(\log t)\sin t, \tag{29}$$

where $p(x,t) \geqq 0$ in $I \times R_+$, $f_1(x) > 0$ in I and $\sigma(>0)$ is the quotient of odd integers. We easily obtain

$$\begin{aligned}\int_T^t\left(1 - \frac{\theta}{t}\right)\left(\int_0^L f_1(x)\theta(\log\theta)\sin\theta\,dx\right)d\theta &= \delta\int_T^t\left(1 - \frac{\theta}{t}\right)\theta(\log\theta)\sin\theta\,d\theta \\ &= -\delta(\log t)\sin t + B(t,T),\end{aligned} \tag{30}$$

where $B(t,T)$ is bounded as t tends to infinity and

$$\delta = \int_0^L f_1(x)\,dx > 0.$$

In view of (30) we see that conditions (24) and (25) are satisfied. By Corollary 3.1 we conclude that every classical solution u of (29) satisfying (SE) is oscillatory in $I \times R_+$.

Remark 2. In the case of sliding ends, specific assumptions are not imposed on the constants β and γ.

4. Various end conditions. In the case of hinged-sliding ends for which

$$\text{(HSE)} \qquad u(0,t)=\frac{\partial^2 u}{\partial x^2}(0,t)=\frac{\partial u}{\partial x}(L,t)=\frac{\partial^3 u}{\partial x^3}(L,t)=0,$$

there exists a positive function $\phi(x)\in C^4(I)$ which satisfies conditions (1), (2) and the following boundary condition

$$\phi(0)=\phi''(0)=\phi'(L)=\phi'''(L)=0.$$

In fact, we may choose $\phi(x)=\sin(\pi/2L)x$. Hence, our results in §2 hold true with (HE) replaced by (HSE), and $\sin(\pi/L)x$ by $\sin(\pi/2L)x$.

We treat the case where $\beta=\gamma=0$, i.e.

$$(**) \qquad \frac{\partial^2 u}{\partial t^2}+\alpha\frac{\partial^4 u}{\partial x^4}+c(x,t,u)=f(x,t).$$

The boundary conditions to be considered are the following:

$$(B_1[u]) \quad u(0,t)=\frac{\partial u}{\partial x}(0,t)=u(L,t)=\frac{\partial u}{\partial x}(L,t)=0 \qquad \text{(clamped-clamped ends)},$$

$$(B_2[u]) \quad u(0,t)=\frac{\partial u}{\partial x}(0,t)=u(L,t)=\frac{\partial^2 u}{\partial x^2}(L,t)=0 \qquad \text{(clamped-hinged ends)},$$

$$(B_3[u]) \quad u(0,t)=\frac{\partial u}{\partial x}(0,t)=\frac{\partial^2 u}{\partial x^2}(L,t)=\frac{\partial^3 u}{\partial x^3}(L,t)=0 \qquad \text{(clamped-free ends)}.$$

THEOREM 4.1. *For each fixed i $(i=1,2,3)$, every classical solution u of* $(**)$ *satisfying the boundary condition $B_i[u]$ is oscillatory in $I\times R_+$ if there exists a function $\psi(x)\in C^4(I)$ such that*

(31) $\psi(x)>0$ *in I,*

(32) $\psi^{(4)}(x)\geqq\varepsilon\psi(x)$ *in I for some $\varepsilon\geqq 0$,*

(33) *the boundary condition $B_i[\psi]$,*

and if the ordinary differential inequalities

$$y''(t)+\alpha\varepsilon y(t)\leqq\int_0^L f(x,t)\psi(x)\,dx,$$

$$y''(t)+\alpha\varepsilon y(t)\leqq-\int_0^L f(x,t)\psi(x)\,dx$$

are oscillatory at $t=\infty$.

Proof. The proof follows by using exactly the same arguments as in the proof of Theorem 2.1 and will be omitted.

We conclude by showing that there exist functions $\psi_i(x)$ $(i=1,2,3)$ which satisfy conditions (31)–(33) (cf. Timoshenko, Young and Weaver [14]).

Defining $\psi_1(x)\equiv x^2(L-x)^2$, we obtain

$$\psi_1^{(4)}(x)=1\geqq\frac{\psi_1(x)}{\sup_{x\in I}\psi_1(x)}=\frac{16}{L^4}\psi_1(x).$$

Hence, $\psi_1(x)$ satisfies (31), (32) and (33) with $i=1$. We define the function $\tilde{\psi}_1(x)$ by

$$\tilde{\psi}_1(x)=(\sinh m_1-\sin m_1)\left(\cosh\frac{m_1}{L}x-\cos\frac{m_1}{L}x\right)$$
$$-(\cosh m_1-\cos m_1)\left(\sinh\frac{m_1}{L}x-\sin\frac{m_1}{L}x\right),$$

where $m_1=4.730\cdots$ is the lowest root of the equation

$$1-(\cosh m)\cos m=0, \qquad m>0.$$

We easily see that $\tilde{\psi}_1(x)$ also satisfies (31), (32) and (33) with $i=1$.

The functions $\psi_i(x)$ $(i=2,3)$ defined below satisfy (31)–(33). We define

$$\psi_2(x)=(\sinh m_2-\sin m_2)\left(\cosh\frac{m_2}{L}x-\cos\frac{m_2}{L}x\right)$$
$$-(\cosh m_2-\cos m_2)\left(\sinh\frac{m_2}{L}x-\sin\frac{m_2}{L}x\right),$$
$$\psi_3(x)=(\sinh m_3+\sin m_3)\left(\cosh\frac{m_3}{L}x-\cos\frac{m_3}{L}x\right)$$
$$-(\cosh m_3+\cos m_3)\left(\sinh\frac{m_3}{L}x-\sin\frac{m_3}{L}x\right),$$

where $m_2=3.927\cdots$ is the lowest root of the equation

$$(\cosh m)\sin m-(\sinh m)\cos m=0, \qquad m>0,$$

and $m_3=1.875\cdots$ is the lowest root of the equation

$$1+(\cosh m)\cos m=0, \qquad m>0.$$

Acknowledgment. The author would like to thank the referee for many helpful suggestions.

REFERENCES

[1] J. M. Ball, *Initial-boundary value problems for an extensible beam*, J. Math. Anal. Appl., 42 (1973), pp. 61–90.

[2] C. Y. Chan and E. C. Young, *Comparison theorems for a coupled system of singular hyperbolic differential inequalities. I. Time-independent uncoupling coefficients*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur., 66 (1979), pp. 250–254.

[3] R. W. Dickey, *Free vibrations and dynamic buckling of the extensible beam*, J. Math. Anal. Appl., 29 (1970), pp. 443–454.

[4] W. E. Fitzgibbon, *Global existence and boundedness of solutions to the extensible beam equations*, this Journal, 13 (1982), pp. 739–745.

[5] C. Kahane, *Oscillation theorems for solutions of hyperbolic equations*, Proc. Amer. Math. Soc., 41 (1973), pp. 183–188.

[6] M. Kopáčková, *On periodic solution of a nonlinear beam equation*, Apl. Mat., 28 (1983), pp. 108–115.

[7] M. Kopáčková and O. Vejvoda, *Periodic vibrations of an extensible beam*, Časopis Pěst. Mat., 102 (1977), pp. 356–363.

[8] K. Kreith, *Oscillation Theory*, Lecture Notes in Mathematics, 324, Springer-Verlag, Berlin, 1973.

[9] T. Kusano and M. Naito, *Oscillation criteria for a class of perturbed Schrödinger equations*, Canad. Math. Bull., 25 (1982), pp. 71–77.

[10] L. A. Medeiros, *On a new class of nonlinear wave equations*, J. Math. Anal. Appl., 69 (1979), pp. 252–262.

[11] G. P. Menzala, *On global classical solutions of a nonlinear wave equation*, Applicable Anal., 10 (1980), pp. 179–195.

[12] T. Narazaki, *On the time global solutions of perturbed beam equations*, Proc. Fac. Sci. Tokai Univ., 16 (1981), pp. 51–71.

[13] G. O. Okikiolu, *Aspects of the Theory of Bounded Integral Operators in L^p-spaces*, Academic Press, New York, 1971.

[14] S. Timoshenko, D. H. Young and W. Weaver, Jr., *Vibration Problems in Engineering*, 4th ed., John Wiley, New York, 1974.

[15] C. C. Travis, *Comparison and oscillation theorems for hyperbolic equations*, Utilitas Math., 6 (1974), pp. 139–151.

[16] C. C. Travis and N. Yoshida, *Oscillation criteria for nonlinear Bianchi equations*, Nonlinear Anal., 6 (1982), pp. 625–636.

[17] S. Woinowsky-Krieger, *The effect of an axial force on the vibration of hinged bars*, J. Appl. Mech., 17 (1950), pp. 35–36.

[18] M. Yamaguchi, *Time-decaying solutions and asymptotically almost periodic solutions of nonlinear evolution equations*, Funkcial. Ekvac., 24 (1981), pp. 281–306.

[19] E. C. Young, *Comparison and oscillation theorems for singular hyperbolic equations*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur., 59 (1975), pp. 383–391.

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

RELATED EVOLUTION EQUATIONS AND LIE SYMMETRIES*

E. G. KALNINS[†] AND WILLARD MILLER, JR.[‡]

Abstract. We show that there is a one-to-one correspondence between the Lie symmetry operators (with nontrivial time dependence) for a given evolution equation and those evolution equations related to the given one by a change of independent and dependent coordinates.

1. Introduction. By an *evolution equation* we mean a partial differential equation of the form

$$(1.1)\quad (*)\qquad \Omega\equiv v_t-K\left(y^1,\cdots,y^n,v,v_{i_1\cdots i_n}\right)=0,$$

where

$$v_t=\partial_t v,\ v_{i_1\cdots i_n}=\partial_{y^1}^{i_1}\cdots\partial_{y^n}^{i_n}v$$

and K depends on only a finite number $m>0$ of the derivatives $v_{i_1\cdots i_n}$. We assume that K is a real local analytic function of its $m+n+1$ variables, and for technical reasons, that it is a polynomial in the derivatives $v_{i_1\cdots i_n}$. A solution of (1.1) is a function $v=v(t,y^1,\cdots,y^n)$, locally analytic in the variables $(t,\mathbf{y})$, such that (1.1) is well defined and identically satisfied for all $(t,\mathbf{y})\in S$ where S is a nonempty open set in $\mathbb{C}^{n+1}$. (In the following, all functions are assumed to be locally real analytic.) A second evolution equation

$$(1.2)\quad (+)\qquad \Phi\equiv u_s-J\left(x^1,\cdots,x^n,u,u_{i_1\cdots i_n}\right)=0$$

is said to be *related* to $(*)$ if there is a coordinate transformation

$$(1.3)\qquad t=t(s,\mathbf{x}),\quad y^j=y^j(s,\mathbf{x}),\quad v=v(s,\mathbf{x},u),$$

$j=1,\cdots,n$, which maps $(*)$ to $(+)$. Here we assume that the Jacobian $\det(\partial(t,\mathbf{y})/\partial(s,\mathbf{x}))$ is locally nonzero and $\partial v/\partial u\neq 0$. It is clear that an arbitrary transformation of the form

$$(1.4)\qquad t=s,\quad y^j=y^j(\mathbf{x}),\quad v=v(\mathbf{x},u)$$

will map $(*)$ to a related evolution equation, so we consider transformations of the form (1.4) to be *trivial*. Our interest is in determining all equivalence classes of evolution equations related to a given equation, where two evolution equations are equivalent if they are related by a trivial coordinate transformation.

It is well known that every generalized Lie symmetry of (1.1) can be expressed in the standard form

$$(1.5)\qquad X(f)=f\partial_v+D_tf\partial_{v_t}+\sum_{i_1+\cdots+i_n\geq 1}D_{y^1}^{i_1}\cdots D_{y^n}^{i_n}f\partial_{v_{i_1\cdots i_n}},$$

* Received by the editors September 6, 1983.

[†] 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.

where $f=f(t,\mathrm{y},v_{i_1\cdots i_n})$ and D_t, D_{y^k} are total derivatives, e.g.,

$$D_{y^k}=\partial_{y^k}+v_{y^k}\partial_v+v_{ty^k}\partial_{v_t}+\sum_{i_1+\cdots+i_n\geqq 1} v_{i_1\cdots i_k+1\cdots i_n}\partial_{v_{i_1\cdots i_k\cdots i_n}}; \tag{1.6}$$

see [1]–[3]. In particular, $X(f)$ is a generalized Lie symmetry provided

$$X(f)\Omega=0, \tag{1.7}$$

whenever $\Omega=0$ and $D_1^{i_1}\cdots D_n^{i_n}\Omega=0$ for all $i_i,\cdots,i_n\geqq 0$. Special Lie symmetries are the point operators

$$Y=\tau(t,\mathrm{y},v)\partial_t+\sum_j \xi^j(t,\mathrm{y},v)\partial_{y^j}+\eta(t,\mathrm{y},v)\partial_v. \tag{1.8}$$

These operators can be identified with the standard form operators $X(f)$ where [1]–[3]

$$f=\eta-\sum_j \xi^j v_{y^j}-\tau K. \tag{1.9}$$

In §2 we show that, under the assumption of a nondegeneracy condition on K, there is a one-to-one association between equivalence classes of evolution equations related to (*) and point symmetry operators Y for (*) such that $\tau\neq 0$. Thus a knowledge of the Lie symmetry algebra of an evolution equation leads directly to a list of all related equations.

In §3 we modify and specialize these ideas to associate each time-dependent Hamilton–Jacobi equation related to a given equation

$$2\lambda p_t-g^{ij}p_{y^i}p_{y^j}-2\lambda\xi^i p_{y^i}-\lambda^2 V=0 \tag{1.10}$$

with a conformal symmetry of (1.10). Further, we associate each time-dependent Schrödinger equation related to the fixed equation

$$2\lambda\partial_t\Psi-\frac{1}{\sqrt{g}}\partial_{y^i}\left(g^{ij}\sqrt{g}\,\partial_{y^j}\right)\Psi-2\lambda\xi^i\partial_{y^i}\Psi-\rho^i\partial_{y^i}\Psi-\left(\lambda^2V+\lambda W+Z\right)\Psi=0 \tag{1.11}$$

with a conformal symmetry of (1.11).

In §4 we examine in some detail the evolution equations related to

$$2\lambda p_t-g^{ij}p_{y^i}p_{y^j}=0,\qquad 2\lambda\Psi_t-\frac{1}{\sqrt{g}}\partial_{y^i}\left(g^{ij}\sqrt{g}\,\partial_{y^j}\Psi\right)=0, \tag{1.12}$$

the Hamilton–Jacobi and Schrödinger equations on an n-dimensional Riemannian manifold. We also show that the conformal symmetry algebras of these two equations are isomorphic. We conclude with some examples.

The results of this paper have applicability in the theory of separation of variables [4, Chapt. 2], in the solution of time-dependent boundary value problems [5], and in the solution of Cauchy problems [6], among others. We note that our coordinate transformation approach is only a special case of the theory of evolution equations related by Bäcklund transformations, e.g., [7]–[9]. However, in this special case we can give rather explicit and complete results.

2. Related evolution equations.

Our basic observation is:

LEMMA 1. *Let*

$$\Omega\equiv v_t-K(\mathrm{y},v,v_{i_1\cdots i_n})=0 \tag{2.1}$$

be an evolution equation and

$$\Phi \equiv u_s - J(\mathbf{x}, u, u_{j_1 \cdots j_n}) = 0 \tag{2.2}$$

an evolution equation related to Ω *by means of the coordinate transformation*

$$t = T(s, \mathbf{x}), \quad y^j = Y^j(s, \mathbf{x}), \quad v = V(s, \mathbf{x}, u). \tag{2.3}$$

Then $X(f)$ *is a standard form point symmetry of* $\Omega = 0$ *where*

$$f = \partial_s V - \sum_j \partial_s Y^j \cdot v_{y^j} - \partial_s T \cdot K. \tag{2.4}$$

Proof. It is obvious from (2.2) that $Y = \partial_s$ is a point symmetry operator for $\Phi = 0$, hence for $\Omega = 0$. From (1.7) and (2.3) we see that Y corresponds to the standard form symmetry $X(f)$. Q.E.D.

The converse of Lemma 1 is false; there may exist point symmetries of $\Omega = 0$ that do not correspond to a related evolution equation. For example,

$$\Omega \equiv v_t - v_{y^1 y^1} - v_{y^2 y^2} = 0, \quad Y = \partial_{y^1}. \tag{2.5}$$

The following result isolates a special class of point symmetries that do correspond to related evolution equations.

LEMMA 2. *Let*

$$Y = \tau(t)\partial_t + \sum_j \xi^j(t, \mathbf{y})\partial_{y^j} + \eta(t, \mathbf{y}, v)\partial_v \tag{2.6}$$

be a point symmetry operator for $\Omega = 0$, *where* $\tau \not\equiv 0$. *Then there exists a transformation to new coordinates* $(s, \mathbf{x}, u)$,

$$t = T(s), \quad y^j = Y^j(s, \mathbf{x}), \quad v = V(s, \mathbf{x}, u) \tag{2.7}$$

such that (in the new coordinates) $Y = \partial_s$ *and the transformed equation can be expressed as an evolution equation*

$$\Phi = u_s - J(\mathbf{x}, u, u_{j_1 \cdots j_n}) = 0.$$

(The coordinate transformation is not unique because of the possibility of trivial transformations (1.4). We are identifying an equivalence class of related equations.)

Proof. It follows directly from (2.6) and Lie's theorem [10, pp. 34, 49, 50], that there exists a new coordinate system $(s, \mathbf{x}, u)$ such that the coordinate transformation takes the form (2.7) and $Y = \partial_s$. Introducing the new coordinates into the equation $\Omega = 0$, we see that only the term v_t contributes a derivative of u with respect to s. Thus this equation can be rewritten as

$$\Phi \equiv u_s - J(s, \mathbf{x}, u, u_{j_1 \cdots j_n}) = 0.$$

However, since $Y = \partial_s$ is a point symmetry operator for $\Phi = 0$, we must have $\partial_s J = 0$. Q.E.D.

The form of the point symmetry operator (2.6) appears somewhat special. However, for a large class of evolution equations it is perfectly general. Given an evolution equation $\Omega = 0$, (1.1), we can express K as a polynomial in the derivatives $v_{i_1 \cdots i_n}$ (with coefficients which are analytic functions of $\mathbf{y}, v$). We say that a given monomial in this polynomial (with nonzero coefficient), $B^{i_1\cdots l_n}(\mathrm{y},v)=v_{i_1\cdots i_n}v_{j_1\cdots j_n}\cdots v_{l_1\cdots l_n}$ has *order* $i_1+i_2+\cdots+i_n+j_1+\cdots+l_n$. Similarly, we say that this monomial has *rank* $(s_1,s_2,\cdots)$, where s_1 is the number of factors $v_{k_1\cdots k_n}$ with $k_1+\cdots+k_n=1$, s_2 is the number of factors $v_{k_1\cdots k_n}$ with $k_1+\cdots+k_n=2$, etc. For example $v_{10}v_{10}v_{01}v_{21}$ has order 6 and rank $(3,0,1)$. Let m be the highest order occurring in K and let

$$S=\sum_{i_1+\cdots+l_n=m} B^{i_1\cdots l_n}(\mathrm{y},v)\,v_{i_1\cdots i_n}\cdots v_{l_1\cdots l_n}$$

be the sum of the monomials of highest order. We can write

$$S=\sum_{s_1+s_2+\cdots=m} S^{(s_1,\cdots)},$$

where $S^{(s_1,s_2,\cdots)}$ is the sum of the monomials with order m and rank $(s_1,s_2,\cdots)$. With respect to the natural basis $v_{y^i},v_{y^iy^j},\cdots$ each term $S^{(s_1,s_2,\cdots)}$ defines a multilinear m-form corresponding to the matrix components

$$(2.8)\quad B^{(i_{klq})}_{(s_1,s_2,\cdots)},\qquad i_{klq}=1,\cdots,n,\quad k=1,\cdots,q,\quad l=1,2,\cdots,s_q,\quad q=1,2,\cdots.$$

We can assume, without loss of generality, that B is symmetric in the indices i_{klq} for fixed l,q and, also, symmetric in l for fixed k,q. We say that Ω is *nondegenerate* if $m>1$ and for each nonzero n-vector ξ_i at least one of the $(m-1)$-forms $B'^{(i_{klq})}_{(s_1,s_2,\cdots)}$ is nonzero (for some choice of $s_1,s_2,\cdots,l,k$), where B' is the contraction of $B_{(s_1,s_2,\cdots)}$ and ξ_i with respect to one of the indices $i_{k'l'q'}$. (Note, however, that the property of nondegeneracy is independent of the choice of basis.) For example, with $n=m=2$ the forms $v_{y^1y^1}+v_{y^2y^2}$, $v^2_{y^1}+v_{y^2y^2}$ and $y^1v_{y^1y^2}$ $(y^1\neq 0)$ are nondegenerate while $v^2_{y^2}$ is degenerate.

THEOREM 1. *Let*

$$(2.9)\qquad \Omega\equiv v_t-K(\mathrm{y},v,v_{i_1\cdots i_n})=0$$

be a nondegenerate evolution equation. There is a one-to-one correspondence between (equivalence classes of) nondegenerate evolution equations related to $\Omega=0$ *and point symmetry operators for* $\Omega=0$ *of the form*

$$(2.10)\qquad Y=\tau(t,\mathrm{y})\,\partial_t+\sum_j\xi^j(t,\mathrm{y})\,\partial_{y^j}+\eta(t,\mathrm{y},v)\,\partial_v$$

with $\tau\neq 0$. *In fact, all such point symmetries have the property that*

$$\partial_{y^i}\tau=0,\qquad i=1,\cdots,n.$$

Proof. Suppose the nondegenerate evolution equation

$$(2.11)\qquad \Phi\equiv u_s-J(\mathrm{x},u,u_{j_1\cdots j_n})=0$$

is related to $\Omega=0$ by means of the coordinate transformation (2.3). Then (2.11) can be obtained from (2.9) by performing the coordinate transformation and solving for u_s in the resulting equation. Since both (2.9) and (2.11) are nondegenerate, the $n\times n$ matrix $(\partial x^i/\partial y^j)$ must be nonsingular. Furthermore, nondegeneracy implies that, unless $\partial s/\partial y^i=0$, $i=1,\cdots,n$, the transformed equation must contain at least one nonzero mth order term with a factor of the form $u_{sx^{i_1}\cdots x^{i_l}}$, $1\leqq l\leqq m-1$. This is impossible! Thus $s=s(t)$ so $\tau=\partial t/\partial s$ is a function of t alone. We must have $\tau\neq 0$ since otherwise there would be no term in the transformed equation involving u_s. It follows from Lemma 1 that $Y=\partial_s$ is a point symmetry of $\Omega=0$.

Conversely, suppose Y, (2.10), is a point symmetry operator for $\Omega=0$ with $\tau\neq 0$. Then $X(\eta-\xi^j v^j_y-\tau K)=X(f)$ is a standard form symmetry operator for this evolution equation. Substituting $X(f)$ into (1.7), comparing coefficients of terms of order $2m-1$ and invoking the nondegeneracy of K, we can conclude that $\partial_{y^i}\tau=0$, $i=1,\cdots,n$. By Lemma 2, Y determines an evolution equation $\Phi=0$, related to (2.9). The induced coordinate transformation takes the form (2.7), so the $n\times n$ matrix $(\partial y^j/\partial x^i)$ is nonsingular. This implies that Φ is nondegenerate. Q.E.D.

As a very simple example, consider the symmetry $Y=t\partial_y+c\partial_t-\partial_v$, $c\neq 0$, for the Korteweg–deVries equation (clearly nondegenerate)

$$v_t-v_{yyy}-vv_y=0. \tag{2.12}$$

(The symmetry algebra for this equation can be found in many references, e.g., [2].) The requirement $\partial_s=Y$ leads to new coordinates (s,x,u) such that

$$t=cs,\quad y=\tfrac{1}{2}cs^2+x,\quad v=-s+u.$$

Equation (2.12) transforms to the new evolution equation

$$u_s=cu_{xxx}+cuu_x+1. \tag{2.13}$$

Note that the group invariant solution corresponding to the operator Y is obtained by requiring that $u_s=0$ and solving the ordinary differential equation so obtained, [2], [11]. (This is a general fact; the group invariant solutions correspond to solutions independent of the "new" time coordinate. However, in many cases one can find additional explicit solutions of the new evolution equation, say by separation of variables, for which $u_s\not\equiv 0$.) To find all evolution equations related to (2.12) it is natural to identify two such equations if one can be transformed to the other by a Lie (group) symmetry of (2.12). The distinct equations correspond to the orbits of symmetry operators (under the adjoint action of the symmetry group) that contain a representative $Y=\tau\partial_t+\xi^j\partial_{y^j}+\eta\partial_v$ with $\tau\not\equiv 0$.

3. The Hamilton–Jacobi and Schrödinger equations. We now modify the results of the preceding section to apply to the Hamilton–Jacobi equation

$$2\lambda p_t-g^{ij}(\mathrm{y})p_{y^i}p_{y^j}-2\lambda\xi^i(\mathrm{y})p_{y^i}-\lambda^2V(\mathrm{y})=0. \tag{3.1}$$

Here $p_t=\partial_t W(t,\mathrm{y})$, $p_{y^i}=\partial_{y^i}W(t,\mathrm{y})$, ξ^i and V are given functions, λ is a parameter and (g^{ij}) is an $n\times n$ nonsingular matrix defining a metric on a pseudo-Riemannian space V^n. We can interpret (3.1) as the (time-dependent) Hamilton–Jacobi equation for a one-particle Hamiltonian system on V^n with (velocity-dependent) potential.

Our interest is in transformations of (3.1) which map this equation into another equation of the same type:

$$2\lambda\tilde{p}_s-\tilde{g}^{ij}(\mathrm{x})p_{x^i}p_{x^j}-2\lambda\tilde{\xi}^i(\mathrm{x})\tilde{p}_{x^i}-\lambda^2\tilde{V}(\mathrm{x})=0. \tag{3.2}$$

Here, $\tilde{p}_s=\partial_s\tilde{W}(s,\mathrm{x})$, λ is unchanged and $\tilde{g}^{ij}$ defines a metric on a pseudo-Riemannian space $\tilde{V}^n$. To determine the form of the permissible transformations it is convenient to consider (3.1) as a zero-potential equation in an $(n+2)$-dimensional pseudo-Riemannian space. In terms of local coordinates, y, t, v we write (3.1) as

$$2p_v p_t-g^{ij}(\mathrm{y})p_{y^i}p_{y^j}-2\xi^i(\mathrm{y})p_v p_{y^i}-V(\mathrm{y})p_v p_v=0,\qquad p_v=\lambda, \tag{3.3}$$

i.e., as the equation $-G^{\alpha\beta}p_\alpha p_\beta=0$ for the metric

$$(3.4)\qquad (G^{\alpha\beta})=\left(\begin{array}{c|c|c} g^{ij}(\mathrm{y}) & 0 & \xi^j(\mathrm{y}) \\ \hline 0 & 0 & -1 \\ \hline \xi^i(\mathrm{y}) & -1 & V(\mathrm{y}) \end{array}\right).$$

Here $p_\nu=\partial_\nu Z(\nu,t,\mathrm{y})$, etc., so the desired solutions take the form

$$(3.5)\qquad Z(\mathrm{y},t,\nu)=\lambda\nu+W(t,\mathrm{y}),$$

where W satisfies (3.1). Comparison of the equations $G^{\alpha\beta}p_\alpha p_\beta=0$ and $\tilde G^{\alpha\beta}\tilde p_\alpha\tilde p_\beta=0$ suggests that the allowable transformations to new coordinates μ,s,x should be of the form

$$(3.6)\qquad t=T(s,\mathrm{x}),\quad y^j=Y^j(s,\mathrm{x}),\quad \nu=H(\mu,s,\mathrm{x}),\quad Z=\tilde Z,\quad p_\nu=\tilde p_\mu$$

so that $\nu=\mu+h(s,\mathrm{x})$. Here $\partial(t,\mathrm{y})/\partial(s,\mathrm{x})$ is nonsingular. Thus, in terms of the variables determining (3.1), the allowable transformations are

$$(3.7)\qquad t=T(s,\mathrm{x}),\quad y^j=Y^j(s,\mathrm{x}),\quad W=\tilde W+\lambda h(x,\mathrm{x}).$$

We must determine which of these transformations will map (3.1) into an equation of the form (3.2).

An analysis similar to the above shows that the allowable point symmetries for (3.1) should be those of the type

$$(3.8)\qquad Y=\tau(t,\mathrm{y})\,\partial_t+\gamma^j(t,\mathrm{y})\,\partial_{y^j}+\lambda k(s,\mathrm{y})\,\partial_W.$$

(It is straightforward to show that the space of all symmetries of this type forms a Lie algebra under the usual operator commutator bracket.)

THEOREM 2. *There is a one-to-one correspondence between* (*equivalence classes of*) *Hamilton–Jacobi equations related to* (3.1) *and point symmetry operators* Y *for* (3.1) *of the form* (3.8) *with* $\tau\neq 0$. *All symmetries of the form* (3.8) *satisfy* $\partial_{y^i}\tau=0$, $i=1,\cdots,n$.

Proof. Suppose the Hamilton–Jacobi equation (3.2) is related to (3.1) via a transformation of the form (3.7). This means that we can obtain (3.2) by substituting the transformation (3.7) into (3.1) and solving for $\lambda\tilde p_s$ in the resulting expression. Since (g^{ij}) and $(\tilde g^{ij})$ are nonsingular the $n\times n$ matrix $(\partial x^i/\partial y^j)$ must also be nonsingular. Furthermore, the coefficient of $\tilde p_s\tilde p_l$ in the resulting expression is $2g^{ij}\partial s/\partial y^i\partial x^l/\partial y^j$. Since this must vanish for each l, we have $\partial s/\partial y^i=0$, so $s=S(t)$ or $t=T(s)$ with $\partial_s T\neq 0$. Clearly, $\tilde Y=\partial_s$ is a point symmetry of (3.2) which implies that $Y=T'(s)\partial_t+\partial y^i/\partial s\,\partial_{y^i}+\lambda\,\partial h/\partial s\,\partial_W$ is a point symmetry of (3.1).

Conversely, suppose Y, (3.8), is a point symmetry operator for (3.1) with $\tau\neq 0$. Then $X(\lambda k-\xi^j p_{y^j}-\tau p_t)$ is a standard form symmetry operator for this equation. This is possible only if $\partial_{y^i}\tau=0$, $i=1,\cdots,n$ (since (g^{jl}) is nonsingular). By Lie's theorem we can introduce new coordinates s,x and a new dependent variable $\tilde W$ such that

$$(3.9)\qquad \partial_s=\tau\partial_t+f^j\partial_{y^j}\quad t=T(s),\quad y^j=Y^j(s,\mathrm{x}),\quad W=\tilde W+\lambda h(s,\mathrm{x}),$$

where $\partial_s h=k$. Clearly, $\partial s/\partial_{y^i}=0$, and $(\partial x^j/\partial y^i)$ is nonsingular. Thus, substituting the new coordinates s,x, $\tilde W'$ into (3.1), we see that the coefficients of $\tilde p_s^2$ and $\tilde p_s\tilde p_i$ vanish in the resulting expression while the quadratic form $g^{ij}\partial x^l/\partial y^i\partial x^m/\partial y^j\tilde p_{x^l}\tilde p_{x^m}$ is nonsingular. The coefficient of $\tilde p_s$ is $\lambda\partial s/\partial t=\lambda f(s)\neq 0$ and, since $Y=\partial_s+\lambda k\partial_W$ is a point symmetry, if we multiply all terms in our equation by $f^{-1}(s)$ we obtain an expression of the form (3.2), where the coefficient of each term is independent of s and $(\tilde{g}^{ij})$ is nonsingular. Q.E.D.

COROLLARY 1. *If the Hamilton–Jacobi equations* (3.1) *and* (3.2) *are related by a transformation* (3.7), *then the tensors* (g^{ij}) *and* $\pm(\tilde{g}^{ij})$ *define metrics on the same pseudo-Riemannian manifold.* (*There is a possible sign ambiguity.*)

There is a similar theory for transformations that map a Schrödinger equation

$$(3.10)\quad 2\lambda\partial_t\Psi - \frac{1}{\sqrt{g}}\partial_{y^i}\left(g^{ij}\sqrt{g}\,\partial_{y^j}\right)\Psi - (2\lambda\xi^i+\rho^i)\,\partial_{y^i}\Psi - (\lambda^2V+\lambda U+W)\Psi=0$$

to another Schrödinger equation

$$(3.11)\quad 2\lambda\partial_s\Theta - \frac{1}{\sqrt{\tilde{g}}}\partial_{x^i}\left(\tilde{g}^{ij}\sqrt{\tilde{g}}\,\partial_{x^j}\right)\Theta - (2\lambda\tilde{\xi}^i+\tilde{\rho}^i)\,\partial_{x^i}\Theta - (\lambda^2\tilde{V}+\lambda\tilde{U}+\tilde{W})\Theta=0.$$

Here, λ is a parameter (which we can roughly identify with $-2\pi\sqrt{-1}/h$ where h is Planck's constant [12]), $(g^{ij}(\mathrm{y}))$ determines a metric on a pseudo-Riemannian n-dimensional manifold and $g^{-1}=\det(g^{ij})$. The allowable coordinate transformations are of the form

$$(3.12)\quad t=T(s,\mathrm{x}),\quad y^j=Y^j(s,\mathrm{x}),\quad \Psi(t,\mathrm{y})=\exp(\lambda R^{(1)}(s,\mathrm{x})+R^{(2)}(s,\mathrm{x}))\Theta(s,\mathrm{x}),$$

where $\partial(t,\mathrm{y})/\partial(s,\mathrm{x})$ is nonsingular. The allowable point symmetries of (3.10) take the form

$$(3.13)\qquad Y=\tau(t,\mathrm{y})\,\partial_t+\gamma^j(t,\mathrm{y})\,\partial_{y^j}+(\lambda k(t,\mathrm{y})+l(t,\mathrm{y}))\Psi\partial_\Psi.$$

The vector space of symmetries of this type is a Lie algebra under the usual operator commutator bracket; we call this the *Lie symmetry algebra* of (3.10).

THEOREM 3. *There is a one-to-one correspondence between* (*equivalence classes of*) *Schrödinger equations related to* (3.10) *and allowable point symmetry operators* Y *for* (3.10) *of the form* (3.13) *with* $\tau\neq 0$. *All symmetries of the form* (3.13) *satisfy* $\partial_{y^i}\tau=0$, $i=1,\cdots,n$.

COROLLARY 2. *If the Schrödinger equations* (3.10) *and* (3.11) *are related by a transformation* (3.12) *then the tensors* (g^{ij}) *and* $\pm(\tilde{g}^{ij})$ *define metrics on the same pseudo-Riemannian manifold.*

The proof of these statements and the connection between the symmetry operator and the coordinate transformation is very similar to that of Theorem 2. Again the time coordinates s,t of related Schrödinger equations satisfy $t=T(s)$, where $T'=\tau$.

4. The zero potential case. In his thesis, [13], Chandler proved the following result:

THEOREM 4. *Let* $\mathcal{G}$ *be the Lie algebra of allowable point symmetry operators for the Schrödinger equation* (3.10) *on an n-dimensional pseudo-Riemannian manifold. Then* $\dim\mathcal{G}\leqq\frac{1}{2}(n+1)(n+2)+3$.

A simple modification of Chandler's proof yields the following:

THEOREM 5. *Let* $\mathcal{G}'$ *be the Lie algebra of allowable point symmetry operators for the Hamilton–Jacobi equation* (3.1) *on an n-dimensional pseudo-Riemannian manifold. Then* $\dim\mathcal{G}'\leq\frac{1}{2}(n+1)(n+2)+3$.

Earlier, Kuwabara [14] showed that for $\xi^i=\rho^i=0$ in (3.10), i.e., no velocity dependent potential, the upper bound on the dimension of the symmetry algebra is actually achieved only for flat spaces and for certain scalar potentials. (In particular the upper bound is achieved for a constant potential in flat space.) Kuwabara's results and proof remain valid for the Hamilton–Jacobi equation (3.1) with $\xi^i=0$.

For arbitrary potential a complete determination of the symmetry algebra is difficult and there are only a few general results available, e.g., [14]–[17]. However, for the zero-potential Hamilton–Jacobi equation

$$(4.1)\qquad 2\lambda p_t - g^{ij}p_{y^i}p_{y^j}=0,$$

we can say quite a lot. The operator Y, (3.8), is an allowable symmetry of this equation if and only if the following conditions are satisfied:

$$(4.2)\quad \partial_t k=0,\quad g^{il}\partial_{y^l}\tau=0,\quad \partial_t\gamma^i=-g^{il}\partial_{y^l}k,\quad g^{il}\partial_{y^l}\gamma^j+g^{jl}\partial_{y^l}\gamma^i-\gamma^l\partial_{y^l}g^{ij}=g^{ij}\partial_t\tau.$$

For an analysis of the solutions of these equations it is convenient to utilize the *Poisson bracket* of two functions $\mathscr{F}_i(\mathbf{y},\mathbf{p})$, $i=1,2$, on a $2n$-dimensional symplectic manifold:

$$(4.3)\qquad \{\mathscr{F}_1,\mathscr{F}_2\}=\sum_{l=1}^{n}\left(\frac{\partial\mathscr{F}_1}{\partial p_l}\frac{\partial\mathscr{F}_2}{\partial y^l}-\frac{\partial\mathscr{F}_2}{\partial p_l}\frac{\partial\mathscr{F}_1}{\partial y^l}\right).$$

Note that $Y_0=\lambda\partial_W$ is always a symmetry of (4.1).

LEMMA 3. *Modulo a change of time coordinate $t\to t+\alpha$ and addition of an arbitrary multiple of Y_0, each nonzero allowable point symmetry operator for the Hamilton–Jacobi equation* (4.1) *is a scalar multiple of exactly one of the following operators*:

I) $Y_1=\partial_t$.

II) $Y_2=-\frac{t^2}{2}\partial_t+t\gamma^i(\mathbf{y})\partial_{y^i}+\lambda k(\mathbf{y})\partial_W$,

$\gamma^j=-g^{jl}\partial_{y^l}k,\quad \{\gamma^l p_l, g^{ij}p_ip_j\}=g^{ij}p_1p_j.$

III) $Y_3=-t\partial_t+\gamma^i(\mathbf{y})\partial_{y^i},\quad \{\gamma^l p_l, g^{ij}p_ip_j\}=g^{ij}p_ip_j.$

IV) $Y_4=t\gamma^i(\mathbf{y})\partial_{y^i}+\lambda k(\mathbf{y}),\ \gamma^j=-g^{jl}\partial_{y^l}k,\quad \{\gamma^l p_l,g^{ij}p_ip_j\}=0.$

V) $Y_5=\gamma^i(\mathbf{y})\partial_{y^i},\ \{\gamma^l p_l,g^{ij}p_ip_j\}=0.$

THEOREM 6. *Let $\mathscr{G}$ be the Lie algebra of allowable point symmetry operators for the Hamilton–Jacobi equation*

$$(4.4)\qquad 2\lambda p_t=g^{ij}p_ip_j$$

on an n-dimensional Riemannian manifold V^n and let m be the dimension of the vector space of type IV *symmetries of this equation. Then there exists a coordinate system $\{x^1,\cdots,x^m,y^{m+1},\cdots,y^n\}$ on V^n, with respect to which* (4.4) *takes the form*

$$(4.5)\qquad 2\lambda p_t=\sum_{i=1}^{m}p_i^2+\sum_{j,k=m+1}^{n}g^{jk}(\mathbf{y})p_jp_k.$$

Furthermore, as a vector space

$$(4.6)\qquad \mathscr{G}=\mathscr{A}_m\oplus\mathscr{B}\oplus\mathscr{C}_q,$$

where:

1) $\dim\mathscr{A}_m=2m$ *and* $\mathscr{A}_m$ *has the basis*

$$\partial_{x^i},\qquad t\partial_{x^i}-\lambda x^i\partial_W,\qquad i=1,\cdots,m.$$

2) *Let $\mathscr{D}$ be the Lie algebra of symmetries of type* V. *Then $\mathscr{B}$ is a subspace of type* V *symmetries such that* $\mathscr{D}=\mathscr{B}\oplus\operatorname{span}(\partial_{x^i}: i=1,\cdots,m)$.

3) $\mathscr{C}_q$ *corresponds to exactly one of the following possibilities*:

a) $\mathscr{C}_2$, *basis* ∂_t, $\lambda\partial_W$.

b) $\mathscr{C}_3$, *basis* ∂_t, $t\partial_t+\frac{1}{2}\Sigma_l x^l\partial_{x^l}+\partial_{y^{m+1}},\lambda\partial_W$.

Here $g^{ij}=\exp(\frac{1}{2}y^{m+1})G^{ij}(y^{m+2},\cdots,y^n)$, $i,j=m+1,\cdots,n$.

c) $\mathscr{C}_4$, *basis* ∂_t, $t\partial_t+\frac{1}{2}(\Sigma_l x^l\partial_{x^l}+y^{m+1}\partial_{y^{m+1}})$,

$$t^2\partial_t+t\Big(\sum_l x^l\partial_{x^l}+y^{m+1}\partial_{y^{m+1}}\Big)-\frac{\lambda}{2}\Big(\sum_l x^lx^l+y^{m+1}y^{m+1}\Big)\partial_W,\lambda\partial_W.$$

Here $g^{m+1,k}=\delta^{m+1,k}$, $1\leqq k\leqq n$, *and*

$$g^{ij}=\left(y^{m+1}\right)^{-2}G^{ij}\left(y^{m+2},\cdots,y^n\right),\qquad i,j=m+2,\cdots,n.$$

Proof. Let c be the maximal number of functionally independent symmetries of type III (with respect to the terms $t\gamma^i\partial_i$) and let $Z_1,\cdots,Z_c$ be c functionally independent type III symmetries. (Clearly, these symmetries pairwise commute.) By Lie's theorem and Lemma 3 there exists a coordinate system $\{x^1,\cdots,x^c,y^{c+1},\cdots,y^n\}$ for V^n such that $Z_i=t\partial_{x^i}-\lambda x^i\partial_W$, $i=1,\cdots,c$ and the Hamilton–Jacobi equation takes the form (4.5). (Here we are using the fact that the metric is positive definite.) If Z is a type III symmetry then by definition of c, Z must be of the form

$$Z=\sum_{i=1}^{c} tf_i(\mathbf{y})\,\partial_{x^i}+\lambda k(\mathbf{x},\mathbf{y})\,\partial_W.$$

However, the condition that Z be a point symmetry operator implies that each f_i is a constant. Thus $c=m$. Furthermore, it is obvious that the operators $\hat{Z}_i=\partial_{x^i}$, $i=1,\cdots,m$ are type V symmetries of (4.5). The remainder of the proof now follows easily from Lemma 3.

The structure of the symmetry algebra of the zero-potential Schrödinger equation

$$2\lambda\partial_t\Psi-\frac{1}{\sqrt{g}}\partial_{y^i}\left(g^{ij}\sqrt{g}\,\partial_{y^j}\Psi\right)=0 \tag{4.7}$$

is similar to that of the Hamilton–Jacobi equation (4.1). The operator Y, (3.13), is an allowable symmetry of (4.7) if and only if conditions (4.2) hold, and if in addition

$$l=l(t),\qquad \frac{1}{\sqrt{g}}\partial_{y^i}\left(g^{ia}\sqrt{g}\,\partial_{y^a}k\right)+\partial_t l=0. \tag{4.8}$$

Since the Schrödinger equation is linear, $\hat{Y}_0=\Psi\partial_\Psi$ is always a symmetry operator.

LEMMA 4. *Modulo an additive change of time coordinate $t\rightarrow t+\alpha$ and addition of an arbitrary multiple $\alpha\lambda+\beta$ of $\hat{Y}_0$, each nonzero allowable point symmetry operator for the Schrödinger equation* (4.7) *is a scalar multiple of exactly one of the following operators*:

I) $\hat{Y}_1=\partial_t$.

II) $\hat{Y}_2=-\dfrac{t^2}{2}\partial_t+t\gamma^i(\mathbf{y})\partial_{y^i}+(\lambda k(\mathbf{y})+l(t))\Psi\partial_\Psi$,

$\gamma^j=-g^{ja}\partial_{y^a}k,\quad \{\gamma^a p_a, g^{ij}p_ip_j\}=g^{ij}p_ip_j,\quad -\dfrac{1}{\sqrt{g}}\partial_{y^i}(\sqrt{g}\,\gamma^i)=\partial_t l.$

III) $\hat{Y}_3 = -t\partial_t + \gamma^i(\mathbf{y})\partial_{y^i}$, $\{\gamma^a p_a, g^{ij}p_i p_j\} = g^{ij}p_i p_j$.

IV) $\hat{Y}_4 = t\gamma^i(\mathbf{y})\partial_{y^i} + \lambda k(\mathbf{y})\Psi\partial_\Psi$ $\gamma^j = -g^{ja}\partial_{y^a}k$, $\{\gamma^a p_a, g^{ij}p_i p_j\} = 0$.

V) $\hat{Y}_5 = \gamma^i(\mathbf{y})\partial_{y^i}$, $\{\gamma^a p_a, g^{ij}p_i p_j\} = 0$.

THEOREM 7. *Let $\hat{\mathscr{G}}$ be the Lie algebra of allowable point symmetry operators for the Schrödinger equation*

$$2\lambda\partial_t\Psi = \frac{1}{\sqrt{g}}\partial_{y^i}\left(g^{ij}\sqrt{g}\,\partial_{y^j}\Psi\right) \tag{4.9}$$

on an n-dimensional Riemannian manifold V^n, and let m be the dimension of the vector space of type IV *symmetries of this equation. Then there exists a coordinate system $\{x^1,\cdots,x^m,y^{m+1},\cdots,y^n\}$ on V^n with respect to which* (4.9) *takes the form*

$$2\lambda\partial_t\Psi = \sum_{a=1}^{m}\partial_{x^a x^a}\Psi + \frac{1}{\sqrt{g}}\sum_{i,j=m+1}^{n}\partial_{y^i}\left(g^{ij}(\mathbf{y})\sqrt{g}\,\partial_{y^j}\Psi\right). \tag{4.10}$$

As a vector space

$$\hat{\mathscr{G}} = \hat{\mathscr{A}}_m \oplus \hat{\mathscr{B}} \oplus \hat{\mathscr{C}}_q \tag{4.11}$$

where

1) $\dim\hat{\mathscr{A}}_m = 2m$ *and* $\hat{\mathscr{A}}_m$ *has the basis* ∂_{x^a}, $t\partial_{x^a} - \lambda x^a\Psi\partial_\Psi$, $a = 1,\cdots,m$.
2) *Let $\hat{\mathscr{D}}$ be the Lie algebra of symmetries of type* V. *Then $\hat{\mathscr{B}}$ is a subspace of type* V *symmetries such that* $\hat{\mathscr{D}} = \hat{\mathscr{B}} \oplus \operatorname{span}(\partial_{x^a} : a = 1,\cdots,m)$.
3) $\hat{\mathscr{C}}_q$ *corresponds to exactly one of the following possibilities*:
   a) $\hat{\mathscr{C}}_2$: *basis* $\partial_t, \Psi\partial_\Psi$.
   b) $\hat{\mathscr{C}}_3$: *basis* $\partial_t, t\partial_t + \frac{1}{2}x^b\partial_{x^b} + \partial_{y^{m+1}}, \Psi\partial_\Psi$.
   Here, $g^{ij} = \exp(\frac{1}{2}y^{m+1})G^{ij}(y^{m+2},\cdots,y^n)$, $i,j = m+1,\cdots,n$.
   c) $\hat{\mathscr{C}}_4$, *basis* ∂_t, $t\partial_t + \frac{1}{2}(\sum_a x^a\partial_{x^a} + y^{m+1}\partial_{y^{m+1}})$,

$$t^2\partial_t + t\left(\sum_a x^a\partial_{x^a} + y^{m+1}\partial_{y^{m+1}}\right) + \left[-\frac{\lambda}{2}\left(\sum_a x^a x^a + y^{m+1}y^{m+1}\right) - nt\right]\Psi\partial_\Psi.$$

Here $g^{m+1,k} = \delta^{m+1,k}$, $1 \leqq k \leqq n$, *and* $g^{ij} = (y^{m+1})^{-2}G^{ij}(y^{m+2},\cdots,y^n)$, $i,j = m+2,\cdots,n$.

To provide examples of related evolution equations we will consider one case in detail: the free particle Hamilton–Jacobi equation

$$2\lambda p_t = p_{y^1}^2 + p_{y^2}^2. \tag{4.12}$$

The symmetry algebra $\mathscr{G}$ of this equation (the Schrödinger algebra) is nine-dimensional, with basis

$$\begin{aligned} &P_i = \partial_{y^i}, \qquad B_i = -t\partial_{y^i} + \lambda y^i\partial_W, \qquad i = 1,2,\\ &M = y^1\partial_{y^2} - y^2\partial_y^1, \qquad E = 2\lambda\partial_W,\\ &K_2 = -t^2\partial_t - t\left(y^1\partial_{y^1} + y^2\partial_{y^2}\right) + \frac{\lambda}{2}\left(y^1y^1 + y^2y^2\right)\partial_W,\\ &K_{-2} = \partial_t, \quad D = y^1\partial_{y^1} + y^2\partial_{y^2} + 2t\partial_t. \end{aligned} \tag{4.13}$$

We will identify two related evolution equations if one can be transformed into the other through an action of the Schrödinger group [4, Chap. 2]. Thus to classify the possible related evolution equations we first determine a complete set of orbit representatives for one-dimensional subalgebras of $\mathcal{G}$ (under the adjoint action of the Schrödinger group). We then compute the related evolution equation, if any, associated

TABLE 1.

Evolution equations related to $2\lambda p_t = p_{y^1}^2 + p_{y^2}^2$.

	Operator	Coordinates	Potential
1.	$K_{-2} - K_2$	$t = \tan s$ $y^i = x^i/\cos s$ $W = \tilde{W} - \frac{\lambda}{2}(x^1x^1 + x^2x^2)\tan s$	$\lambda^2(x^1x^1 + x^2x^2)$
2.	$K_{-2} - K_2 + \beta M$	$t = \tan s$ $y^1 = [x^1 \sin\beta s + x^2 \cos\beta s]/\cos s$ $y^2 = [-x^1 \cos\beta s + x^2 \sin\beta s]/\cos s$ $W = \tilde{W} - \frac{\lambda}{2}(x^1x^1 + x^2x^2)\tan s$	$2\lambda\beta(x^1\tilde{p}_{x^2} - x^2\tilde{p}_{x^1})$ $+\lambda^2(x^1x^1 + x^2x^2)$
3.	$K_{-2} - K_2 + M + \gamma B_1$	$t = \tan s$ $y_1 = x^1 \tan s + x^2 - \frac{\gamma s}{2}\tan s + \frac{\gamma}{2}$ $y_2 = -x^1 + x^2 \tan s + \frac{\gamma s}{2}$ $W = \tilde{W} - \frac{\lambda}{2}\Big[(x^1x^1 + x^2x^2)\tan s$ $-\gamma x^1 s \tan s - \gamma x^2 s$ $+\frac{\gamma^2}{4}s^2 \tan s - \frac{3\gamma^2 s}{4}\Big]$	$2\lambda\left(x^1\tilde{p}_{x^2} - x^2\tilde{p}_{x^1} - \frac{\gamma}{2}\tilde{p}_{x^1}\right)$ $+\lambda^2(x^1x^1 + x^2x^2 - \gamma x^2 - 3\gamma^2/4)$
4.	$D + \beta M$	$t = e^{2s}$ $y^1 = \sqrt{2}\, e^s(x^1 \cos\beta s - x^2 \sin\beta s)$ $y^2 = \sqrt{2}\, e^s(x^1 \sin\beta s + x^2 \cos\beta s)$ $W = \tilde{W}$	$2\lambda(x^1\tilde{p}_{x^1} + x^2\tilde{p}_{x^2})$ $+2\lambda\beta(x^1\tilde{p}_{x^2} - x^2\tilde{p}_{x^1})$
5.	$-K_2 - M$	$t = -s^{-1}$ $y^1 = \frac{x^1}{2}\sin s + \frac{x^2}{s}\cos s$ $y^2 = \frac{x^1}{s}\cos s - \frac{x^2}{s}\sin s$ $W = \tilde{W} + \lambda(x^1x^1 + x^2x^2)/2s$	$2\lambda(x^2\tilde{p}_{x^1} - x^1\tilde{p}_{x^2})$
6.	$-K_2 - P_1$	$t = -s^{-1}$ $y^1 = -\frac{s}{2} + \frac{x^1}{s}$ $y^2 = \frac{x^2}{s}$ $W =$ $\tilde{W} + \lambda\left[\frac{(x^1x^1 + x^2x^2)}{2s} - \frac{s^3}{24} + \frac{x^1 s}{2}\right]$	$-2\lambda^2 x^1$
7.	K_{-2}	$t = s$ $y^j = x^j$ $W = \tilde{W}$	0
8.	M	———	
9.	$P_1 + B_2$	———	
10.	P_1	———	

with the orbit representative. A list of orbit representatives is given on [4, p. 124]. Our final results are presented in Table 1. We express any related equation in terms of Cartesian coordinates:

$$2\lambda\tilde{p}_s=\tilde{p}_{x^1}^2+\tilde{p}_{x^2}^2+2\lambda\left(\tilde{\xi}^1\tilde{p}_{x^1}+\tilde{\xi}^2\tilde{p}_{x^2}\right)+\lambda^2\tilde{V}. \tag{4.14}$$

Thus the related equation can be determined by merely listing its associated potential. Note that the operators corresponding to orbits 8–10 are not associated with an evolution equation since they contain no term in ∂_t. The analogous results for the free-particle Schrödinger equation are virtually identical with those presented here for the Hamilton–Jacobi equation.

REFERENCES

[1] A. S. FOKAS, *Invariants, Lie-Bäcklund operators, and Bäcklund transformations*, Ph.D. Thesis, California Institute of Technology, Pasadena, 1979.

[2] P. J. OLVER, *Applications of Lie groups to differential equations*, mimeographed lecture notes, Mathematical Institute, Oxford, 1980.

[3] ———, *Symmetry groups and group invariant solutions of partial differential equations*, J. Diff. Geom., 14 (1979), pp. 497–542.

[4] W. J. MILLER, *Symmetry and Separation of Variables*, Addison-Wesley, Reading, MA, 1977.

[5] A. MUNIER, J. R. BURGAN, M. FEIX AND E. FIJALKOW, *Schrödinger equation with time-dependent boundary conditions*, J. Math. Phys., 22 (1981), pp. 1219-1223.

[6] S. ROSENCRANS, *Perturbation algebra of an elliptic operator*, J. Math. Anal. Appl., 56 (1976), pp. 317–329.

[7] R. L. ANDERSON AND N. H. IBRAGIMOV, *Lie–Bäcklund Transformations in Applications*, SIAM Studies in Applied Mathematics$_1$, Philadelphia, 1979.

[8] A. S. FOKAS AND R. L. ANDERSON, *Group theoretical nature of Bäcklund transformations*, Lett. Math. Phys., 3 (1979), pp. 117-126.

[9] S. KUMEI AND G. W. BLUMAN, *When nonlinear differential equations are equivalent to linear differential equations*, SIAM J. Appl. Math., 42 (1982), pp. 1157-1173.

[10] L. EISENHART, *Continuous Groups of Transformations*, (reprint), Dover, New York, 1961.

[11] G. W. BLUMAN AND J. D. COLE, *Similarity Methods for Differential Equations*, Applied Mathematics Sci. 13, Springer-Verlag, New York, 1974.

[12] L. LANDAU AND E. LIFSHITZ, *Quantum Mechanics, Non-Relativistic Theory* (from Russian), Addison-Wesley, Reading, MA, 1958.

[13] L. J. CHANDLER, *Separation of variables by the symmetry method for second order linear partial differential equations*, Ph.D. Thesis, Univ. New Mexico, Albuquerque, 1980.

[14] R. KUWABARA, *On the symmetry algebra of the Schrödinger wave equation*, Math. Japonica, 22 (1977), pp. 243-252.

[15] C. BOYER, *The maximal kinematical invariance group for an arbitrary potential*, Helv. Phys. Acta, 47 (1974), pp. 589-605.

[16] D. R. TRUAX, *Symmetry of time-dependent Schrödinger equations*, I., J. Math. Phys., 22 (1981), pp. 1959-1964.

[17] A. S. FOKAS, *Group theoretical aspects of constants of motion and separable solutions in classical mechanics*, J. Math. Anal. Appl., 68 (1979), pp. 347-370.

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

GLOBAL BEHAVIOR FOR A CLASS OF NONLINEAR EVOLUTION EQUATIONS*

PAUL E. SACKS†

Abstract. We derive decay rates and study the asymptotic behavior of solutions of a class of scalar quasilinear reaction diffusion equations of degenerate type.

We study here the initial and boundary value problem

$$\begin{aligned} &v_t=\Delta|v|^{m-1}v+\lambda|v|^{p-1}v, && x\in\Omega,\quad t>0,\\ &v(x,0)=v_0(x), && x\in\Omega,\\ &v(x,t)=0, && x\in\partial\Omega,\quad t>0 \end{aligned} \tag{0.1}$$

where Ω is a bounded domain in $\mathbb{R}^N$, $\lambda\ge 0$, $m>1, p\ge 1$.

If $p<m$, or $p=m$ and Ω is sufficiently small, this problem is known to have a global time solution for $v_0\in L^\infty(\Omega)$ [18] or $v_0^m\in H_0^1(\Omega)$ [12]. In the remaining cases there exists a local time solution if $v_0\in L^\infty(\Omega)$, and $v_0^m\in H_0^1(\Omega)$ provided that p is not too large. These solutions may blow up in finite time [12], [18].

The goal of this article is to prove some decay estimates for solutions of (0.1), to prove solvability of (0.1) for a larger class of initial values, and more specifically to study the behavior of solutions of (0.1) as $t\to 0^+$ and $t\to\infty$.

In some of our considerations a crucial role is played by the first eigenvalue of the Dirichlet problem

$$-\Delta\rho=\lambda\rho,\quad x\in\Omega,\qquad \rho=0,\quad x\in\partial\Omega. \tag{0.2}$$

We denote the first eigenvalue by λ_1 and the corresponding eigenfunction by ρ_1, with the normalization $\rho_1>0$ in Ω and $\|\rho_1\|_{L^1(\Omega)}=1$. It is possible to interpret λ_1 as a measure of the size of the domain Ω.

Here is a summary of the main results.

(i) $p<m$. In this case a certain regularizing effect, known (cf. Aronson and Benilan [2]) in the case $\lambda=0$, continues to hold. A consequence of this is an estimate for $\|v(\cdot,t)\|_{L^\infty(\Omega)}$ which is independent of v_0. We exploit this to prove the existence of solutions to (0.1) with $v_0\in L^1(\Omega)$. We consider the asymptotic behavior when $v_0\ge 0$, $v_0\not\equiv 0$, and show that there is a unique positive steady state solution of (0.1) to which all solutions of (0.1) tend as $t\to\infty$.

(ii) $p=m$, $\lambda<\lambda_1$. We have again a decay estimate which is independent of the initial state. Solvability of (0.1) is proved for $v_0\in L^q(\Omega)$ $q>1$. All solutions of (0.1) tend to zero as $t\to\infty$.

(iii) $p=m$, $\lambda=\lambda_1$. We obtain an estimate for $\|v(\cdot,t)\|_{L^\infty(\Omega)}$ which now depends on v_0. Solvability of (0.1) is again proved for $v_0\in L^q(\Omega) q>1$. As $t\to\infty$, solutions of (0.1) tend to $\theta\rho_1^{1/m}$ for some constant θ depending on v_0.

(iv) $p=m$, $\lambda>\lambda_1$. Previous results [12], [25] show that (0.1) has no nontrivial positive solutions which exist for all time. We do not consider this case.

(v) $p>m$. Here solutions of (0.1) may or may not exist for all time. We restrict attention to those initial values for which the solution of (0.1) remains uniformly

*Received by the editors October 20, 1982, and in revised form August 10, 1983.

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

bounded for all time, and give sufficient conditions that the solution tend to zero as $t \to \infty$. In particular $v \equiv 0$ is an asymptotically stable equilibrium solution of (0.1), while any positive equilibrium solution is unstable.

Besides the papers cited already we mention the works of Gurtin and MacCamy [14] and Aronson and Peletier [4] which include some discussion of the case $p=1$, $m>1$. The problem (0.1) with $\lambda<0$ is studied by Bertsch, Nanbu and Peletier [27], while Aronson, Crandall and Peletier [3] consider (0.1) in one dimension with a different type of nonlinearity on the right-hand side. For the semilinear heat equation $m=1$, $p>1$ there is of course a large literature; see for example Matano [20], Weissler [26], or Lions [19] for some results related to those presented here. See Alikakos and Rostamian [28], [29] for the corresponding problem with Neumann boundary conditions.

1. We begin by making precise some notions of solution of (0.1) and related equations. For simplicity of notation we will always write v^m instead of $|v|^{m-1}v$.

Given $v_0 \in L^1(\Omega)$ and $f \in L^1(Q_T)$, $Q_T = \Omega \times (0,T)$, the problem

$$\begin{aligned} &v_t = \Delta v^m + f, && (x,t) \in Q_T, \\ &v(x,0) = v_0(x), && x \in \Omega, \\ &v(x,t) = 0, && x \in \partial\Omega \end{aligned} \tag{1.1}$$

has a solution in the sense of nonlinear semigroups (see [8], [11], [22]) which is also known as the mild solution of (1.1). We write $v(\cdot,t) = S(t; v_0, f)$ for this solution of (1.1). If $f \equiv 0$ then it is usual to write $S(t; v_0, 0) = S(t)v_0$, and the collection of maps $\{S(t)\}_{t \geq 0}$ is a semigroup of nonlinear nonexpansive operators on $L^1(\Omega)$ with $S(0) =$ identity. More generally, one has

$$\|S(t; v_0, f) - S(t; \hat{v}_0, \hat{f})\|_{L^1(\Omega)} \leq \|v_0 - \hat{v}_0\|_{L^1(\Omega)} + \int_0^t \|f(s) - \hat{f}(s)\|_{L^1(\Omega)} ds. \tag{1.2}$$

If v is a bounded function on some subset $Q \subset Q_T$, then v satisfies (1.1) in the sense of distributions on Q. Also $v \in C([0,T]; L^1(\Omega))$ so the initial condition has a meaning. In general, however, the boundary condition may not be satisfied in any ordinary sense.

For the problem (0.1) we will use two definitions of solution.

DEFINITION 1. A measurable function v is a *mild solution* of (0.1) on $[0,T]$ if

(i) $v^p \in L^1(Q_T)$;

(ii) $v(\cdot,t) = S(t; v_0, \lambda v^p)$, $0 \leq t \leq T$.

DEFINITION 2. A measurable function v is a *weak solution* (0.1) on $[0,T]$ if

(i) $v \in L^\infty(Q_T)$,

$$\text{(ii)} \quad \int_0^t \int_\Omega (v\rho_t + v^m \Delta\rho + \lambda v^p \rho)\, dx\, dt = \int_\Omega v(x,t)\rho(x,t)\, dx - \int_\Omega v_0(x)\rho(x,0)\, dx \tag{1.3}$$

for all $\rho \in C^2(Q_T) \cap C^1(\bar{Q}_T)$, $\rho(x,t) = 0$ for $x \in \partial\Omega$.

We may also define a weak solution of (1.1) by replacing λv^p by f in (1.3). The following facts are then more or less well known: If $v_0 \in L^\infty(\Omega)$ and $f \in L^\infty(Q_T)$, then (1.1) has exactly one weak solution on $[0,T]$ which coincides with $S(t; v_0, f)$. For the existence assertion see, e.g. [3] or [18], and for the uniqueness [3] or [6]. The fact that the weak solution agree with the semigroup solution is seen by observing in the course of the construction of $S(t; v_0, f)$ that it also satisfies (1.3) under these hypotheses.

Using the above remarks, and inequality (1.2), it is easy to prove the following, concerning solutions of (0.1).

PROPOSITION 1.1. (i) *If v is a weak solution of* (0.1) *on* $[0,T]$, *then it is also a mild solution of* (0.1) *on* $[0,T]$.

(ii) *If v is a mild solution of* (0.1) *on* $[0,T]$, *and* $v\in L^\infty(Q_T)$, *then v is a weak solution of* (0.1) *on* $[0,T]$.

(iii) (0.1) *has at most one weak solution on* $[0,T]$.

Regarding the uniqueness of mild solutions of (0.1) we have the following.

PROPOSITION 1.2. *Let* $v_i(x,t)$, $i=1,2$ *satisfy*

(i) v_i *is a weak solution of* (0.1) *on* $[\tau,T]$ *for every* $\tau>0$;

(ii) $\|v_i(\cdot,t)\|_{L^\infty(\Omega)}\leq Ct^{-\alpha}$ *for some* $\alpha\in(0,1/(p-1))$, $0<t\leq T$;

(iii) $\lim_{t\to 0}\|v_1(\cdot,t)-v_2(\cdot,t)\|_{L^1(\Omega)}=0$.

Then $v_1\equiv v_2$.

Proof. By Proposition 1.1

$$v_i(\cdot,t)=S(t-\tau;v_i(\cdot,\tau),\lambda v_i^p),\qquad i=1,2,\ \tau>0;$$

hence by (1.2)

$$\|v_1(\cdot,t)-v_2(\cdot,t)\|_{L^1(\Omega)}\leq\|v_1(\cdot,\tau)-v_2(\cdot,\tau)\|_{L^1(\Omega)}+\lambda\int_\tau^t\|v_1^p(\cdot,s)-v_2^p(\cdot,s)\|_{L^1(\Omega)}ds.$$

Letting $\tau\to 0$ gives

$$\|v_1(\cdot,t)-v_2(\cdot,t)\|_{L^1(\Omega)}\leq\int_0^t f(s)\|v_1(\cdot,s)-v_2(\cdot,s)\|_{L^1(\Omega)}ds$$

with

$$f(s)=p\lambda\max\left(\|v_1(\cdot,s)\|_{L^\infty(\Omega)}^{p-1},\|v_2(\cdot,s)\|_{L^\infty(\Omega)}^{p-1}\right)\leq C\lambda ps^{-\alpha(p-1)}.$$

By hypothesis (ii), $f\in L^1(0,T)$; hence the conclusion follows by Gronwall's inequality. □

In particular Proposition 1.2 implies the uniqueness of mild solutions of (0.1) in the class of functions satisfying the decay estimate in hypothesis (ii). That this condition is not unduly restrictive is seen from the following, which is our main existence theorem for problem (0.1).

THEOREM 1.3. *Assume either*

$$p<m,\qquad v_0\in L^1(\Omega)\tag{1.4}$$

or

$$p=m,\quad \lambda\leq\lambda_1,\quad q>1,\quad v_0\in L^q(\Omega).\tag{1.5}$$

Then (0.1) *has a mild solution* $v(x,t)$ *on* $[0,T]$ *for any* $T>0$, $v\in C((0,T]\times\bar\Omega)$, *and the following estimates are valid.*

(i) *If* $p<m$, *there exists a constant* $\mathcal{C}_1=\mathcal{C}_1(N,\Omega,m,p)$ *such that*

$$\|v(\cdot,t)\|_{L^\infty(\Omega)}\leq\mathcal{C}_1\left(\lambda^{1/(m-p)}+\frac{1}{((m-1)t)^{1/(m-1)}}\right).\tag{1.6}$$

(ii) *If* $p=m$ *and* $\lambda<\lambda_1$, *there exists a constant* $\mathcal{C}_2=\mathcal{C}_2(N,\Omega,m,\sigma)$, $\sigma>\max(1/m,N/4m)$, *such that*

$$\|v(\cdot,t)\|_{L^\infty(\Omega)}\leq\frac{\mathcal{C}_2}{((m-1)t)^{1/(m-1)}}\left(\frac{\lambda^\sigma}{(\lambda_1-\lambda)^{1/(m-1)}}+1\right). \tag{1.7}$$

(iii) *If* $p=m$ *and* $\lambda=\lambda_1$, *there exists a* $\mathcal{C}_3=\mathcal{C}_3(N,\Omega,m,q)$, *such that*

$$\|v(\cdot,t)\|_{L^\infty(\Omega)}\leq\mathcal{C}_3\left(\|v(\cdot,t)\|_{L^q(\Omega)}+\frac{1}{((m-1)t)^{1/(m-1)}}\right). \tag{1.8}$$

(iv) *If* $p=m$, *and* $\lambda\leq\lambda_1$, *there exists a constant* $\mathcal{C}_4=\mathcal{C}_4(N,\Omega,m,q,\lambda,\|v_0\|_{L^q(\Omega)})$ *and a number* $\varepsilon>0$ *such that*

$$\|v(\cdot,t)\|_{L^\infty(\Omega)}\leq\mathcal{C}_4\left(1+\frac{1}{t^{1/(m-1+\varepsilon)}}\right). \tag{1.9}$$

Remarks. (i) Note especially the fact that the estimates (1.6) and (1.7) are independent of v_0. This is already known for $\lambda=0$; see [11] or [22] for example. No such estimate is possible in the case $p=m$, $\lambda=\lambda_1$, as results of §3 will show. If $\Omega=\mathbb{R}^n$ then again there can be no such estimate.

(ii) The estimates (1.6) and (1.9) indicate that one may expect condition (ii) of Proposition (1.2) to be satisfied under certain circumstances. In particular we get uniqueness in the class in which we prove existence.

(iii) It is known [4] that for $\lambda=0$, solutions of (0.1) decay like $t^{1/(m-1)}$ for large t, and this is the best possible exponent. Certainly no better asymptotic decay rate can hold when $\lambda>0$. For small t, this exponent may be improved, e.g. in (1.9), but the constant will depend on some norm of v_0. Precise results of this type may be derived using formula (2.13) in the next section.

(iv) In the estimate (1.9) the constant $\mathcal{C}_4$ tends to zero with $\|v_0\|_{L^q(\Omega)}$. There can be no such result when $p<m$, since, as we shall show in §3, all initial states $v_0\geq0$, $v_0\not\equiv0$ eventually evolve into the same positive equilibrium solution of (0.1). An estimate on the solution involving the size of the initial value is possible on any fixed bounded time interval.

2. From [18] we have the following local existence and continuation theorem.

THEOREM 2.1. *Let* $v_0^m\in H_0^1(\Omega)\cap L^\infty(\Omega)$. *Then there exists a time*

$$T=T(v_0)\geq\frac{1}{\lambda(p-1)\|v_0\|_{L^\infty(\Omega)}^{p-1}}$$

($T=\infty$ *not excluded*) *such that* (0.1) *has a unique weak solution on* $[0,T)$. *If* $T<\infty$, *then*

$$\lim_{t\to T^-}\|v(\cdot,t)\|_{L^\infty(\Omega)}=\infty. \tag{2.1}$$

If $T'<T$, *then* v *satisfies*

$$v\in C((0,T')\times\Omega), \tag{2.2}$$

$$v^m\in L^\infty(0,T';H_0^1(\Omega)), \tag{2.3}$$

$$(v^{(m+1)/2})_t\in L^2(0,T';L^2(\Omega)). \tag{2.4}$$

Remarks. (i) The lower bound on the existence time follows directly by comparison with the initial value problem

$$v' = \lambda v^p, \qquad v(0) = \pm \|v_0\|_{L^\infty(\Omega)}.$$

(ii) There is no restriction on p or λ here. If $\lambda \leq 0$, then $T(v_0) = \infty$ clearly. If $p < m$ or $p = m$ and λ is sufficiently small, the results of [12], [18] again imply that $T(v_0) = \infty$, but this fact will be rederived here, in the course of proving more exact estimates.

(iii) If we do not require $v_0^m \in H_0^1(\Omega)$, then Theorem 2.1 still holds except (2.4) is not valid, and (2.3) must be replaced by

$$v^m \in L^2(0, T; H_0^1(\Omega)).$$

(iv) We actually require a more precise regularity result, corresponding to (2.2). See [10] or [25]. The point is that the modulus of continuity of solutions of (0.1) depends locally only on the L^∞ norm and data.

THEOREM 2.2. *Let $v_0^m \in H_0^1(\Omega) \cap L^\infty(\Omega)$, and v be the corresponding weak solution of* (0.1). *Suppose $|v(x,t)| \leq M$ for $t \in [\tau, T]$, $\tau > 0$. Then the modulus of continuity of v on $[\tau, T] \times \bar{\Omega}$ depends only on*

$$N, \quad m, \quad p, \quad \lambda, \quad M, \quad \tau. \tag{2.5}$$

The following result is the key to all of our estimates.

PROPOSITION 2.3. *Let $p \leq m$, $\lambda \geq 0$, $v_0^m \in H_0^1(\Omega) \cap L^\infty(\Omega)$, $v_0 \geq 0$ and v be the corresponding solution of* (0.1). *Then for $t \in (0, T(v_0))$,*

$$-\Delta v^m(\cdot, t) \leq \lambda v^p(\cdot, t) + \frac{1}{(m-1)t} v(\cdot, t) \tag{2.6}$$

in $\mathcal{D}'(\Omega)$.

Remark. The case $\lambda = 0$ is essentially proved in Aronson and Benilan [2]. The fact that the method generalizes to equations with lower order terms has also been observed by Bertsch and Peletier [30].

Proof. Fix $T' \in (0, T(v_0))$ and suppose $|v| \leq M$ for $t \leq T'$. We may regard v as the solution of (0.1) with λv^p replaced by $\lambda F_M(v)$, where

$$F_M(v) = \begin{cases} v^p, & v \leq 2M, \\ (2M)^p, & v \geq 2M. \end{cases}$$

Consider now the problem

$$\begin{aligned} &w_t = \Delta w^m + \lambda F_M(w), && x \in \Omega, \quad t > 0, \\ &w(x,0) = v_0(x) + \varepsilon, && x \in \Omega, \\ &w(x,t) = \varepsilon, && x \in \partial\Omega. \end{aligned} \tag{2.7(ε)}$$

This problem has a classical solution v_ε, which exists for all $t \geq 0$ (see [17, Chap. V]). By comparison [24], v_ε decreases as $\varepsilon \to 0$ to a limit function $\hat{v}$, and by standard arguments (e.g. [4]) $\hat{v}$ is the weak solution of (0.1) with λv^p replaced by $\lambda F_M(v)$, thus $v = \hat{v}$. (The uniqueness properties discussed in §1 continue to hold for this problem). Since v is continuous on $[\tau, T'] \times \bar{\Omega}$ for every $\tau > 0$, $v_\varepsilon \to v$ uniformly on any such set, by Dini's theorem. Thus there exists $\varepsilon_0 > 0$ such that for $\varepsilon \leq \varepsilon_0$, v_ε solves (2.7(ε)) with $\lambda F_M(w)$ replaced by λw^p.

For $\varepsilon \leq \varepsilon_0$ the following calculation is therefore valid. Set

$$z(x,t)=tv_{\varepsilon t}(x,t)+\frac{v_\varepsilon(x,t)}{m-1}.$$

Then $z(x,0)\geq 0$, $z(x,t)\geq 0$ for $x\in\partial\Omega$, and if we define the elliptic operator

$$\mathfrak{L}w=\Delta m v_\varepsilon^{m-1}w+\lambda p v_\varepsilon^{p-1}w,$$

then $z_t-\mathfrak{L}z\geq 0$ by direct computation. Hence $z\geq 0$ for $x\in\Omega$, $t\leq T'$ by standard comparison theorems [24], and $T'<T$ is arbitrary.

Therefore, using the equation (2.7(ε)),

$$-\Delta v_\varepsilon^m(\cdot,t)\leq \lambda v_\varepsilon^p(\cdot,t)+\frac{v_\varepsilon(\cdot,t)}{t(m-1)} \tag{2.8}$$

for $t\in(0,T(v_0))$ in a pointwise sense. Since $v_\varepsilon(\cdot,t)\to v(\cdot,t)$ uniformly, (2.6) holds. □

Remarks. (i) Since $v^m(\cdot,t)\in H_0^1(\Omega)$ for a.e. $t\in(0,T(v_0))$, the inquality (2.6) actually holds in the usual weak sense.

(ii) From the fact that $v^m(\cdot,t)$ is a positive subsolution of a certain elliptic equation we will derive estimates for $v(\cdot,t)$ in $L^\infty(\Omega)$. In the paper [5] similar types of arguments are used for the study of (0.1) in $\mathbb{R}^N$ with $\lambda=0$ and a very large class of initial values.

PROPOSITION 2.4. *Assume the hypotheses of Proposition* 2.3 *with* $\lambda\leq\lambda_1$ *if* $p=m$, *and let* $t\in(0,T(v_0))$.

(i) *If* $p<m$ *there exists* $\mathcal{C}_1=\mathcal{C}_1(N,\Omega,m,p)$ *such that*

$$\|v(\cdot,t)\|_{L^\infty(\Omega)}\leq \mathcal{C}_1\left(\lambda^{1/(m-p)}+\frac{1}{((m-1)t)^{1/(m-1)}}\right). \tag{2.9}$$

(ii) *If* $p=m$, $\lambda<\lambda_1$ *there exists* $\mathcal{C}_2=\mathcal{C}_2(N,\Omega,m,\sigma)$, $\sigma>\max(1/m,N/4m)$ *such that*

$$\|v(\cdot,t)\|_{L^\infty(\Omega)}\leq \frac{\mathcal{C}_2}{((m-1)t)^{1/(m-1)}}\left(\frac{\lambda^\sigma}{(\lambda_1-\lambda)^{1/(m-1)}}+1\right). \tag{2.10}$$

(iii) *If* $p=m$, $\lambda=\lambda_1$ *there exists* $\mathcal{C}_3=\mathcal{C}_3(N,\Omega,m,q)$, $q>1$, *such that*

$$\|v(\cdot,t)\|_{L^\infty(\Omega)}\leq \mathcal{C}_3\left(\|v(\cdot,t)\|_{L^q(\Omega)}+\frac{1}{((m-1)t)^{1/(m-1)}}\right). \tag{2.11}$$

Proof. It is enough to prove these estimates for those t for which $v^m(\cdot,t)\in H_0^1(\Omega)$.

We will use the following estimate from the regularity theory for divergence from elliptic equations (See, for example, [16, Chap. 2 appendix], for a nice explanation of these matters.)

If $u\in H_0^1(\Omega)$, $u\geq 0$, $-\Delta u\leq f_1+f_2$, $f_i\in L^{r_i}(\Omega)$, $r_i>\max(1,N/2)$, then

$$\|u\|_{L^\infty(\Omega)}\leq \omega(r_1)\|f_1\|_{L^{r_1}(\Omega)}+\omega(r_2)\|f_2\|_{L^{r_2}(\Omega)}, \tag{2.12}$$

where the constant $\omega(r)$ depends also on N and Ω, and is nonincreasing in r.

Set $w=v(\cdot,t)$, $\gamma=1/((m-1)t)$. By Proposition 2.3 and (2.12) we have immediately

$$\|w\|^m_{L^\infty(\Omega)}\leq \omega(r_1)\lambda\|w^p\|_{L^{r_1}(\Omega)}+\omega(r_2)\gamma\|w\|_{L^{r_2}(\Omega)}.$$

Now for $r_1>q/p$, $r_2>q$, $q\geq 1$,

$$\|w^p\|_{L^{r_1}(\Omega)}\leq\|w\|_{L^\infty(\Omega)}^{p-q/r_1}\|w\|_{L^q(\Omega)}^{q/r_1}\leq\delta_1\|w\|_{L^\infty(\Omega)}^m+\frac{\eta(s_1)}{\delta_1^{1/(s_1-1)}}\|w\|_{L^q(\Omega)}^{(mq/r_1)/(m-p+q/r_1)},$$

$$\|w\|_{L^{r_2}(\Omega)}\leq\|w\|_{L^\infty(\Omega)}^{1-q/r_2}\|w\|_{L^q(\Omega)}^{q/r_2}\leq\delta_2\|w\|_{L^\infty(\Omega)}^m+\frac{\eta(s_2)}{\delta_2^{1/(s_2-1)}}\|w\|_{L^q(\Omega)}^{(mq/r_2)/(m-1+q/r_2)}$$

with

$$\delta_1,\delta_2>0,\quad s_1=\frac{m}{p-q/r_1},\quad s_2=\frac{m}{1-q/r_2},\quad \eta(s)=\frac{s-1}{s^{s/(s-1)}}.$$

Choose $\delta_1=1/4\lambda\omega(r_1)$, $\delta_2=1/4\gamma\omega(r_2)$ to find

$$\begin{aligned}(2.13)\qquad \|w\|_{L^\infty(\Omega)}\leq{}&\Gamma(m,p,q,r_1)\lambda^{1/(m-p+q/r_1)}\|w\|_{L^q(\Omega)}^{(q/r_1)/(m-p+q/r_1)}\\&+\Gamma(m,1,q,r_2)\gamma^{1/(m-1+q/r_2)}\|w\|_{L^q(\Omega)}^{(q/r_2)/(m-1+q/r_2)}\end{aligned}$$

where

$$\Gamma(m,p,q,r)=\left(2(4^{1/(s-1)})\eta(s)\right)^{1/m}\omega(r)^{1/(m-p+q/r)},\qquad s=\frac{m}{p-q/r}.$$

If $p<m$, then we may let r_1, $r_2\to\infty$ to obtain (2.9) with

$$\mathcal{C}_1=\max(\Gamma(m,p,1,\infty),\Gamma(m,1,1,\infty)).$$

Next, if $p=m$, $\lambda<\lambda_1$, we first obtain an a priori estimate for $\|w\|_{L^{2m}(\Omega)}$. Multiplying 2.6 by w^m and using Poincaré's inequality gives

$$(2.14)\qquad (\lambda_1-\lambda)\int_\Omega w^{2m}\leq\gamma\int w^{m+1}\leq\gamma\|w\|_{L^{2m}(\Omega)}^{m+1}|\Omega|^{(m-1)/2m}$$

where $|\Omega|$ denotes the Lebesgue measure of Ω. Thus

$$\|w\|_{L^{2m}(\Omega)}\leq\left(\frac{\gamma|\Omega|^{(m-1)/2m}}{\lambda_1-\lambda}\right)^{1/(m-1)}.$$

We now let $r_2\to\infty$ and substitute (2.14) into (2.13), with $q=2m$ to obtain (2.10) with

$$\mathcal{C}_2=\max\left(\Gamma(m,m,2m,r_1)|\Omega|^{1/2m},\Gamma(m,1,1,\infty)\right).$$

Finally, if $p=m$, $\lambda=\lambda_1$ we again let $r_2\to\infty$ and take

$$\mathcal{C}_3=\max\left(\lambda_1^{r_1/q}\Gamma(m,m,q,r_1),\Gamma(m,1,1,\infty)\right)$$

for some fixed $r_1>\max(1,N/2,q/m)$. □

Finally, before giving the proof of Theorem 1.3 we need some control over the behavior of solutions near $t=0$.

PROPOSITION 2.5. *Let the hypotheses of Proposition* 2.3 *be satisfied with* $\lambda\leq\lambda_1$ *if* $p=m$. *Then* $T(v_0)=\infty$.

Furthermore

(i) *If* $p<m$ *there exists a constant* $\mathcal{C}_5=\mathcal{C}_5(N,\Omega,m,p,\lambda,\|v_0\|_{L^1(\Omega)})$ *and a number* $\varepsilon>0$ *such that*

$$\int_0^1\int_\Omega v^{p+\varepsilon}\,dx\,dt\le\mathcal{C}_5.$$

(ii) *If* $p=m$, $\lambda\le\lambda_1$, $q>1$, *then there exists a constant* $\mathcal{C}_6=\mathcal{C}_6(N,\Omega,m,q,\lambda,\|v_0\|_{L^q(\Omega)})$ *and a number* $\varepsilon>0$ *such that*

$$\int_0^1\int_\Omega v^{m+\varepsilon}\,dx\,dt\le\mathcal{C}_6.$$

Proof. (i) In this case the fact that $T(v_0)=\infty$ is clear from (2.1) and (2.9). By (1.2)

$$\begin{aligned}\|v(\cdot,t)\|_{L^1(\Omega)}&\le\|v_0\|_{L^1(\Omega)}+\lambda\int_0^t\|v^p(\cdot,s)\|_{L^1(\Omega)}ds\\&\le\|v_0\|_{L^1(\Omega)}+\lambda\int_0^t\|v(\cdot,s)\|^{p-1}_{L^\infty(\Omega)}\|v(\cdot,s)\|_{L^1(\Omega)}ds\\&\le\|v_0\|_{L^1(\Omega)}+\int_0^t f(s)\|v(\cdot,s)\|_{L^1(\Omega)}ds\end{aligned}$$

where

$$f(s)=\left(\mathcal{C}_1\left(\lambda^{1/(m-p)}+\frac{1}{((m-1)t)^{1/(m-1)}}\right)\right)^{p-1}$$

using (2.9). Since $p<m$, $f\in L^1(0,1)$. By Gronwall's inequality

$$\|v(\cdot,t)\|_{L^1(\Omega)}\le\|v_0\|_{L^1(\Omega)}\left(1+\left(\int_0^t f(s)\,ds\right)\exp\left(\int_0^t f(s)\,ds\right)\right)\le C\|v_0\|_{L^1(\Omega)}$$

for $t\le1$, where C depends only on m,p,λ,N and Ω. Thus, for $\varepsilon\in(0,m-p)$

$$\begin{aligned}\int_0^1\int_\Omega v^{p+\varepsilon}\,dx\,dt&\le\int_0^1\|v(\cdot,s)\|^{p+\varepsilon-1}_{L^\infty(\Omega)}\|v(\cdot,s)\|_{L^1(\Omega)}ds\\&\le C\|v_0\|_{L^1(\Omega)}\int_0^1 f(s)^{(p+\varepsilon-1)/(p-1)}ds\equiv\mathcal{C}_5.\end{aligned}$$

(ii) To begin with we wish to multiply the equation by qv^{q-1} and integrate over Q_t for some $t\in(0,T(v_0))$. This procedure may be justified by approximation, as in Proposition 2.3. We obtain

$$\text{(2.15)}\quad\int_\Omega v^q(x,t)\,dx-\int_\Omega v^q(x,0)\,dx+\frac{4mq(q-1)}{(m+q-1)^2}\int_0^t\int_\Omega|\nabla v^{(m+q-1)/2}|^2dx\,dt$$

$$=\lambda q\int_0^t\int_\Omega v^{m+q-1}\,dx\,dt$$

for $t\in(0,T(v_0))$. By interpolation

$$\int_0^t\int_\Omega v^{m+q-1}\,dx\,dt\le\delta\int_0^t\int_\Omega|\nabla v^{(m+q-1)/2}|^2dx\,dt+C(\delta)\left(\int_0^t\int_\Omega v^q\,dx\,dt\right)^\theta$$

with $\theta=(m+q-1)/q$, $\delta>0$, and $C(\delta)$ depends also on N,Ω,m and q. We choose $\delta=2m(q-1)/(m+q-1)^2$ to find

$$\int_\Omega v^q(x,t)\,dx\le\int_\Omega v^q(x,0)\,dx+C(\delta)\left(\int_0^t\int_\Omega v^q\,dx\,dt\right)^\theta \tag{2.16}$$

and

$$\frac{2mq(q-1)}{(m+q-1)^2}\int_0^t\int_\Omega|\nabla v^{(m+q-1)/2}|^2dx\,dt\le\int_\Omega v^q(x,0)\,dx+C(\delta)\left(\int_0^t\int_\Omega v^q\,dx\,dt\right)^\theta. \tag{2.17}$$

From (2.16) and (2.17) it follows that there exists $\tau_1\in(0,T(v_0))$ depending only on N,Ω,m,q,λ and $\|v_0\|_{L^q(\Omega)}$, and a constant C depending on the same quantities such that

$$\|v(\cdot,t)\|_{L^q(\Omega)}\le C,\qquad 0\le t\le\tau_1, \tag{2.18}$$

$$\int_0^{\tau_1}\int_\Omega v^{m+q-1}\,dx\,dt\le C. \tag{2.19}$$

Now, if $\lambda<\lambda_1$ then the desired conclusion follows from (2.19) and (2.10), with $\varepsilon=q-1$.

If $\lambda=\lambda_1$, then from (2.15) with q replaced by $m+1$ and the lower limit of integration replaced by τ_1,

$$\begin{aligned}\|v(\cdot,t)\|_{L^{m+1}(\Omega)}&\le\|v(\cdot,\tau_1)\|_{L^{m+1}(\Omega)},\qquad t\ge\tau_1\\&\le\|v(\cdot,\tau_1)\|_{L^\infty(\Omega)}|\Omega|^{1/(m+1)}\\&\le\mathcal{C}_3\left(C+\frac{1}{((m-1)\tau_1)^{1/(m-1)}}\right)|\Omega|^{1/(m+1)}\end{aligned}$$

by (2.11) and (2.18).

So by another application of (2.11), with $q=m+1$

$$\|v(\cdot,t)\|_{L^\infty(\Omega)}\le C\quad\text{for } t\ge\tau_1$$

from which the conclusion again follows. □

Proof of Theorem 1.3. Consider the case $p<m$. Let $v_{0n}\to v_0$ in $L^1(\Omega)$ $v_{0n}^m\in H_0^1\cap L^\infty(\Omega)$ for each n. For each n denote the solution of (0.1) with initial value v_{0n}, v_{0n}^+, v_{0n}^- by v_n, v_n^+, v_n^- respectively. (Here $v^+=\max(0,v)$, $v^-=\min(0,v)$.) By comparison

$$v_n^-\le v_n\le v_n^+$$

for $x\in\Omega$ $t\ge0$. The results of this section apply to v_n^+ and $-v_n^-$; in particular

$$\begin{aligned}\|v_n(\cdot,t)\|_{L^\infty(\Omega)}&\le\max\left(\|v_n^+(\cdot,t)\|_{L^\infty(\Omega)},\|v_n^-(\cdot,t)\|_{L^\infty(\Omega)}\right)\\&\le\beta(t),\qquad t>0,\end{aligned} \tag{2.20}$$

where $\beta(t)$ is the function on the right-hand side of (2.9).

Thus, on any compact subset of $\Omega\times(0,\infty)$ the sequence $\{v_n\}$ is uniformly bounded, and therefore equicontinuous by Theorem 2.2. By diagonalization we may then find a subsequence $n_k\to\infty$ and a limit function $v(x,t)$, $\|v(\cdot,t)\|_{L^\infty(\Omega)}\le\beta(t)$, such that $v_{n_k}\to v$ pointwise on $\Omega\times(0,\infty)$ and uniformly on compact subsets, so that $v\in C((0,\infty)\times\Omega)$.

Next, by Proposition 2.5, the sequence $\{v_{n_k}^p\}$ is uniformly bounded in $L^{1+\varepsilon/p}(Q_T)$ for any $T>0$. Since it is pointwise convergent, it is strongly convergent in $L^1(Q_T)$, and in particular $v^p \in L^1(Q_T)$.

Set $\hat{v}(\cdot,t)=S(t;v_0,\lambda v^p)$. By (1.2), since $v_{n_k}(\cdot,t)=S(t;v_{0n_k},\lambda v_{n_k}^p)$,

$$\|\hat{v}(\cdot,t)-v_{n_k}(\cdot,t)\|_{L^1(\Omega)} \le \|v_0-v_{n_k}(\cdot,0)\|_{L^1(\Omega)}+\lambda\int_0^t \|v^p(\cdot,s)-v_{n_k}^p(\cdot,s)\|_{L^1(\Omega)}ds.$$

Letting $n_k\to\infty$ one sees that $v=\hat{v}$, that is $v(\cdot,t)=S(t;v_0,\lambda v^p)$, so v is a mild solution of (0.1) on $[0,T]$.

For the case $p=m$, $\lambda\le\lambda_1$ we pick a sequence v_{0n} so that $v_{0n}\to v_0$ in $L^q(\Omega)$ and $v_{0n}^m\in H_0^1(\Omega)\cap L^\infty(\Omega)$, and define v_n, v_n^+, v_n^- as above.

Arguing as in the conclusion of the proof of Proposition 2.5, we may show that

$$\|v_n^\pm(\cdot,t)\|_{L^q(\Omega)}\le C_0, \qquad t>0 \tag{2.21}$$

for some fixed constant C_0 depending only on N,Ω,m,q,λ and $\|v_0\|_{L^q(\Omega)}$. Therefore, from (2.10) if $\lambda<\lambda_1$ or (2.11) if $\lambda=\lambda_1$, we have the estimate (2.20) in this case also, with a different choice of $\beta(t)$. The remainder of the proof is similar to the case $p<m$ and we omit the details.

Finally to obtain (1.9) we substitute (2.21) into (2.13) to get (1.9) with

$$\mathcal{C}_4=\max\left(\Gamma(m,m,q,r_1)\lambda^{r_1/q}C_0,\Gamma(m,1,q,r_2)C_0^{(q/r_2)/(m-1+q/r_2)}\right), \qquad \varepsilon=\frac{q}{r_2}$$

for any $r_2>\max(q,N/2)$, $r_1>\max(q/m,N/2,1)$. □

Remark. Any solution of (0.1) which can be constructed by the method of this theorem must satisfy the hypotheses of Proposition 1.2. There is then only one possible limit for the sequence $\{v_n\}$, so the entire sequence converges.

3. We turn now to the large time behaviour of solutions of (0.1). Here is some notation.

$$E=\{z: z^m\in C^2(\Omega)\cap C_0^1(\overline{\Omega}), z\ge 0 \text{ and } -\Delta z^m=\lambda z^p\}, \tag{3.1}$$

$$E^*=E\setminus\{0\}. \tag{3.2}$$

E is the set of nonnegative equilibrium solutions of (0.1). Clearly $0\in E$, whatever the choice of λ, m and p. By the strong maximum principle [24], $z>0$ in Ω if $z\in E^*$.

Now suppose v_0 is given in some class for which there is an unambiguously determined solution of (0.1), at least on some time interval. Denote this solution by $v(x,t;v_0)$ and define the semi-orbit

$$\gamma_\tau(v_0)=\{v(\cdot,t;v_0): t\ge\tau\} \tag{3.3}$$

for $\tau\ge 0$. It is possible that $\gamma_\tau(v_0)=\varnothing$ for large enough τ. We will write

$$\gamma_\tau(v_0)\le C$$

if $\gamma_{\tau_1}(v_0)\ne\varnothing$ for every $\tau_1\ge\tau$ and $\|w\|_{L^\infty(\Omega)}\le C$ for every $w\in\gamma_\tau(v_0)$.

Next define the ω-limit set

$$\omega(v_0)=\{z\in C(\overline{\Omega}): \text{there exists } t_n\to\infty \text{ such that } v(\cdot,t_n,v_0)\to z \text{ uniformly as } n\to\infty\}. \tag{3.4}$$

PROPOSITION 3.1. (i) *Suppose* $\gamma_\tau(v_0)\le C$ *for some* $\tau>0$, $C<\infty$. *Then* $\omega(v_0)\ne\varnothing$, *and is compact and connected in* $C(\bar{\Omega})$.

(ii) *Suppose* $\omega(v_0)$ *consists of a single element* z_0. *Then* $z_0\in E$ *and* $\nabla v^m(\cdot,t;v_0)\to\nabla z_0^m$ *strongly in* $L^2(\Omega)$.

Remark. Both parts of this proposition are valid for a much larger class of equations than (0.1). We defer the proof of (ii) until the end of §4; it will not be used anywhere.

Proof. (i) If $\gamma_\tau(v_0)\le C$, then the collection of functions $\{v(\cdot,t;v_0)\}_{t\ge\tau}$ is uniformly bounded, and hence equicontinuous, by Theorem 2.2. Therefore $\omega(v_0)\ne\varnothing$ by the Arzela–Ascoli theorem. The other two properties are standard; see for example [3] or [9]. □

For the remainder of this section we consider the asymptotic behaviour of solutions of (0.1) under the hypotheses of Theorem 1.3. Then $v(\cdot,t;v_0)$ is well defined for $v_0\in L^1(\Omega)$ if $p<m$, or $v_0\in L^q(\Omega)$, $q>1$, if $p=m$ and $\lambda\le\lambda_1$.

THEOREM 3.2. (i) *Let* $1\le p<m$. *Then* E^* *consists of a single element* z_0. *If* $v_0\in L^1(\Omega)$, $v_0\ge0$, $v_0\not\equiv0$, *then* $\omega(v_0)=\{z_0\}$.

(ii) *Let* $p=m$, $\lambda<\lambda_1$, $v_0\in L^q(\Omega)$, $q>1$. *Then* $\omega(v_0)=\{0\}$.

(iii) *Let* $p=m$, $\lambda=\lambda_1$, $v_0\in L^q(\Omega)$, $q>1$. *Then*

$$\omega(v_0)=\{\theta\rho_1^{1/m}\},\qquad \theta=\left(\int_\Omega v_0\rho_1\,dx\right)\Big/\left(\int_\Omega\rho_1^{1+1/m}dx\right).$$

Proof. (i) For a complete discussion of the equilibrium problem, see [4]. The uniqueness of z_0 is actually a consequence of the argument given below.

We first recall that a comparison property is valid for solutions of (0.1); namely, if $v,\hat{v}$ are solutions of (0.1) with $v(x,0)\le\hat{v}(x,0)$, then $v(x,t)\le\hat{v}(x,t)$ for $t>0$. This may be proved as in [3, Thm. 12]. It follows that it is enough to assume that either $v_0\le z_0$ or $v_0\ge z_0$. Suppose first that $v_0\ge z_0$; then $v(x,t)\ge z_0(x)$ for all $x\in\Omega$ and $t\ge0$. By Theorem 1.3 $\gamma_\tau(v_0)\le C$ for all $\tau>0$ and some C; hence v is a weak solution of (0.1) on $[\tau,T]$ for any $\tau>0$, $T<\infty$, by Proposition 1.1. We may take $\rho=z_0^m$ as test function in (1.3) to obtain

$$(3.5)\qquad \int_\Omega v(x,t)z_0^m(x)\,dx-\int_\Omega v(x,\tau)z_0^m(x)\,dx$$
$$=\lambda\int_\tau^t\int_\Omega\left(v^p(x,t)z_0^m(x)-v^m(x,t)z_0^p(x)\right)dx\,dt.$$

Therefore

$$(3.6)\qquad \int_\tau^t\int_\Omega v^p(x,t)z_0^p(x)\left(v^{m-p}(x,t)-z_0^{m-p}(x)\right)dx\,dt\le\frac{1}{\lambda}\int_\Omega v(x,\tau)z_0^m(x)\,dx$$

for any $\tau>0$ and $t<\infty$, and the right-hand side is bounded independently of t.

Now the integrand on the left-hand side of (3.6) is nonnegative, and by Theorem 2.2 the function

$$t\to\int_\Omega v^p(x,t)z_0^p(x)\left(v^{m-p}(x,t)-z_0^{m-p}(x)\right)dx$$

is uniformly continuous on $[\tau,\infty)$. Therefore we must have

$$(3.7)\qquad \lim_{t\to\infty}\int_\Omega v^p(x,t)z_0^p(x)\left(v^{m-p}(x,t)-z_0^{m-p}(x)\right)dx=0.$$

But if $w\in\omega(v_0)$, then from (3.7)

$$\int_\Omega z_0^p(x)w^p(x)(w^{m-p}(x)-z_0^{m-p}(x))\,dx=0.$$

Since the integrand is nonnegative and $w\geq z_0>0$ in Ω, we must have $w\equiv z_0$, i.e. $\omega(v_0)=\{z_0\}$.

Now suppose $v_0\leq z_0$, so that $v(x,t)\leq z_0$ for all $z\in\Omega$ and $t\geq 0$. Then (3.5) is still valid, with $\tau=0$, from which we deduce that

$$(3.8)\qquad \int_0^t\int_\Omega v^p(x,t)z_0^p(x)(z_0^{m-p}(x)-v^{m-p}(x,t))\,dx\,dt\leq\frac{1}{\lambda}\int_\Omega z_0^{m+1}(x)\,dx.$$

As in the first case, it follows that if $w\in\omega(v_0)$, then

$$\int_\Omega w^p(x)z_0^p(x)(z_0^{m-p}(x)-w^{m-p}(x))\,dx=0.$$

Since $z_0>0$ in Ω, it follows that

$$w^p(x)[w^{m-p}(x)-z_0^{m-p}(x)]=0\quad\text{a.e. in }\Omega,$$

but since $z_0>0$ and w is continuous, the only possibilities are $w\equiv 0$ and $w\equiv z_0$.

However from (3.5) again

$$\int_\Omega v(x,t)z_0^m(x)\,dx\geq\int v_0(x)z_0^m(x)\,dx>0.$$

Letting $t\to\infty$ through any subsequence, we see that $w\equiv 0$ is impossible, hence $\omega(v_0)=\{z_0\}$ again.

The proof of (ii) follows directly from estimate (1.7).

For the proof of (iii) and also for later use, we introduce the functional

$$(3.9)\qquad J(v)=\frac{1}{2}\int_\Omega|\nabla v^m|^2dx-\frac{\lambda m}{m+p}\int v^{m+p}\,dx.$$

It is shown in [18, Lemma 4.1] that

$$(3.10)\qquad \frac{4m}{(m+1)^2}\int_0^t\int_\Omega(v^{(m+1)/2})_t^2dx\,dt+J(v(\cdot,t))\leq J(v_0)$$

if v is a weak solution of (0.1) with $v_0^m\in H_0^1(\Omega)$. At the lower limit could be replaced by any time other than 0, we see that $t\to J(v(\cdot,t))$ is nonincreasing, i.e. J is a Lyapunov functional for this problem.

It is also shown in [18] that

$$(3.11)\quad \int_\Omega v^{m+1}(x,t)\,dx-\int_\Omega v^{m+1}(x,0)\,dx=\int_0^t\int_\Omega\left(\lambda v^{m+p}(x,t)-|\nabla v^m(x,t)|^2\right)dx\,dt$$

for weak solutions of (0.1).

In the case at hand, namely $p=m$, $\lambda=\lambda_1$,

$$J(v)=\frac{1}{2}\left(\int_\Omega|\nabla v^m|^2dx-\lambda_1\int_\Omega v^{2m}dx\right).$$

By the variational characterization of λ_1, $J(v)\geq 0$ if $v^m\in H_0^1(\Omega)$, and $J(v)=0$ only if $v^m=\theta\rho_1$ for some constant θ.

Now let $v_0 \in L^q(\Omega)$ $q>1$, and let $v(x,t;v_0)$ be the corresponding solution of (0.1). By Theorem 1.3, $\gamma_\tau(v_0) \leq C$ for every $\tau>0$. By (3.11) with lower limit replaced by τ, $v^m \in L^2(\tau,T;H_0^1(\Omega))$ for any $\tau>0$, $T<\infty$. In particular $v^m(\cdot,t;v_0) \in H_0^1(\Omega)$ for some $t>0$, and hence for any larger t by (3.10); in fact $v^m \in L^\infty((\tau,\infty);H_0^1(\Omega))$ for any $\tau>0$. If $w \in \omega(v_0)$, then certainly $w^m \in H_0^1(\Omega)$ so that $J(w) \geq 0$. We claim that $J(w)=0$.

Supposing that this is not the case, then since J does not increase, and

$$J(w) \leq \varliminf_{t_n \to \infty} J(v(\cdot,t_n;v_0))$$

for any subsequence $t_n \to \infty$, there must exist $\delta>0$ such that

$$J(v(\cdot,t;v_0)) \geq \delta > 0$$

for all $t>0$. From (3.11) it follows that

$$\int_\Omega v^{m+1}(x,t)\,dx - \int_\Omega v^{m+1}(x,\tau)\,dx = -2\int_\tau^t J(v(\cdot,t;v_0))\,dt \leq -2(t-\tau)\delta$$

and this is clearly a contradiction as $t \to \infty$ if $\delta>0$.

Therefore $J(w)=0$, whence $w^m = \theta\rho_1$ for some θ. The constant θ may be determined in the following way. Since $v(\cdot,t;v_0)$ is a weak solution of (0.1) on $[\tau,T]$ for any $\tau>0$, $T<\infty$ we may take $\rho=\rho_1$ as test function in (1.3). This gives

$$\int_\Omega v(x,t)\rho_1(x)\,dx = \int_\Omega v(x,\tau)\rho_1(x)\,dx, \qquad t \geq \tau > 0.$$

Letting $\tau \to 0$, $t \to \infty$ through any subsequence gives the desired conclusion. □

4. Let us finally consider the cases of (0.1) not included in the previous discussion, namely $p>m$, or else $p=m$ and $\lambda>\lambda_1$. We assume that $v_0 \in L^\infty(\Omega)$ so that (0.1) has a local time weak solution, by Theorem 2.1 (or more precisely, by remark (iii) following the statement of the theorem). We also suppose $v_0 \geq 0$.

The case $p=m$, $\lambda>\lambda_1$ may be immediately disposed of; there are no nonnegative solutions of (0.1) which exist for all time except $v \equiv 0$. This is a special case of results in [25, Thm. 5.1] or [12, Thm. 2.1]. We sketch here the short formal argument.

If the equation (0.1) is multiplied by the eigenfunction ρ_1, one integrates to obtain

$$\frac{d}{dt}\int_\Omega v\rho_1 = (\lambda-\lambda_1)\int_\Omega v^m\rho_1 \geq (\lambda-\lambda_1)\left(\int_\Omega v\rho_1\right)^m$$

by Jensen's inequality, since we have assumed that $\|\rho_1\|_{L^1(\Omega)}=1$. Thus the function $t \to \int_\Omega v(x,t)\rho_1(x)\,dx$ satisfies a differential inequality which has no positive global time solutions. The condition $\lambda \leq \lambda_1$ is therefore necessary, in general, for the solvability of (0.1) on arbitrary time intervals. Using arguments as in §2, one can show that for any $v_0 \in L(\Omega)$, $q>1$, there exists $T>0$ such that the problem (0.1) has a mild solution $v \in C([0,T);L^1(\Omega)) \cap L^\infty_{\mathrm{loc}}((0,T)\times\Omega)$ and $\lim_{t\to T}\|v(\cdot,t)\|_{L^1(\Omega)}=\infty$.

We come then to the case $p>m$ which is more complicated than any of the previous cases; our results are correspondingly less complete.

One cause of difficulty is the fact that solution of (0.1) may or may not exist for all time. Here are two conditions which guarantee that $v(\cdot,t;v_0)$ cannot exist for all time.

(4.1) (i) $J(v_0)<0$, J defined by (3.9) (see [12], [18]),

$$\text{(4.2)}\quad \text{(ii)}\quad \int_{\Omega} v_0\rho_1\,dx > \frac{p}{(p-m)^{1/p}}\left(\frac{p\lambda_1}{m\lambda}\right)^{m/(p-m)} \quad \text{(see [12], [25]).}$$

We remark also that nonuniqueness of mild solutions is to be expected in this case, although such a result has only been proved when $m=1$; see Haraux and Weissler [31], Baras [32], Ni and Sacks [33].

From now on we restrict attention to those initial values for which $\gamma_\tau(v_0)\leq C$ for some $\tau>0$, $C<\infty$. In such a case $\omega(v_0)\neq\varnothing$ by Proposition 3.1. We expect that $\omega(v_0)\subset E$ but are unable to prove this (except in the case $N=1$ when it follows by the arguments used in [3]).

In this section we will give three different conditions on v_0 which imply that $\omega(v_0)=\{0\}$. By analogy with the case $m=1$ we expect that 0 is an asymptotically stable equilibrium solution of (0.1), while positive equilibrium solutions are unstable, and indeed our results imply this. Some results concerning the existence of a stable manifold for an unstable equilibrium are given in Ni, Sacks and Tavantzis [34].

First here is a review of some facts about the equilibrium problem

$$\text{(4.3)}\qquad -\Delta z^m=\lambda z^p,\quad x\in\Omega,\qquad z=0\quad x\in\partial\Omega.$$

For arbitrary smooth bounded domains $\Omega\subset\mathbb{R}^N$ and $p/m<(N+2)/(N-2)$, ($p/m<\infty$ for $N=1,2$) it is known that there exists a positive solution of (4.3); see for example [1] or [23]. This solution is the unique positive solution if Ω is a ball [13] or is unique among positive radially symmetric solutions if Ω is an annulus [21]. However if Ω is an annulus, there may exist both radial and nonradial solutions [7]. We know of no multiplicity results for more general domains.

If $p/m\geq(N+2)/(N-2)$ and Ω is starlike with respect to some point, then there is no positive classical solution of (4.3) [23], but for other domains there may be, e.g. if Ω is an annulus [15].

We require two lemmas. The first actually holds without restriction on m,p and λ while the second requires $p>m$.

LEMMA 4.1. *Let $\gamma_\tau(v_0)\leq C$. Then $E\cap\omega(v_0)\neq\varnothing$.*

Proof. By Proposition 3.1 $\omega(v_0)\neq\varnothing$; also there is some constant C so that $J(w)\geq -C$ for $\omega\in\overline{\gamma_\tau(v_0)}$ (closure in the L^∞ norm). As in Theorem 3.2, $v^m\in L^\infty((\tau,\infty);H_0^1(\Omega))$ for any $\tau>0$, and hence $\|\nabla w^m\|_{L^2(\Omega)}$ is uniformly bounded for $w\in\overline{\gamma_\tau(v_0)}\,\tau>0$. Thus J is proper lower semicontinuous and bounded below on the compact metric space $\omega(v_0)$. There exists then $z_0\in\omega(v_0)$ such that

$$J(z_0)=\min_{z\in\omega(v_0)} J(z).$$

But $v(\cdot,t;z_0)\in\omega(v_0)$ for any t (i.e. $\omega(v_0)$ is a positive invariant set) from which it follows that

$$J(v(\cdot,t;z_0))=J(z_0)\quad\text{for all } t>0$$

and this can occur only for $z_0\in E$ by (3.10). □

LEMMA 4.2. *If $0\in\omega(v_0)$ then $\omega(v_0)=\{0\}$.*

Proof. Pick a domain $\hat\Omega$ such that $\overline{\Omega}\subset\hat\Omega$ and let $\hat\rho_1$, $\hat\lambda_1$ be the corresponding first eigenfunction and first eigenvalue, with normalization $\hat\rho_1\geq 0$, $\|\hat\rho_1\|_{L^\infty(\hat\Omega)}=1$. Let $\hat\alpha=\min_{x\in\Omega}\hat\rho_1(x)$; clearly $\hat\alpha>0$. If $w_\varepsilon=(\varepsilon\hat\rho_1)^{1/m}$, then for $\varepsilon\leq(\hat\lambda_1/\lambda)^{m/(p-m)}w_\varepsilon$ is a supersolution of (0.1) and $w_\varepsilon\geq(\varepsilon\hat\alpha)^{1/m}$ for $x\in\overline{\Omega}$.

Now suppose $z \in \omega(v_0)$, $z \not\equiv 0$ and $v(\cdot, t_n, v_0) \to 0$ uniformly. Choose

$$\varepsilon = \min\left(\left(\frac{\hat{\lambda}_1}{\lambda}\right)^{m/(p-m)}, \left(\frac{\|z\|_{L^\infty(\Omega)}}{2}\right)^m\right).$$

There exists n_0 such that for $n \ge n_0$ $\|v(\cdot, t_n; v_0)\|_{L^\infty(\Omega)} \le (\varepsilon\hat{\alpha})^{1/m}$ which implies $v(x, t_n; v_0) \le w_\varepsilon(x)$ for all such n and $x \in \bar{\Omega}$. But then $v(x, t; v_0) \le w_\varepsilon(x) \le \varepsilon^{1/m} \le \|z\|_{L^\infty(\Omega)}/2$ for all $t \ge t_{n_0}$. Thus $z \in \omega(v_0)$ is impossible. □

We introduce now one final notation. Set

$$d = d(\Omega) = \inf_{z \in E^*} J(z) \tag{4.4}$$

with the convention that $d = +\infty$ if $E^* = \varnothing$.

For any $z \in E^*$,

$$J(z) = \left(\frac{1}{2} - \frac{m}{m+p}\right)\int_\Omega |\nabla z^m|^2 dx > 0 \quad \text{for } p > m,$$

so that $d \ge 0$ always. Under certain circumstances we may be sure that $d > 0$. Here are some examples.

(i) If $p/m < (N+2)/(N-2)$ ($p/m < \infty$ if $N = 1, 2$) then for $z \in E^*$

$$\int_\Omega |\nabla z^m|^2 dx = \lambda \int_\Omega z^{p+m} dx \le C\lambda\left(\int_\Omega |\nabla z^m|^2 dx\right)^{(m+p)/2m}$$

by the Sobolev embedding. Hence

$$J(z) \ge \left(\frac{1}{2} - \frac{m}{m+p}\right)\left(\frac{1}{C\lambda}\right)^{2m/(p-m)}.$$

(ii) If $p/m \ge (N+2)/(N-2)$ and Ω is starlike, then $E^* = \varnothing$ [23] so $d = +\infty$.

(iii) If E^* is a finite set, e.g. if Ω is a ball [13], then clearly $d > 0$.

THEOREM 4.3. *Let $p > m$, $v_0 \in L^\infty(\Omega)$, and $v_0 \ge 0$. Then each of the following conditions implies that $\omega(v_0) = \{0\}$.*

(i) *$\gamma_\tau(v_0) \le C$ and $J(v_0) < d$.*

(ii) *There exists $w \in E^*$ such that $v_0 \le w$, $v_0 \not\equiv w$.*

(iii)

$$\|v_0\|_{L^\infty(\Omega)} \le \left(\frac{\hat{\lambda}_1}{\lambda}\right)^{1/(p-m)} \hat{\alpha}^{1/m}$$

for some domain $\hat{\Omega}$ with $\hat{\lambda}_1$, $\hat{\alpha}$ defined as in Lemma 4.2.

Proof. In each case we show that $\gamma_\tau(v_0) \le C$ and $E^* \cap \omega(v_0) = \varnothing$. Then by Lemma 4.1 $0 \in \omega(v_0)$, and the conclusion follows from Lemma 4.2.

(i) If $z \in E^*$, then $J(z) \ge d$, but if $z \in \omega(v_0)$ then $J(z) \le J(v_0) < d$. Hence $E^* \cap \omega(v_0) = \varnothing$.

(ii) By comparison $v(x, t; v_0) \le w(x)$ for all $x \in \Omega$ and $t \ge 0$, so $\gamma_\tau(v_0) \le C$. Also if $z \in \omega(v_0) \cap E^*$, then $0 \le z(x) \le w(x)$ for all x and both functions satisfy (4.3). Multiplying the equation for w by z^m and integrating by parts gives

$$0 = \lambda \int_\Omega w^m z^m (w^{p-m} - z^{p-m})\, dx.$$

Since the integrand is nonnegative, and w,z are positive in Ω, we must have $w\equiv z$. But (3.5), which is still valid here, implies that

$$\int_\Omega v(x,t)w^m(x)\,dx\le\int_\Omega v_0(x)w^m(x)\,dx<\int_\Omega w^{m+1}(x)\,dx.$$

Letting $t\to\infty$ through any subsequence shows that $z\equiv w$ is impossible. Thus $E^*\cap\omega(v_0)=\varnothing$.

(iii) Using the notation of Lemma 4.2 it follows from the given condition that $v(x,t;v_0)\le w_\varepsilon(x)$ for all $x\in\Omega$, $t\ge 0$ and $\varepsilon=(\hat\lambda_1/\lambda)^{m/(p-m)}$. Thus $\gamma_\tau(v_0)\le C$ and if $z\in\omega(v_0)$, then

$$\|z\|_{L^\infty(\Omega)}\le\left(\frac{\hat\lambda_1}{\lambda}\right)^{1/(p-m)}<\left(\frac{\lambda_1}{\lambda}\right)^{1/(p-m)}.$$

But if $z\in E^*$, then multiplying (4.3) by ρ_1 and integrating by parts, we get

$$0=\int_\Omega\rho_1 z^m(\lambda z^{p-m}-\lambda_1)\,dx$$

from which one infers that $\|z\|_{L^\infty}\ge(\lambda_1/\lambda)^{1/(p-m)}$. Therefore $E^*\cap\omega(v_0)=\varnothing$ again. □

Remark. Whenever $v(\cdot,t;v_0)$ tends to zero we may also conclude that a decay estimate of the form

$$\|v(\cdot,t;v_0)\|_{L^\infty(\Omega)}\le\frac{\mathcal{C}}{t^{1/(m-1)}}$$

is valid, for some constant $\mathcal{C}$ depending on the data and the initial value. To see this, we just observe that v is eventually a subsolution of

$$v_t-\Delta v^m=\lambda v^m$$

with $\lambda<\lambda_1$, and then apply the estimate (1.7).

We conclude with the proof of Proposition 3.1(ii).

Proof. Without loss of generality suppose that $v_0^m\in H_0^1(\Omega)\cap L^\infty(\Omega)$. By Lemma 4.1 $z_0\in E$. Let $J_\infty=\lim_{t\to\infty}J(v(\cdot,t;v_0))$; this limit exists since $t\to J(v(\cdot,t;v_0))$ is nonincreasing, and $J_\infty\ge J(z_0)>-\infty$.

By (3.11) and the definition of J

$$\int_\Omega v^{m+1}(x,t)\,dx-\int_\Omega v_0^{m+1}\,dx=\int_0^t\left[\lambda\left(1-\frac{2m}{m+p}\right)\int_\Omega v^{m+p}\,dx-2J(v(\cdot,t;v_0))\right]dt.$$

The term in brackets has a limit as $t\to\infty$ and since its integral from 0 to t is uniformly bounded, this limit can only be zero. Thus

$$J_\infty=\lambda\left(\frac{1}{2}-\frac{m}{m+p}\right)\int_\Omega z_0^{m+p}\,dx=J(z_0)$$

since $z_0\in E$.

If we now pass to the limit in the equation

$$J(v(\cdot,t;v_0))=\frac{1}{2}\int_\Omega|\nabla v^m|^2dx-\frac{m\lambda}{m+p}\int_\Omega v^{m+p}\,dx,$$

we get

$$\overline{\lim_{t\to\infty}}\,\frac{1}{2}\int|\nabla v^m|^2dx\le J(z_0)+\frac{m\lambda}{m+p}\int z_0^{m+p}\,dx=\frac{1}{2}\int|\nabla z_0^m|^2dx.$$

We conclude that $\nabla v^m(\cdot,t)\to\nabla z_0^m$ strongly in $L^2(\Omega)$ since it converges weakly in norm. □

Acknowledgment. I would like to thank M. G. Crandall for some helpful suggestions.

REFERENCES

[1] A. Ambrosetti and P. H. Rabinowitz, *Dual variational methods in critical point theory and applications*, J. Funct. Anal., 14 (1973), pp. 349-381.

[2] D. G. Aronson and P. Bénilan, *Regularité des solutions de l'équation des milieux poreux dans $\mathbb{R}^N$*, C. R. Acad. Sci. Paris Ser. A-B, 288 (1979), pp. 103-105.

[3] D. G. Aronson, M. G. Crandall and L. A. Peletier, *Stabilization of solutions of a degenerate nonlinear diffusion problem*, Nonlinear Anal., 6 (1982), pp. 1001-1022.

[4] D. G. Aronson and L. A. Peletier, *Large time behaviour of solutions of the porous medium equation in bounded domains*, J. Differential Equations, 39 (1981), pp. 378-412.

[5] P. Benilan, M. G. Crandall and M. Pierre, *Solutions of the porous medium equation in $\mathbb{R}^N$ under optimal conditions on initial values*, T. S. R. # 2387, Math. Research Center, Madison, WI.

[6] H. Brezis and M. G. Crandall, *Uniqueness of solutions of the initial value problem for $u_t-\Delta\phi(u)=0$*, J. Math. Pures et Appl., 58 (1979), pp. 153-163.

[7] H. Brezis and L. Nirenberg, *Positive solutions of nonlinear elliptic equations involving critical Sobolev exponents*, Comm. Pure Appl. Math. (1983), to appear.

[8] M. G. Crandall, *An introduction to evolution governed by accretive operators*, in Dynamical Systems—An International Symposium, Academic Press, New York, 1976.

[9] C. M. Dafermos, *Asymptotic behavior of solutions of evolution equations*, in Nonlinear Evolution Equations, Academic Press, New York, 1978.

[10] E. DiBenedetto, *Continuity of weak solutions to a general porous medium equation*, Indiana Univ. Math. J., 32 (1983), pp. 83-118.

[11] L. C. Evans, *Application of nonlinear semigroup theory to certain partial differential equations*, Nonlinear Evolution Equations, Academic Press, New York, 1978.

[12] V. A. Galaktionov, *A boundary value problem for the nonlinear parabolic equation $u_t=\Delta u^{\sigma+1}+u^{\beta}$*, Differential Equations, 17 (1981), pp. 551-555.

[13] B. Gidas, W. M. Ni and L. Nirenberg, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209-243.

[14] M. Gurtin and R. MacCamy, *On the diffusion of biological populations*, Math. Biosciences, 33 (1977), pp. 35-49.

[15] J. Kazdan and F. Warner, *Remarks on some quasilinear elliptic equations*, Comm. Pure Appl. Math., 28 (1975), pp. 567-597.

[16] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[17] O. A. Ladyzenskaja, V. A. Solonnikov and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI 1968.

[18] H. A. Levine and P. E. Sacks, *Some existence and nonexistence theorems for solutions of degenerate parabolic equations*, J. Differential Equations (1984), to appear.

[19] P. L. Lions, *Asymptotic behavior of some nonlinear heat equations*, TSR 2134, Math. Research Center, Madison, WI.

[20] H. Matano, *Asymptotic behavior and stability of solutions of semilinear diffusion equations*, Publ. RIMS, Kyoto Univ., 15 (1979), pp. 401-454.

[21] W. M. Ni, *Uniqueness of solutions of nonlinear Dirichlet problems*, J. Differential Equations, (1983), to appear.

[22] A. Pazy, *The Lyapunov method for semigroups of nonlinear contractions in Banach spaces*, J. Anal. Math. 40 (1982), pp. 239-267.

[23] S. I. Pohozaev, *Eigenfunctions of the equation* $\Delta u+\lambda f(u)=0$, Sov. Math. Dokl., 5 (1965), pp. 1408–1411.

[24] M. H. Protter and H. F. Weinberger, *Maximum Principles in Partial Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[25] P. E. Sacks, *Continuity of solutions of a singular parabolic equation*, Nonlinear Anal., 7 (1983), pp. 387–409.

[26] F. Weissler, *Local existence and nonexistence for semilinear parabolic equations in* L^p, Indiana Univ. Math. J., 29 (1980), pp. 79–102.

[27] M. Bertsch, T. Nanbu and L. A. Peletier, *Decay of solutions of a nonlinear diffusion equation*, Nonlinear Anal., 6 (1982), pp. 539–554.

[28] N. D. Alikakos and R. Rostamian, *Large time behavior of solutions of Neumann boundary value problems for the porous medium equation*, Indiana Univ. Math. J., 30 (1981), pp. 749–785.

[29] ———, *Stabilization of solutions of the equation* $\partial u/\partial t=\Delta\phi(u)-\beta(u)$, Nonlinear Anal., 6 (1982), pp. 637–647.

[30] M. Bertsch and L. Peletier, *A positivity property of solutions of nonlinear diffusion equations*, to appear.

[31] A. Haraux and F. Weissler, *Nonuniqueness for a semilinear initial value problem*, Indiana Univ. Math. J., 31 (1981), pp. 167–189.

[32] P. Baras, *Non-unicité des solutions d'une équation d'evolution non-linéaire*, to appear.

[33] W. M. Ni and P. Sacks, *Singular behavior in nonlinear parabolic equations*, preprint.

[34] W. M. Ni, P. Sacks and J. Tavantzis, *On the asymptotic behavior of solutions of certain quasilinear parabolic equations*, J. Differential Equations, to appear.

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

THE ASYMPTOTIC BEHAVIOUR OF THE SOLUTION OF A NONLINEAR DIFFUSION EQUATION*

CARMEN CORTAZAR† AND MANUEL ELGUETA†

Abstract. We prove that a generalized solution of a filtration equation of the form $\partial u/\partial t = (\partial/\partial x)(\sigma(u)\,\partial u/\partial x)$, where $\lim_{s\to 0}(\sigma(s)/s^\lambda)=1$, behaves asymptotically as the selfsimilar solution of $\partial u/\partial t = (\partial/\partial x)(u^\lambda\,\partial u/\partial x)$.

1. Introduction. The purpose of this note is to study the asymptotic behaviour of the solution of the filtration problem

$$(1.1)\qquad \begin{aligned} &\frac{\partial u}{\partial t}=\frac{\partial^2}{\partial x^2}\Sigma(u) \quad \text{in } \mathbb{R}\times(0,\infty),\\ &u(x,0)=u_0(x) \quad \forall x\in\mathbb{R} \end{aligned}$$

where $\Sigma(s)\in C^2(\mathbb{R}-\{0\})$; $\Sigma'(s)>0$ if $s>0$ and for some $\lambda>0$ $\Sigma'(s)/s^\lambda\to 1$ as $s\to 0$. The initial value $u_0(x)$ is assumed to be bounded and in L^1, $u_0(x)\geq 0$ $\forall x$ and $(\partial^2/\partial x^2)\Sigma(u_0(x))$, which exists in the sense of distributions, belongs to L^1. From now on we denote $\|u_0\|_\infty = M$ and $\|u_0\|_1=E$.

In [9] Oleinick proved existence and uniqueness of generalized solutions of problem (1.1).

Let $Au=(\partial^2/\partial x^2)\Sigma(u)$ defined in a domain of continuous L^1 functions, C^2 at points where $u\neq 0$ and such that $Au\in L^1$ and $(\partial/\partial x)\Sigma(u)\to 0$ if $u\to 0$ or if $x\to\pm\infty$.

It can be shown, using results of Crandall and Liggett [4] about nonlinear semigroups that $\bar{A}$, the closure of A, which is single valued, may be regarded as a dissipative operator in L^1 and generates a contraction semigroup there; i.e. $e^{t\bar{A}}u_0=\lim_{n\to+\infty}(I-t/n\bar{A})^{-n}u_0$ exists in the L^1 norm. See [1], [3] or [5].

It is known that $u(x,t)=e^{t\bar{A}}u_0$ is a generalized solution of problem (1.1) in $S=\mathbb{R}\times[0,\infty)$, that is,

i) $u(x,t)$ is continuous in S,

ii) $(\partial/\partial x)\Sigma(u)$ exists in the sense of distribution and belongs to L^∞,

iii) If $\phi\in C_0^\infty(S)$ then

$$\iint_S\left(u\frac{\partial\phi}{\partial t}-\frac{\partial}{\partial x}\Sigma(u)\frac{\partial\phi}{\partial t}\right)dx\,dt+\int_{-\infty}^{\infty}u_0(x)\phi(x,0)\,dx=0.$$

See [2], and its references. This permits us to use the results for semigroups, as the comparison theorems stated later, to prove results about generalized solutions.

Let $W_E(x,t)$ be the selfsimilar solution (i.e., $kW_E(kx,k^{\lambda+2}t)=W_E(x,t)$ for any $k>0$) of

$$(1.2)\qquad \frac{\partial u}{\partial u}=\frac{1}{\lambda+1}\frac{\partial^2}{\partial x^2}u^{\lambda+1},\qquad u(x,t)=E\delta_0(x),$$

where δ_0 denotes the Dirac delta function at 0 (see [8]).

* Received by the editors July 19, 1982, and in revised form August 18, 1983. This research was supported in party by Dirección de Investigación de la Universidad Católica de Chile through DIUC Grant 45/82.

† Facultad de Matemática, Universidad Católica de Chile, Casilla 114-D Santiago, Chile.

Our main result is an alternative, and hopefully simpler, proof of the following theorem of Kamin [7].

THEOREM 1. *If $u(x,t)$ denotes the generalized solution of* (1.1), *then*

$$\lim_{t\to+\infty} t^{(1/(\lambda+2))}\|u(\cdot,t)-W_E(\cdot,t)\|_\infty=0$$

where $E=\int_{-\infty}^{\infty}u_0(y)\,dy$.

We would like to note that we do not need the hypothesis $\Sigma(s)\to+\infty$ as $s\to+\infty$ and we prove that the uniform convergence takes place in all $\mathbb{R}$. On the other hand Kamin treats a situation more general than $\Sigma'(s)/s^\lambda\to 1$ as $s\to 0$.

Our proof is based on the comparison theorems that appear in Vazquez [10], [11] and Cortazar [3], which we include for the sake of completeness.

THEOREM (comparison).

i) *Let* $f_i=e^{t\bar{A}}g_i$, *with* $g_i\in\overline{D(A)}$, $i=1,2$. *If* $g_1\le g_2$ *then* $f_1\le f_2$ *and if* $\int_{-\infty}^{x}g_1\le\int_{-\infty}^{x}g_2$ *then* $\int_{-\infty}^{x}f_1\le\int_{-\infty}^{x}f_2$.

ii) *Let* $f_i=e^{t_i\bar{A}_i}g_i$, *with* $g_i\in\overline{D(A_i)}$ *symmetric with respect to* $x=0$ *and* g_1 *or* g_2 *increasing for* $x\le 0$.

Let $A_iu=(\partial^2/\partial x^2)\Sigma_i(u)$, $i=1,2$. *If* $\Sigma_1'(s)\le\Sigma_2'(s)$ *for* $s\in[0,\min(\|g_1\|_\infty,\|g_2\|_\infty)]$, $t_1\le t_2$ *and* $\int_{-\infty}^{x}g_1\le\int_{-\infty}^{x}g_2$ *for* $x\le 0$. *Then* f_i *are also symmetric with respect to* $x=0$, $\int_{-\infty}^{x}f_1\le\int_{-\infty}^{x}f_2$ *if* $x\le 0$ *and the corresponding* f_i *is increasing for* $x\le 0$.

We observe that statement ii) is slightly different from the one in [3] but the proof is exactly the same.

In order to prove Theorem 1 we need Theorem 2 below, which we think is of interest by itself.

THEOREM 2. *Let* $u(x,t)$ *be a generalized solution of problem* (1.1) *and let* Ψ: $\mathbb{R}\to[0,\infty]$ *be so that* $\Psi\in C^1(\mathbb{R})\cap C^2(\mathbb{R}-\{0\})$; $\Psi'(0)=0$ *and* $\Psi''(x)\ge 0$ $\forall x\ne 0$. *Then for* $0<T<T'$

$$\int_{-\infty}^{+\infty}\Psi\left(\frac{\partial\Sigma(u)}{\partial x}\right)(x,T')\,dx\le\int_{-\infty}^{+\infty}\Psi\left(\frac{\partial\Sigma(u)}{\partial x}\right)(x,T)\,dx. \tag{1.3}$$

We will use this theorem in the case $\Psi(s)=s^2$.

2. Proof of Theorem 1. Let $u(x,t)$ be the generalized solution of (1.1). Define

$$u_k(x,t)=ku(kx,k^{\lambda+2}t).$$

Then $u_k(x,t)$ is a generalized solution of

$$\frac{\partial u}{\partial t}=\frac{\partial^2}{\partial x^2}\Sigma_k(u),\qquad u(x,0)=ku_0(kx), \tag{2.1}$$

where $\Sigma_k(s)=k^{\lambda+1}\Sigma(s/k)$.

As in Kamenomostskaya [6], Theorem 1 will be an easy consequence of

MAIN LEMMA. *For any* $T>0$

$$\lim_{k\to\infty}\|u_k(\cdot,T)-W_E(\cdot,T)\|_\infty=0.$$

Proof of Theorem 1 *using the main lemma.* Assuming the previous lemma to be true, using the selfsimilarity of $W_E(x,t)$ and setting $T^{1/(\lambda+2)}k=t^{1/(\lambda+2)}$, we obtain $t^{1/(\lambda+2)}\|u(\cdot,t)-W_E(\cdot,t)\|_\infty=T^{1/(\lambda+2)}\|u_k(\cdot,T)-W_E(\cdot,T)\|_\infty$ and hence the theorem.

So our task now is to prove the main lemma. For this we need

LEMMA 1. *Let $u(x,t)$ be a generalized solution of problem* (1.1). *Then*

$$\|u(\cdot,t)\|_\infty \leq B(\lambda)\left[\frac{E^2}{tC}\right]^{1/(\lambda+2)}$$

where $C=\inf_{0\leq s\leq 2M}(\Sigma'(s)/s^\lambda)$ and $B(\lambda)$ depends only on λ.

Proof. Pick $x_0\in\mathbb{R}$, consider the function

$$\tilde{u}_0(x)=u_0(2x_0-x)+u_0(x);$$

then $\tilde{u}_0(x)$ is symmetric with respect to x_0, $u_0(x)\leq\tilde{u}_0(x)$ and $\int\tilde{u}_0(x)=2E$.

Let $C=\inf_{0\leq s\leq 2M}(\Sigma'(s)/s^\lambda)$, then $Cs^\lambda\leq\Sigma'(s)$ for $s\in[0,2M]$ and let $W^C_{2E}(y,t)$ be a selfsimilar solution of

$$\frac{\partial W}{\partial t}=\frac{C}{\lambda+1}\cdot\frac{\partial^2}{\partial x^2}(W^{\lambda+1}),\qquad W(x,0)=2E\delta_{x_0}(x). \tag{2.2}$$

Choose T small enough such that

$$\int_{-\infty}^{x} W^C_{2E}(y,T)\,dy\leq\int_{-\infty}^{x}\tilde{u}_0(y)\,dy\quad \forall x\leq x_0.$$

The comparison theorem ii) implies

$$\int_{-\infty}^{x} W^C_{2E}(y,T+t)\,dy\leq\int_{-\infty}^{x}\tilde{u}(y,t)\,dy,$$

where $\tilde{u}(x,t)$ is the solution of problem (1.1) with initial condition $\tilde{u}_0(x)$.

Since $\int_{-\infty}^{x_0}W^C_{2E}(y,t+T)\,dy=\int_{-\infty}^{x_0}\tilde{u}(y,t)\,dy=2E$, we obtain

$$\int_{x}^{x_0}W^C_{2E}(y,T+t)\,dy\geq\int_{x}^{x_0}u(y,t)\,dy\quad \forall x\leq x_0.$$

This implies

$$\tilde{u}(x_0,t)\leq W^C_{2E}(x_0,T+t)\leq W^C_{2E}(x_0,t).$$

Since $u_0(x)\leq\tilde{u}_0(x)$, by comparison we obtain $u(x_0,t)\leq\tilde{u}(x_0,t)\leq W^C_{2E}(x_0,t)$. Finally

$$W^C_{2E}(x_0,t)=B(\lambda)\left[\frac{E^2}{tC}\right]^{1/\lambda+2}$$

according to [8] and the lemma is proved.

LEMMA 2. *For any fixed $T>0$ and every interval (a,b)*

$$\int_a^b u_k(y,T)\,dy\to\int_a^b W_E(y,T)\,dy\quad \text{as } k\to\infty.$$

Proof. We assume first that u_0 has compact support.

Since $\|u(\cdot,t)\|_\infty\to 0$ as $t\to+\infty$, by Lemma 1, there exists a sequence $\{t_n\}_{n=1}^\infty$ so that $\|u(\cdot,t)\|_\infty\leq 1/n$ for $t\geq t_n$

Let $\Sigma'(u)=u^\lambda+\rho(u)$ and let

$$A_n=\sup_{0\leq u\leq 1/n}\frac{|\rho(u)|}{u^\lambda}.$$

Without loss of generality we assume $1-A_n\geq 0$.

Let W_n^+ be the selfsimilar solution of the initial value problem

$$\frac{\partial W_n^+}{\partial t}=(1+A_n)\frac{\partial}{\partial x}\left((W_n^+)^\lambda\frac{\partial W_n^+}{\partial x}\right),\qquad W_n^+(x,0)=E\cdot\delta_0$$

and W_n^- the corresponding selfsimilar solution with $1-A_n$ instead of $1+A_n$. Since, by [8], [9], $u(\cdot,t_n)$, $W_n^+(\cdot,t_n)$ and $W_n^-(\cdot,t_n)$ have compact support, there exists a sequence $\{x_n\}_{n=0}^\infty(x_n\geq 0)$ such that

$$\text{(2.3)}\quad \int_{-\infty}^x W_n^-(y-x_n,t_n)\,dy\leq\int_{-\infty}^x u(y,t_n)\,dy\leq\int_{-\infty}^x W_n^+(y+x_n,t_n)\,dy\quad\forall x\in\mathbb{R}.$$

Let z_n^+ be the solution of

$$\frac{\partial z_n^+}{\partial t}=\frac{\partial^2}{\partial x^2}\Sigma(z_n^+),\qquad z_n^+(x,t_n)=W_n^+(x,t_n)$$

and z_n^- the corresponding solution with W_n^- instead of W_n^+. Expression (2.3) together with the fact that for $u\leq 1/n$

$$(1-A_n)u^\lambda\leq\Sigma'(u)\leq(1+A_n)u^\lambda$$

given, by the comparison theorem, that

$$\int_{-\infty}^x W_n^-(y-x_n,t)\,dy\leq\int_{-\infty}^x z_n^-(y-x_n,t)\,dy\leq\int_{-\infty}^x u(y,t)\,dy$$
$$\leq\int_{-\infty}^x z_n^+(y+x_n,t)\,dy\leq\int_{-\infty}^x W_n^+(y+x_n,t)\,dy\ \ \forall x\leq -x_n,\ \ \forall t>t_n.$$

A change of variables and the selfsimilarity of the solutions w_n^+ and w_n^- give

$$\int_{-\infty}^x W_n^-\left(y-\frac{x_n}{k},t\right)dy\leq\int_{-\infty}^x u_k(y,t)\,dy\leq\int_{-\infty}^x W_n^+\left(y+\frac{x_n}{k},t\right)dy$$
$$\forall t\geq\frac{t_n}{k^{\lambda+2}},\qquad x\leq -\frac{x_n}{k}.$$

Letting $k\to+\infty$, we get

$$\int_{-\infty}^x W_n^-(y,T)\,dy\leq\lim_{k\to+\infty}\int_{-\infty}^x u_k(y,T)\,dy\leq\overline{\lim_{k\to\infty}}\int_{-\infty}^x u_k(y,T)$$
$$\leq\int_{-\infty}^x W_n^+(y,T)\,dy\qquad\forall x<0$$

and finally letting $n\to\infty$, we obtain that for $x<0$, $\lim_{k\to+\infty}\int_{-\infty}^x u_k(y,T)\,dy$ exists and is equal to $\int_{-\infty}^x W_E(y,T)\,dy$. For $x>0$, the same argument with $\int_x^\infty$ instead of $\int_{-\infty}^x$ shows that $\lim_{k\to\infty}\int_x^\infty u_k(y,T)\,dy$ exists and is equal to $\int_x^\infty W_E(y,T)\,dy$. Since $\int_{-\infty}^\infty u_k=\int_{-\infty}^\infty W_E=E$, we get Lemma 2 for the case when u_0 has compact support.

If u_0 does not have compact support, let $u_{0,n}\in C^\infty$ with compact support be so that $u_0^n\to u_0$ in L^1 as $n\to+\infty$ and $\|u_0^n\|_1=\|u_0\|_1=E$.

Let $u_k^n(x,t)$ be the solution of the initial value problem

$$\text{(2.4)}\qquad \frac{\partial u}{\partial t}=\frac{\partial^2}{\partial x^2}\Sigma_k(u),\qquad u(x,0)=ku_0^n(kx).$$

Now

$$\begin{aligned}
&\left|\int_a^b (u_k(y,T)-W_E(y,T))\,dy\right| \\
&\quad\le\int_{-\infty}^{+\infty}|u_k(y,T)-u_k^n(y,T)|\,dy+\left|\int_a^b (u_k^n(y,T)-W_E(y,T))\,dy\right| \\
&\quad\le\int_{-\infty}^{+\infty}|ku_0(ky)-ku_0^n(ky)|\,dy+\left|\int_a^b (u_k^n(y,T)-W_E(y,T))\,dy\right| \\
&\quad=\int_{-\infty}^{+\infty}|u_0(y)-u_0^n(y)|\,dy+\left|\int_a^b (u_k^n(y,T)-W_E(y,T))\,dy\right|
\end{aligned}$$

and letting $k\to+\infty$ and then $n\to+\infty$, we get Lemma 2.

LEMMA 3. *If $t>0$ then there exists a constant C independent of k such that*

$$\int_{-\infty}^{+\infty}\left|\frac{\partial}{\partial x}\Sigma_k(u_k)\right|^2(y,t)\le\frac{C}{t}. \tag{2.5}$$

Proof. Fix $T>0$ and let u_ε be a sequence of C^∞ functions so that

i) $\|u_\varepsilon(\cdot)-u(\cdot,T)\|_\infty<\varepsilon$, $\|u_\varepsilon(\cdot)-u(\cdot,T)\|_1<\varepsilon$,

ii) $u_\varepsilon(x)>0$ $\forall x\in\mathbb{R}$,

iii) $(\partial^2/\partial x^2)\Sigma(u_\varepsilon)\in L^1$.

Let $u_\varepsilon(x,t)$ be the generalized solution of

$$\frac{\partial u_\varepsilon}{\partial t}=\frac{\partial^2}{\partial x^2}(\Sigma(u_\varepsilon)),\qquad u_\varepsilon(x,T)=u_\varepsilon(x). \tag{2.6}$$

Then $u_\varepsilon(x,t)>0$ $\forall x\in\mathbb{R}$ $\forall t>T$ and hence $u_\varepsilon(x,t)$ is a classical solution of (2.5) ([9], [3]). Therefore

$$\int_{-\infty}^{+\infty}\int_T^{2T}\frac{\partial u_\varepsilon}{\partial t}\Sigma(u_\varepsilon)=\int_{-\infty}^{+\infty}\int_T^{2T}\frac{\partial^2}{\partial x^2}\Sigma(u_\varepsilon)\cdot\Sigma(u_\varepsilon). \tag{2.7}$$

Set $\Phi(v)=\int_0^v\Sigma(s)\,ds$. Integrating in (2.7), we obtain

$$\int_{-\infty}^{+\infty}[\Phi(u_\varepsilon(x,T))-\Phi(u_\varepsilon(x,2T))]\,dx=\int_T^{2T}\int_{-\infty}^{+\infty}\left|\frac{\partial}{\partial x}\Sigma(u_\varepsilon)(x,t)\right|^2dx\,dt.$$

We note that the boundary term in the integration by parts vanishes due to the fact that $\partial\Sigma(u_\varepsilon)/\partial x\in L^\infty$ and $\Sigma(u_\varepsilon)(x,t)\to0$ as $x\to\pm\infty$. Therefore

$$\int_T^{2T}\int_{-\infty}^{+\infty}\left|\frac{\partial}{\partial x}\Sigma(u_\varepsilon)\right|^2dx\,dt\le\int_{-\infty}^{+\infty}\Phi(u_\varepsilon(x,T))\,dx$$

and since

$$\Phi(u_\varepsilon(x,T))\le\Sigma\big(\|u_\varepsilon(\cdot,T)\|_\infty\big)\cdot u_\varepsilon(x,T),$$

we obtain

$$\text{(2.8)}\qquad \int_T^{2T}\int_{-\infty}^{+\infty}\left|\frac{\partial}{\partial x}\Sigma(u_\varepsilon)\right|^2 dx\,dt \le \Sigma\big(\|u_\varepsilon(\cdot,T)\|_\infty\big)\cdot\int_{-\infty}^{+\infty}u_\varepsilon(x,T)\,dx$$
$$\le\Sigma\big(\|u(\cdot,T)\|_\infty+1\big)\cdot\big(\|u(\cdot,T)\|_1+1\big)$$

for $\varepsilon\le 1$.

Since

$$\int_{\mathbb{R}}|\,u_\varepsilon(x,t)-u(x,t)\,dx\le\varepsilon$$

for $t>T$, we have that $(\partial/\partial x)\Sigma(u_\varepsilon)$ converges weakly in $\mathbb{R}\times[T,2T]$ (in the sense of distributions) to $(\partial/\partial x)\Sigma(u)$ as $\varepsilon\to 0$. Hence, by (2.8), converges weakly in L^2 of $\mathbb{R}\times[T,2T]$, and consequently.

$$\iint_{\mathbb{R}\times[T,2T]}\left|\frac{\partial}{\partial x}\Sigma(u)\right|^2\le\Sigma\big(\|u(\cdot,T)\|_\infty+1\big)\cdot\big(\|u(\cdot,T)\|_1+1\big).$$

By the same argument replacing $u(\cdot,T)$ by $u_k(\cdot,T)$ and Σ by Σ_k (and hence $u(x,t)$ by $u_k(x,t)$), one obtains

$$\iint_{\mathbb{R}\times[T,2T]}\left|\frac{\partial}{\partial x}\Sigma_k(u_k)\right|^2\le\Sigma_k\big(\|u_k(\cdot,T)\|_\infty+1\big)\cdot\big(\|u_k(\cdot,T)\|_1+1\big),$$

and using Lemma 1, we get

$$\iint_{\mathbb{R}\times[T,2T]}\left|\frac{\partial}{\partial x}\Sigma_k(u_k)\right|^2\le C$$

where C is a constant independent of k.

Now, if we fix k, there exists $t_k^*\in[T,2T]$ so that

$$\int_{\mathbb{R}}\left|\frac{\partial}{\partial x}\Sigma_k(u_k)\right|^2(y,t_k^*)\,dt\le\frac{C}{T}$$

and, by Theorem 2, one has

$$\int_{\mathbb{R}}\left|\frac{\partial}{\partial x}\Sigma_k(u_k)\right|^2(y,2T)\le\frac{C}{T}.$$

Proof of the main lemma. Fix $T>0$. By Lemmas 1 and 3 the family of functions $\{\Sigma_k(u_k(\cdot,T))\}_{k=0}^{\infty}$ is equicontinuous and uniformly bounded. Hence the same is true for the family $\{u_k(\cdot,T)\}_{k=0}^{\infty}$ and, since it converges weakly to $W_E(\cdot,T)$, it converges uniformly on compact subsets of $\mathbb{R}$ to $W_E(\cdot,T)$.

Let $a\in\mathbb{R}$ be so that $\operatorname{supp}W_E(\cdot,T)\in[-a,a]$; N be so that $\|u_k(\cdot,T)\|_\infty\le N\ \forall k$ and F, $G>0$ be so that $Gs^{\lambda+1}\le\Sigma_\lambda(s)\le Fs^{\lambda+1}\ \forall s\le N$. Pick $x_0\in(-\infty,-a)$. Then

$$\Sigma_k(u_k(x,T))=\Sigma_k(u_k(x_0,T))-\int_x^{x_0}\frac{\partial}{\partial x}\Sigma_k(u_k(y,T))\,dy$$

and hence, by (2.5) and Hölder's inequality,

$$\Sigma_k(u_k(x,T)) \geq \Sigma_k(u_k(x_0,T)) - |x_0-x|^{1/2} \cdot \left(\frac{C}{T}\right)^{1/2}.$$

So, if $0 < x_0 - x < (T/4C)[\Sigma_k(u_k(x_0,T))]^2$, we have

$$\Sigma_k(u_k(x,T)) \geq \frac{1}{2}\Sigma_k(u_k(x_0,T)).$$

Therefore, integrating over this interval, we obtain

$$\frac{T}{2C}[\Sigma_k(u_k(x_0,T))]^3 \leq \int_{-\infty}^{-a} \Sigma_k(u_k(y,T))\,dy,$$

or

$$\sup_{x\leq -a} u_k(x,T) \leq \left[\frac{2C\cdot F\cdot N^\lambda}{T\cdot G}\int_{-\infty}^{-a} u_k(y,T)\,dy\right]^{1/(\lambda+1)}.$$

Finally since, by Lemma 2, $\int_{-\infty}^{-a} u_k(y,T)\,dy \to 0$ as $k \to +\infty$, we get

$$\sup_{x\leq -a} u_k(x,T) \to 0 \quad \text{as } k \to +\infty.$$

Analogously $\sup_{x\geq a} u_k(x,T) \to 0$ as $k \to +\infty$ and the main lemma is proved.

3. Proof of Theorem 2. If

$$\int_{-\infty}^{+\infty} \Psi\left(\frac{\partial \Sigma(u)}{\partial x}\right)(y,T) = \infty$$

there is nothing to prove, so we assume

$$\int_{-\infty}^{+\infty} \Psi\left(\frac{\partial \Sigma(u)}{\partial x}\right)(y,T) < \infty.$$

Let $u_\varepsilon(x)$ and $u_\varepsilon(x,t)$ be as in the proof of Lemma 3 and so that

$$\int_{-\infty}^{+\infty} \Psi\left(\frac{\partial \Sigma(u_\varepsilon)}{\partial x}\right)(y,T)\,dy \to \int_{-\infty}^{+\infty} \Psi\left(\frac{\partial \Sigma(u)}{\partial x}\right)(y,T)\,dy$$

as $\varepsilon \to 0$.

Let $\phi_n(x) \in C^\infty$ so that $\phi_n \equiv 1$ on $[-n,n]$, $\phi_n \equiv 0$ off $[-n-1, n+1]$, $0 \leq \phi \leq 1$ and $|\phi'(x)| \leq 2\ \forall x \in \mathbb{R}$.

Since $u_\varepsilon(x,t)$ is a classical solution, (see [3], [9]), we have

$$\frac{\partial}{\partial t}\left[\phi_n \Psi\left(\frac{\partial \Sigma(u_\varepsilon)}{\partial x}\right)\right] = \phi_n \Psi'\left(\frac{\partial \Sigma(u_\varepsilon)}{\partial x}\right) \cdot \frac{\partial^2 \Sigma(u_\varepsilon)}{\partial x\,\partial t}.$$

Hence

$$\int_{-\infty}^{+\infty}\phi_n(x)\cdot\Psi\left(\frac{\partial\Sigma(u_\varepsilon)}{\partial x}\right)(x,T)\,dx-\int_{-\infty}^{+\infty}\phi_n(x)\Psi\left(\frac{\partial\Sigma(u_\varepsilon)}{\partial x}\right)(x,T')\,dx$$

$$=-\int_{-\infty}^{+\infty}\int_{T}^{T'}\phi_n(x)\cdot\Psi'\left(\frac{\partial\Sigma(u_\varepsilon)}{\partial x}(x,t)\right)\frac{\partial^2\Sigma(u_\varepsilon)}{\partial x\,\partial t}(x,t)\,dx\,dt$$

$$=\int_{T}^{T'}\int_{-\infty}^{+\infty}\phi_n(x)\cdot\Psi''\left(\frac{\partial\Sigma(u_\varepsilon)}{\partial x}(x,t)\right)\cdot\left(\frac{\partial^2\Sigma(u_\varepsilon)}{\partial x^2}(x,t)\right)^2\Sigma'(u_\varepsilon)(x,t)\,dx\,dt$$

$$+\int_{T}^{T'}\int_{-\infty}^{+\infty}\phi_n'(x)\cdot\Psi'\left(\frac{\partial\Sigma(u_\varepsilon)}{\partial x}(x,t)\right)\Sigma'(u_\varepsilon)(x,t)\frac{\partial u_\varepsilon}{\partial t}(x,t)\,dx\,dt$$

$$\geq\int_{-\infty}^{+\infty}\int_{T}^{T'}\phi'(x)\Psi'\left(\frac{\partial\Sigma(u_\varepsilon)}{\partial x}(x,t)\right)\Sigma'(u_\varepsilon)(x,t)\frac{\partial u_\varepsilon}{\partial t}(x,t)\,dx\,dt.$$

Since $\partial\Sigma(u_\varepsilon)/\partial x$, $\Sigma(u_\varepsilon)\in L^\infty(\mathbb{R}\times[T,T'])$ and $\partial u_\varepsilon/\partial t\in L^1(\mathbb{R}\times[T,T'])$, (see [3]), the last term tends to 0 as $n\to+\infty$; therefore letting $n\to+\infty$, we get

$$\int_{-\infty}^{+\infty}\Psi\left(\frac{\partial\Sigma(u_\varepsilon)}{\partial x}\right)(x,T)\geq\int_{-\infty}^{+\infty}\Psi\left(\frac{\partial\Sigma(u_\varepsilon)}{\partial x}\right)(x,T').$$

Since $(\partial\Sigma(u_\varepsilon)/\partial x)(x,T')$ converges weakly to $\partial\Sigma(u)/\partial x$, letting $\varepsilon\to0$, we obtain

$$\int_{-\infty}^{+\infty}\Psi\left(\frac{\partial\Sigma(u)}{\partial x}\right)(x,T)\,dx=\lim_{\varepsilon\to0}\int_{-\infty}^{+\infty}\Psi\left(\frac{\partial\Sigma(u_\varepsilon)}{\partial x}\right)(x,T)\,dx$$

$$\geq\lim_{\varepsilon\to0}\int_{-\infty}^{+\infty}\Psi\left(\frac{\partial\Sigma(u_\varepsilon)}{\partial x}\right)(x,T')\geq\int_{-\infty}^{+\infty}\Psi\left(\frac{\partial\Sigma(u)}{\partial x}\right)(x,T')$$

and the theorem is proved.

REFERENCES

[1] Ph. Benilan, *Equations d'evolution dans un espace de Banach quelconque et applications*, Thesis, Univ. Orsay, 1972.

[2] Ph. Benilan and M. G. Crandall, *Regularizing effects of homogeneous evolution equations*, contribution to Analysis and Geometry (a supplement to Amer. J. Math), D. N. Clark, G. Pecelli, and R. Sacksteder, eds., John Hopkins Univ. Press, Baltimore, MD, 1981, pp. 23–30.

[3] C. Cortazar, *The application of dissipative operators to nonlinear diffusion equations*, J. Differential Equations, 47 (1983), pp. 1–23.

[4] M. Crandall and T. Ligget, *Generation of semi-groups of nonlinear transformations on general Banach spaces*, Amer. J. Math., 93 (1971), pp. 265–298.

[5] L. G. Evans, *Applications of nonlinear semigroup theory to certain partial differential equations*, in Nonlinear Evolution Equations, M. G. Crandall, ed., Academic Press, New York, 1978.

[6] S. Kamenomostskaya, *The asymptotic behaviour of the solution of the filtration equation*, Israel J. Math, 14 (1973), pp. 76–87.

[7] S. Kamin (Kamenomostskaya), *Similar solutions and the asymptotics of filtration equations*, Arch. Rational Mech. Anal., 60 (1976), pp. 171–183.

[8] L. Landau and E. Lifschitz, *Fluid Mechanics*, Addison-Wesley, Reading, MA, 1959.

[9] O. Oleinick, *On some degenerate quasilinear parabolic equations*, Seminari 1962–1963, Edizioni Cremonese, Roma.

[10] J. L. Vazquez, *Large time behaviour of the solution of the one dimensional porous media equation*, Proceedings Symposium on Free Boundary Problems, 1981, Montecatini, Italy; Pitman Research Notes in Mathematics, pp. 78–79.

[11] ———, C. R. Acad. Sci. Paris, 295, Ser. I, Sept. 1982.

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

THE HELMHOLTZ EQUATION WITH L_2 BOUNDARY VALUES*

T. S. ANGELL† AND R. E. KLEINMAN†

Abstract. This paper deals with the Helmholtz equation in exterior domains with Robin or impedance boundary conditions on a smooth ($C_{1,1}$) boundary with data in L_2 on the boundary. Existence and uniqueness of solutions are established and a constructive method for finding the Green's function is presented. This involves modifying the fundamental solution by adding suitable combinations of radiating solutions of the Helmholtz equation. The coefficients of these added terms for which the modified Green's function best approximates the actual Green's function for the problem are shown to be the elements of the T-matrix which arises in the null field or Waterman method for treating such problems. Use is made of complete families in L_2 on the boundary and existing results are extended to the general case considered here.

1. Introduction. In this paper we treat the Robin or third boundary value problem for the Helmholtz equation in an unbounded domain D and with L_2 data on the boundary ∂D. We establish the existence and uniqueness of a solution of the boundary value problem and present a constructive method, utilizing modified Green's functions, for finding the Green's function. The relationship between this constructive method and the null field or T-matrix approach is also described.

The existence and uniqueness of solutions of the Helmholtz equation satisfying a boundary condition of the form

$$\frac{\partial u}{\partial n}+\sigma u=f \quad \text{on } \partial D \tag{1.1}$$

was given by Leis [1] in the case that u is differentiable up to the boundary and the data σ and f are continuous. Later, in [2], we showed that similar results obtain for σ and f in $L_\infty(\partial D)$ and σ complex. These were applied in [3] to extend, to the Robin problem, results on modified Green's functions given by Kleinman and Roach [4]. In particular, we presented a discussion analogous to that in [4] which provided explicit "best approximations" in various senses, to the exact Green's functions for the Dirichlet and Neumann problems.

In the present paper we show how the analysis of [3] essentially ensures existence of solutions of the Robin problem in the "generalized L_2-sense" for $\sigma \in L_\infty(\partial D)$ and for f in $L_2(\partial D)$ rather than L_∞ as required in [2]. We remark that the more familiar existence theorems for generalized solutions (mostly, if not exclusively, concerned with interior problems), see e.g. Lions–Magenes [5] or Berezanskii [6], require that solutions have smoother traces on ∂D than we require here. For example the general results of Berezanskii (p. 206) require, for the present problem, that our function f be at least in $H^{1/2}(\partial D)$ and, in addition, deal only with the case that $\sigma \in C(\partial D)$. Our setting is, therefore, more general. Appropriate definitions of "generalized L_2-solutions" are given in our paper [2] and are discussed in Miranda [7] and, both more recently and in more detail, by Mikhailov in [8] and [9]. The latter author, in particular, shows that, at least

* Received by the editors November 14, 1983, and in revised form February 1, 1984. The research was supported by the National Science Foundation under grant MCS-82-02-033.

† Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19716.

for interior problems, this type of solution, is a generalized solution in $H^1(D_-)$ when the boundary data $f \in H^{1/2}(\partial D)$ (see the remarks following [8, Thm. 6, pp. 197–198]).

After considering the question of existence of solutions we then extend our previous discussion of modifed Green's functions [3] to show how to construct the actual Green's function for the Robin problem. Finally we show how the approximations to the modified Green's function are related to the null field or T-matrix method (e.g. [10] [11]), following the analysis of [12] for the Dirichlet and Neumann problems.

In §2 we establish notation and define the problem; §3 contains the existence and uniqueness theorems which are then used to establish some vital results on completeness and linear independence of a family of functions; §4 shows how the Green's function may be constructed; in §5 the null field equations are derived and the connection with the Green's function results of §4 is shown.

2. Notation and statement of the problem. Let D_- denote a bounded domain in $\mathbb{R}^3$, containing the origin and with boundary ∂D which is a closed Lyapunov surface with index 1. This latter condition implies, among other things that the unit normal is Lipschitz continuous on ∂D. We denote the exterior of ∂D (i.e. the complement of $D_- \cup \partial D$ in $\mathbb{R}^3$) by D_+.

Let $R = R(p,q)$ denote the distance between two typical points p and q in $\mathbb{R}^3$. We will concern ourselves in this work with the following boundary value problem for the Helmholtz equation

$$(2.1) \qquad (\nabla^2+k^2)u(p)=0, \qquad p \in D_+,$$

$$(2.2\text{a}) \qquad u(p)=f(p),$$

$$(2.2\text{b}) \qquad \frac{\partial u}{\partial n_p}+\sigma(p)u(p)=f(p), \qquad p \in \partial D,$$

$$(2.3) \qquad \lim_{r_p \to \infty} r_p\left(\frac{\partial u}{\partial r_p}(p)-iku(p)\right)=0,$$

where $f \in L_2(\partial D)$, $\sigma \in L_\infty(\partial D)$ and $\operatorname{Im} k \geqq 0$, $\operatorname{Im} \bar{k}\sigma \geqq 0$. Here we consider the boundary values to be taken on in the generalized L_2 sense. Thus (2.2) is understood to hold almost everywhere on ∂D (for details see [2], [6], [8] and [9].

We indicate differentiation in the direction of the unit normal $\hat{n}_p$ at the point $p \in \partial D$ by $\partial/\partial n_p$ where $\hat{n}_p$ is directed outward, from ∂D to D_+. Furthermore we write $\partial/\partial n_p^-$ and $\partial/\partial n_p^+$ to denote the normal derivative when p approaches ∂D from D_- and D_+ respectively. Points p and q are assumed to have spherical polar coordinates (r_p,θ_p,ϕ_p) and (r_q,θ_q,ϕ_q) relative to a Cartesian coordinate system centered in D_-.

A fundamental solution of (2.1) is a two point function of position $\gamma_0(p,q)$ which we write as

$$(2.4) \qquad \gamma_0(p,q) := -\frac{e^{ikR}}{2\pi R}.$$

The normalization of the fundamental solution is chosen so that Green's theorem for solutions of the Helmholtz equation in D_+ satisfying a radiation condition (2.3) takes the form

$$(2.5) \qquad \int_{\partial D}\left\{\gamma_0(p,q)\frac{\partial u}{\partial n_q}-\frac{\partial \gamma_0}{\partial n_q}(p,q)u(q)\right\}ds_q=\begin{cases}2u(p), & p \in D_+,\\ u(p), & p \in \partial D,\\ 0, & p \in D_-.\end{cases}$$

That Green's theorem and the divergence theorem may be employed for the class of surfaces considered here is proven in [2].

The fundamental solution has a well-known expansion in spherical wave functions

$$\gamma_0(p,q)=\sum_{|l|=0}^{\infty} v_l^e(p^>)v_l^i(p^<), \tag{2.6}$$

where

$$v_l^{e,i}(p)=\left\{-\frac{ik}{2\pi}\varepsilon_m(2n+1)\frac{(n-m)!}{(n+m)!}\right\}^{1/2} \cdot Z_n^{e,i}(kr)P_n^m(\cos\theta)\{(1-j)\cos m\phi+j\sin m\phi\}, \tag{2.7}$$

$$Z_n^e(kr)=h_n^{(1)}(kr), \qquad Z_n^i(kr)=j_n(kr),$$

$$p^>=\begin{cases} p & \text{if } r_p>r_q, \\ q & \text{if } r_q>r_p, \end{cases} \qquad p^<=\begin{cases} p & \text{if } r_p<r_q, \\ q & \text{if } r_q<r_p, \end{cases}$$

and the multi-index

$$l=(n,m,j), \qquad n\geq 0, \quad 0\leq m\leq n, \quad 0\leq j\leq 1 \tag{2.8}$$

is employed with $|l|=n+m+j$. The functions in the series (2.6) could be reordered so that the summation is carried out over a single index, see [12] for explicit details. Furthermore we define a modified Green's function as

$$\gamma_1^N(p,q)=\gamma_0(p,q)+\sum_{|l|=0}^{N}\sum_{|l'|=0}^{N}\alpha_{ll'}^N v_{l'}^e(p)v_l^e(q), \tag{2.9}$$

where N may be $+\infty$ and l' is again a multi-index as in (2.8). Note that this coincides with the simpler modification used in [3] when

$$\alpha_{ll'}^N=0, \qquad l'\neq l. \tag{2.10}$$

However, even in the present case, single and double layer distributions with a modified kernel may be defined which have the usual properties; that is let

$$(S_j\mu)(p):=\int_{\partial D}\mu(q)\gamma_j^N(p,q)\,ds_q, \qquad p\in\mathbb{R}^3\backslash\{0\}, \tag{2.11}$$

$$(D_j\mu)(p):=\int_{\partial D}\mu(q)\frac{\partial\gamma_j^N}{\partial n_q}(p,q)\,ds_q, \qquad p\in\mathbb{R}^3\backslash\{0\}, \tag{2.12}$$

and

$$(K_j\mu)(p):=\int_{\partial D}\mu(q)\frac{\partial\gamma_j^N}{\partial n_p}(p,q)\,ds_q, \qquad p\in\partial D, \tag{2.13}$$

where $j=0,1$. We remark that K_j and S_j are completely continuous operators on $L_2(\partial D)$ [13] provided of course that the series occurring in (2.9) converges if $N=\infty$. Moreover S_0 and D_0 may be considered as operators with range in $L_2(D_-)$ and, as such, are continuous operators from $L_2(\partial D)$ to $L_2(D_-)$ (see e.g. [7]). This is no longer true for the modified Green's functions which are singular at the origin. Nevertheless S_1 and D_1 are continuous operators from $L_2(\partial D)$ to $L_2(D_-\backslash B_a)$ where B_a is a ball of radius

$a > 0$. The jump conditions are

$$(2.14) \qquad \frac{\partial}{\partial n_p^{\pm}}(S_j\mu) = \pm\mu + K_j\mu, \qquad p \in \partial D$$

and

$$(2.15) \qquad \lim_{p \to p_-^+}(D_j\mu) = \mp\mu + \bar{K}_j^*\mu, \qquad p \in \partial D,$$

where K_j^* is the $L_2(\partial D)$ adjoint of K_j and $^-$ denotes complex conjugate. Note that these relations hold pointwise almost everywhere on ∂D.

3. Existence, uniqueness and complete families. The modified Green's functions described in the previous section may be used to establish the existence of solutions of the standard exterior boundary value problems with L_2-data. It is with these questions that we begin this section. Throughout, we will assume $\operatorname{Re} k > 0$.

We remark, first, that two results established in [4] are of particular interest here. First:

LEMMA 3.1 ([4, Thms. 3.1, 4.1]). *If* $\operatorname{Im} k \geqq 0$ *then*

$$(3.1) \qquad (I + K_j)w = 0 \quad \textit{if and only if} \quad S_j w = 0$$

and

$$(3.2) \qquad (I - \bar{K}_j)w = 0 \quad \textit{if and only if} \quad \frac{\partial}{\partial n}(D_j w) = 0.$$

Second is the result which shows that the homogeneous equation (3.1) has only trivial solutions:

THEOREM 3.1 ([4, Thms. 3.4, 4.4]). *If* $\operatorname{Im} k = 0$, $a_{ll'} = 0$ *if* $l \neq l'$, *and*

$$(3.3) \qquad \left|\alpha_{ll} + \frac{1}{2}\right| < \frac{1}{2} \quad \textit{for all } l,\ |l| \geqq 0,$$

then

$$(3.4) \qquad (I \pm K_1)w = 0 \quad \textit{if and only if} \quad w = 0.$$

The last result is stated for $\operatorname{Im} k = 0$ and it is in this case that the modified Green's function are needed. If $\operatorname{Im} k > 0$ we need consider only the unmodified operator as the following theorem shows.

THEOREM 3.2. *If* $\operatorname{Im} k > 0$ *then*

$$(3.5) \qquad (I \pm K_0)w = 0 \quad \textit{if and only if} \quad w = 0.$$

Proof. The proof proceeds along the usual potential-theoretic lines. If

$$(3.6) \qquad (I + K_0)w = 0$$

then Lemma 3.1 ensures that

$$(3.7) \qquad S_0 w(p) = 0, \qquad p \in \partial D.$$

If we define a function u on D_- by

$$(3.8) \qquad u(p) = S_0 w(p), \qquad p \in D_-$$

then, using Green's theorem, we obtain

$$(3.9)\qquad 0=\int_{\partial D}\left[u\frac{\partial \bar{u}}{\partial n}-\bar{u}\frac{\partial u}{\partial n}\right]ds=[k^2-\bar{k}^2]\int_{D_-}|u|^2dV.$$

But since $k^2-\bar{k}^2\neq 0$ (recall that we are assuming $\operatorname{Re}k>0$) it follows that $u=0$ in D_- and so, with the jump relation (2.14), we have

$$(3.10)\qquad \frac{\partial u}{\partial n^-}=\frac{\partial}{\partial n^-}(S_0w)=(-I+K_0)w=0,$$

which, together with the hypothesis (3.6), implies that $w=0$.

To see that the same result is true for the operator $(I-K_0)$ we observe that, since K_0 is compact, the Fredholm alternative guarantees that the kernels of the operators $I-K_0$ and $I-\bar{K}_0^*$ have the same dimension and hence, to complete the proof, it is sufficient to show that if

$$(3.11)\qquad (I-\bar{K}_0^*)w=0$$

then $w=0$. But if this equation is satisfied, then we may again invoke Lemma 3.1 to see that

$$(3.12)\qquad \frac{\partial}{\partial n^-}(D_0w)=0.$$

We then use the double layer to define u in D_- by

$$(3.13)\qquad u(p)=(D_0w)(p),\qquad p\in D_-$$

and employ the same argument as above with Green's theorem to conclude that

$$(3.14)\qquad u=0\quad \text{in } D_-.$$

The jump relation (2.15) then yields $(I+\bar{K}_0^*)w=0$ which, combined with (3.11), shows that $w=0$ and the proof is complete.

Remark. We should call the reader's attention to the fact that this last theorem is needed at the present juncture because the extension of Theorem 3.1 to the case $\operatorname{Im}k>0$ is not straightforward. The proof of Theorem 3.1 appearing in [4] utilized the usual Wronskian relation for spherical Bessel and Hankel functions and consequently is valid only for real k. The appendix shows how the Wronskian relations may be extended to complex arguments and presents a unified treatment of these results for $\operatorname{Im}k\geqq 0$ and thereby an alternate proof.

Finally we record the analogue of Theorem 3.2 applicable to the Robin problem and which was proven in [3]:

THEOREM 3.3 ([3, Thm. 3.1]). *If* $\operatorname{Im}k\geqq 0$ *and* $\operatorname{Im}\bar{k}\sigma\geqq 0$ *and if* $|\alpha_{ll}+\frac{1}{2}|<\frac{1}{2}$ *for all* l, $|l|\geqq 0$, *then*

$$(3.15)\qquad (I+K_j+\sigma S_j)w=0\quad \textit{if and only if}\quad w=0,$$

where we take $j=1$ *when* $\operatorname{Im}k=0$ *and* $j=0$ *when* $\operatorname{Im}k>0$.

We turn now to a discussion of the boundary value problems for the exterior Helmholtz equation. Notice that, by setting $\sigma=0$, the Robin problem is reduced to the Neumann boundary value problem and hence our proof of existence and uniqueness of solutions of the former problem ensures the same result for the latter. With this observation we can establish the following result:

THEOREM 3.4. *Let ∂D be a closed Lyapunov surface of index* 1 *and let* $\operatorname{Im} k \geqq 0$, $\operatorname{Re} k > 0$. *Then, for every* $f \in L_2(\partial D)$ *there exists a unique solution* $u \in C_2(D_+) \cap L_2(\partial D)$ *of the Helmholtz equation* (2.1) *in the domain* D_+, *exterior to* ∂D, *satisfying the radiation condition* (2.3) *and satisfying either the Dirichlet boundary data* (2.2a)

$$u = f \quad \text{on } \partial D$$

or Robin boundary data (2.2b)

$$\frac{\partial u}{\partial n} + \sigma u = f \quad \text{on } \partial D, \qquad \sigma \in L_\infty(\partial D), \qquad \operatorname{Im} \bar{k} \sigma \geqq 0,$$

in the generalized L_2-sense [2].

Remark. Note, since we require that $u \in C_2(D_+) \cap L_2(\partial D)$, that $\sigma \in L_\infty(\partial D)$ and $f \in L_2(\partial D)$ imply that $\partial u / \partial n \in L_2(\partial D)$.

Proof. Consider, first, the problem with Dirichlet data. The existence of a solution follows from the ansatz

$$u = -D_j \mu \quad \text{in } D_+, \qquad j = \begin{cases} 0, & \operatorname{Im} k > 0, \\ 1, & \operatorname{Im} k = 0. \end{cases} \tag{3.16}$$

This assumption, together with the jump relation (2.15) and the boundary condition, leads to the boundary integral equation for the unknown density μ:

$$\left(I - \bar{K}_j^*\right) \mu = f. \tag{3.17}$$

But Theorems 3.1 and 3.2 show that the null-space of $I - K_j$ and hence that of $I - \bar{K}_j^*$ is trivial and so the integral equation (3.17) is uniquely solvable for all $f \in L_2(\partial D)$. The fact that double layer distributions with L_2-densities assume boundary values in the generalized L_2-sense is established in [2] where it is also shown that the divergence theorem applies even with functions which assume boundary values in this generalized sense. This enables one to essentially repeat the classical uniqueness proof for the Dirichlet problem (e.g. [14]) under the present conditions. It follows that, since the exterior boundary value problem has a unique solution, the function u, defined by (3.16) with density given by the unique solution of the integral equation (3.17), is the solution of the exterior Dirichlet problem.

Turning now to the Robin boundary value problem, we make the ansatz

$$u = S_j w \quad \text{in } D_+, \qquad j = \begin{cases} 0, & \operatorname{Im} k > 0, \\ 1, & \operatorname{Im} k = 0. \end{cases} \tag{3.18}$$

In this case the bounary condition and jump relations lead to the integral equation for w:

$$\left(I + K_j + \sigma S_j\right) w = f. \tag{3.19}$$

Theorem 3.3 guarantees that this integral equation has a unique solution for any $f \in L_2(\partial D)$ and again knowing that single layers and their normal derivatives assume boundary values in $L_2(\partial D)$ as proven in [2], we may invoke the uniqueness theorem for the exterior Robin problem (see [2, Thm. 3.7]) to assert that the function defined in (3.18) is the solution of the exterior boundary value problem.

Having established the existence and uniqueness theorem, we may turn to a discussion of the family of radiating functions defined in (2.7). We first record the important result:

THEOREM 3.5. *The family* $\{v_l^e\}_{|l|=0}^{\infty}$ *defined in* (2.7) *is complete and linearly independent in* $L_2(\partial D)$ *for* $\operatorname{Im} k \geqq 0$.

Proof. The proof of completeness was given by Vekua [15] (see also [16], [17], [18]) whereas linear independence follows if there is no finite linear combination of $\{v_l^e\}$ which vanishes on ∂D. Assume the contrary, that is, assume there are constants C_l, $|l|=0,1,\cdots,N$, such that

$$u=\sum_{|l|=0}^{N} C_l v_l^e=0, \qquad p\in\partial D.$$

Since u is a radiating solution of the Helmholtz equation with zero boundary values, uniqueness of the exterior Dirichlet problem ensures that u vanishes on D_+. The orthogonality of spherical harmonics on spheres then guarantees that

$$\int_{\partial S_A} |u|^2 ds=2|k|A^2\sum_{|l|=0}^{N} |C_l|^2 |h_n^{(1)}(kA)|^2=0 \tag{3.20}$$

where A is the radius of any circumscribing sphere and n is the first component of the multi-index l. Since there are no zeros of $h_n^{(1)}(kA)$ for $\operatorname{Im} k\geqq 0$ [19], it follows that $C_l=0$ for all l which establishes the result.

Finally we have

THEOREM 3.6. *The family*

$$\left\{\frac{\partial v_l^e}{\partial n}+\sigma v_l^e\right\}_{|l|=0}^{\infty}$$

is complete and linearly independent in $L_2(\partial D)$ *for* $\operatorname{Im} k\geqq 0$, $\sigma\in L_\infty(\partial D)$ *and* $\operatorname{Im}\bar{k}\sigma\geqq 0$.

Proof. First we establish completeness. Assume

$$\int_{\partial D}\left(\frac{\partial v_l^e(p)}{\partial n}+\sigma v_l^e(p)\right)f(p)\,ds=0$$

for all l and $f\in L_2(\partial D)$. Let u be the unique solution of the exterior Dirichlet problem with boundary values f (which exists by Theorem 3.4). Then, since u and v_l^e are radiating solutions of the Helmholtz equation,

$$\int_{\partial D}\frac{\partial v_l^e}{\partial n}u\,ds=\int_{\partial D} v_l^e\frac{\partial u}{\partial n}ds; \tag{3.21}$$

hence

(3.22)

$$\int_{\partial D}\left(\frac{\partial v_l^e(p)}{\partial n}+\sigma v_l^e(p)\right)f(p)\,ds=\int_{\partial D} v_l^e(p)\left[\frac{\partial u(p)}{\partial n}+\sigma u(p)\right]ds=0 \quad \text{for all } l.$$

But completeness of $\{v_l^e(p)\}$ (Theorem 3.5) ensures that

$$\frac{\partial u(p)}{\partial n}+\sigma u(p)=0, \qquad p\in\partial D, \tag{3.23}$$

and then uniqueness for the exterior Robin problem shows that

$$u(p)=f(p)=0,\qquad p\in\partial D. \tag{3.24}$$

This establishes completeness. Linear independence follows by assuming that

$$\sum_{|l|=0}^{N} c_l\left(\frac{\partial v_l^e}{\partial n}+\sigma v_l^e\right)=0,\qquad p\in\partial D. \tag{3.25}$$

Defining

$$u=\sum_{|l|=0}^{N} c_l v_l^e,\qquad p\in D_+, \tag{3.26}$$

we see that u is a solution of the homogeneous exterior Robin problem hence vanishes throughout D_+. Then proceeding exactly as in the proof of Theorem 3.5 we see that all of the coefficients c_l must vanish, completing the proof.

Remark. Millar [20] has proved completeness under the more restrictive conditions that k is real and σ is continuous on ∂D. In addition his argument was based on the existence of a unique classical solution with continuous boundary values and the density of trigonometric polynomials in $L_2(\partial D)$. Theorem 2.4 allows for the simpler proof presented here although it is modelled after Millar's argument. An alternative proof of completeness is given by Colton [21] for real k and constant σ.

4. The Green's functions. The Green's function for the Robin problem is a symmetric radiating solution of the Helmholtz equation in each of two points, p and q, satisfying the boundary condition (2.2) with respect to the coordinates of one point for all values of the other in D_+. Following [3], the coefficients in the more complicated modified Green's function (2.9) may be chosen so that the modified Green's function is the best least squares approximation to the actual Green's function for the problem in the sense that the quantity

$$J:=\int_{S_A}\int_{\partial D}\left|\frac{\partial\gamma_1}{\partial n_q}(p,q)+\sigma(q)\gamma_1(p,q)\right|^2 ds_q\,ds_p \tag{4.1}$$

is minimized, where S_A is the boundary of a circumscribing ball of radius A. In what follows, we fix a particular value for A.

With the definition of the modified Green's function (2.9) and the expansion of the free space Green's function (2.6), this may be written explicitly as

(4.2)

$$J=\int_{S_A}\int_{\partial D}\left|\sum_{|l'|=0}^{N} v^e_{l'}(p)\left\{\frac{\partial v^i_{l'}}{\partial n_q}(q)+\sigma(q)v^i_{l'}(q)+\sum_{|l|=0}^{N}\alpha^N_{ll'}\left(\frac{\partial v^e_l}{\partial n_q}(q)+\sigma v^e_l(q)\right)\right.\right.$$

$$\left.\left.+\sum_{|l'|>N}^{\infty} v^e_{l'}(p)\left[\frac{\partial v^i_{l'}}{\partial n_q}(q)+\sigma v^i_{l'}(q)\right]\right\}\right|^2 ds_q\,ds_p,$$

where the superscript N indicates that the coefficient choice may depend on the degree of the modification. The definition of $v^e_{l'}(p)$, equation (2.7), and the orthogonality of spherical harmonics on S_A enable us to explicitly carry out the integration over S_A obtaining

$$(4.3)\quad J=2|k|A^2\sum_{|l'|=0}^{N}\left|h_{n'}^{(1)}(kA)\right|^2$$
$$\cdot\int_{\partial D}\left|\frac{\partial v_{l'}^i}{\partial n_q}(q)+\sigma(q)v_{l'}^i(q)+\sum_{|l|=0}^{N}\alpha_{ll'}^N\left(\frac{\partial v_l^e(q)}{\partial n_q}+\sigma v_l^e(q)\right)\right|^2 ds_q$$
$$+2|k|A^2\sum_{|l'|>N}^{\infty}\left|h_{n'}^{(1)}(kA)\right|^2\int_{\partial D}\left|\frac{\partial v_{l'}^i(q)}{\partial n_q}+\sigma v_{l'}^i(q)\right|^2 ds_q.$$

In this form it is clear that J will be minimized if $\alpha_{ll'}^N$ are chosen so that

$$\sum_{|l|=0}^{N}\alpha_{ll'}^N\left(\frac{\partial v_l^e}{\partial n}+\sigma v_l^e\right)$$

is the best approximation in $L_2(\partial D)$ of $-(\partial v_{l'}^i/\partial n+\sigma v_{l'}^i)$ for $|l'|\leqq N$. Since $\{\partial v_l^e/\partial n+\sigma v_l^e\}_{|l|=0}^{\infty}$ is a complete family and is linearly independent on $L_2(\partial D)$ (Theorem 3.6), this approximation can be made as precise as desired by taking N sufficiently large. An explicit approximation may be obtained by solving the algebraic system

$$(4.4)\qquad \sum_{|l|=0}^{N}\alpha_{ll'}^N\left(\frac{\partial v_l^e}{\partial n}+\sigma v_l^e,\frac{\partial v_t^e}{\partial n}+\sigma v_t^e\right)=-\left(\frac{\partial v_{l'}^i}{\partial n}+\sigma v_{l'}^i,\frac{\partial v_t^e}{\partial n}+\sigma v_t^e\right),$$

where $|l'|$, $|t|\leqq N$.

This equation may be rewritten as a matrix equation if we re-index, eliminating the multi-indices l,l' and t. This may be done in a variety of ways so that (4.4) actually represents $(N+1)^2$ equations with $(N+1)^4$ coefficients $\alpha_{ll'}$. With a slight abuse of notation (4.4) may be written as a matrix equation where α_N, Q_N, Q_N^i are the $(N+1)^2\times(N+1)^2$ matrices defined implicitly in (4.4)

$$(4.5)\qquad \alpha_N Q_N=-Q_N^i.$$

This equation is solvable for the coefficient matrix α_N provided $\det(Q_N)\neq 0$. But this is indeed the case since the set $\{\partial v_l^e/\partial n+\sigma v_l^e\}_{|l|=0}^{N}$ is linearly independent on $L_2(\partial D)$ and (4.5) is always solvable as

$$(4.6)\qquad \alpha_N=-Q_N^iQ_N^{-1}.$$

With $\alpha_{ll'}^N$ so chosen, it follows that

$$(4.7)\qquad \lim_{n\to\infty}\left\|\frac{\partial v_{l'}^e}{\partial n}+\sigma v_{l'}^i+\sum_{|l|=0}^{N}\alpha_{ll'}^N\left(\frac{\partial v_l^e}{\partial n}+\sigma v_l^e\right)\right\|_{L^2(\partial D)}=0.$$

Recall that the exact Green's function for the Robin problem is of the form

$$(4.8)\qquad \gamma_R(p,q)=\gamma_0(p,q)+\Gamma(p,q)$$

where $\Gamma(p,q)$ is a radiating solution of the Helmholtz equation in both p and q. That the modified Green's function described here actually does approximate the Green's function is shown in

THEOREM 4.1. *If $\alpha_{ll'}^N$ are chosen to minimize J, that is as elements of the matrix* (4.6), *then there is a representation*

$$\Gamma(p,q)=\sum_{|l|=0}^{\infty}\sum_{|l|=0}^{\infty}\alpha_{ll'}v_{l'}^e(p)v_l^e(q),\qquad r_p,r_p\geqq A, \tag{4.9}$$

where

$$\alpha_{ll'}=\lim_{N\to\infty}\alpha_{ll'}^N \tag{4.10}$$

and

$$A>\sup_{q\in\partial D}r_q.$$

Proof. Since $\{v_l^e\}_{|l|=0}^{\infty}$ is a complete orthogonal set on any sphere there exists an expansion of the form

$$\Gamma(p,q)=\sum_{|l'|=0}^{\infty}u_{l'}(q)v_{l'}^e(p),\qquad r_p=A, \tag{4.11}$$

where $u_{l'}(q)$ are radiating solutions of the Helmholtz equation. Moreover the expansion of the fundamental solution (2.6) and the boundary condition satisfied by $\gamma_R(p,q)$, (2.2), imply

$$\begin{aligned}0&=\frac{\partial\gamma_R}{\partial n_q}+\sigma(q)\gamma_R\\ &=\sum_{|l'|=0}^{\infty}v_{l'}^e(p)\left[\frac{\partial v_{l'}^i}{\partial n_q}(q)+\sigma v_{l'}^i(q)+\frac{\partial u_{l'}}{\partial n_q}(q)+\sigma u_{l'}(q)\right],\qquad r_p=A,\quad q\in\partial D\end{aligned} \tag{4.12}$$

from which we conclude that

$$\frac{\partial v_{l'}^i}{\partial n_q}(q)+\sigma(q)v_{l'}^i(q)v_{l'}^i(q)+\frac{\partial u_{l'}}{\partial n_q}(q)+\sigma u_{l'}(q)=0\quad\text{a.e. on }\partial D. \tag{4.13}$$

Furthermore since $u_{l'}$ is a radiating solution of the Helmholtz equation, there is an expansion of the form

$$u_{l'}(q)=\sum_{|l|=0}^{\infty}\alpha_{ll'}v_l^e(q),\qquad r_q\geqq A, \tag{4.14}$$

which, with (4.11), shows that Γ is of the form (4.9). Now define

$$u_{l'}^N(q)=\sum_{|l|=0}^{N}\alpha_{ll'}^Nv_l^e(q),\qquad q\in D_+\cup\partial D, \tag{4.15}$$

where $\alpha_{ll'}^N$ are chosen so that (4.7) is satisfied which with (4.13) shows that

$$\lim_{N\to\infty}\left\|\frac{\partial u_{l'}}{\partial n}+\sigma u_{l'}-\frac{\partial u_{l'}^N}{\partial n}-\sigma u_{l'}^N\right\|_{L_2(\partial D)}=0. \tag{4.16}$$

Moreover since both $u_{l'}$ and $u_{l'}^N$ are radiating solutions of the Helmholtz equation we may use Green's representation (2.5) with γ_R in place of γ_0 to obtain

(4.17)

$$u_{l'}(q)-u_{l'}^N(q)=\frac{1}{2}\int_{\partial D}\gamma_R(q,q')\left[\frac{\partial u_{l'}(q')}{\partial n_{q'}}+\sigma u_{l'}(q')-\frac{\partial u_{l'}^N(q')}{\partial n_{q'}}-\sigma u_{l'}^N(q')\right]ds_{q'},$$

$$p\in D_+,$$

and since $\gamma_R(p,q)$ is bounded for $r_p=A$ and $q\in\partial D$, provided $A>\sup_{q\in\partial D}r_q$, there is some constant M such that

$$|u_{l'}(q)-u_{l'}^N(q)|^2\leqq M\left\|\frac{\partial u_{l'}}{\partial n}+\sigma u_{l'}-\frac{\partial u_{l'}^N}{\partial n}-\sigma u_{l'}^N\right\|^2_{L_2(\partial D)}. \tag{4.18}$$

Substituting (4.14) and (4.15) and integrating over S_A yields

$$2|k|\sum_{|l|=0}^{N}|h_n^{(1)}(kA)|^2|\alpha_{ll'}-\alpha_{ll'}^N|^2+2|k|\sum_{|l|>N}^{\infty}|h_n^{(1)}(kA)|^2|\alpha_{ll'}|^2 \tag{4.19}$$
$$\leqq 4\pi M\left\|\frac{\partial u_{l'}}{\partial n}+\sigma u_{l'}-\frac{\partial u_{l'}^N}{\partial n}-\sigma u_{l'}^N\right\|^2_{L_2(\partial D)},$$

where n is the first component of the multi-index l. With (4.16) we see that

$$\lim_{N\to\infty}\alpha_{ll'}^N=\alpha_{ll'}. \tag{4.20}$$

This completes the proof and shows how the modified Green's function approximates the actual Green's function for the problem. The question of whether the series representation for the Green's function (4.9) will remain a valid representation if r_p and r_q are less than A remains unanswered but it seems clear that it will be valid for all p and q in D_+ *only* if some form of Rayleigh criterion [16] is fulfilled.

5. The null field equations. The null field equations for the Robin problem are most easily derived from the Helmholtz representation. If u is the solution of the exterior Robin problem (2.1)–(2.3) then

$$\int_{\partial D}\left(u\frac{\partial v_l^e}{\partial n}-v_l^e\frac{\partial u}{\partial n}\right)ds=\int_{\partial D}\left\{\left(\frac{\partial v_l^e}{\partial n}+\sigma v_l^e\right)u-v_l^e f\right\}ds=0, \tag{5.1}$$

and therefore

$$\int_{\partial D}\left(\frac{\partial v_l^e}{\partial n}+\sigma v_l^e\right)u\,ds=\int_{\partial D}v_l^e f\,ds\quad\text{for all } l,\ |l|=0,1,2,\cdots, \tag{5.2}$$

which are the null field equations. From Theorems 3.4 and 3.6, we know that there is a unique solution of the system (5.2) in $L_2(\partial D)$ for $\operatorname{Re}k>0$, $\operatorname{Im}k\geqq 0$, $\operatorname{Im}\bar{k}\sigma\geqq 0$, $f\in L_2(\partial D)$ and $\sigma\in L_\infty(\partial D)$.

The usual treatment of these equations involves approximating u on ∂D with a linear combination of elements of a complete family and solving the resulting algebraic equations. Of course there are many complete families from which to choose and here we only consider one such expansion which will allow us to relate the results of the previous section to the so-called "null field method" introduced by Waterman [22].

By virtue of Theorems 3.5 and 3.6, the sets $\{v_l^e\}_{|l|=0}^{\infty}$ and $\{\partial v_l^e/\partial n+\sigma v_l^e\}_{|l|=0}^{\infty}$ are complete and linearly independent families. If, in addition, they were *bases* for $L_2(\partial D)$ then we could expand the unknown function u and the known data f in terms of these families:

$$u(p)=\sum_{|l'|=0}^{\infty} c_{l'}v_{l'}^e(p), \qquad p\in\partial D \tag{5.3}$$

and

$$f(p)=\sum_{|l'|=0}^{\infty} d_{l'}\left(\frac{\partial v_{l'}^e}{\partial n}+v_{l'}^e\right), \qquad p\in\partial D, \tag{5.4}$$

substitute in (5.2) obtaining

$$\int_{\partial D}\sum_{|l'|=0}^{\infty} c_{l'}v_{l'}^e\left(\frac{\partial v_l^e}{\partial n}+\sigma v_l^e\right)ds=\int_{\partial D}\sum_{|l'|=0}^{\infty} d_{l'}\left(\frac{\partial v_{l'}^e}{\partial n}+\sigma v_{l'}^e\right)v_l^e\,ds \tag{5.5}$$

which, since

$$\int_{\partial D}\frac{\partial v_{l'}^e}{\partial n}v_l^e\,ds=\int_{\partial D}v_{l'}^e\frac{\partial v_l^e}{\partial n}\,ds, \tag{5.6}$$

may be written as

$$\int_{\partial D}\sum_{|l'|=0}^{\infty}(c_{l'}-d_{l'})v_{l'}^e\left(\frac{\partial v_l^e}{\partial n}+\sigma v_l^e\right)ds=0. \tag{5.7}$$

But the completeness of $\{\partial v_l^e/\partial n+\sigma v_l^e\}_{|l|=0}^{\infty}$ implies that

$$\sum_{|l''|=0}^{\infty}(c_{l'}-d_{l'})v_{l'}^e=0, \tag{5.8}$$

and the linear independence of $\{v_l^e\}_{|l|=0}^{\infty}$ ensures that

$$c_l=d_l. \tag{5.9}$$

That is, the coefficients of an expansion of the unknown function u in $\{v_l^e\}$ are the same as the coefficients of an expansion of the known data f in $\{\partial v_l^e/\partial n+\sigma v_l^e\}$.

However, unless the boundary ∂D is considerably restricted, the families $\{v_l^e\}_{|l|=0}^{\infty}$ and $\{\partial v_l^e/\partial n+\sigma v_l^e\}_{|l|=0}^{\infty}$ will not be bases for $L_2(\partial D)$ (e.g. [23]) hence in place of (5.3) and (5.4), we may only assume approximations of the form

$$u^N(p)=\sum_{|l|=0}^{N} c_l^{(N)}v_l^e(p), \qquad p\in\partial D, \tag{5.10}$$

and

$$f^N(p)=\sum_{|l|=0}^{N} d_l^{(N)}\left(\frac{\partial v_l^e}{\partial n}+\sigma v_l^e\right)(p), \qquad p\in\partial D, \tag{5.11}$$

where the coefficients are chosen so the approximations are best in $L_2(\partial D)$. Since f is known, the coefficients $d_l^{(N)}$ may be determined explicitly, however, since u is as yet unknown, the coefficients $c_l^{(N)}$ are likewise unknown. Nevertheless, the completeness and linear independence of $\{v_l^e\}$ insure that such coefficients exist so that

$$\lim_{N\to\infty}\|u-u^N\|_{L_2(\partial D)}=0 \tag{5.12}$$

and

$$\lim_{N\to\infty}\|f-f^N\|_{L_2(\partial D)}=0. \tag{5.13}$$

Now extend the definition (5.10) of $u^N(p)$ to all $p\in D_+$ and define another function

$$\phi^N(p)=\sum_{|l|=0}^{N} d_l^{(N)}v_l^e(p), \qquad p\in D_+ \tag{5.14}$$

so that

$$\frac{\partial\phi^N}{\partial n}+\sigma\phi^N=f^N. \tag{5.15}$$

We may use the Green's functions for the Dirichlet and Robin problems, denoted by γ_D and γ_R respectively, to obtain the representations

$$u(p)-u^N(p)=-\frac{1}{2}\int_{\partial D}[u(q)-u^N(q)]\frac{\partial}{\partial n_q}\gamma_D(p,q)\,ds_q \tag{5.16}$$

and

$$\begin{aligned} u(p)-\phi^N(p)&=\frac{1}{2}\int_{\partial D}\gamma_R(p,q)\left[\frac{\partial u}{\partial n_q}+\sigma u-\frac{\partial\phi^N}{\partial n_q}-\sigma\phi^N\right]ds_q \\ &=\frac{1}{2}\int_{\partial D}\gamma_R(p,q)[f(q)-f^N(q)]\,ds_q. \end{aligned} \tag{5.17}$$

Since $\{v_l^e\}_{|l|=0}^{\infty}$ is a complete orthogonal set on any circumscribing sphere there is an expansion

$$u(p)=\sum_{|l|=0}^{\infty} c_l v_l^e(p), \qquad r_p>\sup_{q\in\partial D} r_q. \tag{5.18}$$

Moreover for $r_p=A>\sup_{q\in\partial D}r_q$ and q on ∂D both $\gamma_R(p,q)$ and $(\gamma_D/\partial n_q)(p,q)$ are bounded with, for convenience, the common bound M. Then

$$|u(p)-u^N(p)|\leqq\frac{M}{2}\|u-u^N\|_{L_2(\partial D)}, \tag{5.19}$$

and

$$|u(p)-\phi^{(N)}(p)|<\frac{M}{2}\|f-f^N\|_{L_2(\partial D)}. \tag{5.20}$$

Employing the expansions (5.10), (5.14), and (5.18) and integrating over the sphere of radius A we obtain

$$\int_{S_A} |u(p)-u^{(N)}(p)|^2 ds = \int_{S_A} \left| \sum_{|l|}^{N} (c_l - c_l^{(N)}) v_l^e(p) + \sum_{|l|>N}^{\infty} c_l v_l^e \right|^2 ds \tag{5.21}$$

$$\leqq 2\pi M A^2 \|u-u^N\|_{L_2(\partial D)},$$

and, with orthogonality of the $\{v_l^e\}_{|l|=0}^{\infty}$ on S_A

$$|k| \left\{ \sum_{|l|=0}^{N} |c_l - c_l^{(N)}|^2 |h_n^{(1)}(kA)|^2 + \sum_{|l|>N}^{\infty} |c_l|^2 |h_n^{(1)}(kA)|^2 \right\} \leqq \pi M \|u-u^N\|_{L_2(\partial D)}, \tag{5.22}$$

and similarly

(5.23)

$$\int_{S_A} |u(p)-\phi^N(p)|^2 ds = 2|k|A^2 \left\{ \sum_{|l|=0}^{N} |c_l - d_l^{(N)}|^2 |h_n^{(1)}(kA)|^2 + \sum_{|l|>N}^{\infty} |c_l|^2 |h_n^{(1)}(kA)|^2 \right\}$$

$$\leqq 2\pi A^2 \|f-f^N\|_{L_2(\partial D)}.$$

Hence with (5.12) and (5.13) we see that

$$\lim_{N\to\infty} c_l^{(N)} = \lim_{N\to\infty} d_l^{(N)} = c_l. \tag{5.24}$$

Thus even though

$$\{v_l^e\}_{|l|=0}^{\infty} \quad \text{and} \quad \left\{ \frac{\partial v_l^e}{\partial n} + \sigma v_l^e \right\}_{|l|=0}^{\infty}$$

may not be bases for $L_2(\partial D)$, the fact that they are complete and linearly independent ensures convergence of the finite dimensional approximations and the approximate coefficients in (5.10) actually converge to the coefficients in the expansion of the solution in outgoing waves for points outside a circumscribing sphere.

Finally we consider the case when the data f is of the form

$$f = -\frac{\partial u^i}{\partial n} - \sigma u^i, \tag{5.25}$$

where u^i represents an incident field. This is the form in which the boundary data often appears in scattering problems. It is usually convenient to represent u^i in terms of incoming waves

$$u^i(p) = \sum_{|l'|=0}^{\infty} a_{l'} v_{l'}^i(p), \qquad r_p < A, \tag{5.26}$$

where A is the radius of a circumscribing sphere which contains no sources.

Then the null field equations (5.2) become

$$\int_{\partial D} \left(\frac{\partial v_t^e}{\partial n} + \sigma v_t^e \right) u \, ds = -\sum_{|l'|=0}^{\infty} a_{l'} \int_{\partial D} v_t^e \left(\frac{\partial v_{l'}^i}{\partial n} + \sigma v_{l'}^i \right) ds \tag{5.27}$$

where, for notational convenience, we have replaced the multi-index l with t. Now introduce the approximations (cf. (5.10) and (4.4))

$$u = \sum_{|l'|=0}^{N} c_{l'}^{(N)} v_{l'}^{e}, \tag{5.28}$$

and

$$\frac{\partial v_{l'}^{i}}{\partial n} + \sigma v_{l'}^{i} = -\sum_{|l|=0}^{N} \alpha_{ll'}^{N} \left(\frac{\partial v_{l}^{e}}{\partial n} + \sigma v_{l}^{e} \right), \tag{5.29}$$

to obtain the approximate equations

$$\begin{aligned} \sum_{|l'|=0}^{N} c_{l'}^{(N)} \int_{\partial D} \left(\frac{\partial v_{t}^{e}}{\partial n} + \sigma v_{t}^{e} \right) v_{l'}^{e}\, ds &= \sum_{|l|=0}^{N} \sum_{|l'|=0}^{N} a_{l'} \alpha_{ll'}^{N} \int_{\partial D} v_{t}^{e} \left(\frac{\partial v_{l}^{e}}{\partial n} + \sigma v_{l}^{e} \right) ds \\ &= \sum_{|l|=0}^{N} \sum_{|l'|=0}^{N} a_{l'} \alpha_{ll'}^{N} \int_{\partial D} \left(\frac{\partial v_{t}^{e}}{\partial n} + \sigma v_{t}^{e} \right) v_{l}^{e}\, ds \end{aligned} \tag{5.30}$$

where the relation (5.6) was used.

Using matrix notation (5.30) is of the form

$$Qc = Q\alpha_N a \tag{5.31}$$

where α_N is the matrix of coefficients of the Green's function given in (4.6). This equation has the obvious solution

$$c = \alpha_N a \tag{5.32}$$

in which form α_N is often referred to as the T-matrix or transition matrix for the Robin problem since it transforms the coefficients $a = (a_{l'})$ of the incident wave (5.26) into the coefficients $c = (c_{l'}^{(N)})$ of the scattered field (5.28). We see from §4 that the elements of the T-matrix, $\alpha_{ll'}^{N}$ which are chosen to best approximate $-(\partial v_{l'}^{i}/\partial n + \sigma v_{l'}^{i})$ in terms of a finite number of $\partial v_{l}^{e}/\partial n + \sigma v_{l}^{e}$ see (5.29), are exactly the same as the coefficients introduced in section 4 (cf. (4.3)) in order for the modified Green's function to best approximate the actual Green's function for the Robin problem. With Theorem 4.1 then it follows that the elements of the T-matrix become the exact Green's function coefficients as $N \to \infty$.

Appendix. A modified Green's function for the Dirichlet and Neumann problems. In Theorems 3.1 and 3.2 it was shown that the integral equations of the second kind arising in the Dirichlet and Neumann problems could be shown to be uniquely solvable. However for $\operatorname{Im} k = 0$ a modified Green's function was needed whereas for $\operatorname{Im} k > 0$ no modification was necessary. In this appendix it is shown that a modified Green's function may be introduced with coefficients which depend continuously on k in such a way that the modified equations are uniquely solvable without the necessity of treating separately the cases when $\operatorname{Im} k = 0$ and $\operatorname{Im} k > 0$.

First we must extend the Wronskian relation for spherical Bessel and Hankel functions with complex arguments. For any two differentiable functions f and g define

$$w(f,g) := f(kr)\frac{d}{dr}g(kr) - \frac{d}{dr}f(kr)g(kr). \tag{A.1}$$

If Z_n denotes a solution of the spherical Bessel equation

$$r^2\frac{d^2}{dr^2}Z_n(kr)+2r\frac{d}{dr}Z_n(kr)+\left[(kr)^2-n(n+1)\right]_n(kr)=0, \tag{A.2}$$

then we have the following identities.

LEMMA A.1.

$$w(j_n,\overline{j_n})=\frac{k^2-\overline{k}^2}{r^2}\int_0^r t^2|j_n(kt)|^2dt, \tag{A.3}$$

$$w\left(h_n^{(1)},\overline{h_n^{(1)}}\right)=\begin{cases}-\dfrac{2i}{kr^2}, & \operatorname{Im}k=0,\\ \dfrac{\overline{k}^2-k^2}{r^2}\displaystyle\int_r^\infty t^2\left|h_n^{(1)}(kt)\right|^2dt, & \operatorname{Im}k>0.\end{cases} \tag{A.4}$$

Proof. Note that if k is complex $\overline{Z}_n=\overline{Z_n(kr)}=\overline{Z}_n(\overline{k}r)$. Nevertheless the differential equation (A.2) together with its complex conjugate may be used to show that

$$\frac{d}{dr}\left[r^2w(Z_n,\overline{Z_n})\right]=(k^2-\overline{k}^2)r^2|Z_n(kr)|^2. \tag{A.5}$$

Using the fact that $j_n(0)=0$, integration of (A.5) with $Z_n=j_n$ yields (A.3) whereas the known asymptotic behavior of $h_n^{(1)}$ is used to obtain (A.4), again by integrating (A.5) with $Z_n=h_n^{(1)}$ for $\operatorname{Im}k>0$. When k is real (A.4) is merely the Wronskian since

$$\overline{h_n^{(1)}(kr)}=h_n^{(2)}(kr), \qquad \operatorname{Im}k=0. \tag{A.6}$$

If $\operatorname{Im}k>0$ observe that since

$$w(j_n,\overline{j_n})=-w(\overline{j_n},j_n) \tag{A.7}$$

it follows that

$$\frac{w(\overline{j_n},j_n)}{w\left(h_n^{(1)},\overline{h_n^{(1)}}\right)}=\frac{\int_0^r t^2|j_n(kt)|^2dt}{\int_r^\infty t^2\left|h_n^{(1)}(kt)\right|^2dt}>0. \tag{A.8}$$

This enables us to state and prove the following generalization of [4, Thms. 3.4, 4.4].

THEOREM A.1. *If the coefficients in the modified Green's function* (2.9) *satisfy*

$$\alpha_{ll'}=0,\quad l\neq l',\quad \left|\alpha_l+\frac{w\left(j_n,\overline{h_n^{(1)}}\right)}{w\left(h_n^{(1)},\overline{h_n^{(1)}}\right)}\right|<\left\{\left|\frac{w\left(j_n,\overline{h_n^{(1)}}\right)}{w\left(h_n^{(1)},\overline{h_n^{(1)}}\right)}\right|^2+\frac{w(\overline{j_n},j_n)}{w\left(h_n^{(1)},\overline{h_n^{(1)}}\right)}\right\}^{1/2} \tag{A.9}$$

and $\operatorname{Im}k\geqq 0$ *then*

$$(I\pm K_1)w=0 \quad \textit{implies} \quad w=0.$$

Remark. Note that in (A.9) the index n is the first component of the multi-index l and (A.8) ensures that argument of the square root in (A.9) indeed positive.

Proof. If $\operatorname{Im}k=0$ then the Wronskian relations

$$w(j_n,\overline{j_n})=0,\quad w\left(j_n,\overline{h_n^{(1)}}\right)=-\frac{i}{kr^2},\quad w\left(h_n^{(1)},\overline{h_n^{(1)}}\right)=-\frac{2i}{kr^2} \tag{A.10}$$

show that (A.9) reduces to $|\alpha_{ll}+\frac{1}{2}|<\frac{1}{2}$, the case considered in [4], for which the theorem was proven.

If $\operatorname{Im} k>0$ then we proceed along similar lines. Consider first the equation

$$(I+K_1)w=0. \tag{A.11}$$

Lemma 3.1 ensures that

$$S_1 w=0, \qquad p\in\partial D. \tag{A.12}$$

Now define

$$u(p)=S_1 w, \qquad p\in D_-\setminus\{0\}. \tag{A.13}$$

The boundary condition (A.12) together with Green's theorem applied over a domain bounded by ∂D and S_a, the surface of a ball B_a of radius a with center at the origin lying entirely in D_-, are used to write

$$\begin{aligned} 0&=\int_{\partial D}\left[u(p)\frac{\partial}{\partial n_p}\bar{u}-\bar{u}\frac{\partial}{\partial n_p}u(p)\right]ds_p \\ &=(k^2-\bar{k}^2)\int_{D_-\setminus B_a}|u|^2 dv+\int_{S_a}\left[u(p)\frac{\partial}{\partial a}\bar{u}-\bar{u}\frac{\partial}{\partial a}u\right]ds_p. \end{aligned} \tag{A.14}$$

But, for $p\in S_a$ and $\alpha_{ll'}=0$, $l\neq l'$,

$$\begin{aligned} u(p)=(S_1 w)(p)&=\int_{\partial D}\sum_{|l|=0}^{\infty} v_l^e(q)\left[v_l^i(p)+\alpha_{ll}v_l^e(p)\right]w(q)\,ds_q \\ &=\sum_{|l|=0}^{\infty} c_l\left[v_l^i(p)+\alpha_{ll}v_l^e(p)\right], \end{aligned} \tag{A.15}$$

where

$$c_l=\int_{\partial D} v_l^e(q)w(q)\,ds_q. \tag{A.16}$$

Now the orthogonality of spherical harmonics on spheres and the definitions (2.7) of the functions $v_l^{e,i}$ lead to

$$\begin{aligned} &\int_{S_a}\left[u(p)\frac{\partial}{\partial a}\bar{u}-\bar{u}\frac{\partial u}{\partial a}\right]ds_p \\ &=2i\operatorname{Im}\sum_{|l|=0}^{\infty}\sum_{|l'|=0}^{\infty}\int_{S_a} c_l\bar{c}_l\left[v_l^i+\alpha_{ll}v_l^e\right]\frac{\partial}{\partial a}\left[\bar{v}_l^i+\bar{\alpha}_{ll}\bar{v}_l^e\right]ds_p \\ &=\sum_{|l|=0}^{\infty}|c_l|^2 2|k|a^2\left\{\left(j_n+\alpha_{ll}h_n^{(1)}\right)\frac{\partial}{\partial a}\left(\bar{j}_n+\bar{\alpha}_{ll}\overline{h_n^{(1)}}\right)-\left(\bar{j}_n+\overline{\alpha_{ll}h_n^{(1)}}\right)\frac{\partial}{\partial a}\left[j_n+\alpha_{ll}h_n^{(1)}\right]\right\} \\ &=2|k|a^2\sum_{|l|=0}^{\infty}|c_l|^2 w\left(h_n^{(1)},\overline{h_n^{(1)}}\right) \\ &\qquad\cdot\left\{\left|\alpha_{ll}+\frac{w\left(j_n,\overline{h_n^{(1)}}\right)}{w\left(h_n^{(1)},\overline{h_n^{(1)}}\right)}\right|^2-\left|\frac{w\left(j_n,\bar{h}_n^{(1)}\right)}{w\left(h_n^{(1)},\overline{h_n^{(1)}}\right)}\right|^2-\frac{w\left(\bar{j}_n,j_n\right)}{w\left(h_n^{(1)},h_n^{(1)}\right)}\right\}. \end{aligned} \tag{A.17}$$

With Lemma A.1 this becomes

$$\begin{aligned}&\int_{S_a}\left[u(p)\frac{\partial\bar{u}}{\partial a}(p)-\bar{u}(p)\frac{\partial u}{\partial a}(p)\right]ds_p\\ &\quad=2|k|(\bar{k}^2-k^2)\sum_{|l|=0}^{\infty}|c_l|^2\int_a^{\infty}t^2|h^{(1)}(kt)|^2dt\\ &\quad\cdot\left\{\left|\alpha_{ll}+\frac{w\left(j_n,\overline{h_n^{(1)}}\right)}{w\left(h_n^{(1)},\overline{h_n^{(1)}}\right)}\right|-\left|\frac{w\left(j_n,\overline{h_n^{(1)}}\right)}{w\left(h_n^{(1)},\overline{h_n^{(1)}}\right)}\right|^2-\frac{w(\bar{j}_n,j_n)}{w\left(h_n^{(1)},\overline{h_n^{(1)}}\right)}\right\}\end{aligned}\tag{A.18}$$

and substituting in (A.14) we have

$$\begin{aligned}0=(k^2-\bar{k}^2)&\left\{\int_{D_-\setminus B_a}|u|^2dv+2|k|\sum_{|l|=0}|c_l|^2\int_a^{\infty}t^2|h_n^{(1)}(kt)|^2dt\right.\\ &\left.\cdot\left[-\left|\alpha_{ll}+\frac{w\left(j_n,\overline{h_n^{(1)}}\right)}{w\left(h_n^{(1)},\overline{h_n^{(1)}}\right)}\right|^2+\left|\frac{w\left(j_n,\overline{h_n^{(1)}}\right)}{w\left(h_n^{(1)},\overline{h_n^{(1)}}\right)}\right|^2+\frac{w(\bar{j}_n,j_n)}{w\left(h_n^{(1)},\bar{h}_n^{(1)}\right)}\right]\right\},\end{aligned}\tag{A.19}$$

and we see that both terms in brackets on the right will be positive if (A.9) is fulfilled.

Hence

$$u(p)=0,\qquad p\in D_-\setminus B_a\tag{A.20}$$

so the normal derivative on ∂D exists from D_- and vanishes. Using the definition (A.13) and jump relation (2.14) it follows that

$$\frac{\partial u}{\partial n^-}(p)=\frac{\partial}{\partial n}(S_1w)(p)=(-I+K_1)w(p)=0,\qquad p\in\partial D\tag{A.21}$$

which, with the assumption (A.11), establishes that w vanishes, thus proving the first part of the theorem.

If

$$\left(I-\bar{K}_1^*\right)w=0,\qquad p\in\partial D\tag{A.22}$$

then Lemma 3.1 ensures that

$$\frac{\partial}{\partial n^{\pm}}(D_1w)=0,\qquad p\in\partial D.\tag{A.23}$$

Defining

$$u(p)=D_1w(p),\qquad p\in D_-\setminus\{0\},\tag{A.24}$$

it follows that, for $p\in S_a$ and $\alpha_{ll'}=0$, $l\neq l'$,

$$u(p)=\sum_{|l|=0}^{\infty}d_l\left[v^i(p)+\alpha_{ll}v_l^e(p)\right],\tag{A.25}$$

where

$$d_l = \int_{\partial D} \frac{\partial v_l^e}{\partial n_q}(q) w(q)\, ds_q. \tag{A.26}$$

Then the analysis of (A.17)–(A.20) may be repeated with d_l replacing c_l and since in this case

$$u = D_1 w = 0, \qquad p \in D_- \setminus B_a \tag{A.27}$$

the jump relations (2.15) imply that

$$w + \bar{K}_1^* w = 0 \tag{A.28}$$

which with (A.22) ensures that w vanishes. Hence the null-space of $\bar{K}_1^*$ has dimension 0, as does the null-space of K_1 since K_1 is compact. This completes the proof of the theorem.

Remark. Since the modified Green's function is an infinite series, convergence is assured if α_{ll} is additionally restricted to satisfy

$$|\alpha_{ll}| \leqq \frac{1}{2^n \sup_{p\in D_+} |v_l^e(p)|^2}. \tag{A.29}$$

In fact (A.29) ensures that

$$\sum_{|l|=0}^{\infty} \alpha_{ll} v_l^e(p) v_l^e(q) \tag{A.30}$$

converges uniformly and absolutely in $\bar{D}_+$. The relation (A.29) is compatible with (A.9) since α_{ll} may be chosen to vanish if $\operatorname{Im} k > 0$ hence can be chosen as small as we wish. For $\operatorname{Im} k = 0$, α_{ll} may not vanish but may be chosen as small as we wish provided only that $|\alpha_{ll} + \frac{1}{2}| < \frac{1}{2}$.

REFERENCES

[1] R. LEIS, *Über das Neumannsche Randwertproblem für die Helmholtzsche Schwingungsgleichung*, Arch. Rational Mech. Anal, 2 (1958), pp. 101–18.

[2] T. S. ANGELL AND R. E. KLEINMAN, *Boundary integral equations for the Helmholtz equation: The third boundary value problem*, Math. Meth. Appl. Sci., 4 (1982), pp. 64–193.

[3] ______, *Modified Green's functions and the third boundary value problem for the Helmholtz equation*, J. Math. Anal. Appl., 97 (1983), pp. 81–94.

[4] R. E. KLEINMAN AND G. F. ROACH, *On modified Green functions in exterior problems for the Helmholtz equation*, Proc. Roy Soc. London A, 383 (1982), pp. 213–332.

[5] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Springer-Verlag, New York-Heidelberg-Berlin, 1972.

[6] JU. M. BEREZANSKII, *Expansions in Eigenfunctions of Self adjoint Operators*, Amer. Math. Soc. Trans. Math. Monographs, 17, 1968.

[7] C. MIRANDA, *Partial Differential Equations of Elliptic Type*, 2nd rev. ed., Springer-Verlag, New York-Heidelberg-Berlin, 1970.

[8] V. P. MIKHAILOV, *On the boundary values of the solutions of elliptic equations*, Appl. Math. Optim., 6 (1980), 193–199.

[9] ______, *On Dirichlet's problem for elliptic equations of the second order*, Differential Eq., 12 (1976), pp. 1877–1891.

[10] P. A. MARTIN, *Acoustic Scattering and radiation problems and the null field method*, Wave Motion, 4 (1982), pp. 391–408.

[11] V. K. Varadan and V. V. Varadan, eds., *Acoustic, Electromagnetic and Elastic Wave Scattering—Focus on the T-Matrix Approach*, Pergamon, New York, 1980.

[12] R. E. Kleinman, G. F. Roach, and S. E. G. Ström, *The null field method and modified Green's functions*, Proc. Roy. Soc. London A, 394 (1984), pp. 121–136.

[13] S. G. Mikhlin, *Mathematical Physics: An Advanced Course*, North-Holland, Amsterdam, 1970.

[14] I. N. Vekua, *New Methods for Solving Elliptic Equations*, North-Holland, Amsterdam, 1967.

[15] ———, *On completeness of a system of metaharmonic functions*, Dokl. Akad. Nauk. SSSR, 90 (1953), pp. 715–718.

[16] 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.

[17] D. Colton and R. Kress, *The unique solvability of the null field equations of acoustics*, Quart. J. Mech. Appl. Math., 36 Pt. 1 (1983), pp. 87–95.

[18] C. Müller and H. Kersten, *Zwei Klassen Vollständiger Funktionensysteme zur behandlung der Randwertaufgaben der Schwingungsgleichung* $\Delta u + k^2 u = 0$, Math. Meth Appl. Sci., 2, (1980), pp. 47–67.

[19] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, AMS 55, National Bureau of Standards, Washington, 1964.

[20] R. F. Millar, *On the completeness of solutions to the Helmholtz equation*, IMA J. Appl. Math., 30 (1983), pp. 27–38.

[21] D. Colton, *Far field patterns for the impedance boundary value problem in acoustic scattering*, Applic. Anal., 16 (1983), pp. 131–139.

[22] P. C. Waterman, *New formulation of acoustic scattering*, J. Acoust. Soc. Amer., 45 (1969), pp. 1417–1429.

[23] G. Kristennson, A. G. Ramm and S. Ström, *Convergence of the T-matrix approach in scattering theory II*, J. Math. Phys., 24 (1983), pp. 2619–2631.

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

006

APPLICATIONS OF VARIATIONAL INEQUALITIES TO A MOVING BOUNDARY PROBLEM FOR HELE SHAW FLOWS*

BJÖRN GUSTAFSSON[†]

Abstract. We consider a class of two-dimensional moving boundary problems originating from a Hele Shaw flow problem. Concepts of classical and weak solutions are introduced. We show that a classical solution also is a weak solution and, by using variational inequalities, that given arbitrary initial ($t=0$) data there exists a unique weak solution defined on the time interval $0 \leq t < \infty$. We also prove some monotonicity properties of weak solutions and that, under reasonable hypotheses, the moving boundaries consist of analytic curves for $t>0$.

Key words. Hele Shaw flow, moving boundary problem, variational inequalities

Introduction. The aim of the present paper is to prove a global existence and uniqueness theorem for a kind of weak solution to a moving boundary problem arising in two-dimensional Hele Shaw flows. The method used is that of transforming the problem into a series of elliptic variational inequalities.

The problem we shall treat is a slight generalization of the following problem. Let, for D any bounded region in $\mathbb{R}^2$ containing the origin, g_D be the Green's function for D with respect to the origin:

$$g_D(z)=\begin{cases} -\log|z|+\text{harmonic} & \text{in } D, \\ 0 & \text{on } \partial D \end{cases}$$

$$\left(z=x+iy,\ \mathbb{R}^2 \text{ being identified with } \mathbb{C}\right).$$

Then, given an initial domain $D=D_0$, we want to find a family of domains $\{D_t\}$ for $t \geq 0$ ($t=$time) such that ∂D_t moves with the velocity $-(\nabla g_{D_t})|_{\partial D_t}$. (It is assumed here that $\nabla g_{D_t}=$the gradient of g_{D_t} has a continuous extension to ∂D_t.)

This problem (essentially) was introduced by S. Richardson in [12]. The physical interpretation for it as described in [12] is, very briefly, that D_t is the two-dimensional picture of the region of flow in a Hele Shaw flow with a (time-dependent) free boundary and a source point. This means more precisely that an incompressible viscous Newtonian fluid occupies part of the space between two parallel, narrowly separated, infinitely extended surfaces and that more fluid is injected at a constant and moderate rate through a hole in one of the surfaces. The region occupied by fluid then will grow as time increases and, since the gap between the two surfaces is very small, that region can be very well described by its projection D_t onto and $\mathbb{R}^2$-plane lying parallel to the surfaces. The origin of that $\mathbb{R}^2$-plane is taken to correspond to the injection point. For more details and for a derivation of the moving boundary condition above, see [12] and [8]. An incompressible viscous flow in the narrow space between two parallel surfaces is called a Hele Shaw flow. See e.g. [10, p. 581ff.].

The approach in [12] is that of formulating the problem as a differential equation for the Riemann mapping function from the unit disc onto D_t, identifying $\mathbb{R}^2$ with $\mathbb{C}$ and assuming that the D_t are simply connected. No proof of existence or uniqueness of solutions of this differential equation is given in [12]. However, a local (for t in a small,

* Received by the editors June 8, 1983.

[†] Division of Mathematics, Royal Institute of Technology, 10044 Stockholm, Sweden.

two-sided interval about zero) existence and a partial uniqueness proof for the same differential equation have been given in [19]. See also [8].

Since 1972 Richardson's moving boundary problem has been taken up by J. R. Ockendon [11], C. M. Elliott–V. Janovský [6], S. Richardson himself [13], [14], M. Sakai [16], [17], and me [7], [8]. The present paper is, to a large extent, a summary of [7]. It also has much in common with [17] and more detailed references to that paper will be given at relevant places in the text.

The paper is organized as follows. In §1 we define in a precise way what is meant by being a (local) solution of the problem stated above, by introducing a concept of "classical solution." In §2 we also introduce a concept of "weak solution" and prove that a classical solution is a weak solution. In §3 and 4 we prove that being a weak solution is equivalent to satisfying a series of variational inequalities. From this our main result, the existence and uniqueness of weak solutions for arbitrary given initial domains, follows immediately. Section 5 is devoted to proving that a weak solution is equivalently characterized as the solution of what we call "the moment inequality." Finally, in §6, we summarize part of our results in a kind of main theorem (Theorem 8) and also obtain some partial results on the regularity of the boundaries of the domains of a solution.

List of some notation frequently used.

$\mathbb{R}^2$ is identified with $\mathbb{C}$ whenever convenient (by $(x,y)\leftrightarrow z=x+iy$).

$\mathbb{D}(a;r)=\{z\in\mathbb{C}:|z-a|<r\}$.

$\mathbb{D}(r)=\mathbb{D}(0;r)$.

$\mathbb{D}=\mathbb{D}(0;1)$.

$\mathbb{Z}=$ the set of integers.

$(a,b)=\{x\in\mathbb{R}:a<x<b\}$.

$[a,b]=\{x\in\mathbb{R}:a\le x\le b\}$.

$D\subset\subset\Omega$ means: $\bar{D}\subset\Omega$ (if Ω is open and bounded).

$|D|=$ area of D (if $D\subset\mathbb{R}^2$).

$\nabla u=\operatorname{grad} u=(\partial u/\partial x,\partial u/\partial y)$ (if $u=u(x,y)$).

$\Delta u=$ the Laplacian of $u=\partial^2u/\partial x^2+\partial^2u/\partial y^2$.

$$\int_D u=\int_D u(x,y)\,dx\,dy.$$

$\chi_D=$ the characteristic function of $D\subset\mathbb{R}^2=\begin{cases}1 & \text{in } D,\\ 0 & \text{in } \mathbb{R}^2\setminus D.\end{cases}$

$C_c^\infty(\Omega)$: the space of infinitely differentiable functions in Ω with compact support.

$H^{m,p}(\Omega), H^m(\Omega)=H^{m,2}(\Omega), H_0^1(\Omega)$: Sobolev spaces as defined in [18]. $H_0^1(\Omega)$ will always be equipped with the inner product

$$(u,v)=\int_\Omega \nabla u\cdot\nabla v=\int_\Omega(\partial u/\partial x\,\partial v/\partial x+\partial u/\partial y\,\partial v/\partial y)\,dx\,dy$$

(Ω will always be bounded) and norm $\|u\|=\sqrt{(u,u)}$.

$\langle u,v\rangle$: the pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$ ($u\in H_0^1(\Omega)$, $v\in H^{-1}(\Omega)$) when $H^{-1}(\Omega)$ is regarded as the dual space of $H_0^1(\Omega)$ in the usual way. This is consistent with

$$\langle u,v\rangle=\int_\Omega u\cdot v \text{ for } u\in L^p(\Omega),\ v\in L^q(\Omega),\ 1/p+1/q=1,\ 1\le p<\infty.$$

We will often use the fact that the Laplacian Δ is an isomorphism of $H_0^1(\Omega)$ onto $H^{-1}(\Omega)$ with the property that [18, Thm. 23.1]

$$(u,v)=-\langle u,\Delta v\rangle \quad \text{for } u,v\in H_0^1(\Omega). \tag{0.1}$$

$\mathbb{S}_\omega$ is defined in §1 (before Definition 1).

$\mathfrak{R}_{\omega,\Omega}$ is defined in §2 (before Definition 2).

1. A classical formulation of the problem. The purpose of this section is to give an example of how to formulate the moving boundary condition for the problem described in the introduction in a rigorous way. This is done by our concept of a classical solution (Definition 1 below). Actually, we have generalized the problem a little by replacing the Green's function g_D by a domain function p_D depending on a positive measure μ. Our concept of classical solution is to be thought of as just formalizing the rule according to which the moving boundary moves, and we have not tried to make any initial value problem out of it.

Let $\mu\neq 0$ be a finite positive measure with compact support in $\mathbb{R}^2$. For domains $D\subset\mathbb{R}^2$ with $\operatorname{supp}\mu\subset D$ let p_D be the (superharmonic) function in D defined by

$$-\Delta p_D=\mu \quad \text{in } D, \tag{1.1}$$

$$p_D=0 \quad \text{on } \partial D. \tag{1.2}$$

Here (1.2) should be interpreted as follows: for each $\varepsilon>0$ there is a compact set $K\subset D$ such that $|p_D|<\varepsilon$ outside K. For all domains D considered in this section, the problem (1.1)–(1.2) has a unique solution. This solution, moreover, satisfies $p_D\geq 0$ by the minimum principle for superharmonic functions. We are going to consider the problem mentioned in the introduction but with the Green's function g_D there replaced by the more general function p_D. The special case $p_D=g_D$ is obtained by choosing $\mu=2\pi\delta$ in (1.1)–(1.2) (where $\delta=$ the Dirac measure at the origin).

Let ω be a fixed open neighborhood of $\operatorname{supp}\mu$ and set

$\mathbb{S}_\omega=$ the class of all simply connected domains $D\subset\mathbb{R}^2$ with $\omega\subset\subset D$ and such that ∂D is a Jordan curve of class C^2 (i.e. such that there exists a twice continuously differentiable map from the unit circle to ∂D which is bijective and whose derivative never vanishes; such a map will be called a diffeomorphism of class C^2).

For $D\in\mathbb{S}_\omega$ p_D exists and is unique and moreover both p_D and ∇p_D have continuous extensions to ∂D. Let $I\subset\mathbb{R}$ be an open interval.

Definition 1. A map $I\ni t\to D_t\in\mathbb{S}_\omega$ is a *classical solution* of our moving boundary problem if there exists a map ζ: $\mathbb{R}/\mathbb{Z}\times I\to\mathbb{R}^2$ of class C^2 (i.e. twice continuously differentiable) such that

(i) $\zeta(s,t)\in\partial D_t$ for all s,t,

(ii) $\zeta(\cdot,t):\mathbb{R}/\mathbb{Z}\to\partial D_t$ is a diffeomorphism (of class C^2) for each $t\in I$, and

(iii)

$$\frac{\partial\zeta(s,t)}{\partial t}=-\nabla p_{D_t}(\zeta(s,t)) \quad \text{for all } s,t. \tag{1.3}$$

Comment. (i) and (ii) say that for each fixed $t,\zeta(\cdot,t)$ parametrizes ∂D_t. The parameter s (in which ζ has period 1) numbers the points on ∂D_t and (iii) says that each such point moves with the velocity $-\nabla p_{D_t}(\zeta(s,t))$. Here ∇p_{D_t} is the continuous extension of the gradient of p_{D_t} to ∂D_t.

For simplicity we have preferred to let the domains of definition of classical solutions be *open* time intervals. For that reason the concept of a classical solution is not immediately well suited to formalize initial value problems of the kind stated in the introduction. Consider e.g. the following attempt in that direction (where μ and ω are given):

Given $D\in\mathbb{S}_\omega$ find, for some $\varepsilon>0$, a classical solution $(-\varepsilon,+\infty)\ni t\to D_t\in\mathbb{S}_\omega$ such that $D_0=D$. This formulation has the drawback that it requires the existence of the solution for $-\varepsilon<t<0$ and, as it turns out (see below), this can only occur if ∂D is analytic.

A perhaps better attempt would therefore be:

Given $D\in\mathbb{S}_\omega$ find a classical solution $(0,+\infty)\ni t\to D_t\in\mathbb{S}_\omega$ such that $\lim_{t\to 0}D_t=D$ in some specified sense. This formulation does away with the problem of the former formulation, but nevertheless one cannot expect a solution to exist for arbitrary $D\in\mathbb{S}_\omega$. This is because we have not, in our definition of a classical solution, built in any possibility for D_t to change connectivity and it is easy to see that without such a possibility global solutions cannot exist in general.

The above remark shows that the concept of a classical solution has to be fairly complicated in order to be well suited for a formulation of a global initial value problem. We have not thought it to be worth the effort to make such a formulation, since our concept of classical solution is introduced mostly in order to motivate our concepts of a weak solution (and for this purpose we think that Definition 1 is good enough). A global concept of a classical solution (allowing connectivity changes) has, however, been given by Sakai [16].

Definition 1 is, however, well suited for formulating a local problem:

(1.4) Given $D\in\mathbb{S}_\omega$ find, for some $\varepsilon>0$, a classical solution $(-\varepsilon,\varepsilon)$ $\ni t\to D_t\in\mathbb{S}_\omega$ such that $D_0=D$.

The task of proving existence and unicity for solutions of this problem is seemingly hard. In fact, we will prove here (Theorem 10) that a necessary condition for a solution of (1.4) to exist is that ∂D is an analytic curve. Probably this condition is also sufficient. In any case, in the special case that $\mu=2\pi\delta$, Vinogradov and Kufarev [19] have proved local existence of solutions when the problem is formulated as a differential equation for the Riemann mapping function (as in [12]), under the condition that ∂D is analytic (see also [8]). It is, however, not quite easy to prove rigorously that a solution in their sense is also a classical solution in our sense. (The converse is easier.) Vinogradov and Kufarev also prove uniqueness for solutions depending analytically on t (of their problem). Here we prove at least that a solution of (1.4) is unique for $t>0$ (Theorem 10).

Let us next make a remark about the measure μ; namely, as far as classical solutions are concerned, we can always assume that μ is a smooth function. The reason is that nothing but the behaviour of p_D near ∂D comes into Definition 1 and that therefore p_D can be smoothed out in a neighbourhood of $\operatorname{supp}\mu$. To be precise, let h be a smooth (C^∞), positive, radially symmetric (i.e. a function of radius only) function ("mollifier") on $\mathbb{R}^2$ with total mass one ($\int h=1$) and with compact support in the open unit disk $\mathbb{D}$. Define

$$h_\varepsilon(z)=\frac{1}{\varepsilon^2}h\left(\frac{z}{\varepsilon}\right)\qquad(z\in\mathbb{R}^2).\tag{1.5}$$

(Thus $\operatorname{supp} h_\varepsilon \subset \mathbb{D}(0;\varepsilon)$, $\int h_\varepsilon = 1$, $h_\varepsilon \geq 0$.) Then we have

PROPOSITION 1. *Let μ, ω and I be as before Definition 1 and choose $\varepsilon > 0$ such that $2\varepsilon < \operatorname{dist}(\operatorname{supp}\mu, \partial\omega)$. Then $I \ni t \to D_t \in \mathfrak{S}_\omega$ is a classical solution for μ if and only if it is a classical solution for $\mu * h_\varepsilon$ (which is a smooth function). ($*$ denotes convolution.)*

Proof. The proof consists of the observation that the function p_D defined by (1.1)–(1.2) only changes inside ω when μ is replaced by $\mu * h_\varepsilon$ (for $D \in \mathfrak{S}_\omega$). In fact, define

$$q_D = \begin{cases} p_D * h & \text{in } \{z \in D : \operatorname{dist}(z, \partial D) > \varepsilon\}, \\ p_D & \text{in } \{z \in D : \operatorname{dist}(z, \operatorname{supp}\mu) > \varepsilon\}. \end{cases}$$

Then q_D is well defined because p_D is harmonic in a whole ε-neighborhood of any point in the overlap between the two domains above, and therefore $p_D * h_\varepsilon = p_D$ in that overlap by the mean-value property for harmonic functions. Since $\Delta(p_D * h_\varepsilon) = \Delta p_D * h_\varepsilon = -\mu * h_\varepsilon$, it is immediately seen that q_D is the solution of (1.1)–(1.2) with $\mu * h_\varepsilon$ in place of μ. Since $q_D = p_D$ near ∂D, the conclusion of the proposition follows immediately.

2. The weak solution. We now introduce the concept of a weak solution and prove that a classical solution is a weak solution.

The concept of a weak solution is much more flexible than that of a classical solution (e.g. one does not have to bother about boundary regularity or connectivity of the domains), it is much easier to show existence of solutions for, and it is also more apt for numerical treatment (because it is closely related to variational inequalities). These are the main reasons for introducing the concept of a weak solution.

Let $\mu \neq 0$ be a finite positive measure with compact support in $\mathbb{R}^2$, and choose bounded open sets ω and Ω in $\mathbb{R}^2$ such that $\operatorname{supp}\mu \subset \omega \subset\subset \Omega$ and with $\partial\Omega$ smooth, and let $T > 0$. Set

$$\mathfrak{R}_{\omega,\Omega} = \text{the class of all open sets } D \subset R^2 \text{ with } \omega \subset\subset D \subset\subset \Omega.$$

In order for the definition below to make sense, we have to assume that μ belongs to the Sobolev space $H^{-1}(\Omega)$. This is an assumption of purely technical nature and does not mean any restriction of the class of problems considered (in view of Proposition 1 above).

DEFINITION 2. A map of $[0, T] \ni t \to D_t \in \mathfrak{R}_{\omega,\Omega}$ is a *weak solution* of our moving problem if, for each $t \in [0, T]$, the function $u_t \in H_0^1(\Omega)$ defined by

$$\chi_{D_t} - \chi_{D_0} = \Delta u_t + t \cdot \mu \tag{2.1}$$

satisfies

$$u_t \geq 0, \tag{2.2}$$

$$\langle u_t, 1 - \chi_{D_t} \rangle = 0. \tag{2.3}$$

Comments. 1) *Notation.* The subscript t in u_t just indicates that u_t depends on t. We never use subscripts for partial derivatives. (2.1) and (2.2), like all other equalities and inequalities on open sets in this paper, are to be interpreted in the sense of distributions. In (2.3) $1 - \chi_{D_t}$ is regarded as an element of $H^{-1}(\Omega) \cong$ the dual space of $H_0^1(\Omega)$ and $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$. Since, in the present case, $1-\chi_{D_t}\in L^2(\Omega)$ (and $u_t\in H_0^1(\Omega)\subset L^2(\Omega)$), the left member of (2.3) reduces to the Lebesgue integral

$$\langle u_t, 1-\chi_{D_t}\rangle=\int_\Omega u_t\cdot(1-\chi_{D_t})=\int_{\Omega\setminus D_t} u_t,$$

and since $1-\chi_{D_t}\geq 0$, (2.2) and (2.3) together express that $u_t\geq 0$ in Ω and $u_t=0$ a.e. outside D_t.

2) Since $\chi_{D_t}-\chi_{D_0}-t\cdot\mu\in H^{-1}(\Omega)$ and the Laplacian Δ is an isomorphism from $H_0^1(\Omega)$ onto $H^{-1}(\Omega)$, (2.1) really defines u_t uniquely.

3) It is clear from the definition that given $D_0\in\mathfrak{R}_{\omega,R}$, the domains D_t of a weak solution can be unique at most up to two-dimensional Lebesgue measure zero, since (2.1)–(2.3) are not affected if D_t is replaced by another $D_t'\in\mathfrak{R}_{\omega,\Omega}$ such that $\chi_{D_t'}=\chi_{D_t}$ a.e..

4) In order to motivate the concept of a weak solution, we now sketch a proof that a classical solution is a weak solution.

So suppose $t\to D_t$ is a classical solution. Condition (iii) in Definition 1 can be written

$$\frac{\delta n}{\delta t}=-\frac{\partial p_{D_t}}{\partial n}\quad\text{on }\partial D_t,\tag{2.4}$$

where $\delta n/\delta t$ denotes the normal velocity of ∂D_t (in the direction out from D_t), and $\partial/\partial n$ denotes outward normal derivative. (2.4) is equivalent to

$$\int_{\partial D_t}\frac{\delta n}{\delta t}\cdot\varphi\,ds=-\int_{\partial D_t}\frac{\partial p_{D_t}}{\partial n}\cdot\varphi\,ds\quad\text{for all }\varphi\in C_c^\infty(\mathbb{R}^2)\tag{2.5}$$

(ds denotes arc-length measure).

It is not hard to see that

$$\int_{\partial D_t}\frac{\delta n}{\delta t}\varphi\,ds=\frac{d}{dt}\int_{D_t}\varphi\,dx\,dy=\frac{d}{dt}\langle\chi_{D_t},\varphi\rangle.$$

Extend p_{D_t} to all $\mathbb{R}^2$ by setting $p_{D_t}=0$ outside D_t. Then

$$-\Delta p_{D_t}=\mu+\frac{\partial p_{D_t}}{\partial n}ds,$$

where $(\partial p_{D_t}/\partial n)\,ds$ denotes the distribution

$$\varphi\to\int_{\partial D_t}\varphi\frac{\partial p_{D_t}}{\partial n}ds\qquad(\varphi\in C_c^\infty(\mathbb{R}^2)),$$

a nonpositive measure supported by ∂D_t. Hence

$$-\int_{\partial D_t}\frac{\partial p_{D_t}}{\partial n}\cdot\varphi\,ds=\langle\Delta p_{D_t}+\mu,\varphi\rangle,$$

and (2.5) can be written

$$\frac{d}{dt}\langle\chi_{D_t},\varphi\rangle=\langle\Delta p_{D_t}+\mu,\varphi\rangle\quad\text{for all }\varphi\in C_c^\infty(\mathbb{R}^2).\tag{2.6}$$

Now integrate (2.6) with respect to t. With

$$u_t = \int_0^t p_{D_\tau} d\tau \tag{2.7}$$

this gives $\chi_{D_t} - \chi_{D_0} = \Delta u_t + t \cdot \mu$, to hold in the sense of distributions. Thus u_t defined by (2.7) satisfies (2.1) of Definition 2. Since $p_{D_\tau} \geq 0$ in D_τ, $p_{D_\tau} = 0$ outside D_τ, and because $D_{\tau_1} \subset D_{\tau_2}$ for $\tau_1 < \tau_2$ (if $t \to D_t$ is a classical solution) u_t, defined by (2.7), also satisfies

$$\begin{aligned} u_t &\geq 0 \quad \text{in all } \mathbb{R}^2 \quad \text{and} \\ u_t &= 0 \quad \text{outside } D_t, \end{aligned}$$

that is, (2.2) and (2.3) of Definition 2.

This was a sketch of a proof that a classical solution is a weak solution. A formal proof of this will be given later (Theorem 1).

5) A nice feature of the concept of weak solution is that the time variable only occurs as a parameter in it: (2.1)–(2.3) is just a series of uncoupled problems (actually free boundary problems), one for each $t \in [0, T]$. The transformation (2.7) plays a crucial role in this respect. The efficiency of transformations similar to (2.7) on certain kinds of free and moving boundary problems is now well known and has been demonstrated in works by Baiocchi, Duvaut, Elliott and others. (See e.g. [2], [4] and [5].)

Just as for classical solutions, there is no loss of generality in assuming that the measure μ in Definition 1 is a smooth function.

PROPOSITION 2. *Let μ, ω, Ω and T be as in Definition* 2, *choose $\varepsilon > 0$ with* $2\varepsilon <$ dist(supp $\mu, \partial\omega$) *and let h_ε be as before Proposition* 1. *Then $[0, T] \ni t \to D_t \in \mathfrak{R}_{\omega,\Omega}$ is a weak solution for μ if and only if it is a weak solution for $\mu * h_\varepsilon$. In case they are solutions we have*

$$v_t = \begin{cases} u_t * h_\varepsilon & \textit{in } \operatorname{supp} \mu + \mathbb{D}(0; \varepsilon), \\ u_t & \textit{elsewhere in } \Omega, \end{cases} \tag{2.8}$$

where u_t (v_t) is the function occurring in Definition 2 *for μ ($\mu * h_\varepsilon$).* (supp $\mu + \mathbb{D}(0; \varepsilon) = \{z + w \in \mathbb{R}^2 : z \in \operatorname{supp} \mu$ *and* $w \in \mathbb{D}(0; \varepsilon)\}$.)

Proof (*sketch*). Let $t \to D_t$ be a weak solution for μ, let $u_t \in H_0^1(\Omega)$ be defined by (2.1) and define v_t by (2.8). Then u_t is harmonic in $\omega \setminus \operatorname{supp} \mu$ (by (2.1)) and the mean-value property for harmonic functions, together with the properties of h_ε, show that $u_t * h_\varepsilon = u_t$ a distance ε away from $\partial(\omega \setminus \operatorname{supp} \mu)$ in $\omega \setminus \operatorname{supp} \mu$. It follows that v_t is smooth in the join between the two ranges of definition and in particular that $v_t \in H_0^1(\Omega)$. Now it is easy to check that v_t satisfies (2.1)–(2.3) of Definition 2 with $\mu * h_\varepsilon$ in place of μ. This proves the proposition in one direction.

Next, let $t \to D_t$ be a weak solution for $\mu * h_\varepsilon$ and $u_t \in H_0^1(\Omega)$ be defined by (2.1) for μ. In order to prove the proposition in the other direction, we have to prove that u_t also satisfies (2.2) and (2.3).

Define v_t by (2.8). As before we have $v_t \in H_0^1(\Omega)$, and v_t satisfies (2.1) with $\mu * h_\varepsilon$ in place of μ. Since the solution of (2.1) is unique and $t \to D_t$ is a weak solution for $\mu * h_\varepsilon$, v_t also satisfies (2.2) and (2.3). From this it follows immediately that u_t satisfies (2.3). Moreover, $u_t \geq 0$ clearly holds outside supp $\mu + \mathbb{D}(0; \varepsilon)$ ($u_t = v_t$ there), and in fact also in supp $\mu + \mathbb{D}(0; \varepsilon)$ because $u_t \geq u_t * h$ $(= v_t \geq 0)$ there due to the fact that u_t (by (2.1)) is superharmonic in ω ($\supset$ supp $\mu + 2\mathbb{D}(0; \varepsilon)$). Thus u_t also satisfies (2.2), and the proposition is proved.

Now we shall prove that a classical solution is a weak solution. Let $\mu \neq 0$ be a finite positive measure with compact support in $\mathbb{R}^2$ such that $\mu \in H^{-1}(\mathbb{R}^2)$, let $\operatorname{supp} \mu \subset \omega$ and let $a<0<T<b$. Then

THEOREM 1. *Suppose* $(a,b) \ni t \to D_t \in \mathfrak{S}_\omega$ *is a classical solution. Then* $[0,T] \ni t \to D_t \in \mathfrak{R}_{\omega,\Omega}$ *is a weak solution if* Ω *is chosen such that* $D_T \subset\subset \Omega$. *Moreover, the functions* p_{D_t} *and* u_t *occurring in the classical and weak solutions respectively, are related by*

$$u_t = \int_0^t p_{D_\tau} d\tau \tag{2.9}$$

(a $H_0^1(\Omega)$*-valued integral), where* p_{D_τ} *is extended to all* Ω *by setting it equal to zero outside* D_τ.

Proof. Let $(a,b) \ni t \to D_t \in \mathfrak{S}_\omega$ be the classical solution, which we shall prove to be weak. We shall first prove that

$$\frac{d}{dt} \int_{D_t} \varphi \, dx\, dy = -\int_{\partial D_t} \varphi \cdot \frac{\partial p_{D_t}}{\partial n} ds \tag{2.10}$$

for all $\varphi \in C_c^\infty(\mathbb{R}^2)$ and that the right member of (2.10) is a continuous function of t.

Let x,y be the coordinate variables in $\mathbb{R}^2$ and let

$$\xi = \xi(s,t), \qquad \eta = \eta(s,t)$$

denote the components of $\zeta(s,t) \in \mathbb{R}^2$ (see Definition 1). Then (iii) of Definition 1 becomes

$$\frac{\partial \xi(s,t)}{\partial t} = -\frac{\partial p_{D_t}}{\partial x}(\zeta(s,t)), \qquad \frac{\partial \eta(s,t)}{\partial t} = -\frac{\partial p_{D_t}}{\partial y}(\zeta(s,t)).$$

Thus the right member of (2.10) becomes

$$-\int_{\partial D_t} \varphi \frac{\partial p_{D_t}}{\partial n} ds = -\int_{\partial D_t} \varphi \cdot \left(\frac{\partial p_{D_t}}{\partial x} dy - \frac{\partial p_{D_t}}{\partial y} dx \right) = \int_{\partial D_t} \varphi \cdot \left(\frac{\partial \xi}{\partial t} dy - \frac{\partial \eta}{\partial t} dx \right). \tag{2.11}$$

Next we rewrite the left member of (2.10). Choose smooth functions, $a(x,y)$ and $b(x,y)$, on $\mathbb{R}^2$ such that $\partial b/\partial x - \partial a/\partial y = \varphi$ (e.g. $a(x,y)=0$ and $b(x,y)=\int_0^x \varphi(u,y)\,du$). Let $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ (the range of the variable s). Then, using Stokes' formula at the first step, we get

$$\begin{aligned} \frac{d}{dt}\int_{D_t} \varphi\, dx\, dy &= \frac{d}{dt}\int_{\partial D_t} a\,dx + b\,dy \\ &= \frac{d}{dt}\int_{\mathbb{T}} \left[a(\zeta(s,t)) \frac{\partial \xi(s,t)}{\partial s} + b(\zeta(s,t)) \frac{\partial \eta(s,t)}{\partial s} \right] ds \\ &= \int_{\mathbb{T}} \left[a \cdot \frac{\partial^2 \xi}{\partial s \partial t} + b \cdot \frac{\partial^2 n}{\partial s \partial t} \right] ds \\ &\quad + \int_{\mathbb{T}} \left[\left(\frac{\partial a}{\partial x}\frac{\partial \xi}{\partial t} + \frac{\partial a}{\partial y}\frac{\partial \eta}{\partial t} \right) \frac{\partial \xi}{\partial s} + \left(\frac{\partial b}{\partial x}\frac{\partial \xi}{\partial t} + \frac{\partial b}{\partial y}\frac{\partial \eta}{\partial t} \right) \frac{\partial \eta}{\partial s} \right] ds \end{aligned} \tag{2.12}$$

$$=\int_{\mathbb{T}}\left[a\cdot\frac{\partial^2\xi}{\partial s\partial t}+b\cdot\frac{\partial^2\eta}{\partial s\partial t}\right]ds+\int_{\mathbb{T}}\left(\frac{\partial b}{\partial x}-\frac{\partial a}{\partial y}\right)\left(\frac{\partial\xi}{\partial t}\frac{\partial\eta}{\partial s}-\frac{\partial\eta}{\partial t}\frac{\partial\xi}{\partial s}\right)ds$$

$$+\int_{\mathbb{T}}\left[\frac{\partial a}{\partial x}\frac{\partial\xi}{\partial t}\frac{\partial\xi}{\partial s}+\frac{\partial b}{\partial y}\frac{\partial\eta}{\partial t}\frac{\partial\eta}{\partial s}+\frac{\partial a}{\partial y}\frac{\partial\xi}{\partial t}\frac{\partial\eta}{\partial s}+\frac{\partial b}{\partial x}\frac{\partial\eta}{\partial t}\frac{\partial\xi}{\partial s}\right]ds$$

$$=\int_{\mathbb{T}}\left(\frac{\partial b}{\partial x}-\frac{\partial a}{\partial y}\right)\left(\frac{\partial\xi}{\partial t}\frac{\partial\eta}{\partial s}-\frac{\partial\eta}{\partial t}\frac{\partial\xi}{\partial s}\right)ds+\int_{\mathbb{T}}\frac{\partial}{\partial s}\left(a\frac{\partial\xi}{\partial t}+b\frac{\partial\eta}{\partial t}\right)ds$$

$$=\int_{\mathbb{T}}\varphi\cdot\left(\frac{\partial\xi}{\partial t}\frac{\partial\eta}{\partial s}-\frac{\partial\eta}{\partial t}\frac{\partial\xi}{\partial s}\right)ds=\int_{\partial D_t}\varphi\cdot\left(\frac{\partial\xi}{\partial t}dy-\frac{\partial\eta}{\partial t}\right)dx.$$

By (2.11) and (2.12) we have proven (2.10). It is seen from (2.12) also that the right member of (2.10) is a continuous function of t and that hence $\int_{D_t}\varphi\,dx\,dy$ is continuously differentiable with respect to t.

We next prove that $D_t\in\mathfrak{R}_{\omega,\Omega}$ for $t\in[0,T]$ if Ω is chosen such that $D_t\subset\subset\Omega$. Since $\omega\subset\subset D_0$ ($D_0\in\mathbb{S}_\omega$), it suffices to prove that $D_\tau\subset D_t$ for $\tau<t$. Since $p_{D_t}\geq 0$ in D_t we have $-\partial p_{D_t}/\partial\eta\geq 0$ on ∂D_t. Therefore (2.10) shows that $\int_{D_t}\varphi\,dx\,dy$ is a nondecreasing function of t for all $\varphi\in C_c^\infty(\mathbb{R}^2)$ such that $\varphi\geq 0$. This easily implies that $D_\tau\subset D_t$ for $\tau<t$. Thus we have proved that $D_t\in\mathfrak{R}_{\omega,\Omega}$ for $t\in[0,T]$.

Now let $u_t\in H_0^1(\Omega)$ (for $t\in[0,T]$) be defined by (2.1) and extend p_{D_t} to all Ω by setting it equal to zero outside D_t. Then it is easy to see that $p_{D_t}\in H_0^1(\Omega)$. We now want to prove that

$$u_t=\int_0^t p_{D_\tau}d\tau. \tag{2.13}$$

Equation (2.13) means by definition (we are using the "weak" definition of vector-valued integrals as exposed e.g. in [18, p. 73]) that

$$\langle u_t,\rho\rangle=\int_0^t\langle p_{D_\tau},\rho\rangle\,d\tau \tag{2.14}$$

for all $\rho\in H^{-1}(\Omega)$. Since $\Delta\colon H_0^1(\Omega)\to H^{-1}(\Omega)$ is an isomorphism, (2.14) can be written

$$\langle u_t,\Delta\varphi\rangle=\int_0^t\langle p_{D_\tau},\Delta\varphi\rangle\,d\tau$$

for all $\varphi\in H_0^1(\Omega)$, i.e., using the fact that $\langle u,\Delta\varphi\rangle=\langle\varphi,\Delta u\rangle$ for $u,\varphi\in H_0^1(\Omega)$,

$$\langle\varphi,\chi_{D_t}-\chi_{D_0}-t\mu\rangle=\int_0^t\langle p_{D_\tau},\Delta\varphi\rangle\,d\tau. \tag{2.15}$$

We first prove (2.15) for $\varphi\in C_c^\infty(\Omega)$. Green's (second) formula gives (for $\varphi\in C_c^\infty(\Omega)$)

$$-\int_{\partial D_t}\varphi\cdot\frac{\partial p_{D_t}}{\partial n}ds=-\int_{D_t}\varphi\cdot\Delta p_{D_t}+\int_{D_t}\Delta\varphi\cdot p_{D_t}$$

$$=\int_{D_t}\varphi\cdot\mu+\int_{\Omega}\Delta\varphi\cdot p_{D_t}=\langle\varphi,\mu\rangle+\langle p_{D_t},\Delta\varphi\rangle.$$

Combining this with (2.10) gives

$$\langle \varphi, \chi_{D_t} - \chi_{D_0} \rangle = \int_{D_t} \varphi - \int_{D_0} \varphi = \int_0^t \left(- \int_{\partial D_\tau} \varphi \cdot \frac{\partial p_{D_t}}{\partial n} \tau \, ds \right) d\tau$$
$$= \langle \varphi, t \cdot \mu \rangle + \int_0^t \langle p_{D_\tau}, \Delta\varphi \rangle \, d\tau .$$

This proves (2.15) for $\varphi \in C_c^\infty(\Omega)$.

To extend (2.15) to all $\varphi \in H_0^1(\Omega)$ it is enough to prove that the right member of (2.15) depends continuously with respect to the $H_0^1(\Omega)$-norm on φ; for $C_c^\infty(\Omega)$ is dense in $H_0^1(\Omega)$ and the left member of (2.15) obviously depends continuously on φ.

We have

$$\left| \int_0^t \langle p_{D_\tau}, \Delta\varphi \rangle \, d\tau \right| = \left| \int_0^t (p_{D_\tau}, \varphi) \, d\tau \right| \leq \int_0^t \| p_{D_\tau} \| \cdot \| \varphi \| \, d\tau = \| \varphi \| \cdot \int_0^t \| p_{D_\tau} \| \, d\tau ,$$

where $(u,v) = \int_\Omega \nabla u \cdot \nabla v$ is the inner product in $H_0^1(\Omega)$ and $\|u\| = \sqrt{(u,u)}$, and it is not hard to prove that $\int_0^t \| p_{D_\tau} \| d\tau < \infty$. (Details are found in [7, p. 40].) Thus the right-hand side of (2.15) depends continuously on φ and so (2.15) is proven.

Now it only remains to prove that u_t satisfies (2.2) and (2.3). (2.2) follows immediately from (2.13) and $p_{D_\tau} \geq 0$, and (2.3) follows from

$$\langle u_t, 1 - \chi_{D_t} \rangle = \int_0^t \langle p_{D_\tau}, 1 - \chi_{D_t} \rangle \, d\tau$$

by choosing $\rho = 1 - \chi_{D_t}$ in (2.14) and

$$\langle p_{D_\tau}, 1 - \chi_{D_t} \rangle = \int_\Omega p_{D_\tau} (1 - \chi_{D_t}) = 0$$

for $\tau \in [0, t]$, a consequence of $D_\tau \subset D_t$ (already proved) and $p_{D_\tau} = 0$ outside D_τ. This completes the proof of Theorem 1.

3. Complementarity problems and variational equalities. By weakening still more the concept of a solution, we arrive at a series of linear complementarity problems. These are equivalent to a series of variational inequalities which are shown to have unique solutions.

Let μ, ω, Ω and T be as before Definition 2 ($\mu \in H^{-1}(\mathbb{R}^2)$, $\mu \geq 0$, $\mu \neq 0$, ω, Ω open, $\partial\Omega$ nice, $\operatorname{supp} \mu \subset \omega \subset \subset \Omega$ and $T > 0$). Let also $D_0 \in \mathfrak{R}_{\omega,\Omega}$ be given and define $\rho_t = \rho_{t,D_0} \in H^{-1}(\Omega)$ by

$$\rho_t = 1 - \chi_{D_0} - t \cdot \mu . \tag{3.1}$$

Then (2.1)–(2.3) in Definition 2 can be written

$$\Delta u_t - \rho_t = \chi_{D_t} - 1, \quad u_t \geq 0, \quad \langle u_t, 1 - \chi_{D_t} \rangle = 0$$

or

$$\Delta u_t - \rho_t = \chi_{D_t} - 1, \quad u_t \geq 0, \quad \langle u_t, \Delta u_t - \rho_t \rangle = 0.$$

Since $\chi_{D_t}-1\leq 0$, this immediately shows

THEOREM 2. *Suppose* $[0,T]\ni t\to D_t\in\mathfrak{R}_{\omega,\Omega}$ *is a weak solution. Then the functions* $u_t\in H_0^1(\Omega)$ *defined by* (2.1) *also solve the linear complementarity problems*

$$\Delta u_t\leq\rho_t, \tag{3.2}$$

$$u_t\geq 0, \tag{3.3}$$

$$\langle u_t,\Delta u_t-\rho_t\rangle=0 \tag{3.4}$$

$(t\in[0,T])$, *where* ρ_t *are defined by* (3.1).

Remark. Clearly (3.4) (in the presence of (3.2) and (3.3)) expresses that at (almost) every point in Ω equality holds in at least one of the inequalities (3.2) and (3.3).

The complementarity problem (3.2)–(3.4) is equivalent to a variational inequality:

THEOREM 3. *Let* $\rho_t\in H^{-1}(\Omega)$ (*e.g. given by* (3.1)). *Then* $u_t\in H_0^1(\Omega)$ *satisfies* (3.2)–(3.4) *if and only if it satisfies*

$$\Delta u_t\leq\rho_t \text{ and } (u-u_t,u_t)\geq 0 \text{ for all } u\in H_0^1(\Omega) \text{ with } \Delta u\leq\rho_t. \tag{3.5}$$

Remark. Theorems 2 and 3 together are similar to [17, Prop. 2]. However, [17] deals with the equivalent variational inequality (3.11) below in place of (3.5).

Proof. If u_t satisfies (3.2)–(3.4), then $\langle u_t,\Delta u-\rho_t\rangle\leq 0=\langle u_t,\Delta u_t-\rho_t\rangle$ for all $u\in H_0^1(\Omega)$ with $\Delta u\leq\rho_t$. By subtracting $\langle u_t,\rho_t\rangle$ and using (0.1) the theorem immediately follows in one direction.

Conversely, suppose (3.5) holds. Then (by (0.1))

$$\langle u_t,\Delta u-\rho_t\rangle\leq\langle u_t,\Delta u_t-\rho_t\rangle \quad \text{for all } u\in H_0^1(\Omega) \text{ with } \Delta u\leq\rho_t. \tag{3.6}$$

Since $\Delta\colon H_0^1(\Omega)\to H^{-1}(\Omega)$ is an isomorphism, we can choose $u\in H_0^1(\Omega)$ in (3.6) such that $\Delta u=\rho_t$ or such that $\Delta u=2\Delta u_t-\rho_t$ ($\Delta u\leq\rho_t$ is fulfilled in both cases). The first choice shows that the right member of (3.6) is ≥ 0, while the second shows that it is ≤ 0. Thus

$$\langle u_t,\Delta u_t-\rho_t\rangle=0. \tag{3.7}$$

By (3.7), and writing $\varphi=\Delta u-\rho_t$, (3.6) becomes $\langle u_t,\varphi\rangle\leq 0$ for all $\varphi\in H^{-1}(\Omega)$ with $\varphi\leq 0$. This shows that $u_t\geq 0$. Thus (3.2)–(3.4) hold for u_t and so the theorem is proven in the other direction too.

Remark. The variational inequality (3.5) differs somewhat from those variational inequalities most often met with in the literature in that the condition $\Delta u\leq\rho_t$ is of an unusual kind, but it is equivalent to a variational inequality of "obstacle-type." To be precise, define $\psi_t\in H_0^1(\Omega)$ by

$$-\Delta\psi_t=\rho_t. \tag{3.8}$$

Then, in terms of the function

$$v_t=u_t+\psi_t \tag{3.9}$$

the complementarity problem (3.2)–(3.4) can be written

$$\Delta v_t\leq 0,\quad v_t\geq\psi_t,\quad \langle v_t-\psi_t,\Delta v_t\rangle=0. \tag{3.10}$$

By an argument very similar to the proof of Theorem 3 (see [7, pp. 43–45] for details), (3.10) can be shown to be equivalent to the variational inequality

(3.11) find $v_t\in H_0^1(\Omega)$ such that $v_t\geq\psi_t$ and $(v-v_t,v_t)\geq 0$ for all $v\in H_0^1(\Omega)$ with $v\geq\psi_t$.

This variational inequality is of obstacle-type (ψ_t describing the obstacle). See [9, Chap. II, §6].

THEOREM 4. *Let $\rho_t \in H^{-1}(\Omega)$. Then the variational inequality* (3.5) *has a unique solution $u_t \in H_0^1(\Omega)$. If, moreover, $\rho_t \in L^p(\Omega)$ for some $1 < p < \infty$, then $u_t \in H^{2,p}(\Omega)$, in particular u_t is continuous* (*if $p > 2$ even continuously differentiable*).

Proof. The existence and unicity of a solution is immediate from the general theory of variational inequalities or in fact just from ordinary Hilbert space theory since (3.5) just expresses that u_t is the unique element of minimum norm in the closed and convex set $K = \{u \in H_0^1(\Omega)\colon \Delta u \leq \rho_t\}$. The regularity of the solution also follows from the general theory of variational inequalities, e.g. by first rewriting the problem into the form (3.11) as indicated in the Remark above, and then invoking [3, Théorème I.1]. There is, however, also a direct and rather nice proof of the regularity. This goes as follows.

Let $\rho_t \in L^p(\Omega)$ with $1 < p < \infty$. To prove that $u_t \in H^{2,p}(\Omega)$, we shall consider a new variational inequality, namely

find $w_t \in H_0^1(\Omega)$ such that

$$\min(0, \rho_t) \leq \Delta w_t \leq \rho_t \quad \text{and} \tag{3.12}$$

$$(w - w_t, w_t) \geq 0 \text{ for all } w \in H_0^1(\Omega) \text{ with } \min(0, \rho_t) \leq \Delta w \leq \rho_t. \tag{3.13}$$

This variational inequality has a unique solution $w_t \in H_0^1(\Omega)$ for the same reason as for (3.5). Moreover, this solution a priori belongs to $H^{2,p}(\Omega)$ since (3.12) shows that $\Delta w_t \in L^p(\Omega)$ (and $\Delta w_t \in L^p(\Omega)$ implies $w_t \in H^{2,p}(\Omega)$). Thus, to prove that $u_t \in H^{2,p}(\Omega)$, it is enough to prove that $u_t = w_t$. For that purpose we only have to show that w_t satisfies (3.2)–(3.4) (by Theorem 3).

To prove (3.2)–(3.4) for w_t, first rewrite (3.13) as

$$\langle w_t, \Delta w - \rho_t \rangle \leq \langle w_t, \Delta w_t - \rho_t \rangle$$

for all $w \in H_0^1(\Omega)$ with $\min(0, \rho_t) \leq \Delta w \leq \rho_t$. Setting $\varphi = \Delta w - \rho_t$ and using that the bracket $\langle \cdot, \cdot \rangle$ in the present case reduces to a Lebesgue integral, we get

$$\int_\Omega w_t \cdot \varphi \leq \int_\Omega w_t \cdot (\Delta w_t - \rho_t) \tag{3.14}$$

for all $\varphi \in H^{-1}(\Omega)$ with $\min(0, -\rho_t) \leq \varphi \leq 0$. First choose $\varphi = 0$ in (3.14). This gives

$$\int_\Omega w_t \cdot (\Delta w_t - \rho_t) \geq 0. \tag{3.15}$$

Then choose

$$\varphi = \begin{cases} \min(0, -\rho_t) & \text{in } N, \\ \Delta w_t - \rho_t & \text{in } \Omega \setminus N, \end{cases}$$

where $N = \{z \in \Omega : w_t(z) < 0\}$. (Since $w_t \in H^{2,p}(\Omega)$, w_t is a continuous function and so $w_t(z) < 0$ has a natural meaning and N is a well-defined open set.) We get

$$\int_N w_t \cdot \min(0, -\rho_t) \leq \int_N w_t \cdot (\Delta w_t - \rho_t),$$

or

$$\int_N w_t \cdot (\Delta w_t - \rho_t - \min(0, -\rho_t)) \geq 0. \tag{3.16}$$

Since $\Delta w_t - \rho_t - \min(0, -\rho_t) = \Delta w_t - \min(0, \rho_t) \geq 0$ and $w_t < 0$ in N (3.16) shows that

$$\Delta w_t - \min(0, \rho_t) = 0 \quad \text{in } N.$$

In particular,

$$\Delta w_t \leq 0 \quad \text{in } N. \tag{3.17}$$

But now $w_t = 0$ on $\partial N \cup \partial \Omega$ (by the definition of N, and since w_t is continuous and belongs to $H_0^1(\Omega)$). Therefore (3.17) implies $w_t \geq 0$ in N (minimum principle for superharmonic functions). Comparing with the definition of N, we conclude that N is the empty set. Hence

$$w_t \geq 0 \quad (\text{in } \Omega).$$

Thus (3.3) (for w_t) is proven. (3.2) is part of (3.12), and (3.4) follows by combining (3.15) with (3.2) and (3.3). Hence w_t satisfies the complementarity conditions (3.2)–(3.4) which characterize u_t, hence $w_t = u_t$ as we wanted to prove.

The statements about continuity and continuous differentiability of u_t follow from Sobolev's inequalities. See e.g. [18, Thm. 24.2]. This completes the proof.

4. From variational inequalities to a weak solution. Up to now we have performed a series of weakenings of the concept of solution,

$$\begin{array}{ccccccc} \text{classical} & & \text{weak} & & \text{solution of} & & \text{solution of} \\ & \Rightarrow & & \Rightarrow & \text{complementarity} & \Leftrightarrow & \text{variational} \\ \text{solution} & & \text{solution} & & \text{problems} & & \text{inequalities} \end{array}$$

and we have proved existence and uniqueness of solutions at the right end point of this series. On the way from weak solution to complementarity problems we have also lost the domain D_t from the problem.

Now we want to perform the step

$$\begin{array}{ccc} \text{solution of} & & \text{weak} \\ \text{complementarity} & \Rightarrow & \\ \text{problems} & & \text{solution} \end{array}$$

thereby also proving existence of weak solutions with a given initial domain (uniqueness is already proven, by uniqueness of solutions of the variational inequalities). This step involves among other things recovering the regions D_t from the functions u_t (constituting the solution of the complementarity problems). We need two lemmas.

LEMMA 1. *Let $\rho_t \in H^{-1}(\Omega)$ and let $u_t \in H_0^1(\Omega)$ be the solution of the complementarity problem* (3.2)–(3.4). *Then $u_t \leq u$ for all $u \in H_0^1(\Omega)$ which satisfy $\Delta u \leq \rho_t$ and $u \geq 0$.*

Comment. Lemma 1 says that among the functions that satisfy the inequalities (3.2) and (3.3), there is a smallest function, namely that function for which these inequalities hold complementarily.

Lemma 1 is closely related to [9, Thm. 6.4, Chap. II]. In fact, that theorem says that if $v_t \in H_0^1(\Omega)$ is the solution of the variational inequality (3.11) or, equivalently, to the problem (3.10), then $v_t \leq v$ for all $v \in H_0^1(\Omega)$ which satisfy $v \geq \psi_t$ and $\Delta v \leq 0$ ($\psi_t \in H_0^1(\Omega)$). In view of the remark after Theorem 3, this gives a proof of our lemma by setting $u_t = v_t - \psi_t$, $u = v - \psi_t$ and by defining $\psi_t \in H_0^1(\Omega)$ by (3.8).

Let us, however, give an independent proof of Lemma 1. For simplicity we restrict ourselves to the case that $\rho_t \in L^p(\Omega)$ for some $p > 1$ (which suffices for our purposes).

Proof of Lemma 1 in case $\rho_t \in L^p(\Omega)$, $p > 1$. Suppose $\rho_t \in L^p(\Omega)$ where $1 < p < \infty$. Then $u_t \in H^{2,p}(\Omega)$, in particular u_t is continuous, by Theorem 4. Thus

$$I = \{z \in \Omega : u_t(z) = 0\} \quad \text{and} \quad D = \Omega \setminus I = \{z \in \Omega : u_t(z) > 0\}$$

are well-defined (relatively) closed and open sets in Ω respectively.

Put $w = u - u_t$. Thus we want to prove that $w \geq 0$. In view of (3.2) and (3.3) it follows from (3.4), which can be interpreted as a Lebesgue integral in this case, that $\Delta u_t = \rho_t$ in D. Thus

$$\Delta w = \Delta u - \rho_t \leq 0 \quad \text{in } D. \tag{4.1}$$

Take an $\varepsilon > 0$ and define $N = \{z \in \Omega : u_t(z) < \varepsilon\}$. Then N is an open neighborhood of $I \cup \partial\Omega$ in Ω. Now, from $u \geq 0$ and (4.1) we have

$$w + \varepsilon \geq 0 \quad \text{in } N \quad \text{and} \tag{4.2}$$

$$\Delta(w + \varepsilon) \leq 0 \quad \text{in } D. \tag{4.3}$$

Since $\partial D \subset I \cup \partial\Omega$, it essentially follows from (4.2), (4.3) and the minimum principle for superharmonic functions that $w + \varepsilon \geq 0$ in D and hence

$$w + \varepsilon \geq 0 \quad \text{in } \Omega = D \cup N.$$

The only problem is that w is not (necessarily) a nice function but just an element of $H_0^1(\Omega)$, so that some care is needed in applying the minimum principle.

To this end, choose $r > 0$ with $2r < \text{dist}(\partial N, \partial D)$ and let h_r be as before Proposition 1. Then an application of the ordinary minimum principle to $(w+\varepsilon) * h_r$ in $\{z \in D : \text{dist}(z, \partial D) > r\}$ shows that

$$(w + \varepsilon) * h_r \geq 0 \quad \text{in } \{z \in \Omega : \text{dist}(z, \partial\Omega) > r\}. \tag{4.4}$$

Letting first $r \to 0$ and then $\varepsilon \to 0$, (4.4) yields $w \geq 0$ in Ω, as we wanted to prove.

COROLLARY 1. *Let $\rho, \rho' \in H^{-1}(\Omega)$ and let u and u' be the solutions of* (3.2)–(3.4) *for $\rho_t = \rho$ and ρ' respectively. Suppose that $\rho' \leq \rho$. Then $u \leq u'$.*

Proof. This follows from the lemma with $\rho_t = \rho$, $u_t = u$ since $\Delta u' \leq \rho' \leq \rho$ and $u' \geq 0$.

Remark. There is also an inequality in the other direction. Namely, let $\psi, \psi' \in H_0^1(\Omega)$ be defined by $-\Delta\psi = \rho$ and $-\Delta\psi' = \rho'$ respectively. Then, if $\rho' \leq \rho$, we have $u' \leq u + (\psi - \psi')$. This follows by applying the lemma with $u_t = u'$ and "u in the lemma" $= u + (\psi - \psi')$.

COROLLARY 2. *The solution u_t of* (3.2)–(3.4) *is monotonically increasing* (*= non-decreasing*) *as a function of each of μ, D_0 and t* (*more generally, as a function of $\chi_{D_0} + t \cdot \mu$*) *when ρ_t is given by* (3.1), *i.e. if $D_0 \subset D_0'$, $\mu \leq \mu'$ and $t \leq \tau$ then $u_t \leq u_\tau'$* (*self-explanatory notations*).

Proof. This is just a special case of Corollary 1.

LEMMA 2. *Suppose $\mu \in L^p(\Omega)$ for some $1 < p < \infty$ and let $u_t \in H_0^1(\Omega)$ be the solution of* (3.2)–(3.4) *with ρ_t given by* (3.1). *Define*

$$D_t = D_0 \cup \{z \in \Omega : u_t(z) > 0\}. \tag{4.5}$$

Then

$$\Delta u_t = \chi_{D_t} - \chi_{D_0} - t \cdot \mu = \chi_{D_t} - 1 + \rho_t. \tag{4.6}$$

Proof. By Theorems 3 and 4 $u_t \in H^{2,p}(\Omega)$, in particular u_t is continuous. Thus D_t is a well-defined open set. Observe also that the definition (4.5) is consistent for $t=0$ since $u_t=0$ is the solution of (3.2)–(3.4) for $t=0$ (in view of $\rho_0 \geq 0$).

Define $I_t = \{z \in \Omega : u_t(z) = 0\}$. Because $u_t \in H^{2,p}(\Omega)$, all partial derivatives of u_t of order ≤ 2 vanish almost everywhere on I_t. (This follows e.g. from [9, Lemma A.4, p. 53].) In particular

$$\Delta u_t = 0 \quad \text{a.e. on } I_t. \tag{4.7}$$

Hence (3.2) shows that $\rho_t \geq 0$ a.e. on I_t. By (3.1), using that $\operatorname{supp} \mu \subset D_0$ and $\mu \geq 0$, this gives $\rho_t = 1 - \chi_{D_0}$ a.e. on I_t, or, using (4.7),

$$\Delta u_t - \rho_t = \chi_{D_0} - 1 \quad \text{a.e. on } I_t. \tag{4.8}$$

Now, since $\Delta u_t - \rho_t \in L^p(\Omega)$ and u_t is continuous, the left member of (3.4) can be interpreted as a Lebesgue integral, and it follows from (3.2)–(3.4) and the fact that $u_t > 0$ in $\Omega \setminus I_t$ that

$$\Delta u_t - \rho_t = 0 \quad \text{in } \Omega \setminus I_t. \tag{4.9}$$

Equation (4.8) together with (4.9) gives

$$\Delta u_t - \rho_t = (\chi_{D_0} - 1) \cdot \chi_{I_t} = \chi_{(\Omega \setminus D_0) \cap I_t}$$

(a.e. or in the sense of distributions). Since $(\Omega \setminus D_0) \cap I_t = \Omega \setminus D_t$, this shows that $\Delta u_t = \chi_{D_t} - 1 + \rho_t$, which is the desired result.

THEOREM 5. *Let μ, ω, Ω and T be as before Definition 2 with $\mu \in L^\infty$ (for simplicity), let $D_0 \in \mathfrak{R}_{\omega,\Omega}$ and let ρ_t be defined by* (3.1). *Suppose $u_t \in H_0^1(\Omega)$ and solve* (3.2)–(3.4) *for $t \in [0, T]$. Then, if Ω is large enough and D_t is defined by $D_t = D_0 \cup \{z \in \Omega : u_t(z) > 0\}$, the map*

$$[0, T] \ni t \to D_t \in \mathfrak{R}_{\omega,\Omega} \tag{4.10}$$

is well defined and is a weak solution. Further, the function "u_t" appearing in the definition of a weak solution is identical with the u_t above.

Remark. This theorem, showing that solutions of the complementarity problems give rise to weak solutions, is similar to [17, Prop. 3].

Proof. We first show that the map (4.10) is well defined. , i.e. that $\Omega \subset\subset D_t \subset\subset \Omega$ for all $t \in [0, T]$.

$\omega \subset\subset D_t$ is evident since $\omega \subset\subset D_0$ and $D_0 \subset D_t$. Next, choose $M > 0$ and $0 < r < R$ such that $\mu \leq M \cdot \chi_{\mathbb{D}(r)}$ and $D_0 \subset \mathbb{D}(R)$, and define $R_t > 0$ for $t \in [0, T]$ by

$$|\mathbb{D}(R_t)| = tM|\mathbb{D}(r)| + |\mathbb{D}(R)|.$$

Then I claim that $D_t \subset \mathbb{D}(R_t)$ for all t, and hence that it suffices to choose Ω such that $\mathbb{D}(R_T) \subset\subset \Omega$.

To see this, put $\rho_t' = 1 - \chi_{\mathbb{D}(R)} - tM\chi_{\mathbb{D}(r)}$ and define $u_t' \in H_0^1(\Omega)$ by $\Delta u_t' = \chi_{\mathbb{D}(R_t)} - \chi_{\mathbb{D}(R)} - tM\chi_{\mathbb{D}(r)}$. Then it can be checked that $u_t' > 0$ in $\mathbb{D}(R_t)$, $u_t' = 0$ outside $\mathbb{D}(R_t)$. This shows that u_t' is the solution of (3.2)–(3.4) corresponding to ρ_t'.

But now $\rho_t' \leq \rho_t$. Therefore $u_t \leq u_t'$ by Corollary 1 of Lemma 1. Hence $u_t = 0$ outside $\mathbb{D}(R_t)$, showing that $D_t \subset \mathbb{D}(R_t)$.

It remains to prove that u_t and D_t satisfy (2.1)–(2.3) of Definition 2. (2.1) follows from Lemma 2, (2.2) is (3.3) and (2.3) is (3.4) combined with (4.6). This completes the proof of Theorem 5.

THEOREM 6. (corollary of Theorem 5). *Let μ, ω, Ω and T be as before Definition* 2 *and let $D \in \mathfrak{R}_{\omega,\Omega}$ be given. Then, if merely Ω is large enough, there exists a weak solution*

$$(4.11) \qquad [0,T] \ni t \rightarrow D_t \in \mathfrak{R}_{\omega,\Omega}$$

with $D_0 = D$. This solution is unique up to sets of two-dimensional Lebesgue measure zero. Moreover, let $u_t \in H_0^1(\Omega)$ be the function appearing in the definition of a weak solution. Then u_t is unique (as an element of $H_0^1(\Omega)$) and D_t above can be chosen to be

$$(4.12) \qquad D_t = D_0 \cup \{z \in \Omega : u_t(z) > 0\}.$$

(Equation (2.1) *shows that u_t is continuous outside* $\operatorname{supp} \mu$, *in particular outside D_0, so that the right-hand side of* (4.12) *is a well-defined open set.) Further, the weak solution* (4.11) *is monotonically increasing (=nondecreasing) as a function of each of μ, D_0 and t (more generally, as a function of $\chi_{D_0} + t \cdot \mu$) i.e. if $\mu \leq \mu'$, $D_0 \subset D_0'$ and $t \leq \tau$ then $D_t \subset D_\tau'$ up to null sets.*

Proof. Suppose first that $\mu \in L^\infty(\Omega)$. Since the problem (3.2)–(3.4) has a unique solution (Theorems 3 and 4), it follows immediately from Theorem 5 that there exists a weak solution (4.11) such that (4.12) holds. Since $u_0 = 0$ for the solution u_t of (3.2)–(3.4), we also have $D_0 = D$. As to the unicity, suppose we have two solutions, $t \rightarrow D_t$ and $t \rightarrow D_t'$, with $D_0 = D_0' = D$. Then we get, by Theorem 2, two solutions, u_t and u_t', of (3.2)–(3.4) for the same ρ_t. Thus $u_t = u_t'$ and (2.1) shows that $\chi_{D_t} = \chi_{D_t'}$. This is what the unicity statement of the theorem amounts to. The last sentence of the theorem follows immediately from Corollary 2 of Lemma 1. Thus the theorem is proved in case $\mu \in L^\infty(\Omega)$.

If $\mu \notin L^\infty(\Omega)$, we merely apply Proposition 2. Then we are back in the previous case and the theorem follows as before, noting only that the function u_t and v_t in Proposition 2 differ only inside ω, in particular inside D_0, so that (4.12) is not affected by the smoothing process. Note also that the application of Proposition 2 does not destroy the validity of the last sentence in the theorem since the smoothing process is order-preserving (we used positive mollifiers in Proposition 2). This proves the theorem.

5. The moment inequality. We have hitherto shown the equivalence between three concepts of solution for our moving boundary problem, namely the concept of a weak solution, the solution of the linear complementarity problems and the solution of the variational inequalities. There is another equivalent concept of solution which we now want to discuss.

With $\mu \neq 0$ a finite positive measure with compact support, with ω, Ω and T as before Definition 2, and with $D_0 \in \mathfrak{R}_{\omega,\Omega}$ given, let us say that a map $[0,T] \ni t \rightarrow D_t \in \mathfrak{R}_{\omega,\Omega}$ satisfies *the moment inequality* if for each $t \in [0,T]$

$$(5.1) \qquad \int_{D_t} \varphi - \int_{D_0} \varphi \geq t \cdot \int \varphi \, d\mu$$

for every function $\varphi \in H^2(\mathbb{R}^2)$ which is subharmonic in D_t.

The reason for calling this property the moment inequality is that with $\mu = \delta$ (the Dirac measure at the origin) and by choosing $\varphi = \pm \operatorname{Re} z^n$ and $\pm \operatorname{Im} z^n$ $(n \geq 0)$ in a neighborhood of $\overline{D}_0 \cup \overline{D}_t$, (5.1) yields

$$|D_t| = |D_0| + t \; (n=0) \quad \text{and} \quad \int_{D_t} z^n = \int_{D_0} z^n \; (n \geq 1).$$

The quantities $\int_D z^n$ are called the complex (or analytic) moments of the region D. Thus, if (5.1) holds for $t \to D_t$, all complex moments of order ≥ 1 of D_t are preserved under the map $t \to D_t$, while the zeroth order moment ($=$the area of the domain) increases linearly.

The fact that solutions of the Hele Shaw problem have this moment preserving property was discovered by Richardson ([12]). The idea to consider relations such as (5.1) for subharmonic functions φ is due to Sakai ([16] and [17]).

We shall now prove that satisfying the moment inequality is equivalent to being a weak solution. More complete results in the same direction are given in [17] (our Theorem 7 corresponds to [17, Props. 1 and 4]).

THEOREM 7. *Let μ, ω, Ω and T be as above with $\mu \in L^\infty(\Omega)$ (for simplicity). Then a map $[0,T] \ni t \to D_t \in \mathfrak{R}_{\omega,\Omega}$ is a weak solution if and only if it satisfies the moment inequality.*

Proof. Suppose $[0,T] \ni t \to D_t \in \mathfrak{R}_{\omega,\Omega}$ is a weak solution. Since $D_t \subset\subset \Omega$, it is enough to check (5.1) for all $\varphi \in H^2(\mathbb{R}^2)$ which are subharmonic in D_t and vanish on $\partial\Omega$. Thus we assume $\varphi \in H^2(\mathbb{R}^2) \cap H_0^1(\Omega)$.

Let $u_t \in H_0^1(\Omega)$ be the function defined by (2.1). Then $u_t \geq 0$ and $u_t = 0$ a.e. on $\Omega \setminus D_t$ (by (2.2) and (2.3)). Moreover, u_t is continuous and bounded.

Now let $\varphi \in H_0^1(\Omega) \cap H^2(\mathbb{R}^2)$ be subharmonic in D_t. Then $\Delta\varphi \geq 0$ in D_t in the sense of distributions. Moreover, since $\Delta\varphi \in L^2(\Omega)$, the above properties of u_t show that $u_t \cdot \Delta\varphi \in L^2(\Omega)$, $u_t \cdot \Delta\varphi = 0$ a.e. on $\Omega \setminus D_t$ and hence $\int_{\Omega \setminus D_t} u_t \cdot \Delta\varphi = 0$. Using these facts and (2.1), we get

$$\int_{D_t} \varphi - \int_{D_0} \varphi = \langle \chi_{D_t} - \chi_{D_0}, \varphi \rangle = \langle \Delta u_t + t\mu, \varphi \rangle = \langle u_t, \Delta\varphi \rangle + t \langle \mu, \varphi \rangle$$

$$= \int_\Omega u_t \Delta\varphi + t \int \varphi \, d\mu = \int_{D_t} u_t \Delta\varphi + t \int \varphi \, d\mu \geq t \int \varphi \, d\mu.$$

Thus the moment inequality holds.

Conversely, suppose that the moment inequality is satisfied for $[0,T] \ni t \to D_t \in \mathfrak{R}_{\omega,\Omega}$. Again define $u_t \in H_0^1(\Omega)$ by (2.1). In terms of u_t(5.1) takes the form

$$(5.2) \qquad \langle \varphi, \Delta u_t \rangle \geq 0$$

for all $\varphi \in H^2(\mathbb{R}^2)$ subharmonic in D_t. Since the restriction mapping $H^2(\mathbb{R}^2) \to H^2(\Omega)$ is onto [18, Thm. 26.7] the test class, $H^2(\mathbb{R}^2)$ for φ in (5.2) can be replaced by $H^2(\Omega)$. In particular (5.2) holds for all $\varphi \in H^2(\Omega) \cap H_0^1(\Omega)$ which are subharmonic in D_t. For $\varphi \in H^2(\Omega) \cap H_0^1(\Omega)$ we have $\langle \varphi, \Delta u_t \rangle = \langle u_t, \Delta\varphi \rangle$. Therefore, and since Δ maps $H^2(\Omega) \cap H_0^1(\Omega)$ onto $L^2(\Omega)$, (5.2) can be written

$$(5.3) \qquad \langle u_t, \rho \rangle \geq 0$$

for all $\rho \in L^2(\Omega)$ with $\rho \geq 0$ in D_t. (5.3) shows that

$$(5.4) \qquad u_t \geq 0 \qquad (\text{in } \Omega)$$

(because all nonnegative $\rho \in L^2(\Omega)$ are allowed in (5.3)). The choice $\rho = \chi_{D_t} - 1$ is also allowed in (5.3). This gives $\langle u_t, \chi_{D_t} - 1 \rangle \geq 0$ and therefore, since $u_t \geq 0$, $\chi_{D_t} - 1 \leq 0$,

$$(5.5) \qquad \langle u_t, \chi_{D_t} - 1 \rangle = 0.$$

The fact that (5.4) and (5.5) hold for u_t defined by (2.1) shows that our map $t \to D_t$ is a weak solution. Thus the proof of Theorem 7 is completed.

6. Summarizing results and further properties of weak solutions. In this final section we shall first reformulate the main result, Theorem 6, so that it becomes more self-contained and simple, and then we shall prove some modest results on properties of weak solutions.

Theorem 6 and the definition of a weak solution suffer from being a bit complicated because of our desire to work in the Sobolev spaces $H_0^1(\Omega)$ and $H^{-1}(\Omega)$. The following theorem is just Theorem 6 liberated from these complications, and it defines implicitly a more simple concept of a weak solution.

THEOREM 8. *Given a finite positive measure μ with compact support in $\mathbb{R}^2$ and a bounded open set D_0 in $\mathbb{R}^2$ with* $\operatorname{supp}\mu\subset D_0$ *there exist, for each $t>0$, a unique open bounded set D_t containing* $\operatorname{supp}\mu$ *and a unique distribution u_t in $\mathbb{R}^2$ such that*

$$\chi_{D_t}-\chi_{D_0}=\Delta u_t+t\cdot\mu, \tag{6.1}$$

$$u_t\geq 0 \quad \textit{and} \tag{6.2}$$

$$D_t=D_0\cup\{z\in\mathbb{R}^2: u_t(z)>0\}, \tag{6.3}$$

where (6.1) *shows that u_t has a representation in form of a function continuous outside* $\operatorname{supp}\mu$ (*in particular outside D_0*) *and* (6.3) *refers to any such representative.*

Further, D_t is monotonically increasing as a function of each of μ, D_0 and t (*more generally, as a function of $\chi_{D_0}+t\cdot\mu$*), *i.e. if $\mu\leq\mu'$, $D_0\subset D_0'$ and $t\leq\tau$ then $D_t\subset D_\tau'$. Finally*

$$\int_{D_t}\varphi-\int_{D_0}\varphi\geq t\cdot\int\varphi\, d\mu \tag{6.4}$$

holds for every function $\varphi\in H^2(\mathbb{R}^2)$ which is subharmonic in D_t.

Proof (*sketch*). If μ is not sufficiently smooth ($\mu\notin H^{-1}(\mathbb{R}^2)$), we first smooth it out by convolving it with some radially symmetric mollifier (as in Proposition 2) so that $\operatorname{supp}\mu$ is still contained in D_0. Then by choosing appropriate ω, Ω and T and by applying Theorem 6, we obtain functions u_t and domains D_t, related by (4.12) for arbitrary large t. It is easily seen that if we extend the u_t by putting them equal to zero outside Ω, both the u_t and the D_t become independent of the choices of ω, Ω and T, and they satisfy (6.1)–(6.3) (for the smoothed out μ).

Now the D_t will actually provide a solution also for the original μ and the u_t will satisfy (6.1)–(6.3) after a change inside D_0. The details of this are completely similar to the application of Proposition 2 at the end of the proof of Theorem 6 and are therefore omitted. (The details in the case that $\mu=2\pi\delta$ are given in [7, §IVa].)

The unicity and monotonicity properties of D_t also follow easily from Theorem 6. As to the moment inequality (6.4) it follows from Theorem 7 that

$$\int_{D_t}\varphi-\int_{D_0}\varphi\geq t\cdot\int\varphi\, d(\mu*h_\varepsilon)$$

for all $\varphi\in H^2(\mathbb{R}^2)$ subharmonic in D_t, where h_ε is that mollifier, defined by (1.5) for some suitable $\varepsilon>0$, used in the beginning of this proof. But, since $\varphi\leq\varphi*h_\varepsilon$ in a neighborhood of $\operatorname{supp}\mu$ by the sub-mean value property of subharmonic functions, we have

$$\int\varphi\, d(\mu*h_\varepsilon)=\int(\varphi*h_\varepsilon)\, d\mu\geq\int\varphi\, d\mu$$

and so

$$\int_{D_t}\varphi-\int_{D_0}\varphi\geq t\cdot\int\varphi\, d\mu$$

for $\varphi\in H^2(\mathbf{R}^2)$ subharmonic in D_t. This ends the proof of Theorem 8.

Now consider a weak solution $t\to D_t$ in the sense of Theorem 8. There are strong reasons to believe that D_t in some sense becomes nicer as t increases. One for example expects that for any $t>0$, D_t is bounded by analytic curves even if D_0 is not. This we cannot prove (and it is not true for completely arbitary initial domains D_0, as we shall see in a moment). What we can prove is the following.

THEOREM 9. *Let $t\to D_t$ be as in Theorem* 8 *and suppose that, for some fixed t, D_t is connected and finitely connected, and $D_0\subset\subset D_t$. Then ∂D_t is a finite disjoint union of analytic curves and isolated points.*

By an "analytic curve" we mean precisely the following: a subset of $\mathbf{C}$ is an analytic curve if it is the image of $\partial\mathbf{D}$ under some nonconstant function holomorphic in a neighbourhood of $\partial\mathbf{D}$. Thus an analytic curve is allowed to have singular points. For the proof of Theorem 9 we need the following lemma which shows to what extent the term D_0 in (6.3) is necessary.

LEMMA 3. *Let, in the situation of Theorem* 8,

$$U_t=\{z\in\mathbf{C}:u_t(z)>0\}$$

where u_t, being superharmonic in D_0 by (6.1), *is normalized to be lower semicontinuous in D_0 (and continuous outside* supp μ). *Then, for $t>0$,*

(i) *If N is a component of D_0, then either $N\subset U_t$ or $N\cap U_t=\varnothing$, and the latter case can occur only if $N\cap$* supp $\mu=\varnothing$.

(ii) *D_t is the union of U_t and those components of D_0 which do not meet U_t.*

(iii) *If D_0 is connected, then $D_t=U_t$ and D_t is connected.*

Proof of the lemma.

(i) In D_0, and in particular in N, u_t is a superharmonic function. Therefore, since N is connected and $u_t\geq 0$, if u_t attains the value 0 in N, it must be constantly equal to 0 in N. Thus either $u_t>0$ in N or $u_t\equiv 0$ in N. Moreover, it is obvious (from (6.1)) that the latter case can occur only if N does not meet supp μ. This proves (i).

(ii) is an immediate consequence of (i) and the definition (6.3) of D_t.

(iii) Since supp $\mu\subset D_0$, it follows from (i) (with $N=D_0$) that $D_0\subset U_t$. Thus, $D_t=U_t$. It remains to show that U_t is connected. But, since D_0 is connected and $D_0\subset U_t$, if U_t were not connected, there would be some component V of U_t such that $V\cap D_0=\varnothing$. And this is impossible because u_t is subharmonic outside supp μ, in particular outside D_0, $u_t>0$ in V and $=0$ on ∂V. This completes the proof.

Proof of Theorem 9. Let $U_t=\{z\in\mathbf{C}:u_t(z)>0\}$. Then $D_0\subset\subset D_t=D_0\cup U_t$ shows that $\partial D_0\subset U_t$. This implies that each component of D_0 intersects U_t, and so, by (ii) of Lemma 3, $D_t=U_t$.

Now let γ be a component of $\partial D_t=\partial U_t$ and we shall show that γ is an analytic curve or a point. Let $\hat{\mathbf{C}}=\mathbf{C}\cup\{\infty\}$ denote the Riemann sphere. Since U_t is connected, there is exactly one component of $\hat{\mathbf{C}}\setminus\gamma$ which contains U_t. Let V denote that component. Then it is easy to see that $\partial V=\gamma$. Since γ is connected, this also shows that V is simply connected.

Put $W=D_t\setminus\bar{D}_0=U_t\setminus\bar{D}_0\subset V$, and define

$$S(z)=\bar{z}-4\frac{\partial u_t}{\partial z} \tag{6.5}$$

for $z \in W \cup \gamma$ $(\partial/\partial z = \frac{1}{2}(\partial/\partial x - i\partial/\partial y))$. Due to the assumption that D_t is finitely connected, D_t is a neighbourhood of γ in V (the other components of ∂D_t cannot cluster at γ). Since $\bar{D}_0$ is a compact subset of D_t (also by assumption), it follows that also W is a full neighborhood of γ in V.

It follows from (6.1) that u_t is continuously differentiable outside μ. Hence $S(z)$ is a continuous function on $W \cup \gamma$. On $\gamma \subset \mathbb{C} \setminus U_t$, u_t attains its minimum $(u_t = 0)$. Therefore $\partial u_t/\partial z = 0$ on γ, so

(6.6) $$S(z) = \bar{z} \quad \text{on } \gamma.$$

In $W S(z)$ is holomorphic since, by (6.5) and (6.1),

$$\frac{\partial S}{\partial \bar{z}} = 1 - \Delta u_t = 1 - \chi_{D_t} + \chi_{D_0} + t \cdot \mu = 0 \quad \text{in } W.$$

Now it is known (cf. [1, Lemma 6.1] or [9, p. 152]) that the existence of a function with the properties of $S(z)$ above gives the desired conclusion for γ. To be precise, if γ just consists of one point, we are done. Otherwise (since V is simply connected and $\partial V = \gamma$) V can be mapped conformally onto $\mathbb{D}$. Let f: $\mathbb{D} \to V$ be the inverse map.

Then $S(f(\zeta))$ is holomorphic in the neighborhood $f^{-1}(W)$ of $\partial\mathbb{D}$ in $\mathbb{D}$ and (6.6) shows that

$$S(f(\zeta)) - \overline{f(\zeta)} \to 0 \quad \text{as } \zeta \to \partial\mathbb{D} \quad (\zeta \in \mathbb{D}).$$

It can be seen that this implies that $f(\zeta)$ extends analytically across $\partial\mathbb{D}$ by defining $f(\zeta) = \overline{S(f(1/\bar{\zeta}))}$ for ζ in a neighborhood of $\partial\mathbb{D}$ in $\mathbb{C} \setminus \bar{\mathbb{D}}$.

Moreover, it is seen that $f(\partial\mathbb{D}) = \gamma$. This shows that γ is an analytic curve and the theorem is proven.

Remark. Theorem 9 is not quite satisfactory because of its three assumptions a priori on D_t. The first of these, that D_t is connected, is however harmless and is automatically fulfilled if D_0 is connected (by (iii) of Lemma 3).

The second of the assumptions, that D_t is finitely connected, I do not know how to get rid of although I suspect that it is also automatically fulfilled (possibly some weak assumption on D_0 is needed).

The third assumption, that $D_0 \subset\subset D_t$, can be replaced by either one of the following two assumptions.

(i) t is sufficiently large.

(ii) D_0 is connected and is bounded by finitely many disjoint analytic curves.

As to (i), we actually have $D_0 \subset\subset D_t$ for t sufficiently large. This is seen by comparing our solution $t \to D_t$ (corresponding to the measure μ and initial domain D_0) with some suitable known solution $t \to D_t'$ defined by some measure μ' and initial domain D_0'. As indicated by Proposition 2 we can assume that μ is a continuous function. Since $\mu \neq 0$, μ must be strictly positive somewhere, say at the origin. Then we can choose μ' to be a radially symmetric function such that $\mu' \leq \mu$ and D_0' to be a disc, centered at the origin, such that $D_0' \subset D_0$. Then, by the monotonicity properties of D_t as a function of μ and D_0, we have $D_t' \subset D_t$ for all $t > 0$. On the other hand D_t' is a disc (this follows from radial symmetry and uniqueness of solutions of (6.1)–(6.3)) which grows beyond all bounds as t increases (it follows e.g. by integrating both sides of (6.1) over $\mathbb{R}^2$ that $|D_t'| = |D_0'| + t \cdot \int d\mu'$). Therefore $D_0 \subset\subset D_t'$, in particular $D_0 \subset\subset D_t$, for all sufficiently large t, as we wanted to prove.

To prove that (ii) can replace $D_0 \subset\subset D_t$ we first note that, in the proof of Theorem 9, we still have $D_t = U_t$ by (iii) of Lemma 3. The assumption that ∂D_0 is analytic implies that there exists a function $S_0(z)$, defined and continuous in $(D_0 \setminus K) \cup \partial D_0$, where K is some compact subset of D_0, and holomorphic in $D_0 \setminus K$ such that

$$S_0(z) = \bar{z} \quad \text{on } \partial D_0.$$

Now in the proof of Theorem 8 we change the definitions of W and $S(z)$ to $W = D_t \setminus (K \cup \operatorname{supp} \mu)$ and

$$S(z) = \bar{z} - 4\frac{\partial u_t}{\partial z} + \chi_{D_0}(z) \cdot (S_0(z) - \bar{z}) \quad \text{for } z \in W \cup \gamma.$$

Here it is assumed that $S_0(z)$ is extended to $W \cup \gamma$ in some way, e.g. by $S_0(z) = \bar{z}$ for $z \in (W \cup \gamma) \setminus \bar{D}_0$. Then $S(z)$ is continuous on $W \cup \gamma$, holomorphic in W (since (6.1) shows that $\partial S/\partial \bar{z} = 0$ in $W \setminus \partial D_0$ and ∂D_0 is a nice curve) and $S(z) = \bar{z}$ on γ. The rest of the proof of Theorem 8 works as before and so (ii) is proved.

I am sure that the assumption $D_0 \subset\subset D_t$ in Theorem 9 can be replaced by some much weaker assumption on D_0 than (ii). However some assumption is needed as the following example shows. Choose D_0 such that ∂D_0 has positive two-dimensional Lebesgue measure. D_0 could e.g. be a square with a lot of slits (of constant length) along one side, spaced as a Cantor set of positive length. Then it will take a positive time for D_t to move through ∂D_0 (since $|D_t| - |D_0| = t \cdot \int d\mu$) and therefore $D_0 \subset\subset D_t$ cannot hold for small $t > 0$. Moreover ∂D_t cannot be analytic for these t. This shows that the conclusion of Theorem 9 is not valid if the hypothesis $D_0 \subset\subset D_t$ is completely omitted.

Despite its weaknesses Theorem 9 is strong enough to ensure that classical solutions always are bounded by analytic curves.

THEOREM 10. *Let μ, ω and I be as before Definition* 1 *and let $I \ni t \to D_t \in \mathbb{S}_\omega$ be a classical solution. Then*

(i) *If $I \ni t \to D_t' \in \mathbb{S}_\omega$ is another classical solution and $D_\tau' = D_\tau$ for some $\tau \in I$ then $D_t' = D_t$ for all $t \in I$ with $t > \tau$.*

(ii) *∂D_t is an analytic curve for every $t \in I$.*

Proof. We can assume that μ is a nice function, by Proposition 1. To prove (i) assume, without loss of generality since the concept of a classical solution is invariant with respect to time translations, that $\tau = 0 \in I$ and then apply Theorem 1 with $T \geq t$. Combined with the unicity statement of Theorem 6 this gives that $D_t' = D_t$ up to null sets. But it is easy to see that, in view of the regularity assumptions on ∂D_t and $\partial D_t'$, this implies that $D_t' = D_t$ everywhere. This proves (i).

To prove (ii) take $\tau \in I$, $\tau < t$. We may assume that $\tau = 0$. Now we apply Theorem 1 with $T \geq t$. This shows that $t \to D_t$ is a weak solution in the sense of Definition 2. According to Theorem 6 (uniqueness part)

$$D_t = D_0 \cup \{z \in \Omega : u_t(z) > 0\} \tag{6.7}$$

up to null sets, where u_t is the function defined by (2.9).

Now using the regularity of ∂D_t it is not very hard to see that (6.7) actually holds everywhere. In fact we have $D_t = \{z \in \Omega : p_{D_t}(z) > 0\}$ for all $0 < t \leq T$, and the function $p_{D_t}(z)$ increases with t', and from this it follows that $D_t = \{z \in \Omega : u_t(z) > 0\}$ (for $0 < t \leq T$). Moreover, the regularity of ∂D_t also implies that $-\partial p_{D_t}/\partial n > 0$ on ∂D_t for all t. In view of the continuity of $\partial \zeta/\partial t = -\nabla p_{D_t}$ (Definition 1) this easily implies that $D_{t_1} \subset\subset D_{t_2}$ for $t_1 < t_2$, in particular that $D_0 \subset\subset D_t$ for $t > 0$. Now it follows from Theorem 9 that ∂D_t is an analytic curve.

REFERENCES

[1] D. Aharonov and H. S. Shapiro, *Domains on which analytic functions satisfy quadrature identities*, J. d'Anal. Math., 30 (1976), pp. 39–73.

[2] C. Baiocchi, *Free boundary problems in the theory of fluid flow through porous media*, in Proc. International Congress of Mathematicians, Vancouver 1974, Vol. 2, Canadian Mathematical Congress, 1975, pp. 237–243.

[3] H. Brezis, *Problèmes unilatéraux*, J. Math. Pures Appl., 51 (1972), pp. 1–158.

[4] G. Duvaut, *The solution of a two-phase Stefan problem by a variational inequality*, in Moving Boundary Problems in Heat Flow and Diffusion, Ockendon and Hodgkins, eds., Clarendon Press, Oxford, 1975, pp. 173–181.

[5] C. M. Elliott, *On a variational inequality formulation of an electrochemical machining moving boundary problem and its approximation by the finite element method*, J. Inst. Math. Appl., 25 (1980), pp. 121–131.

[6] C. M. Elliott and V. Janovský, *A variational inequality approach to Hele Shaw flow with a moving boundary*, Proc. Royal Soc. Edinburgh, 88A, (1981), pp. 93–107.

[7] B. Gustafsson, *Applications of variational inequalities to a moving boundary problem for Hele Shaw flows*, TRITA-MAT-1981-9, Mathematics, Royal Institute of Technology, S-100 44 Stockholm.

[8] ———, *On a differential equation arising in a Hele Shaw flow moving boundary problem*, TRITA-MAT-1981-36, Mathematics, Royal Institute of Technology, S-100 44 Stockholm.

[9] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[10] H. Lamb, *Hydrodynamics*, 6th ed., Dover, New York, 1932.

[11] J. R. Ockendon, *Numerical and analytic solutions of moving boundary problems*, in Moving Boundary Problems, Wilson, Solomon and Boggs, eds., Academic Press, New York, 1978, pp. 129–145.

[12] S. Richardson, *Hele Shaw flows with a free boundary produced by the injection of fluid into a narrow channel*, J. Fluid Mech., 56 (1972), pp. 609–618.

[13] ———, *Some Hele Shaw flows with time-dependent free boundaries*, J. Fluid Mech., 102 (1981), pp. 263–278.

[14] ———, *Hele Shaw flows with time-dependent free boundaries in infinite and semi-infinite strips*, Quart. J. Mech. and Appl. Math., 35 (1982), pp. 531–548.

[15] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1973.

[16] M. Sakai, *Quadrature domains*, Lecture Notes in Mathematics 934, Springer-Verlag, New York, 1982.

[17] ———, *Applications of variational inequalities to the existence theorem on quadrature domains*, Trans. Amer. Math. Soc., 276 (1983), pp. 267–279.

[18] F. Treves, *Basic Linear Partial Differential Equations*, Academic Press, New York, 1975.

[19] Yu. P. Vinogradov and P. P. Kufarev, *On a seepage problem*, Prikl. Mat. i Makh., XII (1948) bulletin 2. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

GLOBAL EXISTENCE OF CLASSICAL SOLUTIONS IN THE DISSIPATIVE SHALLOW WATER EQUATIONS*

P. E. KLOEDEN†

Abstract. An energy method of Matsumura and Nishida is used to prove the global temporal existence of classical solutions in the dissipative shallow water equations on a spatially periodic horizontal domain. This requires the external forcing, the initial data and the initial first order time derivatives to be sufficiently small. The result is related to the data initialization problem in numerical weather prediction, and extends a local existence result of Bui An Ton for the same equations.

1. Introduction. Inappropriate initial data in numerical weather predication often leads to spurious motions of high frequency and large amplitude which soon dominate and obscure the slower meteorologically significant motions (e.g. [2], [6]). Several different *data initialization* schemes have been proposed to modify the initial data, which is usually incomplete and inaccurate, in order to control these undesired motions and thus to extend the time interval for which a reasonably accurate forecast can be made. In particular, the bounded-derivative method of Kreiss (e.g. see [2]) involves choosing the data so that initial time derivatives are sufficiently small and guarantees that they remain small for a certain period of time.

The shallow water equations [3], [5], [6], [9], [10], with or without dissipative terms, are the simplest type of equations in geophysical fluid dynamics for which this interaction of fast and slow solutions occurs [6]. In this paper we consider the global temporal existence of classical solutions for the dissipative shallow water equations [3], [5], [10]. We use an energy method developed by Matsumura and Nishida [7], [8], which involves extending a solution guaranteed by a local existence theorem by repeatedly applying an a priori estimate. This requires the norms of both the solution and its first order time derivatives to remain small for the time period under consideration. We work with square-integrable Sobolev spaces and then use an embedding theorem to show that the solutions are in fact classical solutions. For simplicity we restrict attention to flows which are periodic in both horizontal spatial variables, as are often considered in meteorological studies, although we note that Matsumura and Nishida have used their method in more complicated boundary value problems. Our result shows that global temporal existence of classical solutions can be achieved provided the first order time derivatives are kept sufficiently small. This provides some mathematical justification for various data initialization schemes, and also extends a local existence result for the dissipative shallow water equations obtained by Bui An Ton [10].

We state our main result in §2 and outline the proof. In §3 we show the existence and uniqueness of a steady solution, about which we take nonlinear perturbations. Then in §4 we state a local existence result, the proof of which is standard, and an a priori estimate for the perturbation equations and show how it is used repeatedly to extend the local solution. We prove the a priori estimate in §5 by successively using eleven estimates, which we prove in §§6 and 7.

* Received by the editors January 27, 1983, and in revised form August 30, 1983.

† School of Mathematical and Physical Sciences, Murdoch University, Murdoch, Western Australia, 6150, and Centre for Environmental Fluid Dynamics, University of Western Australia, Nedlands, Western Australia.

2. The dissipative shallow water equations. The dissipative shallow water equations are a coupled system of nonlinear hyperbolic and parabolic partial differential equations [3], [5], [10]

$$h_t+(hu^j)_{x_j}=0, \tag{2.1}$$

$$h\left(u^i_t+u^j u^i_{x_j}\right)-\nu\left(hu^i_{x_j}\right)_{x_j}+hh_{x_i}=h\phi_i, \tag{2.2}$$

on a horizontal spatial domain Ω, where $i=1,2$ and repeated indices indicate summation. Here gh is the geopotential, $x=(x_1,x_2)$ is the horizontal spatial variable, $u=(u^1,u^2)$ is the horizontal velocity, ν is the viscosity and $\phi=(\phi_1,\phi_2)$ is the external force, which is assumed generated by a potential function Φ with $\Phi_i=-\Phi_{x_i}$. The positivity constraint

$$h(x,t)>0 \quad \text{for all } x\in\Omega, \quad t\geqq 0, \tag{2.3}$$

the boundary condition

$$u^i(x,t)=0 \quad \text{for all } x\in\partial\Omega, \quad t\geqq 0, \tag{2.4}$$

and the initial conditions

$$u^i(x,0)=u^i_0(x) \quad \text{and} \quad h(x,0)=h_0(x)>0 \quad \text{for all } x\in\Omega \tag{2.5}$$

must also be satisfied.

A steady solution $\hat{h}=\hat{h}(x)$, $\hat{u}^i=\hat{u}^i(x)$ of (2.1)–(2.4) satisfies the steady equations

$$(\hat{h}\hat{u}^j)_{x_j}=0, \tag{2.6}$$

$$\hat{h}\hat{u}^j\hat{u}^i_{x_j}-\nu\left(\hat{h}\hat{u}^i_{x_j}\right)+\hat{h}\hat{h}_{x_i}=\hat{h}\phi_i, \tag{2.7}$$

with $\hat{u}^i(x)=0$ on $\partial\Omega$ and $\hat{h}(x)>0$ on Ω.

In this paper we restrict attention to doubly periodic flows, that is, those which are which are periodic in both horizontal spatial variables, and we take the open unit square $(0,1)\times(0,1)$ as the principal flow region Ω. All functions are subsequently doubly periodic on Ω and the spatial average of such a function f is denoted by

$$\bar{f}=\int_\Omega f(x)\,dx.$$

We denote by $W^{k,2}(\Omega)$ the Sobolev space of functions f which, together with their lth-order generalized spatial derivatives $D^l f$ for $l=1,2,\cdots,k$, belong to $L_2(\Omega)$, with norm

$$\|f\|_k=\left(\sum_{l=0}^{k}\int_\Omega |D^l f(x)|^2 dx\right)^{1/2};$$

by $C^j(0,T;\,W^{k,2}(\Omega))$ the space of functions f: $[0,T]\to W^{k,2}(\Omega)$ which are j times continuously differentiable in t; and by $L_2(0,T;\,W^{k,2}(\Omega))$ the space of functions f: $[0,T]\to W^{k,2}(\Omega)$ for which $\|f(t)\|_k$ is square integrable on $0\leqq t\leqq T$.

Our main result is this theorem.

THEOREM. *Let h_0, u^i_0 and Φ be doubly periodic functions with h_0, $u^i_0\in W^{4,2}(\Omega)$, $\Phi\in W^{5,2}(\Omega)$ and $h_0(x)>0$ on Ω.*

Then there exists a positive number ε *such that if*

$$\|h_0-\bar{h}_0\|_4 \leqq \varepsilon, \quad \|u_0^i\|_4<\varepsilon, \quad \|D\Phi\|_4<\varepsilon,$$

the dissipative shallow water equations (2.1)–(2.5) *have a unique doubly periodic solution* (h,u^i) *and a unique doubly periodic steady solution* $(\hat{h},\hat{u}^i)$ *with*

$$\begin{gathered}\hat{h}(x)=\bar{h}_0+\bar{\Phi}-\Phi(x), \qquad \hat{u}^i(x)\equiv 0,\\ h-\hat{h}\in C^0(0,\infty;W^{4,2}(\Omega))\cap C^1(0,\infty;W^{3,2}(\Omega)),\\ u^i\in C^0(0,\infty;W^{4,2}(\Omega))\cap C^1(0,\infty;W^{2,2}(\Omega)),\end{gathered} \tag{2.8}$$

and

$$\sup_{x\in\Omega}\{|h(x,t)-\hat{h}(x)|,|u^i(x,t)|\}\to 0 \quad \text{as } t\to\infty.$$

The solutions (h,u^i) and $(\hat{h},\hat{u}^i)$ in the theorem are in fact *classical solutions*, with all of the derivatives appearing in equations (2.1), (2.2), (2.6) and (2.7) existing in the classical sense and continuous on $\Omega\times[0,\infty)$. This follows from the embedding of the Sobolev space $W^{j+2,2}(\Omega)$ in the Hölder space $C^{j,\lambda}(\Omega)$ for any $0<\lambda<1$ and $j=0,1,2,\cdots$ [1, Thm. 5.4, Part II, Case C''].

To prove the theorem we first show the uniqueness of the steady solution (2.8) (Proposition 1) and then perturb about it to obtain for the perturbations $h'=h-\hat{h}$, $u^{i\prime}=u^i-\hat{u}^i=u^i$ the equations (in which we omit the primes for convenience)

$$h_t+u^jh_{x_j}+\bar{h}_0u^j_{x_j}=F^0, \tag{2.9}$$

$$u^i_t-\nu u^i_{x_jx_j}+h_{x_i}=F^i, \tag{2.10}$$

where

$$F^0=(\bar{h}_0-h-\hat{h})u^j_{x_j}-\hat{h}_{x_j}u^i, \tag{2.11}$$

$$F^i=-u^ju^i_{x_j}+\nu\frac{(h+\hat{h})_{x_j}}{(h+\hat{h})}u^i_{x_j}. \tag{2.12}$$

Next we introduce the bounded closed convex subset.

$$\begin{aligned}\mathscr{X}=\mathscr{X}&(t_1,t_2;E)\\ =\{&(h,u^i);h\in C^0(t_1,t_2;W^{4,2}(\Omega))\cap L_2(t_1,t_2;W^{4,2}(\Omega)),\\ &h_t\in C^0(t_1,t_2;W^{3,2}(\Omega))\cap L_2(t_1,t_2;W^{3,2}(\Omega)),\\ &u^i\in C^0(t_1,t_2;W^{4,2}(\Omega))\cap L_2(t_1,t_2;W^{5,2}(\Omega)),\\ &u^i_t\in C^0(t_1,t_2;W^{2,2}(\Omega))\cap L_2(t_1,t_2;W^{3,2}(\Omega))\\ &\qquad\qquad\qquad \text{with } N(h,u;t_1,t_2)\leqq E\}\end{aligned} \tag{2.13}$$

of a function space $\mathscr{X}(t_1,t_2)$ which is defined as in (2.13) without the restriction on

$N(h,u;t_1,t_2)$,where

(2.14)

$$N^2(h,u;t_1,t_2)=\sup_{t_1\leqq\tau\leqq t_2}\left\{\|h(\tau)\|_4^2+\|h_t(\tau)\|_3^2+\|u(\tau)\|_4^2+\|u_t(\tau)\|_2^2\right\}$$
$$+\int_{t_1}^{t_2}\left\{\|Dh(\tau)\|_3^2+\|Dh_t(\tau)\|_2^2+\|Du(\tau)\|_4^2+\|Du_t(\tau)\|_2^2\right\}d\tau$$

which we abbreviate by $N^2(t_1,t_2)$; then (h,u) is fixed. Next we indicate the proof, which is standard, for the local existence in time of a unique doubly periodic solution in the space $\mathscr{X}$ of the nonlinear perturbation equations (2.9)–(2.12) with boundary and initial conditions

$$(2.15)\qquad u^i|_{\partial\Omega}=0 \quad\text{and}\quad u^i(x,0)=u_0^i(x),\quad h(x,0)=h_0(x)-\hat{h}(x)\quad \text{on } \Omega$$

(Proposition 2). We then obtain the global existence indicated in the Theorem by repeatedly extending the local situation with the aid of a priori estimate (Proposition 3), the proof of which constitutes the bulk of the paper.

The method we use here is due to Matsumara and Nishida [7], [8], who have applied it to the equations of motion for viscous heat conducting compressible fluids. Our restriction to doubly periodic flows simplifies matters in that we essentially only have to establish interior estimates, though like Matsumura and Nishida we could also derive corresponding boundary estimates and thus extend our result to the usual type of initial boundary value problem.

We note that the inclusion of Coriolis terms in equations (2.2) does not alter the results of the Theorem.

3. The steady solution. For any positive constant $\bar{h}_0$ and force potential Φ, the function $\hat{h}$ defined by $\hat{h}(x)=\bar{h}_0+\bar{\Phi}-\Phi(x)$ satisfies

$$(3.1)\qquad |\hat{h}(x)-\bar{h}_0|\leqq\tfrac{1}{4}\bar{h}_0,$$

say, and hence $\hat{h}(x)\geqq\frac{3}{4}\bar{h}_0>0$ on Ω, whenever Φ satisfies

$$(3.2)\qquad |\Phi(x)-\bar{\Phi}|\leqq\tfrac{1}{4}\bar{h}_0.$$

By [1, Thm. 5.4] and a Poincaré inequality [4, p. 157]

$$(3.3)\qquad |\Phi(x)-\bar{\Phi}|\leqq K_p\|D\Phi\|_1,$$

where the constant K_p depends only on the domain Ω. Inequality (3.2) certainly holds whenever

$$(3.4)\qquad \|D\Phi\|_4\leqq E_1(\bar{h}_0)=\bar{h}_0/(4K_p),$$

which implies that $\hat{h}(x)$ is positive on Ω.

Consequently the functions $(\hat{h},\hat{u}^i)$ defined in (2.8) are a steady solution and it remains to show that it is the only steady solution with spatial average

$$(3.5)\qquad \int_\Omega \hat{h}(x)\,dx=\bar{h}.$$

To see this, we let $(\hat{h}, \hat{u}^i)$ be any steady solution. Then from equations (2.7) we have on integrating by parts and using equation (2.6)

$$\begin{aligned}
0 &= \int_\Omega \hat{u}^i \left\{ \hat{h}\hat{u}^j \hat{u}^i_{x_j} - \nu \left(\hat{h}\hat{u}^i_{x_j} \right)_{x_j} + \hat{h}(\hat{h} + \hat{\Phi})_{x_i} \right\} dx \\
&= \int_\Omega \left\{ \hat{h}\hat{u}^j \left(\tfrac{1}{2}\hat{u}^i\hat{u}^i \right)_{x_j} - \nu \hat{u}^i \left(\hat{h}\hat{u}^i_{x_j} \right)_{x_j} + \hat{u}^i \hat{h}(\hat{h} + \Phi)_{x_i} \right\} dx \\
&= -\int_\Omega \left\{ (\hat{h}\hat{u}^j)_{x_j} \left(\tfrac{1}{2}\hat{u}^i\hat{u}^i \right) - \nu \hat{h}\hat{u}^i_{x_j}\hat{u}^i_{x_j} + (\hat{h}\hat{u}^i)_{x_i}(\hat{h} + \Phi) \right\} dx \\
&= \nu \int_\Omega \hat{h}\hat{u}^i_{x_j}\hat{u}^i_{x_j}\, dx.
\end{aligned}$$

As $\hat{h}$ is positive, we have $\hat{u}^i_{x_j}\hat{u}^i_{x_j} = 0$ on Ω and as $u^i = 0$ on $\partial\Omega$, we thus have $\hat{u}^i \equiv 0$ on Ω for any steady solution. Equations (2.7) then reduce to

$$\hat{h}(\hat{h} + \Phi)_{x_i} = 0,$$

for which the $\hat{h}$ in (2.8) is the unique positive solution satisfying (3.5). We have thus proved the following proposition.

PROPOSITION 1 (steady solution). *Let $\bar{h}_0$ be an arbitrary positive constant and let $\Phi \in W^{5,2}(\Omega)$ be doubly periodic and satisfy* (3.4).

Then the functions $(\hat{h}, \hat{u}^j)$ defined by (2.8) *are the unique doubly periodic solution of the dissipative shallow water equations with h satisfying* (3.5).

4. Local and global existence. The positivity constraint (2.3) requires $h(x,t) + \hat{h}(x) > 0$ everywhere for any solution (h, u^i) of the perturbation equations (2.9)–(2.12). This certainly holds if the initial data and the force potential are restricted so that (3.1) and

$$(4.1) \qquad |h(x,t)| \leqq \tfrac{1}{2}\bar{h}_0,$$

say, both hold, for then $|h(x,t) + \hat{h}(x) - \bar{h}_0| \leqq \frac{3}{4}\bar{h}_0$ and consequently $h(x,t) + \hat{h}(x) \geqq \frac{1}{4}\bar{h}_0 > 0$ holds everywhere. We can guarantee that (4.1) holds if the perturbation solution (h, u^i) satisfies

$$(4.2) \qquad N(0,T) \leqq E_2(\bar{h}_0) = \bar{h}_0/(2K_e),$$

where K_e is the constant in the embedding of the Sobolev space $W^{2,2}(\Omega)$ in the Hölder space $C^\lambda(\bar{\Omega})$ [1, Thm. 5.4], for then

$$|h(x,t)| \leqq K_e \cdot \sup_{0 \leqq t \leqq T} \|h(t)\|_2 \leqq K_e N(0,T) \leqq \tfrac{1}{2}\bar{h}_0.$$

It thus suffices to find a solution (h, u^i) of the perturbation equations (2.9)–(2.12) with (2.15) in $\mathscr{X}(0,T;E)$ with $E \leqq E_2(\bar{h}_0)$ and some $T > 0$. In fact we have the next proposition.

PROPOSITION 2 (local existence). *Suppose that $\|D\Phi\|_4 \leqq E_1(\bar{h}_0)$ and that the perturbation equations* (2.9)–(2.12) *with* (2.15) *have a unique doubly periodic solution $(h, u^i) \in \mathscr{X}(0,T;E_2(\bar{h}_0))$ for some $T \geqq 0$.*

Then there exist positive constants τ, ε_0 and C_0 with $\varepsilon_0\sqrt{1 + C_0^2} \leqq E_2(\bar{h}_0)$, which are independent of T, such that the perturbation equations have a unique doubly periodic solution $(h, u^i) \in \mathscr{X}(T, T+\tau; C_0 N(T,T))$ whenever $N(T,T) \leqq \varepsilon_0$.

The proof of Proposition 2 involves the method of successive approximations. We pick any $(\rho, v^i) \in \mathscr{X}(T, T+T_1; E)$ for some positive T_1 and E. Then we use standard methods to solve locally in time $t \geqq T$ the system

$$h_t + u^j h_{x_j} = F^0 - \bar{h} v^j_{x_j}, \tag{4.3}$$

$$u^i_t - \nu u^i_{x_j x_j} = F^i - h_{x_i}, \tag{4.4}$$

with boundary and initial (at $t = T$) conditions like (2.5), where F^0 and F^i are defined as in (2.11) and (2.12) using (ρ, v^i) instead of (h, u^i). The details of the proof are similar to those in Matsumura and Nishida [7] for the initial value problem for the equations of motion of a viscous heat conducting gas, but are simpler. Consequently we omit them here.

We note that Bui An Ton [10] has used Langrangian coordinates and Hölder space estimates to establish the local existence of classical solutions for the unforced dissipative shallow water equations. Here, however, we use Sobolev space estimates in order to extend the solution globally in time by means of the following a priori estimate on the "energy" (2.14) of the system.

Proposition 3. *Suppose that the perturbation problem* (2.9)–(2.12) *with* (2.15) *has a doubly periodic solution* $(h, u^i) \in \mathscr{X}(0, T; E_2(\bar{h}_0))$ *for some* $T > 0$.

Then there exist positive constants ε_1 *and* C_1 *with* $\varepsilon_1 < \varepsilon_0$ *and* $\varepsilon_1 C_1 \leqq E_2(\bar{h}_0)$, *which are independent of* T, *such that inequality*

$$N(0, T) \leqq C_1 N(0, 0) \tag{4.5}$$

holds whenever $N(0, T) \leqq \varepsilon_1$ *and* $\|D\Phi\|_4 \leqq \varepsilon_1$.

We will prove this proposition in the following sections of the paper, but show here how it is combined with Proposition 2 to give the global existence of the theorem. For this we choose the initial data (h_0, u^i_0) and the force potential Φ so small that

$$N(0,0) \leqq \min\left\{\varepsilon_0, \varepsilon_1/C_0, \varepsilon_1/C_1\sqrt{1+C_0^2}, E_1(\bar{h}_0), E_2(\bar{h}_0)\right\}$$

and

$$\|D\Phi\|_4 \leqq \min\left\{\varepsilon_1, E_1(\bar{h}_0), E_2(\bar{h}_0)\right\}.$$

By Proposition 2 with $T = 0$ there exists a local solution (h, u^i) in $\mathscr{X}(0, \tau; C_0 N(0,0))$ and since $C_0 N(0,0) \leqq \varepsilon_1 \leqq \varepsilon_0$ by Proposition 3 with $T = \tau$

$$N(0, \tau) \leqq C_1 N(0, 0). \tag{4.6}$$

Then by Proposition 2 with $T = \tau$ this solution has an extension (h, u^i) in $\mathscr{X}(\tau, 2\tau; C_0 N(\tau, \tau))$, so $N(\tau, 2\tau) \leqq C_0 N(\tau, \tau)$. The solution (h, u^i) on $0 \leqq t \leqq 2\tau$ is then in $\mathscr{X}(0, 2\tau; \sqrt{1+C_0^2}\, N(0, \tau))$ as

$$N^2(0, 2\tau) \leqq N^2(0, \tau) + N^2(\tau, 2\tau) \leqq N^2(0, \tau) + C_0^2 N^2(\tau, \tau) \leqq \left(1 + C_0^2\right) N^2(0, \tau).$$

In view of (4.4) we have $\sqrt{1+C_0^2}\, N(0, \tau) \leqq C_1 \sqrt{1+C_0^2}\, N(0,0) \leqq \varepsilon_1$ and so by Proposition 3 with $T = 2\tau$ we obtain $N(0, 2\tau) \leqq C_1 N(0,0)$. Then by Proposition 2 with $T = 2\tau$ the above solution has an extension in $\mathscr{X}(2\tau, 3\tau;\ C_0 N(2\tau, 2\tau))$ and hence the solution (h, u^i) on $0 \leqq t \leqq 3\tau$ is in $\mathscr{X}(0, 3\tau; \sqrt{1+C_0^2} \cdot N(0, 2\tau))$. Continuing in this way, we obtain global existence of a solution, which is in fact unique for given initial data.

To obtain the asymptotic property stated in the Theorem we note that such a solution (h,u^i) in fact belongs to $\mathscr{X}(0,\infty;\varepsilon_1)$. Hence from definition (2.4) of $N(h,u;0,\infty)$ with $t_1=0$ and $t_2=\infty$,

$$\lim_{t\to\infty}\|Dh(t)\|_3=0 \quad\text{and}\quad \lim_{t\to\infty}\|Du(t)\|_4=0.$$

The result then follows from an application of the Poincaré and Sobolev embedding inequalities. It remains thus to prove the a priori estimate of Proposition 3.

5. Proof of Proposition 3. We need the following a priori estimates for a doubly periodic solution (h,u^i) in some space $\mathscr{X}(0,T;E)$ of the perturbation equations (2.9) and (2.10), where F^0 and F^i are considered as given functions of (x,t) with

$$F^0\in C^0(0,T;\,W^{3,2}(\Omega))\cap L_2(0,T;\,W^{4,2}(\Omega)),$$
$$F^i\in C^0(0,T;\,W^{3,2}(\Omega))\cap L_2(0,T;\,W^{3,2}(\Omega))$$

and

$$F_t^i\in C^0(0,T;\,W^{2,2}(\Omega))\cap L_2(0,T;\,W^{3,2}(\Omega)).$$

The boundary conditions are given by (2.15), but for simplicity we write h_0 for $h_0-\hat{h}$. In each case the constant C is independent of the h, u^i, F^0, F^i and the time $T>0$.

LEMMA 1. $\bar{h}=0$ *and* $\|h(t)\|^2\leqq C\|Dh(t)\|^2$.

LEMMA 2. *For* $k=0,1,2$ *and* 3

$$\|D^{k+1}h(t)\|^2+\int_0^t\|D^{k+1}h(\tau)\|^2d\tau$$
$$\leqq C\Big\{\|D^{k+1}h_0\|^2+\|D^ku_0\|^2+\|D^ku(t)\|^2+N^3(0,t)$$
$$+\int_0^t\Big\{\|D^{k+2}u(t)\|^2+\|D^{k+1}F^0(\tau)\|^2+\|D^kF(\tau)\|^2\Big\}d\tau\Big\}.$$

LEMMA 3. *For* $k=1,2,3$ *and* 4

$$\|D^kh(t)\|^2+\|D^ku(t)\|^2+\int_0^t\|D^{k+1}u(\tau)\|^2d\tau$$
$$\leqq C\Big\{\|D^kh_0\|^2+\|D^ku_0\|^2+N^3(0,t)+\int_0^t\Big\{\|D^kF^0(\tau)\|^2+\|D^{k-1}F(\tau)\|^2\Big\}d\tau\Big\}.$$

LEMMA 4.

$$\|u(t)\|_2^2\leqq C\Big\{\|Dh(t)\|^2+\|u_t(t)\|^2+\|F(t)\|^2\Big\}.$$

LEMMA 5. *For* $k=1,2$ *and* 3

$$\|h_t(t)\|_k^2\leqq C\Big\{\|Du(t)\|_k^2+\|F^0(t)\|_k^2+N^3(0,t)\Big\}$$

and

$$\int_0^t\|h_t(\tau)\|_k^2d\tau\leqq C\int_0^t\Big\{\|Du(\tau)\|_k^2+\|F^0(\tau)\|_k^2\Big\}d\tau+CN^3(0,t).$$

LEMMA 6.

$$\|u_t(t)\|_2^2+\int_0^t\|Du_t(\tau)\|_2^2d\tau\leqq C\left\{\|u_t(0)\|_2^2+\int_0^t\left\{\|Dh_t(\tau)\|_2^2+\|F_t(\tau)\|_1^2\right\}d\tau\right\}.$$

We prove these lemmas in the next section. Using them, we have

$$\begin{aligned}
&\|h(t)\|_4^2+\|h_t(t)\|_3^2+\|u(t)\|_4^2+\|u_t(t)\|_2^2\\
&\quad+\int_0^t\left\{\|Dh(\tau)\|_3^2+\|h_t(\tau)\|_3^2+\|Du(\tau)\|_4^2+\|u_t(\tau)\|_2^2\right\}d\tau\\
&\leqq C\left\{\|Dh_0\|_3^2+\|u_0\|_3^2+N^3(0,t)+\int_0^t\left\{\|DF^0(\tau)\|_3^2+\|F(\tau)\|_3^2\right\}d\tau+N^3(0,\tau)\right\}\\
&\quad+C\left\{\|Du(t)\|_3+\int_0^t\|Du(\tau)\|_4^2d\tau\right\}\\
&\quad+\left\{\|h_t(t)\|_3^2+\int_0^t\|h_t(\tau)\|_3^2d\tau\right\}+\left\{\|u_t(t)\|_2^2+\int_0^t\|u_t(\tau)\|_2^2d\tau\right\}
\end{aligned}$$

by Lemma 1, Lemma 2 with $k=0,1,2$ and 3, and Lemma 4

$$\begin{aligned}
&\leqq C\left\{\cdots+\|u_t(0)\|_2^2+\int_0^t\|F_t(\tau)\|^2d\tau\right\}+C\{\cdots\}\\
&\qquad\qquad+C\left\{\|h_t(t)\|_3^2+\int_0^t\|h_t(\tau)\|_3^2d\tau\right\}
\end{aligned}$$

by Lemma 6, where $\cdots$ indicates terms repeated from the previous inequality

$$\leqq C\left\{\cdots+\|F^0(t)\|_3^2+\int_0^t\|F^0(\tau)\|_3^2d\tau\right\}+C\left\{\|Du(t)\|_3^2+\int_0^t\|Du(\tau)\|_4^2d\tau\right\}$$

by Lemma 5 with $k=3$

$$\begin{aligned}
&\leqq C\left\{\cdots+\|Dh_0\|_3^2+\|Du_0\|_3^2+\int_0^t\left\{\|DF^0(\tau)\|_3^2+\|F(\tau)\|_3^2\right\}d\tau\right\}\\
&\leqq C\left\{\|h_0\|_4^2+\|u_0\|_4^2+\|F(0)\|_2^2+\|F^0(t)\|_3^2+N^3(0,t)\right.\\
&\qquad\qquad\left.+\int_0^t\left\{\|F^0(\tau)\|_4^2+\|F(\tau)\|_3^2+\|F_t(\tau)\|^2\right\}d\tau\right\}
\end{aligned}$$

using the fact that equation (2.10) is satisfied at $t=0$, so

$$\|u_t(0)\|_2^2\leqq C\left\{\|u_0\|_4^2+\|F(0)\|_2^2+\|Dh_0\|_2^2\right\}.$$

Taking the supremum in t over the interval $0\leqq t\leqq T$, we obtain

$$\begin{aligned}
(5.1)\quad N^2(0,T)\leqq C&\left\{\|h_0\|_4^2+\|u_0\|_4^2+\sup_{0\leqq t\leqq T}\left\{\|F^0(t)\|_3^2+\|F(t)\|_2^2\right\}+N^3(0,T)\right.\\
&\left.+\int_0^T\left\{\|F^0(\tau)\|_4^2+\|F(\tau)\|_3^2+\|F_t(\tau)\|_1^2\right\}d\tau\right\}.
\end{aligned}$$

To continue we need to estimate the quantities in (5.1) involving F^0 and F which we now consider as functions of h and u^i defined by (2.11) and (2.12). We obtain Lemmas 7–11.

LEMMA 7.

$$\sup_{0\leqq t\leqq T}\|F^0(t)\|_3^2\leqq C\left\{N^2(0,T)+\|D\Phi\|_4^2\right\}\cdot N^2(0,T).$$

LEMMA 8.

$$\int_0^T\|F^0(\tau)\|_4^2d\tau\leqq C\left\{N^2(0,T)+\|D\Phi\|_4^2\right\}\cdot N^2(0,T).$$

LEMMA 9.

$$\sup_{0\leqq t\leqq T}\|F(t)\|_2^2\leqq C\cdot\sum_{k=1}^{3}\left\{N^{2k}(0,T)+\|D\Phi\|_3^{2k}\right\}\cdot N^2(0,T).$$

LEMMA 10.

$$\int_0^T\|F(\tau)\|_3^2d\tau\leqq C\cdot\sum_{k=0}^{3}\left\{N^{2k}(0,T)+\|D\Phi\|_4^{2k}\right\}\cdot N^2(0,T).$$

LEMMA 11.

$$\int_0^T\|F_t(\tau)\|_1^2d\tau\leqq C\cdot\sum_{k=0}^{2}\left\{N^{2k}(0,T)+\|D\Phi\|_4^{2k}\right\}\cdot N^2(0,T).$$

These lemmas will be proved later. We use them in (5.1) to obtain

$$(5.2)\quad N^2(0,T)\leqq C\left\{\|h_0\|_4^2+\|u_0\|_4^2\right\}$$
$$+C\left\{N(0,T)+\sum_{k=1}^{3}\left\{N^{2k}(0,T)+\|D\Phi\|_4^{2k}\right\}\right\}\cdot N^2(0,T),$$

from which we deduce that there exist positive constants ε_1 and C_1 such that

$$N^2(0,T)\leqq C_1^2\left\{\|h_0\|_4^2+\|u_0\|_4^2\right\}\leqq C_1^2\cdot N^2(0,0)$$

whenever $N(0,T)\leqq\varepsilon_1$ and $\|D\Phi\|_4\leqq\varepsilon_1$. This completes the proof of Proposition 3.

6. Proofs of Lemmas 1–6.

Proof of Lemma 1. Integrating equation (2.1) and using the boundary condition (2.4), we find that the perturbation h satisfies

$$\int_\Omega h(x,t)\,dx-\int_\Omega h(x,0)\,dx-\int_0^t\int_\Omega\left((h+\hat{h})u^j_{x_j}\right)dx=0$$

and as

$$\int_\Omega h(x,t)\,dx=\int_\Omega\left(h_0(x)-\hat{h}(x)\right)dx=0,$$

we have

$$\bar{h}=\bar{h}(t)=\int_{\Omega} h(x,t)\,dx=0.$$

The inequality in Lemma 1 then follows from a Poincaré inequality [7, p. 157].

Proof of Lemma 2. We denote by $L^0=L^0(h,u)$ and $L^i=L^i(h,u)$ the right-hand sides of the perturbation equations (2.9) and (2.10). Then

(6.1)

$$\int_0^t \int_{\Omega} \left\{ (D^k L^0)_{x_i} \cdot D^k h_{x_i} + \frac{\bar{h}_0}{\nu} D^k L^i \cdot D^k h_{x_i} \right\} dx\, d\tau$$

$$= \int_0^t \int_{\Omega} D^k h_{x_i} \cdot \left\{ D^k\left(h_t + u^j h_{x_j} + \bar{h}_0 u^j_{x_j}\right)_{x_i} + \frac{\bar{h}_0}{\nu} D^k\left(u^i_t - \nu u^i_{x_j x_j} + h_{x_i}\right)\right\} dx\, d\tau$$

$$= \tfrac{1}{2} \int_{\Omega} \int_0^t \frac{\partial}{\partial t}\left(D^k h_{x_i}\right)^2 dt\, dx + \frac{\bar{h}_0}{\nu} \int_0^t \int_{\Omega} \left(D^k h_{x_i}\right)^2 dx\, d\tau$$

$$+ \int_0^t \int_{\Omega} D^k h_{x_i} \left\{ D^k\left(u^j h_{x_j}\right)_{x_i} + \bar{h}_0 D^k\left(u^j_{x_j x_i} - u^i_{x_j x_j}\right) + \frac{\bar{h}_0}{\nu} D^k u^i_t \right\} dx\, d\tau$$

$$= \tfrac{1}{2}\left\|D^{k+1} h(t)\right\|^2 - \tfrac{1}{2}\left\|D^{k+1} h_0\right\|^2 + \frac{\bar{h}_0}{\nu} \int_0^t \left\|D^{k+1} h(\tau)\right\|^2 d\tau + \cdots .$$

Replacing L^0 by F^0 and L^i by F^i in (6.1) and rearranging, we obtain

(6.2)

$$\tfrac{1}{2}\left\|D^{k+1} h(t)\right\|^2 + \frac{\bar{h}_0}{\nu} \int_0^t \left\|D^{k+1} h(\tau)\right\|^2 d\tau$$

$$\leqq \tfrac{1}{2}\left\|D^{k+1} h_0\right\|^2 + \int_0^t \int_{\Omega} \left|D^k h_{x_i}\right| \cdot \left\{ \left|D^k F^0_{x_i}\right| + \frac{\bar{h}_0}{\nu} \left|D^k F^i\right| + \bar{h}_0 \left|D^k\left(u^j_{x_j x_j} - u^i_{x_j x_j}\right)\right| \right\} dx\, d\tau$$

$$+ \left| \int_0^t \int_{\Omega} D^k h_{x_i} D^k\left(u^j h_{x_j}\right)_{x_i} \cdot dx\, d\tau \right| + \frac{\bar{h}_0}{\nu} \left| \int_0^t \int_{\Omega} D^k h_{x_i} D^k u^i_t dx\, d\tau \right|$$

$$\leqq \tfrac{1}{2}\left\|D^{k+1} u_0\right\|^2 + K_{\varepsilon}\left\{ \left\|D^{k+1} h(t)\right\|^2 + \int_0^t \left\|D^{k+1} h(\tau)\right\|^2 d\tau \right\}$$

$$+ \frac{K}{\varepsilon} \int_0^t \left\{ \left\|D^{k+1} F^0(\tau)\right\|^2 + \left\|D^k F^i(\tau)\right\|^2 + \left\|D^{k+2} u(\tau)\right\|^2 \right\}$$

$$+ \frac{K}{\varepsilon}\left\{ \left\|D^{k+1} h_0\right\|^2 + \left\|D^k u_0\right\|^2 + \left\|D^k u(t)\right\|^2 \right\}$$

$$+ K \int_0^t \left\{ \left\|D^{k+1} u(\tau)\right\|_1^2 + \left\|D^{k+1} F^0(\tau)\right\|^2 \right\} d\tau + K \cdot N^3(0,t),$$

where we have used Young's inequality in the first integral terms and the estimates for $k=0,1,2$ and 3

$$\left| \int_0^t \int_{\Omega} D^k h_{x_i} D^k\left(u^j h_{x_j}\right)_{x_i} dx\, d\tau \right| \leqq K \cdot N^3(0,t) \tag{6.3}$$

and

(6.4)

$$\left|\int_0^t\int_\Omega D^kh_{x_i}D^ku_t^i\,dx\,d\tau\right| \leqq \varepsilon\|D^{k+1}h(\tau)\|^2+\frac{K}{\varepsilon}\left\{\|D^{k+1}h_0\|^2+\|D^ku_0\|^2+\|D^ku(t)\|^2\right\}$$

$$+K\int_0^t\left\{\|D^ku(\tau)\|_1^2+\|D^{k+1}F^0(\tau)\|^2\right\}d\tau+K\cdot N^3(0,t).$$

Here K is a constant which is independent of time t and the particular functions, and ε is an arbitrary positive constant. By picking ε sufficiently small and rearranging (6.2) we obtain the inequality in Lemma 2. It remains to establish estimates (6.3) and (6.4).

For (6.3) we use Leibniz' rule to obtain for $k=0,1$ and 2

$$(6.5)\quad \left|\int_0^t\int_\Omega D^kh_{x_i}D^k\left(u^jh_{x_j}\right)_{x_i}dx\,d\tau\right|$$

$$\leqq\sum_{l=0}^k\binom{k}{l}\left|\int_0^t\int_\Omega\left\{D^kh_{x_i}D^{k-l}h_{x_j}D^lu^j_{x_i}+D^kh_{x_i}D^{k-l}h_{x_ix_j}D^lu^j\right\}dx\,d\tau\right|$$

$$\leqq\sum_{l=0}^k\binom{k}{l}\int_0^t\left\{\sup_x|D^{l+1}u|\left\{\|D^{k+1}h\|^2+\|D^{k-l+1}h\|^2\right\}\right.$$

$$\left.+\sup_x|D^lu|\left\{\|D^{k+1}h\|^2+\|D^{k-l+2}h\|^2\right\}\right\}d\tau$$

where we have used Young's inequality (with $\varepsilon=1$)

$$\leqq\sum_{l=0}^k\binom{k}{l}\left\{K_e\sqrt{\sup_t\|D^{l+1}u\|_2^2}\cdot\int_0^t\left\{\|D^{k+1}h\|^2+\|D^{k-l+1}h\|^2\right\}d\tau\right.$$

$$\left.+K_e\cdot\sqrt{\sup_t\|D^lu\|_2^2}\cdot\int_0^t\left\{\|D^{k+1}h\|^2+\|D^{k-l+2}h\|^2\right\}d\tau\right\}$$

$$\leqq K\cdot N(0,t)\cdot N^2(0,t).$$

where we have used the embedding of $W^{2,2}(\Omega)$ in $C^\lambda(\bar{\Omega})$.

The above argument is also valid for $k=3$ and $l=1$ or 2. For $(k,l)=(3,0)$ we must first integrate the second term in (6.5) by parts to get

$$\tfrac{1}{2}\left|\int_0^t\int_\Omega\left(D^kh_{x_i}\right)^2u^j_{x_j}\,dx\,d\tau\right|$$

from which we extract $\sup_x|Du|$, and for $(k,l)=(3,3)$ we must extract $\sup_x|Dh|$ from the first integral instead of $\sup_x|D^4u|$.

For the estimate (6.4), we first integrate by parts in t and obtain

$$\left|\int_0^t \int_\Omega D^k h_{x_i} D^k u_t^i dx\, d\tau\right|$$

$$\leqq \left|\int_\Omega D^k h_{x_i}(t) D^k u^i(t)\, dx\right| + \left|\int_\Omega D^k h_{x_i}(0) D^k u^i(0)\, dx\right| + \left|\int_0^t \int_\Omega D^k h_{x_i t} D u^i dx\, d\tau\right|$$

$$\leqq \varepsilon \left\|D^{k+1} h(t)\right\|^2 + \frac{1}{\varepsilon}\left\|D^k u(t)\right\|^2 + \left\|D^{k+1} h_0\right\|^2 + \left\|D^k u_0\right\|^2$$

$$+ \left|\int_0^t \int_\Omega D^k u^i D^k\left(F^0 - u^j h_{x_j} - \bar{h}_0 u^j_{x_j}\right)_{x_j} dx\, d\tau\right|,$$

where we have used Young's inequality twice and equation (2.9). The integral term here is bounded by

$$\int_0^t \left\{\left\|D^k u(\tau)\right\|^2 + \left\|D^{k+1} F^0(\tau)\right\|^2 + \left|\int_\Omega D^k u^i D^k\left(u^j h_{x_j}\right)_{x_i} dx\right| + \bar{h}_0 \left\|D^{k+1} u(\tau)\right\|^2\right\} d\tau$$

where we have used Young's inequality in the first term and integrated by parts in the spatial variable for the third. The middle term is bounded by a constant times $N^3(0,t)$, which is shown in exactly the same way as above after integrating by parts in the spatial variable.

This completes the proof of Lemma 2.

Proof of Lemma 3. This is proved in a similar way to Lemma 2, estimating the identity

$$\int_0^t \int_\Omega \left\{D^k L^0 D^k h + \bar{h}_0 D^k L^i D^k u^i\right\} dx\, d\tau = \int_0^t \int_\Omega \left\{D^k F^0 D^k h + \bar{h}_0 D^k F^i D^k u^i\right\} dx\, d\tau.$$

Similar caution must be taken with the $k=4$ term which is bounded by $KN^3(0,t)$. We omit the details.

Proof of Lemma 4. We write equation (2.10) as in elliptic boundary value problem

$$\nu u^i_{x_j x_j} = F^i - u^i_t - h_{x_i}$$

with homogeneous boundary condition $u^i|_{\partial\Omega} = 0$, for which the estimate [3]

$$\|u^i\|_2 \leqq K\left\|F^i - u^i_t - h_{x_i}\right\| \tag{6.6}$$

is valid. The lemma follows from (6.6).

Proof of Lemma 5.

$$\left\|D^k h_t(t)\right\|^2 = \int_\Omega D^k h_t D^k h_t dx$$

$$= \int_\Omega D^k h_t D^k\left(F^0 - \bar{h}_0 u^j_{x_j} - u^j h_{x_j}\right) dx$$

$$\leqq 2\varepsilon \left\|D^k h_t(t)\right\|^2 + \frac{1}{\varepsilon}\left\|D^k F^0(t)\right\|^2 + \frac{\bar{h}_0}{\varepsilon}\left\|D^{k+1} u(t)\right\|^2$$

$$+ \left|\int_\Omega D^k h_t D^k\left(u^j h_{x_j}\right) dx\right|$$

where the last integral is bounded by $KN^2(0,t)$ for $k = 0, 1, 2$ and 3.

This is proved in a similar way to that in the proof of Lemma 6. The first inequality in Lemma 5 follows by addition, whereas the second is proved similarly with time integration included above.

Proof of Lemma 6. We prove this by estimating the identity

$$\int_0^t \int_\Omega D^k L_t^i \cdot D^k u_t^i dx\,dt = \int_0^t \int_\Omega D^k F_t^i \cdot D^k u_t^i dx\,dt$$

for $k=0,1$ and 2. The details are similar to those in Lemma 2 and are omitted. Note that for $k=1$ and 2 the order of differentiation on F_t^i is reduced by integrating by parts before using Young's inequality.

7. Proofs of Lemmas 7–11. We now prove Lemmas 7–11, where F^0 and F^i are defined by (2.11) and (2.12) for a given solution (h,u^i) of the perturbation equations in some space $\mathscr{X}(0,T;E)$. Here E is sufficiently small that $h(x,t)+\hat{h}(x)\geqq \frac{1}{4}\bar{h}_0>0$. This with the facts that

$$h+\hat{h}\in C^0(0,T;\,W^{4,2}(\Omega))\cap L_2(0,T;\,W^{4,2}(\Omega))$$

and

$$h_t\in C^0(0,T;\,W^{3,2}(\Omega))\cap L_2(0,T;\,W^{3,2}(\Omega))$$

imply that

$$\|(h+\hat{h})^{-1}\|_k \leqq K\sum_{l=0}^{k}\left\{\|D(h+\hat{h})\|_k^{2l}\right\} \tag{7.1}$$

for $k=1,2$ and 3, and

$$\left\|\left(\frac{(h+\hat{h})_{x_j}}{(h+\hat{h})}\right)_t\right\|_1 \leqq K\|h_t\|_2^2\sum_{l=0}^{2}\left\{\|D(h+\hat{h})\|_2^{2l}\right\}, \tag{7.2}$$

where the constant K is independent of h, $\hat{h}$ and t, but depends inversely on $\bar{h}_0>0$.

Proof of Lemma 7. From (2.11)

$$\begin{aligned}\sup_{0\leqq\tau\leqq t}\|F^0(\tau)\|_3^2 &= \sup_{0\leqq\tau\leqq t}\left\|(\bar{h}_0-h-\hat{h})u^j_{x_j}-\hat{h}_{x_j}u^j\right\|_3^2\\ &\leqq K\sup_{0\leqq\tau\leqq t}\left\{\|h\|_3^2+\|\bar{h}_0-\hat{h}\|_3^2+\|D\hat{h}\|_3^2\right\}\|u\|_4^2\end{aligned}$$

by the Banach algebra property of the Sobolev space $W^{3,2}(\Omega)$ [1, Thms. 5, 23]

$$\begin{aligned}&\leqq K\left\{N^2(0,t)+\|\bar{h}_0-\hat{h}\|_3^2+\|D\hat{h}\|_3^2\right\}N^2(0,t)\\ &\leqq K\left\{N^2(0,t)+\|D\Phi\|_3\right\}N^2(0,t)\end{aligned}$$

by the definition of $\hat{h}$ and inequality (3.3).

Proof of Lemma 8.

$$\begin{aligned}\int_0^t\|F^0(\tau)\|_4^2d\tau &= \int_0^t\left\|(\bar{h}_0-\hat{h})u^j_{x_j}-hu^j_{x_j}-\hat{h}_{x_j}u^j\right\|_4^2d\tau\\ &\leqq K\int_0\left\{\|\bar{h}_0-\hat{h}\|_4^2+\|h\|_4^2+\|D\hat{h}\|_4^2\right\}\|u\|_5^2d\tau\end{aligned}$$

by the Banach algebra property of the Sobolev space $W^{4,2}(\Omega)$

$$\leqq K\int_0^t \left\{\|h\|_4^2+\|D\Phi\|_4^2\right\}\|u\|_5^2 d\tau$$

$$\leqq K \sup_{0\leqq\tau\leqq t}\left\{\|h(\tau)\|_4^2+\|D\Phi\|_4^2\right\}\int_0^t \|u(\tau)\|_5^2 d\tau$$

$$\leqq K\left\{N^2(0,t)+\|D\Phi\|_4^2\right\}N^2(0,t).$$

Proof of Lemma 9.

$$\sup_{0\leqq\tau\leqq t}\|F^i(\tau)\|_2^2\leqq \sup_{0\leqq\tau\leqq t}\left\{\left\|u^j u^i_{x_j}\right\|_2^2+\left\|\frac{(h+\hat{h})x_j}{(h+\hat{h})}u^j_{x_j}\right\|_2^2\right\}$$

$$\leqq K\sup_{0\leqq\tau\leqq t}\left\{\|u\|_2^2\|Du\|_2^2+\|(h+\hat{h})^{-1}\|_2^2\|(h+\hat{h})_{x_j}\|_2^2\|Du\|_2^2\right\}$$

by the Banach algebra property of $W^{2,2}(\Omega)$

$$\leqq K\sup_{0\leqq\tau\leqq t}\|Du\|_2^2\left\{\|u\|_2^2+\|D(h+\hat{h})\|_2^2\sum_{l=0}^{2}\left\{\|D(h+\hat{h})\|_2^{2l}\right\}\right\}$$

where we have used inequality (7.1)

$$\leqq KN^2(0,t)\sup_{0\leqq\tau\leqq t}\left\{\|u\|_2^2+\sum_{l=1}^{3}\|Dh\|_2^{2l}+\sum_{l=1}^{3}\|D\Phi\|_2^{2l}\right\}$$

$$\leqq KN^2(0,t)\cdot\sum_{l=1}^{3}\left\{N^{2l}(0,t)+\|D\Phi\|_2^{2l}\right\}.$$

The proof of Lemma 10 is similar, so is omitted.

Proof of Lemma 11. Let $R_j=(h+\hat{h})_{x_j}/(h+\hat{h})$. Then

$$\int_0^t \|F_t^i(\tau)\|_1^2 d\tau=\int_0^t \left\|\left(-u^j u^i_{x_j}+\nu R_j u^i_{x_j}\right)_t\right\|_1^2 d\tau$$

$$\leqq K\int_0^t\left\{\left\|\left(u^j u^i_{x_j}\right)_t\right\|_1^2+\left\|\left(R_j u^i_{x_j}\right)_t\right\|_1^2\right\}d\tau$$

$$\leqq K\int_0^t\left\{\left\|u^j_t u^i_{x_j}\right\|_1^2+\left\|u^j u^i_{x_j t}\right\|_1^2+\left\|R_{jt}u^i_{x_j}\right\|+\left\|R_j u^i_{x_j t}\right\|_1^2\right\}d\tau$$

$$\leqq K\int_0^t\left\{\|u\|_2^2+\|R_j\|_2^2\right\}\|u_t\|_2^2 d\tau+K\int_0^t \left\|R_{jt}u_{x_j}\right\|_1^2 d\tau$$

by the Banach algebra property of $W^{2,2}(\Omega)$

$$\leqq K \sup_{0\leqq\tau\leqq t} \|u_t(\tau)\|_2^2 \int_0^t \left\{ \|u\|_2^2 + \|k_j\|_2^2 \right\} d\tau$$

$$+ K \int_0^t \sup_x \left\{ |u|^2, |Du|^2 \right\} \left\{ \|k\|_2^2 + \|k_t\|_2^2 \right\} d\tau$$

$$\leqq KN^2(0,t) \left\{ N^2(0,t) + \int_0^t \|R_t\|_2^2 d\tau \right\}$$

$$+ K \sup_{0\leqq\tau\leqq t} \|u\|_3^2 \int_0^t \left\{ \|R\|_2^2 + \|R_t\|_2^2 \right\} d\tau$$

$$\leqq KN^2(0,t) \left\{ N^2(0,t) + \int_0^t \left\{ \|R\|_2^2 + \|R_t\|_2^2 \right\} d\tau \right\}$$

$$\leqq KN^2(0,t) \sum_{k=0}^{2} \left\{ N^{2k}(0,t) + \|D\Phi\|_4^{2k} \right\} \quad \text{by (7.1) and (7.2).}$$

This completes the proof of Lemma 11.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] G. Browning, A. Kasahara and H.-O. Kreiss, *Initialization of the primitive equations by the bounded derivative method*, J. Atmos. Sci., 37 (1980), pp. 1424–1436.

[3] M. J. P. Cullen, *The application of finite element methods to the primitive equations of fluid motion*, in Finite Elements in Water Resources, W. G. Gray and G. F. Pinder, eds., Pentech, London, 1977.

[4] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.

[5] E. N. Lorenz, *Attractor sets and quasigeostrophic equilibrium*, J. Atmos. Sci., 37 (1980), pp. 1685–1699.

[6] B. Machenhauer, *On the dynamics of gravity oscillations in a shallow water model, with application to normal mode initialization*, Beitr. Phys. Atmos., 50 (1977), pp. 253–271.

[7] A. Matsumura and T. Nishida, *The initial value problem for the equations of motion of viscous and heat-conductive gases*, J. Math. Kyoto Univ., 20 (1980), pp. 67–104.

[8] ———, *Initial boundary value problems for the equations of motion of general fluids* in Computing Methods in Applied Sciences and Engineering, V, R. Glowinski and T. L. Lions, eds., North-Holland, Amsterdam, 1982.

[9] J. Pedlosky, *Geophysical Fluid Dynamics*, Springer-Verlag, Berlin, 1979.

[10] B. A. Ton, *Existence and uniqueness of a classical solution of an initial boundary value problem of the theory of shallow waters*, this Journal, 12 (1981), pp. 229–241.

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

008

ON COHERENT GROWTH OF CONFIGURATIONS*

STEPHEN J. WILLSON†

Abstract. A configuration X is an m-tuple of subsets of Euclidean n-space. A transition rule F assigns to any X a new configuration FX. If X is bounded but sufficiently large, the sequence X, FX, $F(FX)$, $\cdots$ is studied by computing upper and lower bounds for the collection of limit points of the sequence $(F^pX)/p$. If F is coherent, these upper and lower bounds coincide with a certain convex polytope depending only on the rule f and not on X. These results may be interpreted to study patterns of growth in certain cellular automata and may be applied to study the growth of crystals.

AMS-MOS subject classifications (1980). Primary 68D20; secondary 54H20

1. Introduction. Considerable work has been done on cellular automata, homogeneous structures, and their applications to biological or physical phenomena. (See, for example, Aladyev [1] or Wolfram [13]). The idea is that one is given a "configuration" X—perhaps a pattern of zeros and ones on an infinite checkerboard, or some generalization—and a "transition rule" F which acts on X to give a new configuration FX. The basic problem is to study the iterates X, FX, $F^2X=F(FX),\cdots,F^pX,\cdots$.

In the literature can be found some very detailed analyses of particular kinds of rules F and configurations X. For example, Butler and Ntafos [3] are able to characterize those squares in an infinite checkerboard which ultimately contain a one if F is a particular rule studied by S. Ulam and X is a finite configuration. As another example, Greenberg and Hastings [4] address the question as to when a pattern will disappear ultimately—i.e., F^pX will be entirely zeros for some p—for a particular collection of rules F.

Our questions are analogous, but our approach is somewhat different. Instead of dealing with a particular F we shall only put abstract hypotheses on F and obtain approximations to F^pX.

This paper is a generalization of Willson [11]. In that paper, X was any finite "sufficiently large" configuration of ones on an n-dimensional checkerboard otherwise filled with zeros, while F was assumed "ordered". The main result asserted that there was a convex rational polytope W depending only on F so that for large p, F^pX was well approximated by placing ones at the lattice points inside the polytope pW.

The generalization in this paper is of two sorts. Primarily, the intention is to study structures where cells may have more than two states. (If I try merely to encode the other states in terms of zeros and ones, I do not get systems susceptible to the analysis in [11].) It appears that the tools of [11] generalize best when cells have 2^m states. Accordingly, we shall treat the situation where the state of each cell is given by an m-tuple of zeros and ones.

The second mode of generalization is that the notion of approximation in [11] is weakened. This weakening allows us to avoid many technicalities which unfortunately encumber [11]. Instead of approximating F^pX directly, we study $\lim X$, the collection of limit points, in Kuratowski's sense, of the sequence $(F^pX)/p$. Thus, a weakening of the

* Received by the editors March 26, 1980, and in final revised form September 15, 1983. This research was supported in part by the National Science Foundation under grant MCS-78-00687.

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

main result of [11] discussed above would assert that for all sufficiently large finite X, $\lim X = W$.

The analysis is facilitated by reinterpreting the configurations. Let n and m be positive integers. For us, a configuration X will be an m-tuple $(X_1,\cdots,X_m)$ where each X_i is a subset of R^n. Suppose we are given a configuration Y in the sense of cellular spaces—thus Y is a map from Z^n to the set of m-tuples of zeros and ones, where Z^n is the lattice of integer points in R^n. We interpret Y as a configuration X in our sense in the following way: for $i=1,\cdots,m$ let $X_i=\{v\in Z^n\colon\ Y(v)_i=1\}$, so X_i is the set of locations v where the ith coordinate of $Y(v)$ is one and not zero. Then the corresponding configuration X is $X=(X_1,X_2,\cdots,X_m)$.

Our major result is analogous to the major result in [11]: Suppose the transition rule F can be expressed in a certain kind of local form. For $i=1,\cdots,m$ we should like to compute the set of limit points of the sequence $(F^pX)_i/p$, where $/p$ indicates division by the scalar p. In the case where F is "coherent", it will turn out that the limit points will be independent of i and X unless X is degenerate. This common limiting shape will then be a convex polytope depending only on F. We call it $\lim X$.

This result has a natural physical interpretation in the context of crystal growth. Regard X as describing the state of an n-dimensional crystal in which each fundamental cell may contain m atoms; X_i describes the locations of the atoms of the ith type. The rule F corresponds to passage of time, so that FX describes the crystal one unit of time later. It is observed in real life that such crystals tend to grow toward a preferred shape, usually polyhedral, characteristic of the type of crystal. This preferred shape corresponds to $\lim X$. Whether a vacancy for one type of atom is filled at a certain moment depends on what is happening to all types; yet still the macroscopic crystal tends toward one shape common to all types of atom in that particular crystal. The author finds it intriguing that in this situation (Theorem 5.4) $\lim X$ will be a convex rational polytope just as physical crystals tend toward a rational polytope preferred shape.

The notation is simplified if we allow each X_i to be an arbitrary subset of R^n, rather than just of Z^n; we are then able to apply F to disks, for example, rather than just to sets of lattice points. For this reason most of the paper seems to apply to a slightly different situation from that of cellular spaces, as indicated above. The return to cellular spaces is made in §7, where the results are restated in that context.

This paper is organized as follows: Basic notations and definitions are presented in §2. In §3 we obtain two growth rate functions $g_l(v)$ and $g_u(v)$ which give respectively lower and upper bounds on growth rates in the direction v. We show they are continuous functions. In §4 we use g_l and g_u to construct configurations W_l and W_u, and we show that these are respectively lower and upper bounds on $\lim X$. In §5 we characterize *coherent* growth and prove that in this case $\lim X = W_l = W_u$ is a convex polytope. In §6 we show that the notion of coherence is not void by giving a simple sufficient condition for coherence. Finally in §7 we re-express our results in terms of cellular spaces.

For more details on the motivation for this paper and applications to crystal growth, the reader should consult Example 2 of §2. In particular, this example should be regarded as an extension of this introduction.

2. Notation and basic definitions.

We denote n-dimensional Euclidean space by R^n; it possesses the standard inner product denoted by $\langle x,y\rangle$ for $x, y\in R^n$, and it has norm given by $|x|=\langle x,x\rangle^{1/2}$. The set of points $(x_1,\cdots,x_n)\in R^n$ such that each x_i is an integer is denoted Z^n and is called the integer lattice of R^n. The set of unit vectors in R^n is denoted Σ^{n-1}.

DEFINITION. Throughout this paper, m and n will be fixed positive integers. A *configuration* X is an ordered m-tuple $(X_1,\cdots,X_m)$ where each X_i is a subset of R^n. The ith component may be variously denoted by $X_i=(X)_i=(X_1,\cdots,X_m)_i$. The set of all such configurations (for fixed n and m) is denoted $\mathcal{P}^n_m$, or more commonly just $\mathcal{P}$ when there is no risk of confusion. If X, $Y\in\mathcal{P}$ we write $X\subseteq Y$ provided that we have $X_i\subseteq Y_i$ for $i=1,\cdots,m$.

Heuristically, we interpret X in terms of crystals as giving the locations in R^n of atoms of m different types: If $x\in X_i$, then there is an atom of type i at position x. We allow the same $x\in R^n$ to be in several different X_i simultaneously; this is justified because in the applications x will locate a translate of the primitive cell (see Kittel [6, p. 11]), and this cell may contain many atoms.

Certain special configurations are particularly important for this paper. Given any unit vector $v\in R^n$ and any $r\in R$, we let $H(v;r)$ be the closed half-space $\{x\in R^n: \langle x,v\rangle\leqq r\}$. If $a_1,\cdots,a_m\in R$, we define the configuration $H(v;a_1,\cdots,a_m)\in\mathcal{P}$ by letting its ith component be $H(v;a_1,\cdots,a_m)_i=H(v;a_i)$ for $i=1,\cdots,m$. Thus the components of $H(v;a_1,\cdots,a_m)$ are closed half-spaces with parallel bounding hyperplanes; we shall refer to $H(v;a_1,\cdots,a_m)$ informally as a *half-space*.

If L is a positive constant, we denote the ball of radius L by $D_L=\{x\in R^n: |x|\leqq L\}$. We will also use D_L to mean the configuration $(D_L,D_L,\cdots,D_L)\in\mathcal{P}$. No confusion should arise from this abuse of notation. We will say X is *bounded* if for some L we have $X\subseteq D_L$.

If $X\in\mathcal{P}$ and $r\in R^n$, we will frequently consider the *translate* $X+r$ of X by r, defined as the configuration whose ith component is $(X+r)_i=\{x+r: x\in X_i\}$.

Our principal object of study is defined next:

DEFINITION. An *ordered transition rule* F on $\mathcal{P}$ is a map F: $\mathcal{P}\to\mathcal{P}$ such that all the following conditions (2.1) through (2.4) hold:

(2.1) F is *monotone*, in the sense that $X\subseteq FX$ for all $X\in\mathcal{P}$.

(2.2) F *preserves inclusions*, in the sense that, whenever X, $Y\in\mathcal{P}$ and $X\subseteq Y$ it follows that $FX\subseteq FY$.

(2.3) F is *invariant under translations*; i.e., for all $w\in R^n$ and for all $X\in\mathcal{P}$ we have $F(X+w)=(FX)+w$.

(2.4) If $\varnothing$ denotes the configuration each of whose components is empty, then $F\varnothing=\varnothing$.

(2.5) F is *finitely determined*. This means that there are finitely many vectors $v^1,\cdots,v^N\in R^n$ called the *neighborhood vectors*, such that for each $y\in R^n$, for each $i=1,\cdots,m$ and for each $X\in\mathcal{P}$, the fact as to whether $y\in(FX)_i$ is completely determined by which of the vectors $y+v^j$ lie in the various components X_k. Still more explicitly, for each y,i, and X the truth of the assertion "$y\in(FX)_i$" shall be determined entirely by which of the Nm statements "$y+v^j\in X_k$" are true, where $1\leqq j\leqq N$ and $1\leqq k\leqq m$.

It will be important to have explicit descriptions of ordered transition rules F, and the notation needed is unfortunately cumbersome. In this paragraph we shall show how objects which we shall call "generators" can be used to define an ordered transition rule F. Fix the allowed neighborhood vectors $v^1,\cdots,v^N$. Let $E=\{(j,k): j=1,\cdots,N; k=1,\cdots,m\}$. For each $i=1,\cdots,m$ suppose there are given finitely many nonempty subsets $S^i_1,\cdots,S^i_t$ of E, called *i-generators*. Make the further assumption that

(2.6) $$v^1=(0,0,\cdots,0)\in Z^n \quad \text{and for each } i, \quad \{(1,i)\}=S^i_1.$$

We can then define a map F: $\mathcal{P}\to\mathcal{P}$ as follows: for each $i=1,\cdots,m$, $y\in(FX)_i$ if and only if there exists an i-generator S_l^i such that $y+v^j \in X_k$ for all $(j,k) \in S_l^i$. In symbols,

$$(2.7) \qquad (FX)_i = \bigcup_l \bigcap_{(j,k)\in S_l^i} (X_k - v^j)$$

where $X_k - v^j$ denotes the subset X_k translated by the vector $-v^j$. It is trivial that (2.1) through (2.5) hold for the map F so defined; for example, (2.4) is a consequence of the assumption that the sets S_l^i are all nonempty, while (2.1) is a consequence of assumption (2.6). Thus F is an ordered transition rule.

Conversely, one may easily prove that any ordered transition rule may be described as in the previous paragraph, using generators. Explicitly, suppose F is an ordered transition rule. From (2.5) we may obtain neighborhood vectors $v^1, \cdots, v^N$ and we may ensure $v^1 = (0,0,\cdots,0)$, if necessary by inserting it into the list of neighborhood vectors.

For each $i = 1, \cdots, m$ let A be any subset of E such that $y \in (FX)_i$ is true whenever $y + v^j \in X_k$ for each $(j,k) \in A$ yet $y + v^j \notin X_k$ for each $(j,k) \in E - A$. List all such sets A for each i as $A_1^i, A_2^i, \cdots, A_t^i$; we shall see these constitute the i-generators for F. First of all, from (2.1) if $X_i = \{(0,0,\cdots,0)\}$ but $X_j = \varnothing$ for $j \neq i$ it still follows $(0,0,\cdots,0) \in (FX)_i$. Hence $A = \{(1,i)\}$ is some A_l^i which we may take to be A_1^i; and (2.6) holds.

Next, suppose $X \in \mathscr{P}$; we show that if for some l and some $y \in R^n$ we have $y + v^j \in X_k$ for all $(j,k) \in A_l^i$, then $y \in (FX)_i$. The difficulty we encounter is the possibility that $y + v^j \in X_k$ for some $(j,k) \in E - A_l^i$; if there were no such (j,k), then y would be in $(FX)_i$ by the choice of A_l^i. Let $B = \{(j,k) \in E\colon\ y + v^j \in X_k\}$, so $A_l^i \subseteq B$. Define configurations X^A and $X^B \in \mathscr{P}$ by $(X^A)_k = \{v^j\colon\ (j,k) \in A_l^i\}$, $(X^B)_k = \{v^j\colon\ (j,k) \in B\}$ for $k = 1, \cdots, m$. Then $X^A \subseteq X^B$ since $A_l^i \subseteq B$. Moreover, $(0,0,\cdots,0) \in (FX^A)_i$ by the choice of A_l^i. From (2.2) it follows that $(0,0,\cdots,0) \in (FX^B)_i$, whence by (2.5) we see that $y \in (FX)_i$.

Thus each set A_l^i is an i-generator for F, and it is clear that the list of i-generators for each i gives a complete description of F. Thus any ordered transition rule is described in terms of generators.

DEFINITION. If F is an ordered transition rule, its *neighborhood parameter M* is defined as $M = \max|v^j|$, where v^j ranges over the neighborhood vectors of F.

The neighborhood parameter will be a useful measure of how far the influence of a point may extend. It is easy to obtain the following results:

(2.8) If L is a positive real number, then $F(D_L) \subseteq D_{L+M}$.

(2.9) For each positive integer p, suppose F^p is the result of a p-fold iteration of F. Then F^p is an ordered transition rule on $\mathscr{P}$ with neighborhood parameter pM.

Example 1. Let $n = m = 2$. Define the ordered transition rule F on $\mathscr{P}$ as follows: The neighborhood vectors are $v^1 = (0,0)$, $v^2 = (-1,0)$, $v^3 = (0,-1)$ and $N = 3$ since there are three neighborhood vectors, while the neighborhood parameter $M = 1$. We choose F to have two 1-generators, namely $S_1^1 = \{(1,1)\}$ and $S_2^1 = \{(2,1)\}$. We choose three 2-generators: $S_1^2 = \{(1,2)\}$, $S_2^2 = \{(3,2)\}$, $S_3^2 = \{(1,1),(2,1),(2,2)\}$.Thus by virtue of S_3^2, we know that $y \in (FX)_2$ if $y + v^1 \in X_1$, $y + v^2 \in X_1$, and $y + v^2 \in X_2$. By virtue of S_2^2 we know $y \in (FX)_2$ if $y + v^3 \in X_2$. Suppose the initial configuration is X where $X_1 = X_2 = \{(0,0)\}$. The reader can verify the following:

$$\begin{aligned} (FX)_1 &= \{(0,0),(1,0)\}, & (FX)_2 &= \{(0,0),(0,1)\},\\ (F^2X)_1 &= \{(0,0),(1,0),(2,0)\}, & (F^2X)_2 &= \{(0,0),(0,1),(0,2),(1,0)\},\\ (F^3X)_1 &= \{(0,0),(1,0),(2,0),(3,0)\}, & (F^3X)_2 &= \{(0,0),(0,1),(0,2),\\ &&&\quad (0,3),(1,0),(1,1),(2,0)\}. \end{aligned}$$

Example 2. This example illustrates the relationship between this current paper and crystal growth and may be regarded as motivation for this paper. Let $n=3$, $m=2$. For the remainder of this section we interpret a configuration $Y=(Y_1, Y_2)\in\mathscr{P}_2^3$ as a description of a crystal. The locations of the cells of the crystal lattice will be indexed by $v\in Z^3$. We shall assume that the primitive cell (see Kittel [6, p. 11]) may contain $m=2$ atoms of different types The "cell at $v\in Z^3$" is a translate of the primitive cell, but in an incomplete crystal lattice not all atoms will be present in each cell. We interpret $Y\in\mathscr{P}$ as follows: If $v\in Y_1$, then the cell at $v\in Z^3$ contains an atom of type 1; and if $v\notin Y_1$, then the cell at v does not contain an atom of type 1. Similarly $v\in Y_2$ if and only if the cell at v contains an atom of type 2. Thus if $v\notin Y_1\cup Y_2$, the cell at v is empty.

We shall say v and v' are *adjacent* if $v-v'$ is one of the vectors $(\pm 1,0,0)$, $(0,\pm 1,0)$, $(0,0,\pm 1)$; we say two atoms are adjacent if the indices v and v' of their cells are adjacent. Let us assume that there is a binding energy e_i between adjacent atoms of type i, for $i=1,2$; a binding energy e_{12} between adjacent atoms of different type; and a binding energy e'_{12} between two atoms in the same cell. Assume e_1, e_2, e_{12}, e'_{12} are positive, and let λ_1, λ_2 be positive numbers (regarded as threshold energies).

We define an ordered transition rule F on $\mathscr{P}_2^3$ as follows: If $X=(X_1,X_2)\in\mathscr{P}_2^3$, then $v\in(FX)_1$ if and only if either $v\in X_1$ or $e_1n_1+e_{12}n_2+e'_{12}n'_{12}\geqq\lambda_1$, where n_1 is the number of $w\in Z^3$ adjacent to v such that $w\in X_1$, n_2 is the number of $w\in Z^3$ adjacent to v such that $w\in X_2$, and $n'_{12}=1$ if $v\in X_2$ while $n'_{12}=0$ if $v\notin X_2$. Thus v appears in $(FX)_1$ if it is already in X_1 or if the total energy released by the appearance of an atom of type 1 in the cell at v exceeds the threshold energy λ_1. Similarly we let $v\in(FX)_2$ if and only if $v\in X_2$ or $e_2n_2+e_{12}n_1+e'_{12}n'_{21}\geqq\lambda_2$, where n_2 and n_1 are as above and $n'_{21}=1$ if $v\in X_1$ while $n'_{21}=0$ if $v\notin X_1$.

It is not hard to see that F is an ordered transition rule. There will be numerous generators, depending on the numbers e_1, e_2, e_{12}, e'_{12}, λ_1, λ_2.

Hopefully, the transition rule F corresponds to the growth process of the physical crystal. Thus, if X describes the initial seed crystal, then F^pX describes the crystal after p units of time. With reasonable restrictions on e_1, e_2, e_{12}, e'_{12}, λ_1, λ_2 the theorems in the remainder of this paper will apply to this F. In particular, (7.2) will show why the crystal should grow into a polyhedral shape, provided that the initial seed crystal X is "sufficiently large."

The description of F above is very close to descriptions of crystal growth made in the 1920's but never exploited mathematically. (See, for example, Stranski [9].) Instead of developing those ideas, the physics literature has emphasized approaches using differential equations. Recently, some researchers have returned to discrete methods, especially computer simulation using Monte Carlo probabilistic techniques. (See Swendsen et al. [10] or Jackson [5].) In [11] the author obtained abstract theorems rather than examples of simulation, but that paper applied only to the case $m=1$. The present paper extends the methods to crystals where the fundamental cell contains more than one atom.

3. Growth rates on half-spaces. In this section we study the result of applying an ordered transition rule F to a half-space $H(v; a_1,\cdots,a_m)$. It turns out that the image of such a half-space is another half-space with different parameters $a_1,\cdots,a_m$ but the same v. This face permits us to analyze which parameters $a_1,\cdots,a_m$ lead to slow growth and which lead to fast growth. Our ultimate problem—the study of the sequence X, FX, $F^2X,\cdots$ for bounded X—will be reduced later to the special case studied in this section, where X has form $H(v; a_1,\cdots,a_m)$.

Throughout this section, F will be an ordered transition rule, and v will be a unit vector. The fundamental result is the following:

THEOREM 3.1. *The image under F of a half-space is another half-space. More specifically, if v is a unit vector and $(a_1,\cdots,a_m)\in R^m$, then there exists $(b_1,\cdots,b_m)\in R^m$ such that $FH(v;a_1,\cdots,a_m)=H(v;b_1,\cdots,b_m)$. The numbers b_i are given by*

$$b_i=\max_l \min_{(j,k)\in S_l^i}\left(a_k-\langle v,v^j\rangle\right)$$

where l indexes the different i-generators S_l^i.

Proof. By (2.7),

$$\begin{aligned}[FH(v;a_1,\cdots,a_m)]_i&=\bigcup_l \bigcap_{(j,k)\in S_l^i}\left[H(v;a_1,\cdots,a_m)_k-v^j\right]\\ &=\bigcup_l \bigcap_{(j,k)\in S_l^i}\left[H(v;a_k)-v^j\right]\\ &=\bigcup_l \bigcap_{(j,k)\in S_l^i}\left[H\left(v;a_k-\langle v,v^j\rangle\right)\right]\\ &=H\left(v;\max_l \min_{(j,k)\in S_l^i}\left(a_k-\langle v,v^j\rangle\right)\right),\end{aligned}$$

and the conclusion follows. □

If $H(v;b_1,\cdots,b_m)$ were a translate of $H(v;a_1,\cdots,a_m)$, say $H(v;b_1,\cdots,b_m)=H(v;a_1,\cdots,a_m)+cv$ for some constant c, then iteration of F would be easily described; we would have $F^pH(v;a_1,\cdots,a_m)=H(v;a_1,\cdots,a_m)+pcv$ for all natural numbers p, since F is assumed to be invariant under translation. Unfortunately, this nice state of affairs need not occur. In any event, for each v and $(a_1,\cdots,a_m)$ there is a *lower growth rate* $l(v;a_1,\cdots,a_m)$ such that

$$FH(v;a_1,\cdots,a_m)\supseteq H(v;a_1,\cdots,a_m)+l(v;a_1,\cdots,a_m)v \tag{3.2}$$

while (3.2) would be false if $l(v;a_1,\cdots,a_m)$ were replaced by any larger number. From (3.1) it is easy to see $l(v;a_1,\cdots,a_m)=\min_i(-a_i+b_i)$, or more explicitly,

$$l(v;a_1,\cdots,a_m)=\min_i\left[-a_i+\max_l \min_{(j,k)\in S_l^i}\left(a_k-\langle v,v^j\rangle\right)\right]. \tag{3.3}$$

Similarly there will always be an *upper growth rate* $u(v;a_1,\cdots,a_m)$ such that

$$FH(v;a_1,\cdots,a_m)\subseteq H(v;a_1,\cdots,a_m)+u(v;a_1,\cdots,a_m)v \tag{3.4}$$

while (3.4) would be false if $u(v;a_1,\cdots,a_m)$ were replaced by any smaller number. Explicitly, $u(v;a_1,\cdots,a_m)=\max_i(-a_i+b_i)$, or

$$u(v;a_1,\cdots,a_m)=\max_i\left[-a_i+\max_l \min_{(j,k)\in S_l^i}\left(a_k-\langle v,v^j\rangle\right)\right]. \tag{3.5}$$

It is clear that both u and l are continuous functions of v; $a_1,\cdots,a_m$. Moreover, they are invariant under translation in the sense that for any constant k,

$$(3.6)\qquad \begin{aligned} l(v;a_1-k,\cdots,a_m-k)&=l(v;a_1,\cdots,a_m) \quad\text{and}\\ u(v;a_1-k,\cdots,a_m-k)&=u(v;a_1,\cdots,a_m). \end{aligned}$$

What is the fastest guaranteed growth rate in the direction v under application of F? Since $l(v;a_1,\cdots,a_m)$ measures the minimal growth rate in direction v of $H(v;a_1,\cdots,a_m)$ by (3.2), the fastest growth we can guarantee is obtained by choosing $(a_1,\cdots,a_m)$ to make $l(v;a_1,\cdots,a_m)$ as large as possible. Similarly, the lowest guaranteed bound on the growth in direction v is found by making $u(v;a_1,\cdots,a_m)$ as small as possible. Hence we make the following definitions:

DEFINITION. The *lower growth rate in direction* v is $g_l(v)=\sup l(v;a_1,\cdots,a_m)$, and the *upper growth rate in direction* v is $g_u(v)=\inf u(v;a_1,\cdots,a_m)$, where in each case $(a_1,\cdots,a_m)$ ranges over all of R^n.

The functions g_l and g_u will be crucial to this paper. The remainder of this section will be devoted to showing that g_l and g_u are well-behaved functions on the set of unit vectors in R^n.

PROPOSITION 3.7. *For any unit vector v, $g_l(v)$ and $g_u(v)$ are well defined. If M is the neighborhood parameter for F, then $0\leqq g_l(v)\leqq g_u(v)\leqq M$.*

Proof. Since F is monotone, (2.1) and (3.4) show $u\ (v;a_1,\cdots,a_m)\geqq 0$; hence u is bounded below and g_u is well defined. Moreover,

$$g_u(v)\leqq u(v;0,0,\cdots,0)=\max_l \min_{(j,k)\in S_l^j}\left(-\langle v,v^j\rangle\right)\leqq M.$$

Thus $0\leqq g_u(v)\leqq M$.

Similarly, since F is monotone, (3.2) shows $l(v;a_1,\cdots,a_m)\geqq 0$. We shall prove $l(v;a_1,\cdots,a_m)\leqq g_u(v)$ for each $(a_1,\cdots,a_m)$. It follow that $g_l(v)$ exists and indeed the proof of the proposition will be complete, since then $0\leqq g_l(v)\leqq g_u(v)$.

To see that $l(v;a_1,\cdots,a_m)\leqq g_u(v)$, we choose any $(b_1,\cdots,b_m)\in R^m$ and show $l(v;a_1,\cdots,a_m)\leqq u(v;b_1,\cdots,b_m)$. Let $c=\max(a_i-b_i)$ where the maximum occurs when $i=j$, so $c=a_j-b_j$. Then $a_i-c\leqq b_i$ for $i=1,\cdots,m$; hence $H(v;a_1-c,\cdots,a_m-c)\subseteq H(v;b_1,\cdots,b_m)$. It now follows that $H(v;a_1-c,\cdots,a_m-c)+l(v;a_1-c,\cdots,a_m-c)v\subseteq FH(v;a_1-c,\cdots,a_m-c)$ [by (3.2)] $\subseteq FH(v;b_1,\cdots,b_m)$ [by (2.2)] $\subseteq H(v;b_1,\cdots,b_m)+u(v;b_1,\cdots,b_m)v$ [by (3.4)].

The jth component of this inclusion says $a_j-c+l(v;a_1-c,\cdots,a_m-c)\leqq b_j+u(v;b_1,\cdots,b_m)$. Since $c=a_j-b_j$, it follows $l(v;a_1-c,\cdots,a_m-c)\leqq u(v;b_1,\cdots,b_m)$. Finally, the invariance under translation (3.6) shows $l(v;a_1,\cdots,a_m)\leqq u(v;b_1,\cdots,b_m)$. The proposition follows. □

The next result shows that the parameters $(a_1,\cdots,a_m)$ may without loss of generality be restricted to a compact subset of R^m. Hence the values $g_u(v)$ and $g_l(v)$ are in fact attained.

PROPOSITION 3.8. *There exists a compact subset $K\subseteq R^m$ such that, for each unit vector $v\in R^n$ there exist $(c_1,\cdots,c_m)\in K$ and $(d_1,\cdots,d_m)\in K$ for which $g_l(v)=l(v;c_1,\cdots,c_m)$ and $g_u(v)=u(v;d_1,\cdots,d_m)$.*

Proof. Let K be the set of $(a_1,\cdots,a_m)\in R^m$ such that $|a_i|\leqq 3(m-1)M$ for each i, where M is the neighborhood parameter of F. We prove this K works by showing that if v is a unit vector of R^n and $(a_1,\cdots,a_m)\in R^m$, then there exists $(b_1,\cdots,b_m)\in K$ for which $l(v;a_1,\cdots,a_m)=l(v;b_1,\cdots,b_m)$ and $u(v;a_1,\cdots,a_m)\geqq u(v;b_1,\cdots,b_m)$. The supremum of $l(v;b_1,\cdots,b_m)$ for $(b_1,\cdots,b_m)\in K$ is attained since K is compact; from the preceding sentence, this will be the supremum for $(a_1,\cdots,a_m)\in R^m$. Similarly, the infimum of $u(v;b_1,\cdots,b_m)$ for $(b_1,\cdots,b_m)\in K$ is attained, whence this same value will be the infimum for $(a_1,\cdots,a_m)\in R^m$.

Suppose then that $(a_1,\cdots,a_m)\in R^m$. We seek $(b_1,\cdots,b_m)\in K$ so $l(v;a_1,\cdots,a_m)=l(v;b_1,\cdots,b_m)$ and $u(v;a_1,\cdots,a_m)\geqq u(v;b_1,\cdots,b_m)$. Without loss of generality we may renumber the components so $a_1\leqq a_2\leqq\cdots\leqq a_m$, and by (3.6) we may translate by a_1 and thereby assume $a_1=0$. If for every $p=1,\cdots,m-1$ we have $a_{p+1}-a_p\leqq 3M$, then $(a_1,\cdots,a_m)$ is obviously in K and we are done. If on the other hand for some p we have $a_{p+1}-a_p>3M$, let $L=a_{p+1}-a_p-3M>0$. Write $d_i=a_i$ for $i\geqq p$ and $d_i=a_i-L$ for $i\geqq p+1$. Then I claim $l(v;a_1,\cdots,a_m)=l(v;d_1,\cdots,d_m)$ and $u(v;a_1,\cdots,a_m)\geqq u(v;d_1,\cdots,d_m)$.

To see this, note that by (3.3),

$$l(v;a_1,\cdots,a_m)=\min_i\max_l\min_{(j,k)\in S_l^i}\left(-a_i+a_k-\langle v,v^j\rangle\right).$$

Suppose the extrema are realized at parameters $\hat{i},\hat{l},\hat{j},\hat{k}$. Then since $0\leqq -a_{\hat{i}}+a_{\hat{k}}-\langle v,v^{\hat{j}}\rangle\leqq M$ by (3.7), we cannot have $|-a_{\hat{i}}+a_{\hat{k}}|>3M$. Hence either (1) $\hat{i}\leqq p$ and $\hat{k}\leqq p$ or (2) $\hat{i}\geqq p+1$ and $\hat{k}\geqq p+1$. In either case $-a_{\hat{i}}+a_{\hat{k}}-\langle v,v^{\hat{j}}\rangle=-d_{\hat{i}}+d_{\hat{k}}-\langle v,v^j\rangle$. By a similar argument since $0\leqq l(v;a_1,\cdots,a_m)\leqq u(v;0,0,\cdots,0)\leqq M$ by the proof of (3.7), we see that for no parameters i,l,j,k will the values of $-a_i+a_k-\langle v,v^j\rangle$ be relevant to the computation of $l(v;a_1,\cdots,a_m)$ unless $0\leqq -a_i+a_k-\langle v,v^j\rangle\leqq M$; and for such parameters $-d_i+d_k-\langle v,v^j\rangle=-a_i+a_k-\langle v,v^j\rangle$ since either $i\leqq p$, $k\leqq p$ or else $i\geqq p+1$, $k\geqq p+1$. Hence $l(v;a_1,\cdots,a_m)=l(v;d_1,\cdots,d_m)$. The reader may supply further details.

For the case of

$$u(v;a_1,\cdots,a_m)=\max_i\max_l\min_{(j,k)\in S_l^i}\left(-a_i+a_k-\langle v,v^j\rangle\right),$$

one argues similarly. If for any parameters i,l,j,k we have $-d_i+d_k-\langle v,v^j\rangle>-a_i+a_k-\langle v,v^j\rangle$, then clearly $i\geqq p+1$ while $k\leqq p$, so $a_i>a_k+3M$ and $-a_i+a_k-\langle v,v^j\rangle<-3M-\langle v,v^j\rangle\leqq -2M$. Since $u(v;a_1,\cdots,a_m)\geqq 0$, such parameter values will not be relevant to the computation of $u(v;a_1,\cdots,a_m)$. Hence $u(v;d_1,\cdots,d_m)\leqq u(v;a_1,\cdots,a_m)$, where again the reader may provide further details.

In both cases the claim has been proved. If for every $p=1,\cdots,m-1$ we have $d_{p+1}-d_p\leqq 3M$, then $(d_1,\cdots,d_m)\in K$ and we are done. Otherwise there exists p so $d_{p+1}-d_p>3M$ and we may repeat the above argument. After finitely many repetitions we clearly reduce to the case where $d_{p+1}-d_p\leqq 3M$ for all p, and the proof is complete. □

Recall that Σ^{n-1} denotes the set of unit vectors in R^n.

PROPOSITION 3.9. *g_u and g_l are continuous functions on Σ^{n-1}.*

Proof. Note that by (3.8), there is a compact set K so $g_l(v)=\sup_{y\in K}l(v;y)$ and $g_u(v)=\inf_{y\in K}u(v;y)$. We prove (3.9) only for g_l. Suppose for some sequence v_i in Σ^{n-1}, $v_i\to v_0$; we must show $g_l(v_i)\to g_l(v_0)$. Since Σ^{n-1} is compact, we need only show that $g_l(v_0)$ is the unique limit point of the sequence $g_l(v_i)$. A choice of subsequences lets us assume $g_l(v_i)\to L$ for some L, and we must prove only that $L=g_l(v_0)$.

Since K is compact, we may choose $y_i\in K$ so $g_l(v_i)=l(v_i;y_i)$. By another choice of subsequence, using the compactness of K again, we may assume there exists $y_0\in K$ so $y_i\to y_0$. Since $l(v_i;y_i)\to l(v_0;y_0)$ by the continuity of l, $L=l(v_0;y_0)$; hence by the definition of $g_l(v_0)$, $g_l(v_0)\geqq L$.

On the other hand, pick $z \in K$ so $g_l(v_0) = l(v_0; z)$. Since $l(v_i; z) \leqq g_l(v_i) = l(v_i; y_i)$, we may let $i \to \infty$ and obtain $l(v_0; z) \leqq l(v_0; y_0) = L$, so that $g_l(v_0) \leqq L$. Hence $g_l(v_0) = L$ and the proof is complete. □

4. Growth on bounded configurations. In this section we consider a bounded configuration X and study certain limit configurations associated with the sequence X, FX, $F^2X, \cdots$. The idea is to compare X in each direction v (where v is a unit vector) with some half-space $H(v; a_1, \cdots, a_m)$. Since the behavior of the half-spaces under iteration of F is understood via §3, we can infer information about the behavior of X.

The main definition of this section is that of $\lim X$, which will now be given. Suppose $X \in \mathscr{P}$ and F is an ordered transition rule on $\mathscr{P}$. Define $\limsup X$ to be the configuration such that $(\limsup X)_i$ is the set of limit points from the sets $(F^pX)_i/p \subset R^n$ as $p \to \infty$, where $/p$ indicates division of each vector by the scalar p. More specifically, $(\limsup X)_i$ is the set of $y \in R^n$ such that for some sequence $p_k \to \infty$ there is a choice $y_k \in (F^{p_k}X)_i$ such that $y_k/p_k \to y$ as $k \to \infty$. Similarly define $\liminf X$ as the configuration such that $(\liminf X)_i$ is the set of all $y \in R^n$ such that for all sufficiently large p there exists a point $y_p \in (F^pX)_i$ such that $y_p/p \to y$ as $p \to \infty$. Clearly $\liminf X \subseteq \limsup X$. If there is equality, we call this common set $\lim X$ and say the *limiting shape* for X *exists*.

Intuitively, $\liminf X$ and $\limsup X$ describe the rough behavior of the sequence X, $FX, \cdots, F^pX, \cdots$ as $p \to \infty$. If, for example, $\lim X$ exists and $(\lim X)_i$ is an equilateral triangle of side L, then for large p, $(F^pX)_i$ is approximately an equilateral triangle of side pL. Our major objective it to tell when $\lim X$ exists and, if it does exist, what it looks like, because this summarizes much information about F^pX for large p.

The above definitions are related to the notion of Kuratowski convergence of sequences of sets ([7] or [8]). In the notation of Salinetti and Wets [8, p. 19], $(\liminf X)_i = \liminf (F^pX)_i/p$.

It is not hard to verify that both $(\liminf X)_i$ and $(\limsup X)_i$ are closed subsets of R^n for $i = 1, \cdots, m$. If X is a bounded configuration, then $\limsup X$ is compact; this is because, if M is the neighborhood parameter for F and for some L we know $X \subseteq D_L$, then $F^pX \subseteq D_{L+pM}$ by (2.2) and (2.8), whence $\limsup X \subseteq D_m$. Moreover, if X is bounded, $\liminf X$ is compact also, since $\liminf X \subseteq \limsup X$.

One can also easily see the following facts:

(4.1) $\liminf X = \liminf FX$; $\limsup X = \limsup FX$.

(4.2) If $X \subseteq Y$, then $\liminf X \subseteq \liminf Y$ and $\limsup X \subseteq \limsup Y$.

(4.3) $\liminf(X + a) = \liminf X$ and $\limsup(X + a) = \limsup X$ for $a \in R^n$.

Most of this section will be spent using the functions g_l and g_u from §3 to compute a lower bound W_l for $\liminf X$ and an upper bound W_u for $\limsup X$. We shall first define these shapes W_l and W_u and then show that they are bounds in the appropriate sense:

DEFINITION. The *upper limiting shape* $W_u \subseteq R^n$ is $W_u = \bigcap H(v; g_u(v))$, and the *lower limiting shape* W_l is $W_l = \bigcap H(v, g_l(v))$. Here the intersections are over all unit vectors v in R^n. (Recall that we may also regard W_u and W_l as elements of $\mathscr{P}^n_m$ in the standard way.)

It is clear that W_u and W_l, regarded as subsets of R^n, are convex and closed since they are the intersections of closed convex sets. By (3.7), $0 \in W_l \subseteq W_u \subseteq D_M = \bigcap H(v, M)$, so it follows that, in addition, W_l and W_u are compact and nonempty.

PROPOSITION 4.4. *Let $X\in\mathcal{P}^n_m$ be bounded. Then W_u is an upper bound on* $\limsup X$ *in the sense that* $\limsup X\subseteq W_u$.

Proof. Let $v\in R^n$ be a unit vector. By (3.8) there exists $(a_1,\cdots,a_m)\in R^m$ so $g_u(v)=u(v;a_1,\cdots,a_m)$. Since X is bounded, we may choose a constant c so $X\subseteq H(v;a_1+c,\cdots,a_m+c)$. For $p\geqq 1$ it follows $F^pX\subseteq F^pH(v;a_1+c,\cdots,a_m+c)\subseteq H(v;a_1+c,\cdots,a_m+c)+pu(v;a_1,\cdots,a_m)v$ by (3.4) and invariance under translation. Hence, if $x_p\in(F^pX)_i$, then $\langle x_p,v\rangle\leqq a_i+c+pg_u(v)$ and $\langle x_p/p,v\rangle\leqq g_u(v)+(a_i+c)/p$. It follows that any limit point y of $(F^pX/p)_i$ satisfies $\langle y,v\rangle\leqq g_u(v)$. Since v was arbitrary, this completes the proof. □

DEFINITION. If $Y\subset R^n$ and $\mu>0$, let $\operatorname{Int}_\mu Y=\{y\in R^n\colon y+D_\mu\subset Y\}$. This is a subset of the interior of Y, $\operatorname{Int} Y$; indeed $\operatorname{Int} Y=\bigcup\operatorname{Int}_\mu Y$ where the union is over all $\mu>0$. We remark that the closure of Y is denoted $\mathrm{Cl}(Y)$.

The hard work in showing W_l is a lower bound on $\liminf X$ is obtained in the following lemma:

LEMMA 4.5. *Let F be an ordered transition rule on $\mathcal{P}$. Let $\mu>0$ and assume $\operatorname{Int}_\mu W_l\neq\varnothing$. Then there exist a natural number p and a positive number J so that whenever $R\geqq J$ it follows that $F^p(D_R)\supseteq D_R+pw$ simultaneously for each $w\in\operatorname{Int}_\mu W_l$.*

Proof. Let the neighborhood parameter of F be denoted M, and let $L=3M(m-1)$. Choose a positive number t and a natural number p so large that $p\mu-L\geqq t>0$, and then define $J=p^2M^2/(2t)$. We show that these p and J work. To do this, let $w_0\in\operatorname{Int}_\mu W_l$, choose $R\geqq J$, choose i so $1\leqq i\leqq m$, and suppose $y\in D_R+pw_0$. We must show that $y\in(F^pD_R)_i$.

Since $D_R\subseteq F^p(D_R)$ by (2.1), we may assume $y\notin D_R$. Write $w_0=|w_0|v_0$, $y=|y|v$ for unit vectors v_0 and v. By the proof of (3.8) choose $(c_1,\cdots,c_m)\in R^n$ with $\max|c_j-c_k|\leqq L$ such that $g_l(v)=l(v;c_1,\cdots,c_m)$.

Iterate (3.2) p times, using invariance under translation, to obtain

$$F^pH(v;c_1,\cdots,c_m)\supseteq H(v;c_1,\cdots,c_m)+pl(v;c_1,\cdots,c_m)v.$$

By (2.9), F^p is an ordered transition rule, and we may apply (3.1) to F^p (not F) to show for each $i=1,\cdots,m$.

$$pl(v;c_1,\cdots,c_m)\leqq -c_1+\max_l\ \min_{(j,k)\in S^i_l}\left(c_k-\langle v,v^j\rangle\right),$$

where l indexes all i-generators for F^p (*not for* F). Let S^i_l be the i-generator realizing the maximum. Then for all $(j,k)\in S^i_l$ we have $pg_l(v)\leqq -c_i+c_k-\langle v,v^j\rangle\leqq L-\langle v,v^j\rangle$. Yet each such v^j is a neighborhood vector for F^p, so $|v^j|\leqq pM$ by (2.9).

Hence, it suffices to show that for all $r\in R^n$ such that $|r|\leqq pM$ and $\langle v,r\rangle\leqq L-pg_l(v)$ we have $|y+r|\leqq R$. Once this is shown, the proof of the lemma will be complete, because then $y\in(F^pD_R)_i$ since S^i_l is an i-generator for F^p, and for each $(j,k)\in S^i_l$ we have $|y+v^j|\leqq R$.

Suppose $|r|\leqq pM$ and $\langle v,r\rangle\leqq L-pg_l(v)$. Then

$$\begin{aligned}|y+r|^2&=\langle y+r,y+r\rangle=\langle y,y\rangle+2\langle y,r\rangle+\langle r,r\rangle\\&=\langle y-pw_0,y-pw_0\rangle+2p\langle y,w_0\rangle-p^2\langle w_0,w_0\rangle+2\langle y,r\rangle+\langle r,r\rangle\\&\leqq R^2+2p\langle y,w_0\rangle-p^2\langle w_0,w_0\rangle+2\langle y,r\rangle+p^2M^2\end{aligned}$$

$$\left[\text{since } |y-pw_0|\leqq R \text{ and } |r|\leqq pM\right]$$

$$\begin{aligned}
&\leqq R^2+2p\langle y,w_0\rangle+2|y|\langle v,r\rangle+p^2M^2\\
&\leqq R^2+2p\langle y,w_0\rangle+2|y|(L-pg_l(v))+p^2M^2\\
&\leqq R^2+2p\langle y,w_0\rangle+2|y|L-2|y|p\langle v,w_0\rangle-2|y|p\mu+p^2M^2\\
&\qquad\qquad\qquad\qquad [\text{since } g_l(v)\geqq\langle v,w_0\rangle+\mu \text{ since } w_0\in \operatorname{Int}_\mu W_l]\\
&=R^2+2p\langle y,w_0\rangle-2p\langle y,w_0\rangle-2|y|(p\mu-L)+p^2M^2\\
&\leqq R^2-2Rt+p^2M^2 \qquad [\text{since } |y|\geqq R \text{ and } p\mu-L\geqq t\geqq 0]\\
&\leqq R^2-p^2M^2+p^2M^2 \qquad [\text{since } 2Rt\geqq 2Jt\geqq p^2M^2]\\
&=R^2.
\end{aligned}$$

The lemma follows. □

Now we can show that W_l is a lower bound for $\liminf X$ in the sense that $W_l\subseteq \liminf X$, provided that x meets some technical requirements:

PROPOSITION 4.6. *Let F be an ordered transition rule on $\mathcal{P}$. Suppose* $\operatorname{Int} W_l\neq\varnothing$. *There exists a positive real number J such that for any configuration X for which $X\supseteq D_J$, we have* $W_l\subseteq\liminf X$.

Proof. We use (4.5) to bootstrap our way to (4.6). Since $\operatorname{Int} W_l\neq\varnothing$, there is a positive number μ so $\operatorname{Int}_\mu W_l\neq\varnothing$. Obtain p and J from (4.5) so that for $R\geqq J$, $F^pD_R\supseteq D_R+pw$ for each $w\in\operatorname{Int}_\mu W_l$. We show that this J works in (4.6). Note that $F^{2p}D_J=F^p(F^pD_j)\supseteq F^p(D_J+pw)\supseteq D_j+2pw$, and in general $F^{kp}D_J\supseteq D_J+kpw$ for $k=1,2,\cdots$. Hence $kpw\in(F^{kp}D_J)_i$ for $i=1,\cdots,m$. If a is a positive integer, write $a=q(a)p+r(a)$ where $q(a)$ and $r(a)$ are integers and $0\leqq r(a)<p$. Then $q(a)pw\in(F^{q(a)p}D_J)_i\subseteq(F^aD_J)_i$ for each i. Since $q(a)pw/a\to w$ as $a\to\infty$, it follows $w\in(\liminf D_J)_i$.

Hence, if $X\supseteq D_J$ it follows from (4.2) that $\operatorname{Int}_\mu W_l\subseteq\liminf D_J\subseteq\liminf X$.

On the other hand, for any $\varepsilon>0$ if $\operatorname{Int}_\varepsilon W_l\neq\varnothing$ we may again by (4.5) find $\hat p$ and K; so for $R\geqq K$, $F^{\hat p}D_R\geqq D_R+\hat pw$ for each $w\in\operatorname{Int}_\varepsilon W_l$. Clearly there exists q such that $D_J+qp\operatorname{Int}_\mu W_l\supseteq D_K$. Hence $F^{qp}D_J\supseteq D_K$. By a repetition of the argument in the first paragraph, $\operatorname{Int}_\varepsilon W_l\subseteq\liminf D_K$. Hence $\operatorname{Int}_\varepsilon W_l\subseteq\liminf F^{qp}D_J$ [by (4.2)] $=\liminf D_J$ [by (4.1)] $\subseteq\liminf X$ [since $D_J\subseteq X$].

Since each point of $\operatorname{Int} W_l$ is in $\operatorname{Int}_\varepsilon W_l$ for some $\varepsilon>0$, we see $\operatorname{Int} W_l\subseteq\liminf X$. Finally, since W_l is closed and convex with nonempty interior, $W_l=\operatorname{Cl}(\operatorname{Int} W_l)$; because $\liminf X$ is closed it follows that $W_l\subseteq\liminf X$. □

5. Coherent growth. The bounds found in §4 may be summarized by saying that $W_l\subseteq\liminf X\subseteq\limsup X\subseteq W_u$ for appropriate X. In general, these inclusions may be strict. This section treats the special case where all the inclusions are equalities. We say an ordered transition rule is *coherent* if $g_l(v)=g_u(v)$ for each unit vector $v\in R^n$. If F is coherent, then $W_l=W_u$ by their definitions, so that all the above inclusions are equalities. We thus obtain the following result from (4.4) and (4.6):

THEOREM 5.1. *Suppose F is a coherent transition rule, and suppose W_l has nonempty interior. There exists an $R>0$ such that whenever X is a bounded configuration containing a translate of D_R, then* $\lim X$ *exists and* $\lim X=W_l=W_u$.

We shall see in §6 that coherent rules are common. Our objective in this section is to show that if F is coherent, then W_l is particularly nice. More specifically, we prove in

(5.4) that W_l is a convex polytope. In order to obtain this result, we first need some other descriptions of coherence:

PROPOSITION 5.2. *Suppose F is an ordered transition rule and v is a unit vector. The following are equivalent:*

(i) $g_l(v)=g_u(v)$;

(ii) *there exist* $(c_1,\cdots,c_m)\in R^m$ *and* $d\in R$ *so* $FH(v;c_1,\cdots,c_m)=H(v;c_1,\cdots,c_m)+dv$;

(iii) *there exists* $(c_1,\cdots,c_m)\in R^m$ *so* $l(v;c_1,\cdots,c_m)=u(v;c_1,\cdots,c_m)$.

Note that once (5.2) is proved, it is obvious that $d=l(v;c_1,\cdots,c_m)=u(v;c_1,\cdots,c_m)=g_l(v)=g_u(v)$. Thus the result gives several descriptions of $g_l(v)$.

Proof. It is immediate that (ii) and (iii) are equivalent. If we assume (ii), then clearly $g_u(v)\leqq d$ and $g_l(v)\geqq d$ by the definitions of g_u and g_l, while $g_l(v)\leqq g_u(v)$ by (3.7). Hence $g_l(v)=g_u(v)=d$ and (i) follows. There remains to prove only that (i) implies (ii).

So suppose $g_u(v)=g_l(v)$. By (3.8) we may choose $(a_1,\cdots,a_m)$ and $(b_1,\cdots,b_m)$ so $g_l(v)=l(v;a_1,\cdots,a_m)$ and $g_u(v)=u(v;b_1,\cdots,b_m)$. Subtracting a constant k from each a_i and applying (3.6), we may assume $a_i\leqq b_i$ for $i=1,\cdots,m$. Let

$$E=\{(c_1,\cdots,c_m)\in R^m\colon a_i\leqq c_i\leqq b_i \text{ for } i=1,\cdots,m\}.$$

Define a map $f\colon E\to E$ by

$$f(c_1,\cdots,c_m)=\max_{l}\min_{(j,k)\in S_l^i}\left(c_k-\langle v,v^j\rangle\right)-g_l(v).$$

We verify that $f(c_1,\cdots,c_m)\in E$ as follows: By (3.1),

$$FH(v;c_1,\cdots,c_m)=H(v;f(c_1,\cdots,c_m))+g_l(v)v.$$

But since $H(v;a_1,\cdots,a_m)\subseteq H(v;c_1,\cdots,c_m)\subseteq H(v;b_1,\cdots,b_m)$, it follows $FH(v;a_1,\cdots,a_m)\subseteq FH(v;c_1,\cdots,c_m)\subseteq FH(v;b_1,\cdots,b_m)$. Applying (3.2) and (3.4), we obtain

$$\begin{aligned}H(v;a_1,\cdots,a_m)+g_l(v)v&\subseteq FH(v;a_1,\cdots,a_m)\subseteq H(v;f(c_1,\cdots,c_m))+g_l(v)v\\&\subseteq FH(v;b_1,\cdots,b_m)\subseteq H(v;b_1,\cdots,b_m)+g_u(v)v=H(v;b_1,\cdots,b_m)+g_l(v)v.\end{aligned}$$

Hence $f(c_1,\cdots,c_m)\in E$.

Since E is homeomorphic with a Euclidean cell, Brouwer's fixed point theorem implies f has a fixed point $(c_1,\cdots,c_m)$. For this point it follows $FH(v;c_1,\cdots,c_m)=H(v;c_1,\cdots,c_m)+g_l(v)v$. □

We can now show that, when F is coherent, the function $g_l(v)$ must have a very special form:

LEMMA 5.3. *Suppose F is an ordered transition rule on $\mathscr{P}_m^n$. Suppose $v\in S^{n-1}$ and $g_u(v)=g_l(v)$. Then there exists an integer q so $1\leqq q\leqq m$ and there exist distinct neighborhood vectors $x_1,\cdots,x_q\in R^n$ for F such that*

$$g_l(v)=\langle -(x_1+\cdots+x_q)/q,v\rangle.$$

Remark. This says that point $g_l(v)v$ lies on the sphere through the origin on the diameter $-(x_1+\cdots+x_q)/q$.

Proof. By (5.2) there exists $(c_1,\cdots,c_m)$ such that $FH(v;c_1,\cdots,c_m)=H(v;c_1,\cdots,c_m)+g_l(v)v$. By (3.1) it follows that for each $i=1,\cdots,m$,

$$g_l(v)=\max_l \min_{(j,k)\in S_l^i} \left(c_k-\langle v,v^j\rangle\right)-c_i.$$

For each i let $l(i), j(i), k(i)$ realize the extrema, so

$$(*) \qquad g_l(v)=c_{k(i)}-\langle v,v^{j(i)}\rangle-c_i.$$

By considering the iterates 1, $k(1)$, $k(k(1)),\cdots$ in the finite set $\{1,\cdots,m\}$ we see there exist $q\leqq m$ and $i_1, i_2,\cdots,i_q$ such that $k(i_1)=i_2$, $k(i_2)=i_3,\cdots,k(i_{q-1})=i_q$, $k(i_q)=i_1$. Adding the expressions (∗) obtained by replacing i successively by $i_1,\cdots,i_q$, we obtain

$$qg_l(v)=\langle v,-v^{i_1}\rangle+\langle v,-v^{i_2}\rangle+\cdots+\langle v,-v^{i_q}\rangle.$$

The result follows by writing $x_j=v^{i_j}$. □

DEFINITION. A *convex polytope* is the convex hull of finitely many vectors. A convex polytope is *rational* if each of those vectors has rational coordinates.

THEOREM 5.4. *Suppose F is a coherent transition rule on $\mathscr{P}_m^n$. Then $W_l=W_u$ is a convex polytope. If, in addition, each neighborhood vector lies in Z^n, then W_l is a convex rational polytope.*

Proof. There are only finitely many neighborhood vectors $x_i\in R^n$ and finitely many q so $1\leqq q\leqq m$. By the lemma then there are finitely many spheres through the origin so that all points $g_l(v)v$ lie on the union of these spheres. Since $g_l(v)$ is a continuous function, the result follows from Willson [12, Prop. 6.6]. □

6. A sufficient condition for coherence. In this section we show that there are many coherent transition rules.

DEFINITION. Let F be an ordered transition rule on $\mathscr{P}_m^n$. Let S be a nonempty subset of $\{1,\cdots,m\}$. Let $\mathscr{P}_{m,S}^n=\{X\in\mathscr{P}_m^n: X_i=\varnothing \text{ for all } i\notin S\}$. We say S is *invariant* under F if F restricted to $\mathscr{P}_{m,S}^n$ has range in $\mathscr{P}_{m,S}^n$; i.e., if F induces a map $\mathscr{P}_{m,S}^n\to\mathscr{P}_{m,S}^n$.

LEMMA 6.1. *The following conditions are equivalent*:

(1) *S is invariant under F.*

(2) *If $1_S\in\mathscr{P}_m^n$ is defined by $(1_S)_i=R^n$ if $i\in S$ and $(1_S)_i=\varnothing$ if $i\notin S$, then $F(1_S)=1_S$.*

(3) *If $i\notin S$, then for every i-generator S_l^i there exists $(j,k)\in S_l^i$ with $k\notin S$.*

Proof. Trivial. □

THEOREM 6.2. *Let F be an ordered transition rule on $\mathscr{P}_m^n$. Assume $\{1,\cdots,m\}$ has no proper invariant subsets. Then F is coherent.*

Remark. Since the third condition in (6.1) is easy to check, this theorem gives a practical sufficient condition for coherence. One can then construct numerous examples of coherent transition rules for any m. Note that if $m=1$, any ordered transition rule is coherent.

Proof. For v an arbitrary unit vector in R^n we show there exists $(c_1,\cdots,c_m)\in R^m$ so $FH(v;c_1,\cdots,c_m)=H(v;c_1,\cdots,c_m)+dv$ for some constant d. Coherence will then follow by (5.2). If $m=1$, the result is trivial; so we assume $m>1$.

Define f: $R^{m-1}\to R^{m-1}$ as follows: If $(c_1,\cdots,c_{m-1})\in R^{m-1}$ define $c_m=0$;

$$b_i=\max_l \min_{(j,k)\in S_l^i} \left(c_k-\langle v,v^j\rangle\right),$$

and $f(c_1,\cdots,c_m)_i=b_i-b_m$ for $i=1,\cdots,m-1$. Thus by (3.1) $FH(v;c_1,\cdots,c_{m-1},0)=H(v;b_1,\cdots,b_m)=H(v;f(c_1,\cdots,c_{m-1}),0)+b_mv$. Hence, in order to show coherence, it suffices to show f has a fixed point.

We establish notation as follows: If $(a_1,\cdots,a_m)\in R^m$ define $a_i^1=a_i$ and for $p\geqq 2$ define a_i^p by $H(v;a_1^p,\cdots,a_m^p)=F^{p-1}H(v;a_1,\cdots,a_m)$. (A formula may be obtained from (3.1).) Let $a_{i_1}=\max a_j$. Let M be the neighborhood parameter of F.

CLAIM. *For all i, $a_{i_1}-(m-1)M\leqq a_i^m\leqq a_{i_1}+(m-1)M$.*

We prove the claim by induction on a parameter r. Assume inductively we know distinct $i_1,\cdots,i_r$ so $a_{i_j}^r\geqq a_{i_1}-(r-1)M$ for $j=1,\cdots,r$; and also $a_i^r\leqq a_{i_1}+(r-1)M$ for $i=1,\cdots,m$. (The case $r=1$ is immediate.) We show if $r<m$, then we can obtain the case $(r+1)$. Since $r<m$, the set $\{i_1,\cdots,i_r\}$ is not invariant. Hence by (6.1) there exists $i_{r+1}\notin\{i_1,\cdots,i_r\}$ and an i_{r+1}-generator T with $k\in\{i_1,\cdots,i_r\}$ for every $(j,k)\in T$. Then $a_{i_{r+1}}^{r+1}\geqq\min\{a_{i_j}^r: j=1,\cdots,r\}-M$; hence by the inductive hypothesis, $a_{i_{r+1}}^{r+1}\geqq a_{i_1}-rM$. Trivially $a_{i_j}^{r+1}\geqq a_{i_j}^r\geqq a_{i_1}-(r-1)M\geqq a_{i_1}-rM$ for $j=1,\cdots,r$. And similarly $a_i^{r+1}\leqq(\sup_j a_j^r)+M\leqq a_{i_1}+rM$. Hence induction continues until $r=m$, at which point the claim is proved.

It follows immediately from the claim that for all $(a_1,\cdots,a_m)\in R^m$ and for all i and j we have $|a_i^m-a_j^m|\leqq 2(m-1)M$.

We may now complete the proof of the theorem. Let $E=\{(c_1,\cdots,c_{m-1})\in R^{m-1}: |c_i|\leqq 2(m-1)M$ for all $i\}$. If $(c_1,\cdots,c_{m-1})\in R^{m-1}$ is arbitrary, define $(a_1,\cdots,a_m)\in R^m$ by $(a_1,\cdots,a_m)=(c_1,c_2,\cdots,c_{m-1},0)$. Then $H(v;a_1^m,\cdots,a_m^m)=F^{m-1}H(v;c_1,\cdots,c_{m-1},0)=H(v;f^{m-1}(c_1,\cdots,c_{m-1}),0)+kv$ for some constant k by the definition of f. It follows that for each i, $|f^{m-1}(c_1,\cdots,c_{m-1})_i|\leqq 2(m-1)M$ and hence $f^{m-1}(R^{m-1})\subset E$. Since E is compact, by a fixed point theorem of Felix Browder [2, p. 292] we conclude f has a fixed point. The theorem follows. □

7. Configurations supported on the integer lattice. Recall that the integer lattice Z^n consists of the set of points $(x_1,\cdots,x_n)\in R^n$ such that each x_i is an integer. In this section we extend the results from §4 and §5 to the case of configurations X supported only on Z^n; i.e., we assume $X_i\subseteq Z^n$ for $i=1,\cdots,m$. The reason for making this extension is that, as in Example 2 of §2, the configurations which describe crystals are supported on Z^n. Thus, this extension is needed for applications. The difficulty we must face is that, for example, (5.1) cannot apply to configurations supported on Z^n since it is impossible for such configurations to contain a translate of D_R. It turns out, however, that the hypotheses in (5.1) can be easily modified to deal with this case.

DEFINITION. Let m and n be positive integers. An *integral configuration* X is an m-tuple $(X_1,\ldots,X_m)$ where each X_i is a subset of Z^n. We may alternatively regard X as a map from Z^n into the set $(Z/2)^m$ of m-tuples of zeros and ones, by identifying X with the map ω so $\omega(v)_i=1$ if and only if $v\in X_i$.

A transition rule F on $\mathscr{P}_m^n$ is called *integral* if each neighborhood vector lies in Z^n. An ordered integral transition rule F may be regarded as a special type of global transformation rule for a cellular automaton or homogeneous space. (See Aladyev [1] or Wolfram [13].)

If $K\subset R^n$, then $K\cap Z^n$ denotes the lattice points inside K; and if $X=(X_1,\cdots,X_m)\in\mathscr{P}_m^n$, then $X\cap Z^n=(X_1\cap Z^n,\cdots,X_m\cap Z^n)$.

The major results obtained in this paper apply, with minor changes to the context of integral configurations:

THEOREM 7.1. *Let F be an integral ordered transition rule on $\mathscr{P}_m^n$, and suppose* $\operatorname{Int} W_l\neq 0$. *There exists a positive real number R such that any integral configuration X containing all the integer lattice points of some translate of D_R satisfies* $W_l\subseteq\liminf X$.

Note that the hypothesis means that we assume there exists $v\in R^n$ so $(D_R+v)\cap Z^n\subseteq X_i$ for $i=1,\cdots,m$.

Proof. Suppose $\operatorname{Int}_\mu W_l \neq \varnothing$ for some $\mu > 0$. Choose R to be the larger of the two numbers $\sqrt{n}/2$ and the number J obtained in the proof of (4.5). As in the proof of (4.6), the theorem will follow once we show $\operatorname{Int}_\mu W_l \subseteq \liminf((D_R + v) \cap Z^n)$.

Suppose $w \in \operatorname{Int}_\mu W_l$. By (4.5) there exists p so $F^p(D_R + v) \supseteq D_R + pw + v$. Since F is integral, it is clear the integer lattice points in $F^p(D_R + v)$ are present only by virtue of the lattice points in $D_R + v$; i.e., $F^p((D_R + v) \cap Z^n) \supseteq (D_R + pw + v) \cap Z^n$. As in the proof of (4.6), for each q, $F^{qp}((D_R + v) \cap Z^n) \supseteq (D_R + qpw + v) \cap Z^n$.. Writing $a = q(a)p + r(a)$ as in (4.6),

$$(D_R + q(a)pw + v) \cap Z^n \subseteq F^a((D_R + v) \cap Z^n).$$

Finally, since $R \geqq \sqrt{n}/2$, some lattice point z_a lies in $(D_R + q(a)pw + v) \cap Z^n$, whence $\lim_{a \to \infty} z_a/a = w \in \liminf((D_R + v) \cap Z^n)$. □

THEOREM 7.2. *Let F be a coherent integral transition rule on $\mathcal{P}_m^n$. Suppose W_l has a nonempty interior. There exists an $R > 0$ such that whenever X is a bounded configuration containing all integer lattice points of some translate of D_R, then* $\lim X$ *exists and* $\lim X = W_l = W_u$. *Moreover* $\lim X$ *is a convex rational polytope.*

Proof. Analogous to (5.1). For the second assertion, see (5.4). □

Acknowledgment. The author is indebted to the referees for suggesting many clarifications and simplifications in the original version of this paper.

REFERENCES

[1] V. ALADYEV, *Survey of research in the theory of homogeneous structures and their applications*, Math. Biosc., 22 (1974), pp. 121–154.

[2] F. E. BROWDER, *On a generalization of the Schauder fixed point theorem*, Duke Math. J., 26 (1959), pp. 291–303.

[3] J. T. BUTLER AND S. C. NTAFOS, *The vector string descriptor as a tool in the analysis of cellular automata systems*, Math. Biosc., 35 (1977), pp. 55–84.

[4] J. M. GREENBERG AND S. P. HASTINGS, *Spatial patterns for discrete models of diffusion in excitable media*, SIAM J. Appl. Math., 34 (1978), pp. 515–523.

[5] K. A. JACKSON, *The present state of the theory of crystal growth from the melt*, Journal of Crystal Growth, 24/25 (1974), pp. 130–136.

[6] C. KITTEL, *Introduction to Solid State Physics*, 4th ed., Wiley, New York, 1971.

[7] C. KURATOWSKI, *Topologie*, I, (2nd ed.), Monografie Matematyczne vol. 20, Warsaw, 1948, pp. 241–250.

[8] C. SALINETTI AND R. J. B. WETS, *On the convergence of sequences of convex sets in finite dimensions*, SIAM Rev., 21 (1979), pp. 18–33.

[9] I. STRANSKI, *Zur Theorie des Kristallwachstums*, Z. für Physikal. Chemie, Leipzig, 136 (1928), pp. 259–278.

[10] R. H. SWENDSEN, P. J. KORTMAN, D. P. LANDAU AND H. MÜLLER-KRUMBHAAR, *Spiral growth of crystals; simulations on a stochastic model*, J. Crystal Growth, 35 (1976), pp. 73–78.

[11] S. J. WILLSON, *On convergence of configurations*, Discrete Math., 23 (1978), pp. 279–300.

[12] ———, *A semigroup on the space of compact convex sets*, this Journal, 11 (1980), pp. 448–457.

[13] S. WOLFRAM, *Statistical mechanics of cellular automata*, Rev. Modern Physics, 55 (1983), pp. 601–644.

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

LIÑÁN'S PROBLEM FROM COMBUSTION THEORY, PART II*

S. P. HASTINGS[†] AND A. B. POORE[‡]

Abstract. Liñán's problem is a boundary value problem which governs the thin reaction-diffusion zone in many diverse problems in combustion and chemical reactor theory. Mathematically, the problem as it arises from matching in activation energy asymptotics is that of the existence of a unique positive solution of $u''=\frac{1}{2}u\exp(\alpha x-u)$, $u'(-\infty)=-1$ and $u'(+\infty)=0$. In earlier work we established the existence and uniqueness for positive α, and in the current work for $-\frac{1}{2}<\alpha\leqq 0$. Also we show nonexistence for $\alpha\leqq -\frac{1}{2}$.

1. Introduction. With the advent and success of activation energy asymptotics as an effective analytical technique for dealing with the Arrhenius rate function in combustion and chemical reactor theory [1], [6], a host of mathematical problems has arisen. One such problem that has a demonstrated permanence is Liñán's problem which first appeared in Liñán's paper in 1974 on counterflow diffusion flames [7]. Since that time, this problem and minor variations have been found to govern the thin reaction-diffusion zone in such diverse problems as the burning of monopropellant drops [9], [10], detonations and fast deflagration waves [8], the flame-front region problem [2], and the nonadiabatic tubular chemical reactor problem [6]. No doubt this Liñán's problem will continue to arise as a problem which is fundamental to the success of matching in activation energy asymptotics for a wide variety of combustion phenomena.

The canonical form of Liñán's problem [8] is that of the existence of a unique positive solution of

$$(1)\qquad u''(x)=\tfrac{1}{2}u(x)\exp(\alpha x-u(x)),\qquad u'(-\infty)=-1,\quad u'(+\infty)=0$$

where all values of the real parameter α are of interest. In our previous work [5], we established the existence of a unique positive solution of

$$(2)\qquad u''=\tfrac{1}{2}u\exp(\alpha x-u),\quad u'(-\infty)=-\theta,\quad u'(+\infty)=0$$

for all positive α and θ. (The inclusion of a positive θ for positive α was motivated by the work of Bush and Fink [2] on the flame-front-region problem.) Also, it was shown that a condition equivalent to $u'(+\infty)=0$ in (2) is $u(+\infty)=0$ for positive α.

In our present work we show that problem (1) has a unique positive solution for $-\frac{1}{2}<\alpha\leqq 0$. Nonexistence of a solution of (1) for $-1<\alpha\leqq-\frac{1}{2}$ was established by Ludford, Yannitell and Buckmaster [11]. We also establish nonexistence for $\alpha\leqq -1$, thereby completing our analysis of this problem.

There is a difference in both the physics and the mathematical analysis of Liñán's problem for positive and negative α. For example, in the burning of a monopropellant drop as treated by Ludford, Yannitell and Buckmaster [9], [10], a positive α corresponds to a heat gain from the atmosphere and negative α to a heat loss, the latter being more physically realistic. Also, in chemical reactor theory [6], α is negative.

* Received by the editors June 21, 1983. This research was partially supported by the National Science Foundation under grant MCS8101891.

[†] Department of Mathematics, SUNY at Buffalo, Buffalo, New York 14222.

[‡] Department of Mathematics, Colorado State University, Fort Collins, Colorado 80523.

In our previous analysis of Liñán's problem [5] for positive α we used the monotonicity of $u\exp(\alpha x-u)$ for $u\leqq 1$ and the fact that $u(+\infty)=0$. For negative α, $u(+\infty)>0$ and is in fact greater than one for α close to $-\frac{1}{2}$. This necessitates a more involved analysis. Finally, we observe that when $\alpha=0$, the only value of θ for which (2) has a positive solution is $\theta=1$ as in problem (1). (This can be seen by integrating the equation in (2).)

The paper is divided into three additional sections. In §2 we establish the existence of at least one positive solution for $-\frac{1}{2}<\alpha\leqq 0$ using a topological "shooting" method. Uniqueness is contained in §3 and nonexistence for $\alpha\leqq -\frac{1}{2}$ in §4.

2. Existence.

2. Existence. The main objective of this section is to establish Theorem 1.

THEOREM 1. *The boundary value problem*

$$u''=\tfrac{1}{2}ue^{\alpha x-u}, \tag{3}$$

$$u'(-\infty)=-1, \qquad u'(+\infty)=0 \tag{4}$$

has at least one positive solution for each α *in* $(-\frac{1}{2},0]$.

For $\alpha=0$ the proof can be accomplished by integrating the equations and using a phase plane argument. Thus we consider only the case in which α is in $(-\frac{1}{2},0)$. The proof is based on a topological "shooting" method in which we consider the equation (3) together with the initial conditions

$$u(0)=c, \qquad u'(0)=d \tag{5}$$

where (c,d) lies in the strip S of the plane $\mathbb{R}^2$ defined by

$$S=\{(c,d):c>0,\ -1<d<0\}.$$

Let u denote the solution of (3) and (5) and define four subsets A,B,C, and D of S by

$$\begin{aligned}
A&=\{(c,d)\in S\text{: there is an } x>0 \text{ such that } u'(x)>0\},\\
B&=\{(c,d)\in S\text{: there is an } x>0 \text{ such that } u(x)<0\},\\
C&=\{(c,d)\in S\text{: there is an } x<0 \text{ such that } u'(x)<-1\},\\
D&=\{(c,d)\in S\text{: } -1<u'(x)<0 \text{ for all } x<0 \text{ and } \lim_{x\to-\infty}u'(x)>-1\}.
\end{aligned}$$

Since any (c,d) which lies in $S-(A\cup B\cup C\cup D)$ corresponds to a solution of (3) and (4), our goal is to show that $A\cup B\cup C\cup D\neq S$. The basis of our proof is Lemma 1.

LEMMA 1 [11, p. 112]. *Let M and N be open subsets of the plane with connected components $M_1\subset M$ and $N_1\subset N$ such that $M_1\cap N_1$ is disconnected. Then $M\cup N$ is not homeomorphic to $\mathbb{R}^2$.*

However, a corollary that is more directly suited to the present setting is Lemma 2.

LEMMA 2 [4]. *Suppose A,B,C and D are open subsets of a set $S\subset\mathbb{R}^2$ which is homeomorphic to $\mathbb{R}^2$, with $A\cap B$ and $C\cap D$ both empty. Assume further that these sets have connected components $A_1\subset A$, $B_1\subset B$, $C_1\subset C$ and $D_1\subset D$ such that $A_1\cap C_1$, $A_1\cap D_1$, $B_1\cap C_1$ and $B_1\cap D_1$ are all nonempty. Then $A\cup B\cup C\cup D\neq S$.*

It is convenient in discussing C and D to let $y=-x$, $v(y)=u(x)$ and $\beta=-\alpha$. Then v satisfies

$$v''(y)=\tfrac{1}{2}v(y)e^{\beta y-v(y)}, \tag{6}$$

$$v(0)=c, \qquad v'(0)=-d. \tag{7}$$

Assuming for the moment that A, B, C and D are nonempty, we first establish Lemma 3.

LEMMA 3. *The sets A, B, C and D are open subsets of S and $A\cap B$ and $C\cap D$ are empty.*

Proof. The sets A, B and C are open by continuity with respect to initial data. To show that D is open, we consider the solution v of (6) and (7) with $(c,d)\in D$. Then $v'(+\infty)<1$ by the definition of D. If $v'(+\infty)\leqq\beta$, then $v'(y)\leqq\beta$ for all $y\geqq 0$ so that $\beta y - v\geqq -c$ and $v''\geqq\frac{1}{2}ve^{-c}$ which implies $v'(+\infty)=+\infty$, a contradiction. Thus $\beta< v'(+\infty)<1$.

Next, let v_0 be the solution of (6) with the initial conditions $v_0(0)=c_0$ and $v_0'(0)= -d_0$ with $(c_0,d_0)\in D$. Let $\gamma=v_0'(+\infty)$ and choose y_1 so large that

$$\int_{y_1}^{\infty} e^{-(\gamma+\beta)/2}(c+y)e^{-(\gamma-\beta)(y-y_1)/2}dy<1-\gamma, \tag{8}$$

$$v_0(y_1)>\frac{(\gamma+\beta)}{2}y_1, \tag{9}$$

$$v_0'(y_1)>\frac{(\gamma+\beta)}{2} \tag{10}$$

for all c in a neighborhood of c_0. If (c,d) is sufficiently close to (c_0,d_0), then

$$\frac{(\gamma+\beta)}{2}y_1<v(y_1)<c+y_1, \tag{11}$$

$$v'(y_1)<\gamma, \tag{12}$$

$$v'(y)>\frac{(\gamma+\beta)}{2}\quad \text{if}\, y\geqq y_1. \tag{13}$$

It then follows from (8) and (11)–(13) that $v'(+\infty)<1$, which shows that D is open.

Finally, that $A\cap B$ and $C\cap D$ are empty follows easily from the definition of the sets A, B, C and D, and the observation that $uu''>0$ if $u\neq 0$. Q.E.D.

Next we establish in Lemma 4 the existence of continuous functions which define the subsets A_1, B_1, C_1 and D_1 of A, B, C and D, respectively.

LEMMA 4. *There exist continuous functions $d_1(\cdot)$, $c_2(\cdot)$, $d_3(\cdot)$ and $c_4(\cdot)$ with the following properties*:

i) d_1: $[0,\infty)\to(-1,0)$, *and if* $c>0$ *and* $0>d>d_1(c)$, *then* $(c,d)\in A$.

ii) c_2: $[-1,0)\to(0,\infty)$ *and if* $-1<d<0$ *and* $c>c_2(d)$, *then* $(c,d)\in B$.

iii) d_3: $[0,\infty)\to(-1,0]$, $d_3(c)=0$ *for sufficiently small* $c>0$, *and if* $c>0$ *and* $-1<d<d_3(c)$, *then* $(c,d)\in C$.

iv) c_4: $(-1,0]\to[0,\infty)$, *and if* $c>c_4(d)$, *then* $(c,d)\in D$.

In particular, the sets A, B, C and D are nonempty.

For part (i) we first observe that $u\geqq c+dx\geqq c/2$ for x on $[0,\varepsilon]$ where $\varepsilon=\min(1, -c/2d)$. Then for x on $[0,\varepsilon]$, $u''=\frac{1}{2}ue^{\alpha x-u}\leqq(e^{-1}/2)e^{\alpha x}\leqq 1$ implies $u\leqq c+dx+x^2/2\leqq c+\varepsilon^2/2$. Next, $u''=\frac{1}{2}ue^{\alpha x-u}\geqq(c/4)e^{-c+\alpha\varepsilon-\varepsilon^2/2}$ implies $u'(\varepsilon)-d\geqq\varepsilon(c/4)\exp(-c+\alpha\varepsilon-\varepsilon^2/2)$. If we choose $d_1(c)=-\varepsilon(c/4)\exp(-c+\alpha\varepsilon-\varepsilon^2/2)$, $d_1(c)$ has the desired properties, since $d>d_1(c)$ implies $u'(\varepsilon)\geqq d-d_1(c)>0$.

To establish part (ii), we first need some estimates. For a given $d\in(-1,0)$, pick an $x_0>0$ such that

$$\int_{x_0}^{\infty} e^{\alpha x-1}dx=2k \quad \text{for some}\, k<-\frac{d}{2} \tag{14}$$

and a positive ε so that

$$(15)\qquad \varepsilon\int_0^{x_0} e^{\alpha x}\,dx < -\frac{d}{2}.$$

Further choose $u_1>0$ so large that $\frac{1}{2}ue^{-u}<\varepsilon$ if $u\geqq u_1$. Finally choose $c>u_1-dx_0$ and observe that $x_0<(c-u_1)/(-d)<-c/d$ and $u(x)\geqq c+dx>0$ on $[0,x_0]$. Then, since $u'(x)\geqq d$ on $[0,x_0]$, $u(x)\geqq u_1$ on $[0,x_0]$. Hence, by (12), $u'(x_0)<d/2$. Also $u''(x)\leqq e^{\alpha x-1}/2$ and so by (14), $u'(x)<d/2+k<0$ on $[0,\infty)$. Therefore, $u(x)$ must eventually become negative so that $(c,d)\in B$ whenever $c>c_2(d):=u_1-dx_0$.

For parts (iii) and (iv) we turn to equation (6). To verify there is a function d_3 as described in (iii) we first construct d_3 for large c and then show how it can be modified for small c in such a way that $d_3(c)=0$ for small c.

Let $H(v)=(1+v)e^{-v}$. If $-1<d<0$, $c>0$ and

$$(16)\qquad d^2+H(c)>1,$$

then we claim that $(c,d)\in C$. To see this, we first observe that $v''=\frac{1}{2}ve^{\beta y-v}>\frac{1}{2}ve^{-v}$ for $y>0$ implies that $(v')^2+H(v)$ is increasing with respect to y. If $v(0)=c>0$ and $v'(0)=-d>0$, the differential equation implies that $v(+\infty)=+\infty$ so that $(v'(+\infty))^2\geqq d^2+H(c)>1$, as desired. We define for the moment $d_3(c)=-\sqrt{1-H(c)}$. Observe that $d_3(0)=0$. Next, we modify this d_3 for small c.

We now show that if $c>0$ and $-d>0$ are sufficiently small, then $(c,d)\in C$. First pick a y_0 sufficiently large so that

$$\int_{y_0}^{y_0+1/2} e^{\beta y-1}\,dy>1.$$

By continuity with respect to initial data one can then choose $\varepsilon>0$ and $\delta>0$ sufficiently small so that if $0<c\leqq\varepsilon$ and $0<-d\leqq\delta$, then $0<v(y)<\frac{1}{2}$ on $[0,y_0]$. Thus there is a $y_1>y_0$ for which $v(y_1)=\frac{1}{2}$. If $v'(y)\leqq 1$ for all $y\geqq 0$, then $\frac{1}{2}\leqq v\leqq 1$ on $[y_1,y_1+\frac{1}{2}]$ and

$$v'(y_1+\tfrac{1}{2})\geqq v'(y_1)+\int_{y_1}^{y_1+1/2} e^{\beta y-1}\,dy>1,$$

a contradiction. Thus if $0<c\leqq\varepsilon$ and $0<-d\leqq\delta$, then $(c,d)\in C$. Thus we may choose $0<c_1<c_2\leqq\varepsilon$ and m so that

$$d_3(c)=\begin{cases}0, & 0\leqq c\leqq c_1,\\ m(c-c_1), & c_1\leqq c\leqq c_2,\\ -\sqrt{1-H(c)}, & c\geqq c_2,\end{cases}$$

is continuous and the line segment $m(c-c_1)$ lies in $0\leqq c\leqq\varepsilon$, $0\leqq -d\leqq\delta$ for $c_1\leqq c\leqq c_2$.

Finally, we come to (iv). We relegate to the Appendix (Lemma A.1) the fact that $v'(+\infty)>\beta$ and exists as a finite number. Now, we multiply the differential equation for v in (6) by v', integrate from 0 to y and let $y\to+\infty$ to obtain

$$(17)\quad (v'(+\infty)+v'(0)-2\beta)(v'(+\infty)-v'(0))=(1+v(0))e^{-v(0)}+\beta\int_0^\infty e^{\beta t-v}\,dt.$$

(We make the observation that since the right-hand side is positive and $v'(+\infty)-v'(0)>0$, $v'(+\infty)+v'(0)>2\beta$. If $v'(+\infty)\leqq 1$ and $v'(0)\in(0,1)$ is arbitrary, then we see that $\beta<\frac{1}{2}$ is required.) Our objective now is to show that $0<\beta<\frac{1}{2}$ and $v(0)$ sufficiently large imply $v'(+\infty)<1$, but first we need to estimate

$$(18)\qquad \int_0^\infty e^{\beta y-v}dy=2\int_0^\infty \frac{v''}{v}dy$$
$$=2\left(\frac{-v'(0)}{v(0)}+\int_0^\infty\left(\frac{v'}{v}\right)^2dy\right)$$
$$<2\left(\frac{-v'(0)}{v(0)}+(v'(+\infty))^2\int_0^\infty\frac{1}{(v(0)+v'(0)y)^2}dy\right)$$
$$<-2\frac{v'(0)}{v(0)}+2\frac{(v'(+\infty))^2}{v(0)v'(0)}.$$

The combination of (17) and (18) yields

$$(19)\qquad \left(1-\frac{2}{v(0)v'(0)}\right)(v'(+\infty))^2-\left(1-\frac{2}{v(0)v'(0)}\right)(v'(0))^2$$
$$-2\beta(v'(+\infty)-v'(0))-(1+v(0))e^{-v(0)}<0.$$

From this equation we have

$$(20)\qquad v'(0)<(\beta+\sqrt{D})\Big/\left(1-\frac{2}{v(0)v'(0)}\right)$$

where

$$D=\beta^2-\left(1-\frac{2}{v(0)v'(0)}\right)\left(2\beta v'(0)-(v'(0))^2-(1+v(0))e^{-v(0)}+\frac{2v'(0)}{v(0)}\right).$$

The right-hand side of (20) can be shown to be positive and less than one if $v_0>c_4(d)$ where

$$c_4(d)=\max\left\{\frac{-2}{(1-\beta)d},v_1,\frac{8[(1-\beta)-d(d+\beta)]}{d\left[(d+\beta)^2-(1-\beta)^2\right]}\right\}$$

and

$$(1+v_1)e^{-v_1}=\frac{(1-\beta)^2-(d+\beta)^2}{2}.$$

This completes the proof of Lemma 4.

The four subsets A_1, B_1, C_1 and D_1 are defined by

$$A_1=\{(c,d)\in A\colon d_1(c)<d<0,0<c<\infty\},$$
$$B_1=\{(c,d)\in B\colon c>c_2(d),\ -1<d<0\},$$
$$C_1=\{(c,d)\in C\colon -1<d<d_3(c),0<c<\infty\},$$
$$D_1=\{(c,d)\in D\colon c>c_4(d),\ -1<d<0\}.$$

Figure 1 depicts the conclusion of Lemma 4, the sets A_1, B_1, C_1 and D_1 and the completion of the proof of Theorem 1.

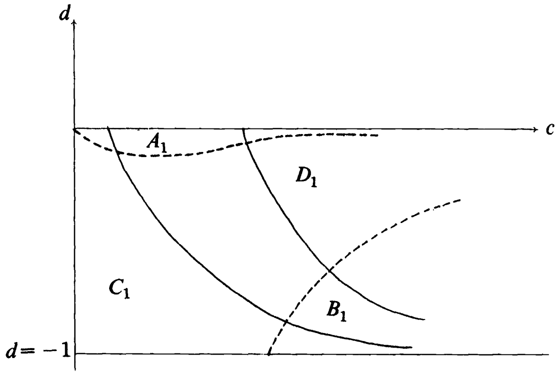


FIG. 1.

3. Uniqueness. In this section we establish the uniqueness of the solution to problem (3) (4) in Theorem 2.

THEOREM 2. *If $\beta>0$, then the problem*

$$v''=\tfrac{1}{2}ve^{\beta y-v}, \tag{21}$$

$$v'(-\infty)=0, \qquad v'(+\infty)=1 \tag{22}$$

has at most one solution.

The case $\beta=0$ can be treated by simply integrating the equations. We concentrate on $\beta>0$. The following technical lemmas are required for the proof.

LEMMA 5. *For each $v_0 \geqq 0$ there is a unique solution $v(y,v_0)$ of (21) such that*

$$v(-\infty)=v_0, \qquad v'(-\infty)=0. \tag{23}$$

This solution exists on $(-\infty,\infty)$, and $v(y,v_0)$ is continuous in v_0 for $v_0\geqq 0$ and all y. Also, $v(y,0)=0$.

LEMMA 6. *The derivative $\partial v(y,v_0)/\partial v_0 := z(y,v_0)$ exists for each y and each $v_0>0$, and satisfies the equation*

$$z''=f'(v(y,v_0))e^{\beta y}z \tag{24}$$

where $f(v)=\frac{1}{2}ve^{-v}$ and $z''=\partial^2 z/\partial y^2$. Also,

$$z(-\infty,v_0)=1 \quad and \quad z'(-\infty,v_0)=0. \tag{25}$$

LEMMA 7. *For each $v_0>0$ the limits $v'(+\infty,v_0)$ and $z(+\infty,v_0)$ exist. Also, both limits are continuous in v_0, $v'(+\infty,v_0)>\beta$, and $z'(+\infty,v_0)=(d/dv_0)v'(+\infty,v_0)$.*

The proofs of these lemmas involve straightforward techniques and are outlined in the Appendix. Theorem 2 is proved by showing that $(d/dv_0)v'(+\infty,v_0)\neq 0$ for each $v_0>0$. Fix v_0 and let $v(y)=v(y,v_0)$ and $z(y)=z(y,v_0)$. By Lemma 3 we must show that $z'(+\infty)\neq 0$.

Suppose, on the contrary, that $z'(+\infty)=0$. To obtain a contradiction, we first differentiate (21) to get

$$v'''=f'(v(y))e^{\beta y}v'+\beta e^{\beta y}f(v(y)). \tag{26}$$

Now multiply (26) by z and (24) by v' and subtract to obtain

$$v'''z-v'z''=\beta e^{\beta y}f(v(y))z. \tag{27}$$

Since $v'>0$, $f'(v)>0$ for $0<v<1$, and $f'(v)<0$ if $v>1$, equations (24) and (25) show that one of the following must hold:

a) $v_0 \geqq 1$ and $z'(y)<0$ for all y;

b) $v_0 \geqq 1$ and $z''(y)=0$ for at least one y;

c) $0<v_0<1$ and $z'(y)>0$ for all y;

d) $0<v_0<1$ and $z'(y)=0$ for at least one y.

We now proceed to show that each of these cases (with $z'(\infty)=0$) is impossible, thereby establishing the contradiction. If (a) is the case, then there must be a y_0 such that $z>0$ and $z''<0$ on $(-\infty, y_0)$ while $z<0$ and $z''>0$ on (y_0,∞). Also,

$$v''(\pm\infty)=z'(+\infty)=v'(-\infty)=0, \qquad 0<v'(+\infty)<\infty. \tag{28}$$

Next, we need Lemma 8.

LEMMA 8. *The limit $z(+\infty)$ exists, and neither z nor z' oscillates infinitely often on* $(-\infty,\infty)$.

Proof. From Lemma 3 it follows that there is a $\delta>0$ such that

$$|f'(v)e^{\beta y}| \leqq e^{-\delta y} \quad \text{for large } y. \tag{29}$$

The result follows from (24). Q.E.D.

We can now integrate (27) from $-\infty$ to $+\infty$, getting a convergent integral, and by integrating by parts on the left and using (28) we find that

$$\int_{-\infty}^{\infty} f(v(y))e^{\beta y}z(y)\,dy=0. \tag{30}$$

On the other hand, since $z'(\pm\infty)=0$, it follows from (24) that

$$\int_{-\infty}^{\infty} f'(v(y))e^{\beta y}z(y)\,dy=0. \tag{31}$$

Since $f(v)+f'(v)=\frac{1}{2}e^{-v}$, the addition of (30) and (31) yields

$$\int_{-\infty}^{\infty} e^{\beta y-v(y)}z(y)\,dy=0. \tag{32}$$

However, $v>0$ and $v'>0$ on $(-\infty,\infty)$, so that the definition of y_0 and (30) show that

$$\begin{aligned} 0=\int_{-\infty}^{\infty} ve^{\beta y-v}z\,dy &= \int_{-\infty}^{y_0} ve^{\beta y-v}z\,dy+\int_{y_0}^{\infty} ve^{\beta y-v}z\,dy \\ &< v(y_0)\int_{-\infty}^{y_0} e^{\beta y-v}z\,dy+v(y_0)\int_{y_0}^{\infty} e^{\beta y-v}z\,dy \\ &= v(y_0)\int_{-\infty}^{\infty} e^{\beta y-v}z\,dy=0. \end{aligned}$$

This contradiction eliminates possibility (a).

To discuss (b), we let y_1 be the largest zero of z', which exists by Lemma 8. Either (i) $z'>0$ on (y_1,∞) or (ii) $z'<0$ on (y_1,∞). Consider case (i). The assumption that $v_0>1$ implies $z'<0$ near $-\infty$. It follows that $z<0$ on some interval (y_1,y_2) and $z>0$ on (y_2,∞). As before we arrive at (27) and integration by parts yields

$$(zv''-v'z')\big|_{y_1}^{\infty}=\int_{y_1}^{\infty} \beta f(v)e^{\beta y}z\,dy. \tag{33}$$

This time $z(+\infty)>0$, $v''(+\infty)=0$, $v'(+\infty)=1$ and $z'(+\infty)=0$ while $v''(y_1)>0$, $z(y_1)<0$, $v'(y_1)>0$ and $z'(y_1)=0$, so that

$$\int_{y_1}^{\infty} f(v)e^{\beta y}z\,dy>0.$$

Since $z'(y_1)=z'(+\infty)=0$, it again follows from (24) that $\int_{y_1}^{\infty}f'(v)e^{\beta y}z\,dy=0$. Combining these as before, we obtain $0<\int_{y_1}^{\infty}e^{\beta y-v}z\,dy$, while

$$0=\int_{y_1}^{\infty} ve^{\beta y-v}z\,dy>v(y_2)\int_{y_1}^{\infty} e^{\beta y-v}z\,dy,$$

again a contradiction. The case (b) with $z'<0$ on (y_1,∞) is similar.

If for case (c), $z'>0$ on $(-\infty,\infty)$, then one again obtains (27) and (28) which leads to $\int_{-\infty}^{\infty}f(v)e^{\beta y}z\,dy=0$. However, in this situation $z>0$ on $(-\infty,\infty)$, which shows that (c) cannot occur.

If case (d) holds, then there is a y where $z'(y)=0$, $z(y)>1$ and $z''(y)<0$, so that $f'(v)<0$ and $v(y)>1$. The proof then proceeds as for case (b).

Since none of (a)–(d) can occur, we reach the desired contradiction and the completion of the proof of our uniqueness theorem.

4. Nonexistence. The results of Theorems 1 and 2 combined with our previous work [5] show that for each $\alpha>-\frac{1}{2}$, Liñán's problem as posed in (1) has a unique positive solution. The nonexistence for $-1<\alpha\leqq-\frac{1}{2}$ is due to Ludford, Yannitell and Buckmaster [10]. One can multiply the equation (1) through by $u'(x)$, integrate from $-\infty$ to $+\infty$, and arrive at a contradiction in the sign of the sides of the resulting equation for $-1<\alpha\leqq-\frac{1}{2}$. The same type of argument is used in the proof of part (iv) of Lemma 4 (see the remark following equation (17)).

To treat the case $\alpha\leqq-1$, we prefer to work with the equivalent problem (4) where $\beta=-\alpha$. Suppose there is a positive solution v of equation (4) with $v'(+\infty)=1$ and $v'(-\infty)=0$ for $\beta\geqq1$. Then

$$\begin{aligned} v''=\tfrac{1}{2}ve^{\beta y-v}&\geqq\tfrac{1}{2}(v(0)+v'(0)y)e^{\beta y-v(0)-y}\\ &\geqq\tfrac{1}{2}(v(0)+v'(0)y)e^{-v(0)}\end{aligned}$$

so that

$$v'(y)-v'(0)\geqq\int_0^y \tfrac{1}{2}(v(0)+v'(0)s)e^{-v(0)}ds,$$

which tends to $+\infty$ as $y\to\mathrm{r}+\infty$. This, of course, contradicts $\lim_{y\to+\infty}v'(y)=1$. Thus there is no positive solution v for $\beta\geqq1$ or no positive solution u of (1) for $\alpha\leqq-1$.

Appendix. We give here the proof of sketches of some of the facts and lemmas used in the text.

LEMMA A.1. *Given the initial value problem*

$$v''=\tfrac{1}{2}ve^{\beta y-v},\quad v(0)=c,\quad v'(0)=-d,$$

where $c>0$ *and* $-1<d<0$, $\lim_{y\to+\infty}dv/dy$ *exists as a finite number and is greater than* β.

Proof. Since v'' is positive, v' is increasing on the maximal interval of existence for $y\geqq0$ with $v'(0)=-d>0$. If $-d\geqq\beta$, then $v'(y)>\beta$ for $y>0$. Suppose $-d<\beta$ and $v'(y)\leqq\beta$ for all $y\geqq0$ for which v exists. Then the maximal interval of existence is $[0,\infty)$

and

$$\begin{aligned} v'(y) &= v'(0) + \int_0^y \tfrac{1}{2} v e^{\beta s - v}\, ds \\ &\geqq v'(0) + \int_0^y \tfrac{1}{2}(c - ds)\exp(\beta s - (c + \beta s))\, ds \\ &= v'(0) + \int_0^y \tfrac{1}{2}(c - ds) e^{-c} ds, \end{aligned}$$

which for large y contradicts $v'(y) \leqq \beta$ for all $y \geqq 0$. Thus there is $y_1 > 0$ which $v'(y_1) > \beta$, and this is valid, of course, whether $-d < \beta$ or $-d \geqq \beta$. Then we have

$$\begin{aligned} \text{(A.1)} \qquad v'(y) &= v'(y_1) + \tfrac{1}{2}\int_{y_1}^y v(s) e^{\beta s - v(s)} ds \\ &\leqq v'(y_1) + \tfrac{1}{2}\int_{y_1}^y [v(y_1) + v'(s)(s - y_1)] \\ &\qquad \cdot \exp[\beta s - (v'(y_1)s - y_1) + v(y_1)]\, ds \end{aligned}$$

where we have used $v(s) \leqq v(y_1) + v'(s)(s - y_1)$ and $v(s) \geqq v(y_1) + v'(y_1)(s - y_1)$ for $s \geqq y_1$. Equation (A.1) can be written as

$$v'(y) \leqq \alpha(y) + \int_{y_1}^y \gamma(s) v'(s)\, ds$$

where

$$\alpha(y) = v'(y_1) + \frac{v(y_1)}{2}\int_{y_1}^y \exp[(\beta - v'(y_1))s + v'(y_1)y_1 - v(y_1)]\, ds,$$

$$\gamma(s) = \frac{(s - y_1)}{2}\exp[(\beta - v'(y_1))s + v'(y_1)y_1 - v(y_1)].$$

Gronwall's inequality [3] may be applied to yield

$$v'(y) \leqq \alpha(y) + \int_{y_1}^y \gamma(s)\alpha(s)\exp\left(\int_s^y \gamma(u)\, du\right) ds.$$

Since $\beta - v'(y_1) < 0$, the right-hand side is bounded as $y \to +\infty$ so that $v'(y)$ is bounded above for all $y \geqq 0$. Thus the maximal interval of existence is $[0, \infty)$ and the conclusions of the lemma follow. Q.E.D.

Proof of Lemma 5. Consider the integral equation

$$\text{(A.2)} \qquad v(y) = v_0 + \int_{-\infty}^y (y - s) f(v(s)) e^{\beta s} ds,$$

which can be solved by successive approximations via $v^0(y) \equiv v_0$ and

$$v^{n+1}(y) = v_0 + \int_{-\infty}^y (y - s) f(v^n(s)) e^{\beta s} ds.$$

Since $f(v) \leqq f(1)$ for all $v \geqq 0$, it is easily seen that each v^n is well defined on $(-\infty, \infty)$.

Let $L = \max_{v \geqq 0} |f'(v)|$. Choose y_0 so that $Le^{\beta y_0}/4\beta^2 \leqq \frac{1}{2}$. Then by induction

$$|v^{n+1}(y) - v^n(y)| \leqq f(v_0) e^{\beta y}/2^n \beta^2$$

for $-\infty < y < y_0$ and $n = 0, 1, 2, \cdots$. The existence of a unique solution of problem (21), (23) follows routinely. This solution obviously exists on $[-\infty, y_0]$, and existence for all y follows because $f(v)$ is bounded on $0 \leqq v < \infty$. It is also clear that $v \equiv 0$ if $v_0 = 0$, while if $v_0 > 0$, then v, v' and v'' are positive, and we have the estimate

$$|v(y,v_0)-v_0| \leqq 2f(v_0)e^{\beta y}/\beta^2 \tag{A.3}$$

for $y \leqq y_0$. Q.E.D.

Proof of Lemma 6. From (A.2) it is seen that for small $|h|$,

$$\frac{v(y,y_0+h)-v(y,y_0)}{h} = 1 + \int_{-\infty}^{y} (y-s)e^{\beta s}\frac{f(v(s,v_0+h))-f(v(s,v_0))}{h}ds.$$

Since f, f', f'' are bounded on $0 \leqq v < \infty$, the proof can be completed by standard analysis, with the help of (A.3). Q.E.D.

Proof of Lemma 7. As in the proof of Lemma A.1 there is a $y_1 > 0$ such that $v'(y) \geqq \beta + 2\delta$ for some $\delta > 0$ and all $y \geqq y_1$. This implies $|f(v)e^{\beta y}| \leqq e^{-\delta y}$ for y sufficiently large, which leads to the conclusion that $v'(+\infty)$ exists. The remaining conclusions of Lemma 7 are easily established using the equation $v'(y, v_0) = \int_{-\infty}^{y} f(v(s, v_0))e^{\beta s} ds$ and the estimate (29). Q.E.D.

REFERENCES

[1] J. BUCKMASTER AND G. S. S. LUDFORD, *Theory of Laminar Flames*, Cambridge Univ. Press, Cambridge, 1982.

[2] W. B. BUSH AND S. F. FINK, *Planar premixed-flame end-wall interaction: the jump conditions across the thin flame*, Quart. Appl. Math., 38 (1981), pp. 427–438.

[3] J. K. HALE, *Ordinary Differential Equations*, Wiley-Interscience, New York, 1969.

[4] S. P. HASTINGS, *An existence theorem for a problem from boundary layer analysis*, Arch. Rat. Mech. Anal., 33 (1969), pp. 103–109.

[5] S. P. HASTINGS AND A. B. POORE, *A nonlinear problem arising from combustion theory: Liñán's problem*, this Journal, 14 (1983), pp. 425–430.

[6] A. K. KAPILA AND A. B. POORE, *The steady response of a nonadiabatic tubular reactor: new multiplicities*, Chem. Engrg. Sci., 37 (1982), pp. 57–68.

[7] A. LIÑÁN, *The asymptotic structure of counterflow diffusion flames for large activation energy*, Acta Astronaut., 1 (1974), pp. 1007–1039.

[8] G. S. S. LUDFORD AND D. S. STEWART, *Mathematical questions from combustion theory*, Trans. of the Twenty-Sixth Conf. of Army Mathematicians, USARO Report 81-1, January, 1981, pp. 53–66.

[9] G. S. S. LUDFORD, D. W. YANNITELL AND J. D. BUCKMASTER, *The decomposition of a hot monopropellant in an inert atmosphere*, Combustion Sci. Tech., 14 (1976), pp. 125–131.

[10] ———, *The decomposition of a cold monopropellant in an inert atmosphere*, Combustion Sci. and Tech., 14 (1976), pp. 133–145.

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

010

EXTREMAL PROBLEMS FOR EIGENVALUES WITH APPLICATIONS TO BUCKLING, VIBRATION AND SLOSHING*

DAVID C. BARNES†

Abstract. Let λ_n denote the nth eigenvalue of the equation $[R(x)y']'+[\lambda P(x)+Q(x)]y=0$ subject to self-adjoint boundary conditions. Many applications of this equation involve calculating extremal values of λ_n when the coefficients are subjected to some kind of additional constraints. For example the shape of the strongest column can be determined by maximizing an eigenvalue λ_n.

In this work we will give a new method for solving extremal problems for λ_n which unifies many previous works and provides (in some cases for the first time) a mathematically rigorous approach to these extremal properties. As an example of our method we determine the shape of the strongest "Profile" column. Our method can also be used to study fourth (and higher) order equations.

We reduce the problem of maximizing or minimizing λ_n to an elementary problem of minimizing or maximizing a real valued function of one real variable. One salient feature of this work is that it is completely independent of the theory of Rayleigh quotients.

1. Statement of the problem.

Consider the eigenvalue problem

$$[R(x)y']'+[\lambda P(x)+Q(x)]y=0, \qquad 0\leqq x\leqq l \tag{1.1}$$

where we assume self-adjoint boundary conditions are prescribed at $x=0$ and l. We will allow the possibility that $R(x)=0$ at $x=0$ or l which gives a singular eigenvalue problem.

Let us now suppose that each coefficient function R, P, Q depends on some other function, say $\rho(x)$, so that

$$R(x)=f(x,\rho(x)), \quad P(x)=g(x,\rho(x)), \quad Q(x)=q(x,\rho(x))$$

and (1.1) becomes

$$[f(x,\rho(x))y']'+[\lambda g(x,\rho(x))+q(x,\rho(x))]y=0. \tag{1.2}$$

We assume that $\rho(x)$ is piecewise continuous and that $h\leqq\rho(x)\leqq H$. Suppose also that $f(x,\rho)$, $g(x,\rho)$ and $q(x,\rho)$ are piecewise continuous on $[0,l]\times[h,H]$ and that $f(x,\rho(x))$, $g(x,\rho(x))$ are positive for $0<x<l$. The eigenvalues of (1.2) are then real valued functionals of $\rho(x)$ and we denote them accordingly by $\lambda_n(\rho)$. We will now consider some examples of (1.2).

The buckling problem for a slender column leads to the equation (see [11], [12])

$$y''+\lambda[\rho(x)]^{-m}y=0. \tag{1.3}$$

The interesting values of m are 1, 2 and 3. Assuming the column has given mass leads to a condition on $\rho(x)$ of the form

$$\int_0^l \rho(x)\,dx=V. \tag{1.4}$$

*Received by the editors January 14, 1983, and in revised form August 10, 1983.

†Associate Professor of Mathematics, Washington State University, Pullman, Washington 99164-2930.

The characteristic frequencies of a vibrating string are determined by the equation

$$(1.5) \qquad y''+\lambda\rho(x)y=0.$$

Assuming the string has given mass leads to a condition of the form (1.4).

Troesch [14] has studied the "sloshing" frequencies of liquid in certain containers using the equation

$$[x\rho(x)y']'+[\lambda x-m^2\rho(x)x^{-1}]y=0$$

where m is a given constant. Assuming a given container volume leads to a condition on $\rho(x)$ of the form

$$\int_0^l x\rho(x)\,dx=V.$$

We will also consider fourth order problems of the general form

$$-[r(x,\rho(x))y'']''+[f(x,\rho(x)y)y']'+[\lambda g(x,\rho(x))+q(x,\rho(x))]y=0.$$

Niordson [10] has studied the vibrations of a tapered beam using the equation

$$-[\rho^2(x)y'']''+\lambda\rho(x)y=0.$$

Assuming a beam has given mass leads to a condition on $\rho(x)$ of the form (1.4).

In these examples it is of considerable interest to maximize or to minimize $\lambda_n(\rho)$ over all functions $\rho(x)$ satisfying certain constraints. For example Keller [6] found the shape of the strongest column, pinned at each end, by maximizing $\lambda_n(\rho)$ over all $\rho(x)\geqq 0$ satisfying (1.4). Later E. R. Barnes [3] showed how to solve the same problem with the additional constraint $\rho(x)\geqq h>0$.

To generalize these examples consider a class $\mathfrak{C}$ of admissible functions $\rho(x)$ defined by the following conditions:

1. Constants h, H are given and

$$(1.6) \qquad h\leqq\rho(x)\leqq H \quad \forall x\in[0,l].$$

We allow the possibility that $H=+\infty$ or $h=-\infty$.

2. A finite constant V is given and $w(x)\geqq 0$ is a given weight function and

$$(1.7) \qquad \int_0^l w(x)\rho(x)\,dx=V.$$

The constants h, H and V satisfy

$$(1.8) \qquad h\int_0^l w(x)\,dx<V<H\int_0^l w(x)\,dx.$$

3. All of the functions $\rho(x)$, $f(x,\rho(x))$, $g(x,\rho(x))$ and $q(x,\rho(x))$ are piecewise continuous and if $x\in(0,l)$ then

$$f(x,\rho(x))>0, \qquad g(x,\rho(x))>0.$$

4. The eigenvalue problem is self-adjoint so

$$(1.9) \qquad f(x,\rho(x))[yz'-zy']_{x=0}^{x=l}=0$$

for all functions y, z which satisfy the boundary conditions.

The major purpose of this work is to provide a mathematically rigorous way of solving some problems of the following type:

Problem I. Find a function $\rho_n^+ \in \mathfrak{C}$ for which

$$\lambda_n(\rho) \leqq \lambda_n(\rho_n^+) \quad \forall \rho \in \mathfrak{C}.$$

Problem II. Find a function $\rho_n^- \in \mathfrak{C}$ for which

$$\lambda_n(\rho) \geqq \lambda_n(\rho_n^-) \quad \forall \rho \in \mathfrak{C}.$$

E. R. Barnes [3] has studied a problem similar to our problem I in the special case $n=1$. However that work considered a dual kind of problem in which the differential equation was $y''+\lambda P(x)y=0$ and constraints of the form

$$h^* \leqq P(x) \leqq H^*, \qquad \int_0^l F(x,P(x))\,dx = V^*$$

were used. If we take $P(x)=g(x,\rho(x))$ and assume that we can solve this equation for $\rho(x)$ to get $\rho(x)=F(x,P(x))$ then we see that [3, problem 1.2] is equivalent to ours in this special case.

It is common practice to assume the existence of the extremal functions $\rho_n^\pm(x)$. For the most part we must also assume this. However in a few special cases (see [1], [4], [9]) it is possible to prove $\rho_n^\pm(x)$ exists. More work needs to be done on this existence question.

2. The basic method. In this section we will introduce our method and give some ways of solving problems I and II for the simple equation

$$y''+\lambda g(x,\rho(x))y=0 \tag{2.1}$$

where the boundary conditions are of the form

$$B_1(y)=\alpha_1 y(0)+\alpha_2 y'(0)+\alpha_3 y(l)+\alpha_4 y'(l)=0, \tag{2.2}$$

$$B_2(y)=\beta_1 y(0)+\beta_2 y'(0)+\beta_3 y(l)+\beta_4 y'(l)=0. \tag{2.3}$$

Such a problem is self-adjoint if and only if the constants α_i, β_i satisfy $\alpha_2\beta_4-\alpha_4\beta_2=\alpha_1\beta_3-\alpha_3\beta_1$.

One of the classical problems in calculus of variations is that of finding extremal values for functionals of the form $\int_0^l F(x,y,y')\,dx$. One way to do this is to use the Euler equation $\partial F/\partial y-(d/dx)(\partial F/\partial y')=0$. We will now develop an analogous procedure for finding extremals of the functionals $\lambda_n(\rho)$. Equations (2.5), (2.6) or (2.8), (2.9) below can be thought of as the analogue of the Euler equation.

We will need to impose some convexity conditions on $g(x,\rho)$. We say that $g(x,\rho)$ is convex in ρ if $g_{\rho\rho}(x,\rho)\geqq 0$ and $g(x,\rho)$ is concave in ρ if $g_{\rho\rho}(x,\rho)\leqq 0$.

THEOREM I. *Let $\lambda_n(\rho)$ denote the nth eigenvalue of* (2.1), (2.2), (2.3) *and suppose there exists some $\rho^* \in \mathfrak{C}$ for which $\lambda_n(\rho^*)>0$. Suppose also that there exists a function $\rho_n^+(x) \in \mathfrak{C}$ which solves problem* I *so*

$$\lambda_n(\rho) \leqq \lambda_n(\rho_n^+) \quad \forall \rho \in \mathfrak{C}.$$

Let $g(x,\rho)$ be convex in ρ, $(g_{\rho\rho} \geqq 0)$.

Then there exists a constant μ such that for each $x \in [0,l]$ the minimum over all $\rho \in [h,H]$ of the function $\mathfrak{B}(x,\rho)$ defined by

$$\mathfrak{B}(x,\rho)=y_n^2 g(x,\rho)+\mu w(x)\rho \tag{2.4}$$

is attained when $\rho=\rho_n^+(x)$. *That is,*

$$\min_{h\leqq\rho\leqq H}\mathfrak{B}(x,\rho)=\mathfrak{B}(x,\rho_n^+(x)). \tag{2.5}$$

Here $y_n=y_n(x)$ *is an nth eigenfunction corresponding to* $\rho_n^+(x)$, *so*

$$y_n''+\lambda_n(\rho_n^+)g(x,\rho_n^+(x))y_n=0, \tag{2.6}$$

$$B_1(y_n)=B_2(y_n)=0. \tag{2.7}$$

THEOREM II. *Let* $\lambda_n(\rho)$ *denote the nth eigenvalue of* (2.1), (2.2), (2.3) *and suppose there exists a function* $\rho_n^-(x)\in\mathfrak{C}$ *which solves problem* II *so*

$$\lambda_n(\rho)\geqq\lambda_n(\rho_n^-)\quad\forall\rho\in\mathfrak{C}.$$

Suppose also that $\lambda_n(\rho_n^-)>0$. *Let* $g(x,\rho)$ *be concave in* ρ *so that* $g_{\rho\rho}\leqq 0$.

Then there exists a constant μ *such that for each* $x\in[0,l]$ *the maximum over all* $\rho\in[h,H]$ *of the function* $\mathfrak{B}(x,\rho)$ *defined by*

$$\mathfrak{B}(x,\rho)=y_n^2g(x,\rho)+\mu w(x)\rho$$

is attained when $\rho=\rho_n^-(x)$. *That is*

$$\max_{h\leqq\rho\leqq H}\mathfrak{B}(x,\rho)=\mathfrak{B}(x,\rho_n^-(x)). \tag{2.8}$$

Here $y_n=y_n(x)$ *is an nth eigenfunction corresponding to* $\rho_n^-(x)$, *so*

$$y_n''+\lambda_n(\rho_n^-)g(x,\rho_n^-(x))y_n=0, \tag{2.9}$$

$$B_1(y_n)=B_2(y_n)=0. \tag{2.10}$$

For a proof of these theorems see §4 below. We will first make a few remarks about them and give some examples.

In order to use Theorem I to calculate a maximal function $\rho_n^+(x)$ we may, in principle anyway, proceed as follows. First pick $x\in[0,l]$ and then treating y_n and μ as unknown parameters we solve the elementary minimum problem (2.5). This gives a function, ρ_n^+, which depends on x as well as the parameters y_n and μ. Call it $\rho_n^+(x,y_n,\mu)$. Substituting this function into (2.6) we obtain a (generally nonlinear) differential equation for y_n,

$$y_n''+\lambda_n(\rho_n^+)g(x,\rho_n^+(x,y_n,\mu))y_n=0.$$

This equation, together with the boundary conditions and the constraint

$$\int_0^l w(x)\rho_n^+(x,y_n,\mu)\,dx=V,$$

are then solved to find y_n, $\lambda_n(\rho_n^+)$ and $\rho_n^+(x)$. The eigenfunction y_n of (2.6) could be normalized in any way which is convenient. Multiplying y_n by a constant C simply multiplies the constant μ in (2.4) by C^2.

If these equations have a unique solution it must by Theorem I be the extremal function. Even if the solution is not unique the equations severely restrict the form $\rho_n^+(x)$ may take.

Of course similar remarks apply to Theorem II and $\rho_n^-(x)$. We will now look at some applications.

3. Some applications. As usually formulated [2], [6], [11], [13] the buckling problem for a slender column leads to the equation

$$y''+\lambda[\rho(x)]^{-2}y=0, \tag{3.1}$$

where various boundary conditions may be used:

Pinned at $x=0$ and $x=l$,

$$y(0)=y(l)=0. \tag{3.2}$$

Clamped at $x=0$, pinned at $x=l$,

$$y(l)=0, \qquad ly'(0)+y(0)=0. \tag{3.3}$$

Clamped at $x=0$, free at $x=l$,

$$y'(0)=0, \qquad y(l)=0. \tag{3.4}$$

Clamped at $x=0$ and $x=l$,

$$y'(0)-y'(0)=0, \qquad ly'(0)-y(l)+y(0)=0. \tag{3.5}$$

They are all self-adjoint boundary conditions. The critical buckling load is proportional to the smallest nonzero eigenvalue.

Comparing (2.1) and (3.1) we see $g(x,\rho)=\rho^{-2}$ which is a convex function of ρ. Thus we may use Theorem I to maximize $\lambda_n(\rho)$ and find the shape of the strongest column for any given set of boundary conditions.

We will first reproduce some results of Keller [6] by taking $h=0$, $H=+\infty$ and using boundary conditions (3.2). Theorem I requires that we minimize $\mathfrak{B}(x,\rho)=y_n^2\rho^{-2}+\mu\rho$ over $\rho\geqq 0$.

Setting $\partial\mathfrak{B}/\partial\rho=0$ gives $2y_n^2\rho^{-3}=\mu$. Normalizing the eigenfunction appropriately yields $y_n^2=(\rho_n^+)^3$. This optimality condition was used by Keller [6] and by Tadjbakhsh and Keller [12] to solve for $\rho_n^+(x)$ using the boundary conditions (3.2)–(3.5). Assuming the existence of ρ_n^+, Theorem I provides a rigorous proof the optimality of those solutions which is valid for all boundary conditions. In this regard see [2, p. 176].

As a further example of our method we will now determine the shape of the strongest "Profile" column which is pinned at each end. A profile column is one which is formed from a flat slab of material having constant thickness say T. We assume that only the width, call it $\rho(x)$, varies. Let $h>0$ be a given constant and suppose also that $\rho(x)$, h and T satisfy

$$\rho(x)\geqq h\geqq T \quad \forall x\in[0,l].$$

Since we assume the width $\rho(x)$ is always larger than the thickness T when such a column buckles it will buckle in a plane parallel to its axis but perpendicular to its flat side (the z-x plane in Fig. 1 below). Actually if h/T is large then the "Column" might be called a "Plate".

Since we have a profile column the governing equation (see [11, p. 136]) becomes

$$y''+\lambda[\rho(x)]^{-1}y=0, \qquad y(0)=y(l)=0. \tag{3.6}$$

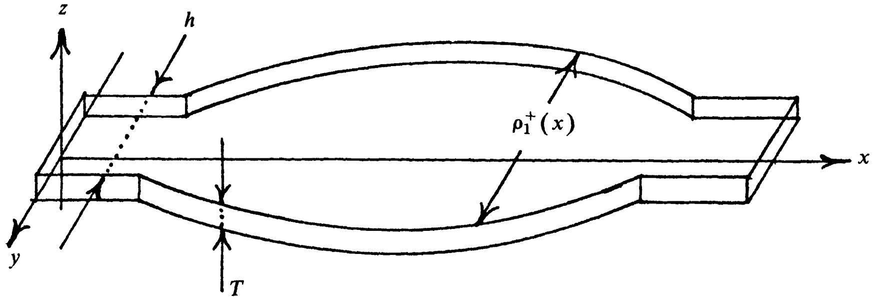


FIG. 1. *The shape of the strongest profile column having both ends pinned and constrained by* $\rho(x) \geqq h \geqq T$. *The curved parts are parabolic arches and* $\rho_1^+(x)$ *is defined by* (3.9).

Assuming a given volume V we see

$$\int_0^l \rho(x)\,dx = VT^{-1}. \tag{3.7}$$

Thus we take $w(x)=1$ and $g(x,\rho)=\rho^{-1}$ which is a convex function of ρ. Theorem I requires that we minimize $\mathfrak{B}(x,\rho)=y_1^2\rho^{-1}+\mu\rho$ over $\rho \geqq h$.

It is easy to see that if y_1^2 is small (as it is when $x \simeq 0$ or $x \simeq l$) then the minimum of $\mathfrak{B}$ occurs when $\rho = h$. However as x increases from 0 or decreases from l there will come points at which the minimum of $\mathfrak{B}$ occurs when $\partial\mathfrak{B}/\partial\rho = 0$ which implies $y_1^2 = \mu[\rho_1^+]^2$. Assuming, as we may, that the eigenfunction y_1 is positive and properly normalized we obtain $\rho_1^+ = y_1$ for values of x near the center of the column and $\rho_1^+ = h$ for values of x near the ends of the column. Substituting $\rho_1^+ = y_1$ into (3.6) shows that y_1 must be a quadratic polynomial near the center. Using (3.2) we see that y_1 must be of the general form

$$y_1 = \begin{cases} C_1 \sin(\lambda h)^{1/2} x, & 0 \leqq x \leqq s, \\ h + h\alpha(x-s)(t-x), & s \leqq x \leqq t, \\ C_2 \sin(\lambda h)^{1/2}(l-x), & t \leqq x \leqq l, \end{cases} \tag{3.8}$$

where C_1, C_2, α, s and t are parameters to be determined. The function $\rho_1^+(x)$ must be of the form

$$\rho_1^+(x) = \begin{cases} h, & 0 \leqq x \leqq s, \\ h + h\alpha(x-s)(t-x), & s \leqq x \leqq t, \\ h, & t \leqq x \leqq l. \end{cases} \tag{3.9}$$

Using (3.9) and (3.7) we find

$$\alpha = 6[VT^{-1}h^{-1} - l](t-s)^{-3}. \tag{3.10}$$

Now y_1 must be continuous at $x=s$ and $x=t$. This implies $C_1^{-1} = h\sin(\lambda h)^{1/2}s$ and $C_2^{-1} = h\sin(\lambda h)^{1/2}(l-t)$. Furthermore y_1' must also be continuous at $x=s$ and at $x=t$ which yields the equations

$$\begin{aligned} &h(\lambda h)^{1/2}\cot(\lambda h)^{1/2}s = h\alpha(t-s), \\ &-h(\lambda h)^{1/2}\cot(\lambda h)^{1/2}(l-t) = -h\alpha(t-s). \end{aligned} \tag{3.11}$$

It follows that $\cot(\lambda h)^{1/2}s=\cot(\lambda h)^{1/2}(l-t)$. Since we are solving for the smallest eigenvalue λ_1 this implies $(\lambda h)^{1/2}s=(\lambda h)^{1/2}(l-t)$ so $s=l-t$. Thus the column is symmetric about $l/2$. From (3.8) and (3.6) we see that at $x=l/2, y_1''=-\lambda_1(\rho_1^+)=-2h\alpha$. Using this relationship and (3.10) and (3.11) we see that s is defined (as a function of h, l, V, T) to be the smallest positive root of the equation

$$(3.12)\qquad h^{3/2}(l-2s)^2\cot s(l-2s)^{-1}[3h(VT^{-1}-hl)]^{1/2}=3(VT^{-1}-hl).$$

Using s determined by (3.12) we obtain α using (3.9) and $\rho_1^+(x)$ using (3.8). We have proved:

THEOREM III. *The strongest profile column pinned at each end and having length l, volume V and thickness T satisfying $\rho(x)\geqq h\geqq T>0$ will have a width function $\rho_1^+(x)$ defined by* (3.9). *Equation* (3.12) *defines s and for any $\rho\in\mathfrak{C}$ we have*

$$\lambda_1(\rho)\leqq\lambda_1(\rho_1^+)=12h[VT^{-1}-hl][l-2s]^{-3}.$$

As a further example of our method we will determine the shape of the strongest profile column which is clamped at $x=0$ and pinned at $x=l$ and is unconstrained in width so $\rho(x)\geqq 0$. The critical buckling load is determined by the second eigenvalue of the system

$$(3.13)\qquad \begin{gathered} y''+\lambda[\rho(x)]^{-1}y=0,\\ y(l)=0,\ ly'(0)+y(0)=0.\end{gathered}$$

The first eigenvalue of this system is zero and the eigenfunction is $y_1(x)=l-x$. We must therefore maximize $\lambda_2(\rho)$ over all $\rho(x)$ satisfying $\rho(x)\geqq 0$ and (3.7).

THEOREM IV. *The strongest unconstrained profile column which is clamped at $x=0$ and pinned at $x=l$ and has volume V and thickness T will have a width function $\rho_2^+(x)$ defined by*

$$(3.14)\qquad \rho_2^+(x)=\begin{cases} C(s-x)(x+a), & 0\leqq x\leqq s,\\ C(x-s)(l-x), & s\leqq x\leqq l.\end{cases}$$

The constants s, C and a are given by

$$(3.15)\qquad s=l(1-\sqrt{2}/2)\simeq .293l,$$

$$(3.16)\qquad C=3(3\sqrt{2}+4)V(2Tl^3)^{-1}\simeq 12.36V(Tl^3)^{-1},$$

$$(3.17)\qquad a=l(\sqrt{2}-1)\simeq .414l,$$

and for any $\rho(x)\in\mathfrak{C}$ it follows that

$$\lambda_2(\rho)\leqq\lambda_2(\rho_2^+)=3(3\sqrt{2}+4)V(Tl^3)^{-1}.$$

The proof is based on Theorem I, which we may use since $g(x,\rho)=\rho^{-1}$ is convex in ρ. We must minimize $y^2\rho^{-1}+\mu\rho$. Setting its derivative to zero and normalizing y_2 yields the condition $y_2^2(x)=[\rho_2^+(x)]^2$ so $\rho_2^+(x)=|y_2(x)|$ for all $x\in[0,l]$. Putting this into (3.13) yields

$$y_2''=-\lambda_2(\rho_2^+)\operatorname{sign}(y_2).$$

From this it follows that in each nodal domain the function $y_2(x)$ is a quadratic polynomial. Now $y_2(x)$ has two nodal domains, say $(0,s)$ and (s,l), so for any choice of normalizing constant C we obtain

$$y_2(x)=\begin{cases} C(s-x)(x+a), & 0\leqq x\leqq s, \\ C(s-x)(l-x), & s\leqq x\leqq l. \end{cases} \tag{3.18}$$

This function satisfies $y_2(s)=0$ and the boundary condition $y_2(l)=0$. We select the constant a so that the other boundary condition, $ly'(0)+y(0)=0$ is satisfied. This implies that

$$a=sl(l-s)^{-1}. \tag{3.19}$$

Now $y_2'(x)$ is continuous at $x=s$ so $y_2'(s-0)=y_2'(s+0)$ which (using (3.18)) yields $a=l-2s$. Combining this with (3.19) yields (3.15) and (3.17). Using $\rho_2^+(x)=|y_2(x)|$ and (3.18) we obtain the formula (3.14) for $\rho_2^+(x)$. Finally to determine C we integrate (3.14) directly and use (3.7). After some manipulation we find C is given by (3.16). This proves the theorem.

Many generalizations of these results are possible. For example the other boundary conditions (3.3)–(3.5) could be used in the constrained or unconstrained mode. In any of these cases it would be possible to maximize $\lambda_n(\rho)$ for any $n=1,2,3,\cdots$.

The generalized equation

$$y''+\lambda[\rho(x)]^m y=0$$

has been studied by Tadjbakhsh and Keller [12] and by E. R. Barnes [3]. Our theorems could be used to maximize $\lambda_n(\rho)$ if $m<0$ or if $m\geqq 1$ since then ρ^m is convex. If, however, $0<m\leqq 1$ then ρ^m is concave and we could use Theorem I to minimize $\lambda_n(\rho)$. In all of these cases arbitrary self-adjoint boundary conditions may be used.

If $m=1$ then ρ^m is both concave and convex so we could get both upper and lower bounds and generalize the result of Krein [9] to include arbitrary self-adjoint boundary conditions. To pursue that idea consider the differential equation

$$y''+\lambda\rho(x)y=0,$$

where $\rho(x)$ is subject to the constraints

$$h\leqq\rho(x)\leqq H, \qquad \int_0^l \rho(x)\,dx=M.$$

We will consider the lower bound $\lambda_n(\rho_n^-)$. Take $g(x,\rho)=\rho$ and $\mathfrak{B}(x,\rho)=\rho[y_n^2+\mu]$. The maximum of $\mathfrak{B}$ is at $\rho=h$ if $y_n^2+\mu<0$ but is at $\rho=H$ if $y_n^2+\mu>0$. Let x_i, $i=0,1,2,\cdots,m$ be the zeros of y_n. In each nodal domain $x_{i-1}\leqq x\leqq x_i$ which does not include $x=0$ or $x=l$ the extremal function will be a symmetric step function having two jumps. It will also be periodic in these interior nodal domains. In each of the two nodal domains $0\leqq x\leqq x_1$ and $x_{n-1}\leqq x\leqq l$ the function $\rho_n^-(x)$ will be a step function having at most two jumps. Depending on the boundary conditions used the solution can be quite involved from this point on and we will not give the details.

4. Proof of Theorems I and II. We will give a detailed proof for Theorem I and then indicate modifications necessary to prove Theorem II. Along the way we will need to use standard theorems to solve problems of the following type:

Problem III. Minimize

$$J(\rho)=\int_0^l F_0(x,\rho(x))\,dx$$

subject to constraints

$$\int_0^l F_i(x,\rho(x))\,dx=V_i, \qquad i=1,2,\cdots,N,$$

and

$$h\leqq\rho(x)\leqq H.$$

The book by Hestenes [5, p. 215] provides the following theorem used for solving problem III. There is a dual theorem used for maximizing $J(\rho)$ which we do not state here.

THEOREM V. *Let $F_i(x,\rho)$ be continuous and let $\rho_0(x)$ be a solution of problem* III. *Then there exist constants $\eta_0\geqq 0$ and $\eta_1,\eta_2,\cdots,\eta_N$ not all zero such that for each $x\in[0,l]$*

$$(4.1)\qquad \min_{h\leqq\rho\leqq H}[\eta_0F_0(x,\rho)+\eta_1F_1(x,\rho)+\cdots+\eta_NF_N(x,\rho)]$$
$$=\eta_0F_0(x,\rho_0(x))+\eta_1F_1(x,\rho_0(x))+\cdots+\eta_NF_N(x,\rho_0(x)).$$

Conversely if a function $\rho_0(x)$ and constants $\eta_0>0$, $\eta_1,\cdots,\eta_N$ exist which satisfy (4.1) *and if the conditions*

$$\int_0^l F_i(x,\rho_0(x))\,dx=V_i$$

hold then $\rho_0(x)$ solves problem III.

We now proceed to the proof of Theorem I. Let $\rho_n^+(x)\in\mathfrak{C}$ be a solution of problem I and let y_n be an eigenfunction of (2.1), (2.2), (2.3) corresponding to $\lambda_n(\rho_n^+)$. Under certain conditions there may be two linearly independent eigenfunctions. If this happens simply select y_n to be any one of them. Normalize it so that

$$(4.2)\qquad \int_0^l y_n^2g(x,\rho_n^+(x))\,dx=1.$$

For any $\rho\in\mathfrak{C}$ define the functional $J(\rho)$ by

$$(4.3)\qquad J(\rho)=\int_0^l y_n^2g(x,\rho(x))\,dx.$$

In order to prove Theorem I we will first show that $\rho_n^+(x)$ minimizes $J(\rho)$ over all $\rho\in\mathfrak{C}$,

$$(4.4)\qquad J(\rho)\geqq J(\rho_n^+)\quad \forall\rho\in\mathfrak{C}.$$

Assuming this has been done we see that Theorem I will follow directly by applying Theorem V to the functional $J(\rho)$ using $N=1$ and

$$F_0(x,\rho)=y_n^2g(x,\rho), \qquad F_1(x,\rho)=\omega(x)\rho$$

and taking $\eta_0=1$. Actually we need to first show $\eta_0\neq 0$ in order to justify dividing it out of (4.1). Suppose then that $\eta_0=0$. Then $\eta_1\neq 0$ and (4.1) becomes

$$\min_{h\leqq\rho\leqq H}[\eta_1w(x)\rho]=\eta_1w(x)\rho_n^+(x).$$

Thus it follows that $\rho_n^+(x)=h$ for all x if $\eta_1>0$ or else $\rho_n^+(x)=H$ for all x if $\eta_1<0$. Either case contradicts (1.8) so $\eta_0>0$. We only need to prove (4.4) to finish the proof of Theorem I.

Let ε be a parameter satisfying $0\leqq\varepsilon\leqq 1$ and let $\rho\in\mathfrak{C}$. Define functions $\rho(x,\varepsilon)$ and $G(\varepsilon)$ by

$$\rho(x,\varepsilon)=(1-\varepsilon)\rho_n^+(x)+\varepsilon\rho(x), \tag{4.5}$$

$$G(\varepsilon)=J(\rho(x,\varepsilon))=\int_0^l y_n^2 g(x,(1-\varepsilon)\rho_n^+(x)+\varepsilon\rho(x))\,dx. \tag{4.6}$$

Now we may recast inequality (4.4) in the form $G(0)\leqq G(1)$. To prove this inequality we will show $G(\varepsilon)$ is an increasing function of ε. Denoting derivatives with respect to ε by $(\dot{\ })$, we easily obtain

$$\dot{G}(\varepsilon)=\int_0^l y_n^2 g_\rho(x,\rho(x,\varepsilon))[\rho(x)-\rho_n^+(x)]\,dx, \tag{4.7}$$

and

$$\ddot{G}(\varepsilon)=\int_0^l y_n^2 g_{\rho\rho}(x,\rho(x,\varepsilon))[\rho(x)-\rho_n^+(x)]^2 dx. \tag{4.8}$$

Since $g(x,\rho)$ is convex in ρ it follows that $G(\varepsilon)$ is convex in ε so we only need to show $G'(0)\geqq 0$ in order to prove $G(\varepsilon)$ is increasing.

Let z_n be an eigenfunction of (2.1) corresponding to $\lambda_n(\rho(x,\varepsilon))$. Thus

$$z_n''+\lambda_n(\rho(x,\varepsilon))g(x,\rho(x,\varepsilon))z_n=0 \tag{4.9}$$

and

$$y_n''+\lambda_n(\rho_n^+(x))g(x,\rho_n^+(x))y_n=0. \tag{4.10}$$

We multiply (4.9) by y_n and (4.10) by z_n then subtract the two equations then integrate by parts. Using the self-adjoint condition (1.9) (with $f(x,\rho)=1$) we find

$$\int_0^l[\lambda_n(\rho(x,\varepsilon))g(x,\rho(x,\varepsilon))-\lambda_n(\rho_n^+(x))g(x,\rho_n^+(x))]\,y_n z_n dx=0. \tag{4.11}$$

After some manipulation we transform this equation into

$$-\Delta\lambda=\lambda_n(\rho_n^+)\int_0^l y_n^2\Delta g\,dx+O(\varepsilon^2), \tag{4.12}$$

where to simplify notation we have collected together all terms which are $O(\varepsilon^2)$ and used

$$\Delta\lambda=\lambda_n(\rho(x,\varepsilon))-\lambda_n(\rho_n^+), \tag{4.13}$$

$$\Delta g=g(x,\rho(x,\varepsilon))-g(x,\rho_n^+(x)). \tag{4.14}$$

Now in the special case we are treating here we have $f(x,\rho)=1$ for all x, ρ. It follows that the class $\mathfrak{C}$ is convex. This implies that

$$\rho(x,\varepsilon)=(1-\varepsilon)\rho_n^+(x)+\varepsilon\rho(x)\in\mathfrak{C}\quad\forall\varepsilon\in[0,1]. \tag{4.15}$$

Since $\rho_n^+(x)$ maximizes $\lambda_n(\rho)$ we see $\Delta\lambda \leqq 0$ for all $\varepsilon \in (0,1]$. We now divide (4.12) by ε and then let $\varepsilon \to 0^+$ to obtain

$$\lambda_n(\rho_n^+)\int_0^l y_n^2 g_\rho(x,\rho(x,0))\dot{\rho}(x,0)\,dx \geqq 0. \tag{4.16}$$

Since $\lambda_n(\rho_n^+)>0$ and $\rho(x,0)=\rho_n^+(x)$ and $\dot{\rho}(x,0)=\rho(x)-\rho_n^+(x)$ we see that (4.16) implies $G'(0)\geqq 0$ which was to be proved.

The proof of Theorem II is much the same. Now however y_n is an eigenfunction corresponding to $\lambda_n(\rho_n^-)$, we have $\Delta\lambda \geqq 0$, inequality (4.4) is reversed and $G(\varepsilon)$ turns out to be a concave decreasing fucntion since $G'(0)\leqq 0$.

Incidentally (4.12) implies that the first variation $\delta\lambda_n$ of $\lambda_n(\rho)$ is given by

$$\delta\lambda_n = -\lambda_n(\rho_n^+)\int_0^l y_n^2 g_\rho(x,\rho_n^+(x))\,\delta\rho_n^+\,dx$$

where $\delta\rho_n^+ = \varepsilon[\rho(x)-\rho_n^+(x)]$. This generalizes a result of Banks [4] (where $g(x,\rho)=\rho$ and $g_\rho(x,\rho)=1$) based on an idea due originally to Nehari. Keller [8] has also used similar ideas to minimize ratios $\lambda_1(\rho)/\lambda_2(\rho)$.

5. Generalizations. Our method can be generalized to handle equations of the form

$$(f(x,\rho(x))y')' + [\lambda g(x,\rho(x)) + q(x,\rho(x))]y = 0 \tag{5.1}$$

coupled with arbitrary, self-adjoint boundary conditions of the form (2.2), (2.3). Now, however, we will also allow the coefficients α_i and β_i in (2.2), (2.3) to depend on the values of $\rho(x)$ at $x=0$ and $x=l$. Thus we include boundary conditions like $\rho(x)y'(x) \to 0$ as $x \to l$ which were used by Troesch [14]. It is also possible for $f(x,\rho(x))$ to vanish at $x=0$ and/or $x=l$ which then leads to a singular eigenvalue problem. The self-adjoint condition (1.9) might also impose a nonlinear constraint on the function $\rho(x)$ so the class $\mathfrak{C}$ defined by (1.6)–(1.9) may not even be convex.

In order to avoid these difficulties at the end points we will define a new subclass of $\mathfrak{C}$. Let $\delta>0$ be a small number and let $\rho_n^+(x)$ be the solution of problem I for (5.1), (2.2), (2.3). Let $\mathfrak{C}^+(\delta)$ be the class of all functions $\rho \in \mathfrak{C}$ which agree with $\rho_n^+(x)$ near the ends,

$$\rho(x) = \rho_n^+(x) \qquad \text{if } x\in[0,\delta] \quad \text{or} \quad x \in [l-\delta, l]. \tag{5.2}$$

Similarly we define $\mathfrak{C}^-(\delta)$ using $\rho_n^-(x)$ instead of $\rho_n^+(x)$. Clearly $\mathfrak{C}^+(\delta)$ is a convex subclass of $\mathfrak{C}$ and

$$\max_{\rho\in\mathfrak{C}^+(\delta)} \lambda_n(\rho) = \max_{\rho\in\mathfrak{C}} \lambda_n(\rho) = \lambda_n(\rho_n^+).$$

Thus we may do our analysis on the interval $[\delta, l-\delta]$ and afterwards let $\delta \to 0^+$. We will first state our theorems and then show how to carry out this process.

THEOREM VI. *Given self-adjoint boundary conditions let* $\lambda_n(\rho)$ *denote the nth eigenvalue of* (5.1). *Suppose there exists a function* $\rho_n^+ \in \mathfrak{C}$ *which solves problem* I *so that*

$$\lambda_n(\rho) \leqq \lambda_n(\rho_n^+) \quad \forall \rho \in \mathfrak{C}.$$

Let $g(x,\rho)$ *and* $q(x,\rho)$ *be convex in* ρ *and let* $f(x,\rho)$ *be concave in* ρ.

Then there exists a constant μ *such that for each* $x\in[0,l]$ *the minimum over all* $\rho\in[h,H]$ *of the function* $\mathfrak{B}(x,\rho)$ *defined by*

$$(5.3)\qquad \mathfrak{B}(x,\rho)=\lambda_n(\rho_n^+)y_n^2g(x,\rho)+q(x,\rho)y_n^2-f(x,\rho)(y_n')^2+\mu w(x)\rho$$

is attained when $\rho=\rho_n^+(x)$, *that is*

$$(5.4)\qquad \min_{h\leqq\rho\leqq H}\mathfrak{B}(x,\rho)=\mathfrak{B}(x,\rho_n^+(x)).$$

Here $y_n=y_n(x)$ *is an nth eigenfunction of* (5.1) *corresponding to* $\rho_n^+(x)$, *so*

$$(5.5)\qquad [f(x,\rho_n^+(x))y_n']'+[\lambda_n(\rho_n^+)g(x,\rho_n^+(x))+q(x,\rho_n^+(x))]y_n=0.$$

Theorem VII. *Given self-adjoint boundary conditions let* $\lambda_n(\rho)$ *denote the nth eigenvalue of* (5.1). *Suppose that there exists a function* $\rho_n^-(x)\in\mathfrak{C}$ *which solves problem* II, *so*

$$\lambda_n(\rho)\geqq\lambda_n(\rho_n^-)\quad\forall\rho\in\mathfrak{C}.$$

Let $g(x,\rho)$ *and* $q(x,\rho)$ *be concave in* ρ *and let* $f(x,\rho)$ *be convex in* ρ.

Then there exists a constant μ *such that for each* $x\in[0,l]$ *the maximum over all* $\rho\in[h,H]$, *of the function* $\mathfrak{B}(x,\rho)$ *defined by*

$$(5.6)\qquad \mathfrak{B}(x,\rho)=\lambda_n(\rho_n^-)y_n^2g(x,\rho)+q(x,\rho)y_n^2-f(x,\rho)(y_n')^2+\mu w(x)\rho$$

is attained when $\rho=\rho_n^-(x)$, *that is*,

$$(5.7)\qquad \max_{h\leqq\rho\leqq H}\mathfrak{B}(x,\rho)=\mathfrak{B}(x,\rho_n^-(x)).$$

Here $y_n=y_n(x)$ *is an nth eigenfunction of* (5.1) *corresponding to* $\rho_n(x)$, *so*

$$(5.8)\qquad [f(x,\rho_n^-(x))y_n']'+[\lambda_n(\rho_n^-)g(x,\rho_n^-(x))+q(x,\rho_n^-(x))]y_n=0.$$

The proof of Theorem VI proceeds much like that of Theorem I with a few modifications. In this case we replace $J(\rho)$ as defined by (4.3) with

$$(5.9)\quad J(\rho)=\int_0^l\left[\lambda_n(\rho_n^+)y_n^2g(x,\rho(x))+y_n^2q(x,\rho(x))-(y_n')^2f(x,\rho(x))\right]dx.$$

We will first show that $\rho_n^+(x)$ minimizes $J(\rho)$ over all $\rho\in\mathfrak{C}^+(\delta)$. Assuming this has been done, we then let $\delta\to0$ and we see that $\rho_n^+(x)$ minimizes $J(\rho)$ over all $\rho\in\mathfrak{C}$. We now apply Theorem V using $N=1$ and

$$F_0(x,\rho)=\lambda_n(\rho_n^+)y_n^2g(x,\rho)+y_n^2q(x,\rho)-(y_n')^2f(x,\rho),$$
$$F_1(x,\rho)=w(x)\rho.$$

This will prove the theorem.

Let $\rho\in\mathfrak{C}^+(\delta)$. We need to show $J(\rho)\geq J(\rho_n^+)$. Define $\rho(x,\varepsilon)$ for $\varepsilon\in[0,1]$ by (4.5) and let

$$G(\varepsilon)=J[\rho(x,\varepsilon)]=J[(1-\varepsilon)\rho_n^+(x)+\varepsilon\rho(x)].$$

We see

$$\dot{G}(\varepsilon)=\int_0^l\Big[\lambda_n(\rho_n^+)y_n^2g_\rho(x,\rho(x,\varepsilon))\\ +y_n^2q_\rho(x,\rho(x,\varepsilon))-(y_n')^2f_\rho(x,\rho(x,\varepsilon))\Big]\dot{\rho}(x,\varepsilon)\,dx \tag{5.10}$$

and

$$\ddot{G}(\varepsilon)=\int_0^l\Big[\lambda_n(\rho_n^+)y_n^2g_{\rho\rho}(x,\rho(x,\varepsilon))\\ +y_n^2q_{\rho\rho}(x,\rho(x,\varepsilon))-(y_n')^2f_{\rho\rho}(x,\rho(x,\rho(x,\varepsilon)))\Big][\dot{\rho}(x,\varepsilon)]^2dx. \tag{5.11}$$

Thus $G(\varepsilon)$ is convex in ε so we only need to show $\dot{G}(0)\geqq 0$.

To do this we proceed along the same lines as (4.9), (4.10), (4.11), (4.12). The generalization of (4.12) now reads

$$-\Delta\lambda=\int_0^l\big[\lambda_n(\rho_n^+)y_n^2\Delta g+y_n^2\Delta q-(y_n')^2\Delta f\big]dx+\Delta f y_n y_n'\Big]_{x=0}^{x=l}+O(\varepsilon^2) \tag{5.12}$$

where the boundary term results from integration by parts and $\Delta f=f(x,\rho(x,\varepsilon))-f(x,\rho_n^+(x))$. Now however $\rho\in\mathfrak{C}^+(\delta)$ and so $\Delta f=0$ for all x in the intervals $0\leqq x\leqq\delta$ or $l-\delta\leqq x\leqq l$. The boundary term in (5.12) drops out and we now divide (5.12) by $\varepsilon>0$ and let $\varepsilon\to 0^+$. Since $\Delta\lambda\leqq 0$ for a maximum we obtain $G'(0)\geqq 0$ which finishes the proof. The proof of Theorem VII is similar and will not be given.

Theorems VI and VII have converses which, interestingly enough, do not require convexity conditions on the coefficients.

THEOREM VIII. *Suppose* $\rho_n^+(x)\in\mathfrak{C}$ *is any solution of* (5.4), (5.5) *and* $\rho_n^-(x)\in\mathfrak{C}$ *is any solution of* (5.7), (5.8). *Then* $\rho_n^+(x)$ *is a very weak maximum of* $\lambda_n(\rho)$ *meaning that* $\forall\rho\in\mathfrak{C}$

$$\frac{d}{d\varepsilon}\lambda_n\,[(1-\varepsilon)\rho_n^+(x)+\varepsilon\rho(x)]\big|_{\varepsilon=0}\leqq 0 \tag{5.13}$$

and $\rho_n^-(x)$ *is a very weak minimum of* $\lambda_n(\rho)$, *meaning that* $\forall\rho\in\mathfrak{C}$,

$$\frac{d}{d\varepsilon}\lambda_n\,[(1-\varepsilon)\rho_n^-(x)+\varepsilon\rho(x)]\big|_{\varepsilon=0}\geqq 0. \tag{5.14}$$

We will prove (5.13). We first let $\delta\to 0$ in (5.12) and rewrite the result as

$$-\Delta\lambda=J(\rho(x,\varepsilon))-J(\rho_n^+(x))+O(\varepsilon^2). \tag{5.15}$$

Now $\rho_n^+(x)$ solves the minimum problem (5.4) so by the converse part of Theorem V we see that $\rho_n^+(x)$ minimizes $J(\rho)$. Thus (5.15) implies $-\Delta\lambda\geqq O(\varepsilon^2)$. Dividing by $\varepsilon>0$ and letting $\varepsilon\to 0$ gives (5.13).

As an example of Theorem VI we will take $h=0$, $H=\infty$ and reproduce some of the results of Troesch [14] on the sloshing frequency of liquid in a container. The governing equation is

$$[x\rho(x)y']'+[\lambda x-m^2\rho(x)x^{-1}]y=0 \tag{5.16}$$

with boundary conditions

$$y(0)=0,\qquad \rho(x)y'=0\quad \text{at } x=l.$$

Assuming the container has given volume the function $\rho(x)$ will satisfy

$$\int_0^l x\rho(x)\,dx = V,$$

and m is a given constant describing the "symmetry class". Thus we will take

$$w(x)=x, \quad f(x,\rho)=x\rho, \quad g(x,\rho)=x, \quad q(x,\rho)=m^2\rho x^{-1},$$

and we must minimize

$$\mathfrak{B}(x,\rho)=\lambda_n(\rho_n^+)y_n^2x-y_n^2m^2\rho x^{-1}-(y_n')^2x\rho+\mu x\rho$$

over all $\rho \geqq 0$. Ignoring the first term which is independent of ρ, we see that we must minimize the expression $\rho\mathfrak{D}$ where $\mathfrak{D}$ is defined by

$$\mathfrak{D}=\left[-m^2y_n^2-x^2(y_n')^2+\mu x^2\right]x^{-1}.$$

Now if $\mathfrak{D}>0$ at any point then it is positive in a neighborhood. Thus the minimum of $\rho\mathfrak{D}$ is at $\rho=0$. However this and (5.15) implies that $y_n=0$ for all x in the neighborhood which is impossible for an eigenfunction. Thus $\mathfrak{D}\leqq 0$. If however $\mathfrak{D}<0$ at any point then it is negative in a whole neighborhood so the minimum of $\rho\mathfrak{D}$ is at $\rho=+\infty$. This however contradicts (1.7). It must therefore be the case that $\mathfrak{D}=0$ for all x which yields

$$m^2y_n^2+x^2(y_n')^2=\mu x^2.$$

When y_n is properly normalized this is Troesch's optimality condition [14, eq. 12]. We could now use Theorems VI and VII to obtain maximum and minimum values for sloshing frequencies with depth functions $\rho(x)$ constrained by $h \leqq \rho(x) \leqq H$. We will not pursue that matter here.

6. The fourth order problem. We will now give theorems which can be used to solve extremal problems for the fourth order equation

$$(6.1) \qquad -[r(x,\rho(x))y'']''+[f(x,\rho(x))y']'+[\lambda g(x,\rho(x))+q(x,\rho(x))]y=0,$$

where we assume that self-adjoint boundary conditions are also specified.

THEOREM IX. *Given self-adjoint boundary conditions let $\lambda_n(\rho)$ denote the nth eigenvalue of* (6.1) *and suppose there exists a function $\rho_n^+ \in \mathfrak{C}$ which solves problem* I, *so*

$$\lambda_n(\rho)\leqq\lambda_n(\rho_n^+) \quad \forall\rho\in\mathfrak{C}.$$

Let $g(x,\rho)$, $q(x,\rho)$ be convex in ρ and let $f(x,\rho)$, $r(x,\rho)$ be concave in ρ.

Then there exists a constant μ such that for each $x\in[0,l]$, the minimum over all $\rho\in[h,H]$ of the function $\mathfrak{B}(x,\rho)$ defined by

(6.2)

$$\mathfrak{B}(x,\rho)=\lambda_n(\rho_n^+)y_n^2g(x,\rho)+q(x,\rho)y_n^2-f(x,\rho)(y_n')^2-r(x,\rho)(y_n'')^2+\mu w(x)\rho$$

is attained when $\rho=\rho_n^+(x)$. That is,

$$(6.3) \qquad \min_{h\leqq\rho\leqq H}\mathfrak{B}(x,\rho)=\mathfrak{B}(x,\rho_n^+(x)).$$

Here $y_n = y_n(x)$ is an nth eigenfunction of (6.1) *corresponding to $\lambda_n(\rho_n^+)$, so*

$$(6.4) \qquad -[r(x,\rho_n^+(x))y_n'']'' + [f(x,\rho_n^+(x))y_n']' + [\lambda_n(\rho_n^+)g(x,\rho_n^+(x)) - q(x,\rho_n^+(x))]y_n = 0.$$

THEOREM X. *Given self-adjoint boundary conditions let $\lambda_n(\rho)$ denote the nth eigenvalue of* (6.1) *and suppose there exists a function $\rho_n^- \in \mathfrak{C}$ which solves problem* II, *so*

$$\lambda_n(\rho) \geqq \lambda_n(\rho_n^-) \quad \forall \rho \in \mathfrak{C}.$$

Suppose that $g(x,\rho)$ and $q(x,\rho)$ are concave in ρ and that $f(x,\rho)$ and $r(x,\rho)$ are convex in ρ.

Then there exists a constant μ such that for each $x \in [0,l]$ the maximum of the function $\mathfrak{B}(x,\rho)$ defined by

(6.5)

$$\mathfrak{B}(x,\rho) = \lambda_n(\rho_n^-)y_n^2 g(x,\rho) + q(x,\rho)y_n^2 - f(x,\rho)(y_n')^2 - r(x,\rho)(y_n'')^2 + \mu w(x)\rho$$

is attained when $\rho = \rho_n^-(x)$, that is

$$(6.6) \qquad \max_{h \leqq \rho \leqq H} \mathfrak{B}(x,\rho) = \mathfrak{B}(x,\rho_n^-(x)).$$

Here y_n is an nth eigenfunction of (6.1) *corresponding to $\lambda_n(\rho_n^-)$, so*

$$(6.7) \qquad -[r(x,\rho_n^-(x))y_n'']'' + [f(x,\rho_n^-(x))y_n']' + [\lambda_n(\rho_n^-)g(x,\rho_n^-(x)) - q(x,\rho_n^-(x))]y_n = 0.$$

The proofs of Theorems IX and X are much like those of the second order case using

$$J(\rho) = \int_0^l [\lambda_n(\rho_n^+)y_n^2 g(x,\rho(x)) + q(x,\rho(x))y_n^2 - f(x,\rho(x))(y_n')^2 - r(x,\rho(x))(y_n'')^2]\,dx.$$

We leave the details to the reader.

The analogue of the converse Theorem VIII also holds for the fourth order case.

As an example we will consider Niordson's problem [10] of maximizing the fundamental frequency of a vibrating beam by selecting an appropriate tapering. This leads to the problem

$$-[\rho^2(x)y'']'' + \lambda\rho(x)y = 0.$$

We see that $\mathfrak{B}(x,\rho)$ is given by

$$\mathfrak{B}(x,\rho) = \lambda_n y_n^2 \rho - (y_n'')^2 \rho^2 + \mu\rho.$$

Thus $\mathfrak{B}(x,\rho)$ is concave in ρ so, assuming the existence of ρ_n^-, we could use Theorem X to minimize $\lambda_n(\rho)$. But alas, Theorem IX cannot be used to reproduce Niordson's optimality condition for a maximum.

It turns out that Niordson's problem leads to strange and unexpected difficulties. Suppose for a moment that we assume only that $\rho(x) \geqq 0$. Then $\lambda_1(\rho) > 0$, $\forall \rho \in \mathfrak{C}$. If, however, we define $\rho_h(x)\in\mathfrak{C}$ by

$$\rho_h(x)=\begin{cases} h, & 0\leqq x\leqq s,\\ H, & s<x\leqq l,\end{cases}$$

then one can show that $\lambda_1(\rho_h)\to 0$ as $h\to 0$. Thus in this case there is no function $\rho_1^-(x)\in\mathfrak{C}$ which minimizes $\lambda_1(\rho)$. It is probably true that if we assume $H\geqq\rho(x)\geqq h>0$ then there does exist a minimizing function $\rho_1^-(x)\in\mathfrak{C}$ in which case Theorem X implies that if $\rho_1(x)$ is not equal to h or H then it satisfies

$$(6.8)\qquad 2\rho_1^-(x)(y_1'')^2=\lambda_1 Y_1^2+\mu.$$

Thus the calculation of $\rho_1^-(x)$ seems to be much more difficult than appears at first glance. We will not pursue this problem at this time.

Strangely (6.8) is the same optimality condition employed by Niordson to solve for the maximum of $\lambda_1(\rho)$ subject only to the condition $\rho(x)\geqq 0$ and $\int_0^l\rho(x)\,dx=1$. Niordson [10, p. 53] says:

> The variational method employed yields a *stationary* value of λ. Although there is little doubt that this value is the true maximum, a strict proof of this is lacking. Also, there is no proof concerning the supposed convergence of the sequence of functions obtained by the iteration formulas. We may hope that such proofs will eventually be given.

We will now take $h=0$, $H=+\infty$ and assume the existence of a maximizing function $\rho_1^+(x)$ for Niordson's problem. Remember that the minimizing function $\rho_1^-(x)$ does not exist in this case. We will prove that $\rho_1^+(x)$ and its eigenfunction Y_1 satisfy Niordson's optimality condition [10], eq. 3.6

$$(6.9)\qquad 2\rho_1^+(x)(Y_1'')^2=\lambda_1 Y_1^2+a^2$$

where $a=$ some constant and

$$-\left[[\rho_1^+(x)]^2Y_1''\right]''+\lambda\rho_1^+(x)Y_1=0.$$

We will do this by proving that the maximizing function $\rho_1^+(x)$ also *maximizes* $J(\rho)$ $\forall\rho\in\mathfrak{C}$,

$$J(\rho)=\int_0^l\left[\lambda\rho(x)Y_1^2-[\rho(x)]^2Y_1''\right]dx.$$

To do so we examine $G(\varepsilon)$ defined by

$$G(\varepsilon)=J\left[(1-\varepsilon)\rho_1^+(x)+\varepsilon\rho(x)\right].$$

We see $G(\varepsilon)$ is a concave function of ε. Now since we are dealing with the unconstrained problem we see that $\delta\lambda_1=0$ at $\rho=\rho_1^+$. It follows that $G'(0)=0$ so $G(\varepsilon)$ is decreasing. Thus $G(0)\geqq G(1)$ which gives

$$J(\rho_1^+)\geqq J(\rho)\quad\forall\rho\in C.$$

An application of Theorem V shows that $\rho=\rho_1^+(x)$ maximizes

$$\mathfrak{B}(x,\rho)=\lambda_1 Y_1^2\rho-(Y_1'')^2\rho^2+\mu\rho.$$

This yields (6.9). To prove that $\mu(=a^2)$ is positive we multiply (6.9) by $\rho_1^+(x)$ and integrate over $0 \leqq x \leqq 1$ to find

$$a^2 = \int_0^1 [\rho_1^+(x)]^2 [Y_1'']^2 dx.$$

We remark that if we look at the constrained problem $h \leqq \rho(x) \leqq H$ then we can only conclude that $\delta\lambda_1 = G'(0) \geqq 0$ and the above proof fails.

7. Extensions. Many extensions of these results seem to be possible. One extension would be to allow the coefficient functions to depend not only on $\rho(x)$ but on derivatives and integrals of $\rho(x)$. This would include problems like the Tallest Column considered by Keller and Niordson [7]. Another possible direction would be to consider the extremal values of functions of eigenvalues similar to those problems considered by Keller [8]. These questions are under consideration and will be published later.

Note added in proof. The new book by T. L. Troutman, *Variational Calculus with Elementary Convexity*, Springer-Verlag, New York, 1983, appeared while this article was in press. The methods and notation given there could be used to shorten and simplify some of the proofs given here.

REFERENCES

[1] D. C. Barnes, *Some isoperimetric inequalities for the eigenvalues of vibrating strings*, Pacific J. Math., 29 (1969), pp. 43–61.

[2] ———, *Buckling of columns and rearrangements of functions*, Quart. Appl. Math., to appear.

[3] E. R. Barnes, *The shape of the strongest column and some related extremal eigenvalue problems*, Quart. Appl. Math., 34 (1977), pp. 393–409.

[4] D. O. Banks, *Upper bounds for the eigenvalues of some vibrating systems*, Pacific J. Math., (4) 11 (1961), pp. 1183–1203.

[5] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[6] J. B. Keller, *The shape of the strongest column*, Arch. Rat. Mech. Anal., 5 (1960), pp. 275–285.

[7] J. B. Keller and F. I. Niordson, *The tallest column*, J. Math. Mech., 16 (1966), pp. 433–446.

[8] J. B. Keller, *The minimum ratio of two eigenvalues*, SIAM J. Appl. Math., 31 (1976), pp. 485–491.

[9] M. G. Krein, *On certain problems on the maximum and minimum of characteristic values and on Lyapunov zones of stability*, Trans. Amer. Math. Soc., 2 (1955), pp. 163–187.

[10] F. I. Niordson, *On the optimal design of a vibrating beam*, Quart. Appl. Math., 23 (1965), pp. 47–53.

[11] G. J. Simitses, *An Introduction to the Elastic Stability of Structures*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[12] I. Tadjbakhsh and J. B. Keller, *Strongest columns and isoperimetric inequalities for eigenvalues*, J. Appl. Mech., 29 (1962), pp. 159–164.

[13] S. P. Timoshenko and J. N. Gere, *Theory of Elastic Stability*, McGraw-Hill, New York, 1961.

[14] B. A. Troesch, *An isoperimetric sloshing problem*, Comm. Pure Appl. Math., 18 (1965), pp. 319–338.

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

n-SERIES PROBLEMS AND THE COUPLING OF ELECTROMAGNETIC WAVES TO APERTURES: A RIEMANN–HILBERT APPROACH*

RICHARD W. ZIOLKOWSKI[†]

Abstract. An effective approach to the solution of a large class of mixed boundary value problems (those reducible to an n-series problem) is developed. The method is based on the deduction of the equivalent Riemann–Hilbert problem and its solution. This generalized n-series approach leads to analytical descriptions of the coupling of electromagnetic waves through apertures in canonical structures into open or enclosed regions. In particular, it is applied to the canonical problem of plane wave coupling to an infinite circular cylinder with multiple infinite axial slots. Numerical results for currents induced by an H-polarized plane wave on a circular cylinder with a single slit are given.

1. Introduction. Mixed boundary value problems occur in many areas of physics and engineering. A particular class, the electromagnetic and acoustic coupling problems as they apply to an enclosed region, an external source and a coupling aperture, are of major importance, both theoretically and from a practical point of view. Nonetheless, the separable geometries in which one might expect to obtain an analytic solution have not been amenable to treatment until recently, and purely numerical techniques present difficulties largely due to the edge at the rim. Moreover, approximate solutions, such as the one developed by Bethe [1], are limited in their range of applicability.

Techniques borrowed from the analysis of the Riemann–Hilbert problem of complex variable theory and recent developments [2]–[5] in the theory and applications of dual series equations have made it possible to obtain analytical solutions to families of canonical problems descriptive of electromagnetic and acoustic coupling via apertures into enclosed and open regions. Examples of the canonical problems amenable to solution by these techniques include a plane wave incident (with an arbitrary angle of incidence) on a perfectly conducting diffraction grating, on a perfectly conducting circular cylinder with an infinite axial slot, and on a perfectly conducting spherical shell with a circular aperture. They all involve a scattering body with a single aperture (the unit cell of the grating corresponds to the slitted cylinder). Canonical problems involving structures with $(n-1)$-apertures $(n \geqq 2)$ require the solution of n-series problems. For instance, the coupling to a cylinder with two axial slots is described by a triple series equations problem.

These coupling problems constitute only a small subset of a large class of mixed boundary value problems that can be reduced to equivalent n-series problems. Standard techniques available from potential theory, such as the ones described in connection with the dual and triple series equations in [6], are cumbersome and are tailored to specific problems. On the other hand, the Riemann–Hilbert problem techniques provide a unified, systematic approach to these equations. The resultant general n-series approach is applicable to all separable geometries. Therefore, it represents a generalization of the Wiener-Hopf method.

* Received by the editors March 22, 1983.

[†] Electronics Engineering Department, Lawrence Livermore National Laboratory, Livermore, California 94550. This work was supported under the auspices of the U. S. Department of Energy by the Lawrence Livermore National Laboratory under contract W-7405-ENG-48.

It has been brought to the author's attention recently that the Riemann–Hilbert problem techniques have actually been applied in this manner to the dual series equations problems of the diffraction grating [7] and the slitted cylinder [8]. Nonetheless, the parallel approach to the general classes of n-series problems that will be discussed in this paper does not appear to have been reported. Several of the notations in this paper were chosen to resemble those employed in [7] for convenient reference.

The connections between the Riemann–Hilbert problem, n-series problems and the electromagnetic coupling through an aperture will be made in this paper. In particular, in §2 the solution of a general class of n-series problems is developed with Riemann–Hilbert problem techniques. A brief review of the Riemann–Hilbert problem itself is included in the appendix for completeness. The application of the resultant generalized n-series approach to the electromagnetic aperture coupling problem is discussed in §3. Analytic solutions for the coupling of E-polarized and H-polarized plane waves to a perfectly conducting infinite circular cylinder with multiple infinite axial slots are derived. Typical numerical results for the currents induced by an H-polarized plane wave on a circular cylinder with a single axial slit are described. Various comments concerning the main aspects of the generalized n-series approach are given in §4.

2. The Riemann–Hilbert approach to n-series problems. As shown in [6], there are many generic problems of the dual and triple series equations type. Only the n-series canonical problems encompassing those related to the slitted cylinder examples to be discussed below will be considered. They are sufficient to illustrate the proposed Riemann–Hilbert approach. The solutions to other generic classes of problems can be inferred from these results.

The Riemann–Hilbert problem, as described in the appendix, is concerned with finding the analytic function that satisfies a prescribed transition condition across an open or a closed curve. Let the unit circle S^1 be divided into two sets, Γ and L, the closure of Γ being the complement of L in S^1, and let each of these sets consist of $(n-1)$, $n \geqq 2$, open nonintersecting segments: $\Gamma = \{\Gamma_1, \cdots, \Gamma_{n-1}\}$ and $L = \{L_1, \cdots, L_{n-1}\}$. Also let $I(\Gamma) = \{I(\Gamma_1), \cdots, I(\Gamma_{n-1})\}$ and $I(L) = \{I(L_1), \cdots, I(L_{n-1})\}$ be the angular decomposition of the interval $[0, 2\pi]$ corresponding to those sets. In particular, set

$$\Gamma_j = \left\{ e^{i\phi} \mid \phi \in I(\Gamma_j) = (\theta_{2j-2}, \theta_{2j-1}) \right\} \quad (j = 1, \cdots, n-1), \tag{2.1a}$$

$$L_j = \left\{ e^{i\phi} \mid \phi \in I(L_j) = (\theta_{2j-1}, \theta_{2j}) \right\} \quad (j = 1, \cdots, n-1). \tag{2.1b}$$

Consider first the basic n-series problem $(n \geqq 2)$:

$$\sum_{m=-\infty}^{\infty} a_m e^{im\phi} = 0, \qquad \phi \in I(L), \tag{2.2a}$$

$$\sum_{m=-\infty}^{\infty} \varepsilon_m |m| a_m e^{im\phi} = \xi a_0 + f(\phi), \qquad \phi \in I(\Gamma). \tag{2.2b}$$

Depending on the specific problem, $\varepsilon_m = \operatorname{sgn}(m)$ or $\varepsilon_m = [\operatorname{sgn}(m)]^2 \equiv +1$, where it is assumed that

$$\operatorname{sgn}(m) = \begin{cases} +1 & \text{for } m \geqq 0, \\ -1 & \text{for } m < 0. \end{cases} \tag{2.3}$$

It can be reduced to a Riemann–Hilbert problem as follows. Differentiating (2.2) with respect to ϕ and substituting $x_m = ma_m (m \neq 0)$ in both (2.2a) and (2.2b), one obtains the modified n-series problem

$$\begin{cases} \sum\limits_{m\neq 0} x_m e^{im\phi} = 0, & \phi \in I(L), \quad (2.4a) \\ \sum\limits_{m\neq 0} \varepsilon_m x_m \dfrac{|m|}{m} e^{im\phi} = \xi a_0 + f(\phi), & \phi \in I(\Gamma). \quad (2.4b) \end{cases}$$

The symbol $\Sigma_{m\neq 0}$ indicates the sum over all terms from $m = -\infty$ to $m = +\infty$ except the term with $m = 0$. Now, introduce the functions

$$x_+(z) = \sum_{m>0} x_m z^m, \tag{2.5a}$$

$$x_-(z) = -\sum_{m<0} x_m z^m \tag{2.5b}$$

which are assumed to be analytic, respectively, on the interior and the exterior of the unit circle S^1. The n-series equations (2.4) can then be rewritten as

$$\begin{cases} x_+(\lambda) - x_-(\lambda) = 0, & \lambda \in L, \quad (2.6a) \\ x_+(\gamma) - T(\gamma) x_-(\gamma) = F(\gamma), & \gamma \in \Gamma, \quad (2.6b) \end{cases}$$

where

$$T(e^{i\phi}) = \begin{cases} +1 & \text{for } \varepsilon_m = \operatorname{sgn}(m), \\ -1 & \text{for } \varepsilon_m = +1, \end{cases} \tag{2.7}$$

and

$$F(e^{i\phi}) = \xi a_0 + f(\phi), \qquad \phi \in I(\Gamma). \tag{2.8}$$

Equation (2.6a) means that $x_+(z)$ and $x_-(z)$ coincide on L, i.e., they continue analytically across L and thus become the same analytic function,

$$x(z) = \begin{cases} x_+(z), & |z| < 1, \\ x_-(z), & |z| > 1. \end{cases} \tag{2.9}$$

Similarly, the functions $x_+(\gamma)$ and $x_-(\gamma)$ in (2.6b) represent, respectively, the limiting values on Γ from the interior and the exterior of S^1 of the same analytic function (2.9); hence (2.6b) describes a discontinuity in that function across the open curve Γ.

It is assumed that the solution $x(z)$ has singularities of order $+\frac{1}{2}$ at each of the endpoints $\alpha_j = \exp(i\theta_{2j-2})$, $\beta_j = \exp(i\theta_{2j-1})$ of $\Gamma_j (j = 1, \cdots, n-1)$ and is zero at infinity. This properly models the behavior of the solution in the electromagnetics case near the edges of the aperture and at infinity. Moreover, for the moment, let the transition function F be a least Hölder continuous on S^1. As indicated in [9], the Riemann–Hilbert problem techniques can actually handle solutions with other singularities, e.g., any of those whose order lies in the interval $(0,1)$, with a nonzero behavior at infinity and with a transition function satisfying a relaxed continuity condition.

Rewriting (2.6b) as the transition condition

$$x_+(\gamma) = T(\gamma) x_-(\gamma) + F(\gamma), \qquad \gamma \in \Gamma, \tag{2.10}$$

an inhomogeneous Riemann–Hilbert problem with discontinuous coefficients on an open curve is realized. The factors T and F are called, respectively, the coefficient and the free term of this Riemann–Hilbert problem. Its solution, $x(z)$, is developed in [9, Chapter VI, §42]. This problem is first reduced to one with discontinuous coefficients on the closed curve S^1 by setting

$$T_0(\zeta)=\begin{cases} T(\zeta) & \text{for } \zeta\in\Gamma, \\ +1 & \text{for } \zeta\in L \end{cases} \tag{2.11}$$

and

$$F_0(\zeta)=\begin{cases} F(\zeta) & \text{for } \zeta\in\Gamma, \\ 0 & \text{for } \zeta\in L, \end{cases} \tag{2.12}$$

so that (2.10) becomes

$$x_+(\zeta)=T_0(\zeta)x_-(\zeta)+F_0(\zeta), \qquad \zeta\in S^1. \tag{2.13}$$

Next the problem is reduced to one with continuous coefficients by introducing the characteristic function $[1/G(z)]$ of the problem, i.e., the function that has the same singular behavior as $x(z)$ at the endpoints (α_j,β_j) of the segments $\Gamma_j(j=1,\cdots,n-1)$, and which makes the product xG nonsingular at those points, and satisfies the homogeneous Riemann–Hilbert problem

$$1/G_+(\zeta)=T_0(\zeta)/G_-(\zeta), \qquad \zeta\in S^1, \tag{2.14a}$$

Note that (2.14a) also means

$$T_0(\zeta)=G_-(\zeta)/G_+(\zeta), \qquad \zeta\in S^1. \tag{2.14b}$$

Thus, mulitplying (2.13) by $G_+(\zeta)$ and defining the functions

$$\Phi(z)=x(z)G(z), \tag{2.15}$$
$$\Psi(z)=G_+(z)F_0(z), \tag{2.16}$$

one obtains

$$\Phi_+(\zeta)=\Phi_-(\zeta)+\Psi(\zeta), \qquad \zeta\in S^1. \tag{2.17}$$

This represents the transition condition of a Riemann–Hilbert problem with continuous coefficients on a closed curve. Its solution is simply [9, pp. 96–99]

$$\Phi(z)=\frac{1}{2\pi i}\int_{S^1}\frac{\Psi(\tau)\,d\tau}{\tau-z}+P_{n-2}(z), \tag{2.18}$$

where $P_{n-2}(z)$ is a polynomial of degree $(n-2)$ in z:

$$P_{n-2}(z)=c_0+c_1z^1+\cdots+c_{n-2}z^{n-2}. \tag{2.19}$$

Consequently, the desired solution of (2.13) is defined as

$$x(z)=\frac{1}{2\pi i}\frac{1}{G(z)}\int_\Gamma\frac{G_+(\tau)F(\tau)\,d\tau}{\tau-z}+\frac{1}{G(z)}P_{n-2}(z). \tag{2.20}$$

The procedure to obtain the characteristic function or equivalently the function G is given in [9, §42]. It depends on the index of the coefficient $T_0(\zeta)$, the index of the problem. It is readily shown that the index is $(n-1)$ for the present problem and that

$$G(z)=\begin{cases}\prod_{j=1}^{n-1}\left|(z-\alpha_j)(z-\beta_j)\right|^{1/2} & \text{for } \varepsilon_m=\operatorname{sgn}(m),\\ \prod_{j=1}^{n-1}\left[(z-\alpha_j)(z-\beta_j)\right]^{1/2} & \text{for } \varepsilon_m+1.\end{cases}\tag{2.21}$$

The results for the $\varepsilon_m=+1$ case are presented explicitly in [9, §42.2]. For that case the branches of G will be chosen so that as $z\to\gamma\in\Gamma$ from the interior of S^1: $G(z)\to G_+(\gamma)$, and from its exterior: $G(z)\to G_-(\gamma)=-G_+(\gamma)$. This choice satisfies the restriction of (2.14b) to Γ:

$$G_-(\gamma)=T(\gamma)G_+(\gamma),\qquad \gamma\in\Gamma.\tag{2.22}$$

The polynomial term, $P_{n-2}(z)$, in (2.18) and (2.20) is introduced to account for the assumed behavior of x at infinity. In particular, as $|z|\to\infty$ (2.21) yields

$$|G(z)|\sim|z|^{n-1}.\tag{2.23}$$

Therefore, in that limit the magnitude of the solution

$$|x(z)|\sim\frac{|P_{n-2}(z)|}{|G(z)|}\sim\frac{|c_{n-2}|}{|z|}\to 0\tag{2.24}$$

as desired.

The solution (2.20) provides a means to generate another relation between the limiting values x_+ and x_- on Γ. Let

$$\Omega(z)=\frac{1}{2\pi i}P\int_\Gamma\frac{G_+(\tau)F(\tau)\,d\tau}{\tau-z},\tag{2.25}$$

where $P\int$ means to take the Cauchy principal value of the integral. The Plemelj–Sokhotskii formulas [see (A.3) in the appendix] together with (2.20) and (2.22) give

$$x_+(\gamma)+T(\gamma)x_-(\gamma)=2[\Omega(\gamma)+P_{n-2}(\gamma)]/G_+(\gamma),\qquad \gamma\in\Gamma.\tag{2.26}$$

The coefficients $x_m(m\neq 0)$ and the constants a_0, $c_0,\cdots,c_{n-2}$ can now be obtained as follows.

First consider the case in which $\varepsilon_m=\operatorname{sgn}(m)$. Combining (2.5), (2.6) and (2.12), one obtains for all $\zeta=e^{i\phi}\in S^1$:

$$x_+(e^{i\phi})-x_-(e^{i\phi})=\sum_{m\neq 0}x_m e^{im\phi}=F_0(e^{i\phi}).\tag{2.27}$$

Fourier inversion of this expression gives the terms

$$x_m=\frac{1}{2\pi}\int_0^{2\pi}d\phi e^{-im\phi}F_0(e^{i\phi})=\frac{1}{2\pi}\int_\Gamma d\phi e^{-im\phi}F(e^{i\phi})\qquad(m\neq 0),\tag{2.28a}$$

$$a_0=\frac{-\int_\Gamma f(\phi)\,d\phi}{\xi\int_\Gamma d\phi}.\tag{2.28b}$$

If the solution (2.20) is desired, the constants $c_0,\cdots,c_{n-2}$ are then obtained from a system of $(n-1)$ equations:

$$\sum_{m\neq 0}\frac{|m|}{m}x_m e^{im\gamma_j}=2\left[\Omega(e^{i\gamma_j})+P_{n-2}(e^{i\gamma_j})/G_+(e^{i\gamma_j})\right] \qquad (j=1,\cdots,n-1), \tag{2.29a}$$

derived by evaluating the relation (2.26) at the midpoints

$$\gamma_j=\frac{1}{2}(\theta_{2j-1}+\theta_{2j-2}) \tag{2.29b}$$

of the intervals $I(\Gamma_j)(j=1,\cdots,n-1)$.

On the other hand, for the case in which $\varepsilon_m=+1$ the combination of (2.5), (2.6a) and (2.26) yields for all $\zeta=e^{i\phi}\in S^1$:

$$x_+(e^{i\phi})-x_-(e^{i\phi})=\sum_{m\neq 0}x_m e^{im\phi}=2g(e^{i\phi})\left[\Omega(e^{i\phi})+P_{n-2}(e^{i\phi})\right], \tag{2.30}$$

where

$$g(\zeta)=\begin{cases}1/G_+(\zeta) & \text{for } \zeta\in\Gamma,\\ 0 & \text{for } \zeta\in L.\end{cases} \tag{2.31}$$

Defining the terms

$$v(\zeta)=\frac{1}{i\pi}P\int_\Gamma\frac{G_+(\tau)\,d\tau}{\tau-\zeta}, \tag{2.32a}$$

$$V(\zeta)=\frac{1}{i\pi}P\int_\Gamma\frac{G_+(\tau)f(\tau)\,d\tau}{\tau-\zeta}, \tag{2.32b}$$

$$v_m=\frac{1}{2\pi}\int_0^{2\pi}v(e^{i\phi})g(e^{i\phi})e^{-im\phi}d\phi=\frac{1}{2\pi}\int_\Gamma\frac{v(e^{i\phi})e^{-im\phi}d\phi}{G_+(e^{i\phi})}, \tag{2.32c}$$

$$V_m=\frac{1}{2\pi}\int_0^{2\pi}V(e^{i\phi})g(e^{i\phi})e^{-im\phi}d\phi=\frac{1}{2\pi}\int_\Gamma\frac{V(e^{i\phi})e^{-im\phi}d\phi}{G_+(e^{i\phi})}, \tag{2.32d}$$

$$R_m=\frac{1}{2\pi}\int_0^{2\pi}g(e^{i\phi})e^{-im\phi}d\phi=\frac{1}{2\pi}\int_\Gamma\frac{e^{-im\phi}d\phi}{G_+(e^{i\phi})}, \tag{2.32e}$$

Fourier inversion of (2.30) yields a linear system of equations for the coefficients a_0 and $x_m(m\neq 0)$ of the form:

$$x_m=\xi a_0 v_m+V_m+2\sum_{j=0}^{n-2}c_jR_{m-j} \qquad (m\neq 0), \tag{2.33a}$$

$$0=\xi a_0 v_0+V_0+2\sum_{j=0}^{n-2}c_jR_{-j} \qquad (m=0). \tag{2.33b}$$

This system is completed by the $(n-1)$ relations

$$a_0=-\sum_{m\neq 0}\frac{x_m}{m}e^{im\psi_l} \qquad (l=1,\cdots,n-1) \tag{2.34a}$$

obtained from (2.2a) by setting ϕ equal to ψ_l, the midpoint of the interval $I(L_l)$:

$$\psi_l=\frac{1}{2}(\theta_{2l-1}+\theta_{2l}). \tag{2.34b}$$

With (2.33a) this constraint system becomes

$$-a_0=\xi a_0 w_l+W_l+2\sum_{j=0}^{n-2} c_j S_l^j \qquad (l=1,\cdots,n-1), \tag{2.35}$$

where

$$w_l=\sum_{m\neq 0}\frac{v_m}{m}e^{im\psi_l}, \tag{2.36a}$$

$$W_l=\sum_{m\neq 0}\frac{V_m}{m}e^{im\psi_l}, \tag{2.36b}$$

$$S_l^j=\sum_{m\neq 0}\frac{R_{m-j}}{m}e^{im\psi_l}. \tag{2.36c}$$

Note that the introduction of the $(n-1)$ constraint relations (2.29) and (2.34) is necessitated by the appearance of the $n-1$ constants $c_0,\cdots,c_{n-2}$ in the Riemann–Hilbert solution (2.20). They have, however, a direct effect only on the solution of the n-series problem (2.2) with $\varepsilon_m=+1$. Furthermore, the choice of those particular relations is somewhat arbitrary. Their evaluation at any one point in each of the intervals $I(\Gamma_j)$ and $I(L_j)(j=1,\cdots,n-1)$ instead of the angles γ_j and $\psi_j(j=1,\cdots,n-1)$ would equally suffice. Nonetheless, the midpoint rule is systematic and computationally convenient.

These general results are considerably simplified if the forcing function f has the Fourier expansion:

$$f(\phi)=\sum_{n=-\infty}^{\infty} f_n e^{in\phi}. \tag{2.37}$$

Defining the additional coefficient

$$Q_m=\frac{1}{2\pi}\int_\Gamma d\phi e^{-im\phi}, \tag{2.38}$$

the solution system (2.28) becomes

$$ma_m=\xi a_0 Q_0+\sum_{n=-\infty}^{\infty} f_n Q_{m-n}, \tag{2.39a}$$

$$a_0=-\sum_{n=-\infty}^{\infty} f_n Q_{-n}/\xi Q_0. \tag{2.39b}$$

Similarly, defining the coefficients

$$V_n(e^{i\phi})=\frac{1}{i\pi}P\int_\Gamma\frac{G_+(\tau)\tau^n d\tau}{\tau-e^{i\phi}}, \tag{2.40a}$$

$$V_m^n=\frac{1}{2\pi}\int_0^{2\pi}V_n(e^{i\phi})g(e^{i\phi})e^{-im\phi}d\phi=\frac{1}{2\pi}\int_\Gamma\frac{V_n(e^{i\phi})e^{-im\phi}d\phi}{G_+(e^{i\phi})}, \tag{2.40b}$$

$$W_l^n=\sum_{m\neq 0} V_m^n\frac{e^{im\psi_l}}{m}, \tag{2.40c}$$

the solution system (2.33) and (2.35) becomes

$$ma_m=\xi a_0V_m^0+\sum_{n=-\infty}^{\infty} f_nV_m^n+2\sum_{l=0}^{n-2} c_lR_{m-l} \qquad (m\neq 0), \tag{2.41a}$$

$$0=\xi a_0V_0^0+\sum_{n=-\infty}^{\infty} f_nV_0^n+2\sum_{l=0}^{n-2} c_lR_{-l} \qquad (m=0), \tag{2.41b}$$

$$0=(1+\xi W_l^0)a_0+\sum_{n=-\infty}^{\infty} f_nW_l^n+2\sum_{j=0}^{n-2} c_jS_l^j \qquad (l=1,\cdots,n-1). \tag{2.41c}$$

Note that (2.41b) and (2.41c) can be solved simultaneously to obtain the n constants a_0, $c_0,\cdots,c_{n-2}$ and then the values $a_m(m\neq 0)$ follow immediately from (2.41a) or (2.41a), (2.41b) and (2.41c) can be solved simultaneously as an entire system. However, because (2.41b) and (2.41c) are decoupled from (2.41a), the former would be numerically superior to the latter.

The solution to the basic n-series problem

$$\sum_{m=-\infty}^{\infty} b_me^{im\phi}=h(\phi), \qquad \phi\in I(L), \tag{2.42a}$$

$$\sum_{m=-\infty}^{\infty} \varepsilon_mb_m|m|e^{im\phi}=0, \qquad \phi\in I(\Gamma), \tag{2.42b}$$

which is complementary to the one defined by (2.2), follows in an analogous manner. Let

$$y_+(z)=\sum_{m=0}^{\infty} b_mz^m, \tag{2.43a}$$

$$y_-(z)=\sum_{m<0} \varepsilon_mb_mz^m. \tag{2.43b}$$

Integrating (2.42b) and setting

$$b_0=-\sum_{m\neq 0}\varepsilon_m\frac{|m|}{m}b_me^{im\gamma_l} \qquad (l=1,\cdots,n-1). \tag{2.44}$$

Equations (2.42) can then be rewritten as

$$y_+(\lambda)-T(\lambda)y_-(\lambda)=h(\phi), \qquad \lambda=e^{i\phi}\in L, \tag{2.45a}$$

$$y_+(\gamma)-y_-(\gamma)=0, \qquad \gamma\in\Gamma. \tag{2.45b}$$

The $(n-1)$ relations (2.44) are analogous to the constraint system (2.34). Moreover, the system (2.45) has the same form as (2.6) except that the line of discontinuity is now L rather than Γ. Consequently, the characteristic function $\tilde{G}(z)$ of the corresponding Riemann-Hilbert problem is (2.21) with α_j and β_j, the endpoints of the arcs Γ_j, replaced with $\tilde{\alpha}_j=\exp(i\theta_{2j-1})$ and $\tilde{\beta}_j=\exp(i\theta_{2j})$, the endpoints of the complementary arcs L_j. However, since S^1 is a closed curve, $\theta_{2(n-1)} \equiv \theta_0$. Therefore, $\tilde{G} \equiv G$. Assuming the forcing function h has the Fourier expansion

$$h(\phi) = \sum_{n=-\infty}^{\infty} h_n e^{in\phi}, \tag{2.46}$$

the preceding Riemann–Hilbert techniques then yield for the $\varepsilon_m = \operatorname{sgn}(m)$ case the solution coefficients

$$b_m = \frac{1}{2\pi} \int_L h(\phi) e^{-im\phi} d\phi = \sum_{n=-\infty}^{\infty} h_n \tilde{Q}_{m-n} \qquad (\text{all } m), \tag{2.47a}$$

where

$$\tilde{Q}_m = \frac{1}{2\pi} \int_L e^{-im\phi} d\phi, \tag{2.47b}$$

and for the $\varepsilon_m = +1$ case the solution system

$$\operatorname{sgn}(m) b_m = \sum_{n=-\infty}^{\infty} h_n \tilde{V}_m^n + 2 \sum_{j=0}^{n-2} c_j \tilde{R}_{m-j} \qquad (\text{all } m), \tag{2.48a}$$

$$-b_0 = \sum_{n=-\infty}^{\infty} h_n \tilde{W}_l^n + 2 \sum_{j=0}^{n-2} c_j \tilde{S}_l^j \qquad (l = 1, \cdots, n-1), \tag{2.48b}$$

where

$$\tilde{V}_n(e^{i\phi}) = \frac{1}{i\pi} P \int_L \frac{G_+(\tau) \tau^n d\tau}{\tau - e^{i\phi}}, \tag{2.49a}$$

$$\tilde{V}_m^n = \frac{1}{2\pi} \int_L \frac{\tilde{V}_n(e^{i\phi}) e^{-im\phi} d\phi}{G_+(e^{i\phi})}, \tag{2.49b}$$

$$\tilde{R}_m = \frac{1}{2\pi} \int_L \frac{e^{-im\phi} d\phi}{G_+(e^{i\phi})}, \tag{2.49c}$$

$$\tilde{W}_l^n = \sum_{m \neq 0} \tilde{V}_m^n e^{im\gamma_l}, \tag{2.49d}$$

$$\tilde{S}_l^j = \sum_{m \neq 0} \tilde{R}_{m-j} e^{im\gamma_l}. \tag{2.49e}$$

Note that the solution of the general n-series problem

$$\sum_{m=-\infty}^{\infty} c_m e^{im\phi} = h(\phi), \qquad \phi \in I(L), \tag{2.50a}$$

$$\sum_{m=-\infty}^{\infty} \varepsilon_m c_m |m| e^{im\phi} = \xi a_0 + f(\phi), \qquad \phi \in I(\Gamma) \tag{2.50b}$$

can now be obtained. Solving independently the problems defined by (2.2) and (2.42), the solution of (2.50) follows immediately by setting $c_m = a_m + b_m$.

More complicated *n*-series systems of the form

$$\sum_{m=-\infty}^{\infty} a_m e^{im\phi}=0, \qquad \phi\in I(L), \tag{2.51a}$$

$$\sum_{m=-\infty}^{\infty} \varepsilon_m a_m |m| \tau_m e^{im\phi}=\xi a_0+f(\phi), \qquad \phi\in I(\Gamma) \tag{2.51b}$$

and

$$\sum_{m=-\infty}^{\infty} b_m e^{im\phi}=h(\phi), \qquad \phi\in I(L), \tag{2.52a}$$

$$\sum_{m=-\infty}^{\infty} \varepsilon_m b_m |m| \tau_m e^{im\phi}=0, \qquad \phi\in I(\Gamma) \tag{2.52b}$$

are encountered in mixed boundary value problems such as those describing aperture coupling. Assuming that the coefficient function τ_m has the decomposition

$$\tau_m=1+\chi_m, \tag{2.53}$$

where the function χ_m satisfies the limiting condition

$$\lim_{|m|\to\infty} \chi_m \sim o\left(\frac{1}{|m|}\right), \tag{2.54}$$

these *n*-series problems can be reduced to the basic problems (2.2) and (2.42) by treating the χ_m dependent terms as forcing functions. In particular, define the functions $\tilde{y}_+(z)$ and $\tilde{y}_-(z)$ by (2.43a) and (2.43b) respectively, with b_m replaced by the modified coefficient

$$\tilde{b}_m=b_m\tau_m, \tag{2.55}$$

and define the modified forcing functions

$$\tilde{F}(e^{i\phi})=\xi a_0+\sum_{n=-\infty}^{\infty} \tilde{F}_n e^{in\phi} \tag{2.56a}$$

and

$$\tilde{H}(e^{i\phi})=\sum_{n=-\infty}^{\infty} \tilde{H}_n e^{in\phi}, \tag{2.57a}$$

where the Fourier coefficients

$$\tilde{F}_n=f_n-|n|a_n\chi_n\equiv f_n-\frac{|n|}{n}x_n\chi_n \tag{2.56b}$$

and

$$\tilde{H}_n=h_n+b_n\chi_n=h_n+\frac{\tilde{b}_n\chi_n}{1+\chi_n}. \tag{2.57b}$$

Systems (2.51) and (2.52) can then be replaced by the equations

$$x_+(\lambda)-x_-(\lambda)=0, \qquad \lambda\in L, \tag{2.58a}$$

$$x_+(\gamma)-T(\gamma)x_-(\gamma)=\tilde{F}(\gamma), \qquad \gamma\in\Gamma, \tag{2.58b}$$

and

$$\tilde{y}_+(\lambda)-T(\lambda)\tilde{y}_-(\lambda)=\tilde{H}(\lambda), \qquad \lambda\in L, \tag{2.59a}$$

$$\tilde{y}_+(\gamma)-\tilde{y}_-(\gamma)=0, \qquad \gamma\in\Gamma. \tag{2.59b}$$

The associated constraint relations in the $\varepsilon_m=+1$ case are (2.34) and (2.44), the latter having each b_m replaced with $\tilde{b}_m$. The Riemann–Hilbert technique can now be applied to (2.58) and (2.59). The solutions for the $\varepsilon_m=\operatorname{sgn}(m)$ cases:

$$\begin{aligned} x_m&=\xi a_0Q_0+\sum_{n=-\infty}^{\infty} f_nQ_{m-n}-\sum_{n=-\infty}^{\infty}\frac{|n|}{n}x_n\chi_nQ_{m-n} && (m\neq 0),\\ a_0&=\left[\sum_{n=-\infty}^{\infty}\frac{|n|}{n}x_n\chi_nQ_{-n}-\sum_{n=-\infty}^{\infty}f_nQ_{-n}\right]\Big/\xi Q_0 && (m=0), \end{aligned} \tag{2.60}$$

$$\tilde{b}_m=\sum_{n=-\infty}^{\infty}h_n\tilde{Q}_{m-n}+\sum_{n=-\infty}^{\infty}\frac{\tilde{b}_n\chi_n\tilde{Q}_{m-n}}{1+\chi_n} \qquad (\text{all } m), \tag{2.61}$$

and for the $\varepsilon_m=+1$ cases:

(2.62)

$$\begin{aligned} x_m&=\xi a_0V_m^0+\sum_{n=-\infty}^{\infty}f_nV_m^n-\sum_{n=-\infty}^{\infty}\frac{|n|}{n}x_n\chi_nV_m^n+2\sum_{j=0}^{n-2}c_jR_{m-j} && (m\neq 0),\\ 0&=\xi a_0V_0^0+\sum_{n=-\infty}^{\infty}f_nV_0^n-\sum_{n=-\infty}^{\infty}\frac{|n|}{n}x_n\chi_nV_0^n+2\sum_{j=0}^{n-2}c_jR_{-j} && (m=0),\\ 0&=(1+\xi W_l^0)a_0+\sum_{n=-\infty}^{\infty}f_nW_l^n-\sum_{n=-\infty}^{\infty}\frac{|n|}{n}x_n\chi_nW_l^n+2\sum_{j=0}^{n-2}c_jS_l^j && (l=1,\cdots,n-1), \end{aligned}$$

$$\begin{aligned} \operatorname{sgn}(m)\tilde{b}_m&=\sum_{n=-\infty}^{\infty}h_n\tilde{V}_m^n+\sum_{n=-\infty}^{\infty}\frac{\tilde{b}_n\chi_n\tilde{V}_m^n}{1+\chi_n}+2\sum_{j=0}^{n-2}c_j\tilde{R}_{-j} && (\text{all } m),\\ -\tilde{b}_0&=\sum_{n=-\infty}^{\infty}h_n\tilde{W}_l^n+\sum_{n=-\infty}^{\infty}\frac{\tilde{b}_n\chi_n\tilde{W}_l^n}{1+\chi_n}+2\sum_{j=0}^{n-2}c_j\tilde{S}_l^j && (l=1,\cdots,n-1) \end{aligned} \tag{2.63}$$

follow immediately from (2.39), (2.41), (2.47) and (2.48).

Note that all of these solution systems have the general form

$$u_m=\sum_{n=-\infty}^{\infty}\Lambda_{mn}u_n+v_m, \qquad m=0,\pm1,\pm2,\cdots \tag{2.64}$$

where the matrix Λ_{mn} and the vector v_m are known quantities. This infinite system of equations represents a Fredholm equation of the second kind and may be treated with a variety of methods. The technique utilized in the slitted cylinder examples will be described in the next section. It also should be noted that the assumption (2.54) is made because it causes, for example, the combination $x_n \chi_n$ to behave as $a_n n^{-\varepsilon}$, $\varepsilon > 0$, as $n \to \infty$, thereby insuring the convergence of the associated sums in (2.60) and (2.62). If desired, this condition could be relaxed. It suffices for the slitted cylinder examples.

Finally, as a further generalization of the above results, consider, for instance, the n-series problem

$$(2.65) \qquad \begin{cases} \sum_{j=-\infty}^{\infty} a_j e_j(\phi) = 0, & \phi \in I(L), \\ \sum_{j=-\infty}^{\infty} a_j \tilde{e}_j(\phi) = f(\phi), & \phi \in I(\Gamma), \end{cases}$$

where the functions $\{e_m(\phi), m = 0, 1, \cdots\}$ form an orthonormal basis of the Hilbert space $\mathscr{L}_2([0, 2\pi])$. Since this system is constructed from the mixed boundary conditions, the functions $\{\tilde{e}_m(\phi)\}$ must be linear combinations of the basis functions $\{e_m(\phi)\}$. Moreover, because the set $\{e^{im\phi}, m = 0, \pm 1, \cdots\}$ is also a basis of $\mathscr{L}_2([0, 2\pi])$, each function $e_m(\phi)$; hence, each $\tilde{e}_m(\phi)$ can be expanded in terms of those basis functions. In particular, set

$$(2.66a) \qquad e_j(\phi) = \sum_{m=-\infty}^{\infty} U_{jm} e^{im\phi},$$

$$(2.66b) \qquad \tilde{e}_j(\phi) = \sum_{m=-\infty}^{\infty} \eta(m) U_{jm} e^{im\phi}.$$

Thus, defining

$$(2.67) \qquad x_m = \sum_{j=-\infty}^{\infty} U_{jm} a_j,$$

the n-series system (2.65) becomes, for example,

$$(2.68a) \qquad \sum_{m=-\infty}^{\infty} x_m e^{im\phi} = 0, \qquad \phi \in I(L),$$

$$(2.68b) \qquad \sum_{m=-\infty}^{\infty} x_m \operatorname{sgn}(m) e^{im\phi} = F(\phi), \qquad \phi \in I(\Gamma),$$

where the Fourier coefficients of the forcing function F are

$$(2.69) \qquad F_n = f_n + [\operatorname{sgn}(n) - \eta(n)] x_n.$$

The solution to the system (2.68) follows immediately from the preceding results.

3. Electromagnetic coupling to a slitted cylinder. A variety of problems including those describing the coupling of electromagnetic waves to an enclosed region can be reduced to an n-series problem. For instance, if the shape of the scattering body coincides with a constant coordinate surface in one of the coordinate systems for which the vector field equations are separable, the incident and scattered fields are first expanded in terms of the corresponding eigenfunctions. The n-series equations are then realized by enforcing on that surface the boundary conditions for the tangential electric and magnetic fields over the *aperture* and on the *perfect conductor*.

In particular, consider the electromagnetic coupling of a plane wave to a thin infinite perfectly conducting circular cylinder with $(n-1)$ infinite axial slots. The magnetic field vector of the plane wave is taken to be parallel to the axis of the cylinder. This H-polarized plane wave is assumed normally incident on the cylinder; hence, the problem is two-dimensional. A cylindrical coordinate system (ρ,ϕ,z) is centered on the axis of the cylinder; the z-axis coincides with the cylinder's axis. The angle of incidence, ϕ^{inc}, of the plane wave is arbitrary. The radius of the cylinder is $\mathbf{a}$. The angular extent of the metallic portions of the cylinder coincides with the interval $I(\Gamma)$, the apertures with $I(L)$. This geometry is illustrated in Fig. 1 for a cylinder with a single axial slot $(n=2)$. The currents induced by the plane wave on the metallic portions of the cylinder are desired. This problem is reduced to an n-series problem as follows.

For the given polarization Maxwell's equations decouple and only the E_ρ, E_ϕ and H_z components of the field are excited. The components of the field tangential to the surface of the aperture and the cylinder are of particular importance. They are related by

$$E_\phi = \frac{j}{\omega\varepsilon}\partial_\rho H_z \tag{3.1}$$

where, as throughout this paper, a $e^{j\omega t}$ time dependence is assumed.

The incident magnetic field has the Fourier mode expansion:

$$H_z^{\text{inc}} = A_0 e^{jk\rho\cos(\phi-\phi^{\text{inc}})} = A_0 \sum_{n=-\infty}^{\infty} \left[j^{|n|} J_{|n|}(k\rho) e^{-jn\phi^{\text{inc}}} \right] e^{jn\phi}. \tag{3.2}$$

From (3.1) it follows that

$$E_\phi^{\text{inc}} = jZ_0 A_0 \sum_{n=-\infty}^{\infty} \left[j^{|n|} J'_{|n|}(k\rho) e^{-jn\phi^{\text{inc}}} \right] e^{jn\phi} \tag{3.3}$$

where $J'_m(x) = dJ_m/dx$ and $Z_0 = k/\omega\varepsilon$ is the free-space characteristic impedance. The corresponding Fourier expansions of the scattered fields are:

$$H_{z>}^s = A_0 \sum_{n=-\infty}^{\infty} a_n J'_{|n|}(k\mathbf{a}) H_{|n|}(k\rho) e^{jn\phi} \qquad (\rho > \mathbf{a}), \tag{3.4a}$$

$$H_{z<}^s = A_0 \sum_{n=-\infty}^{\infty} a_n J_{|n|}(k\rho) H'_{|n|}(k\mathbf{a}) e^{jn\phi} \qquad (\rho < \mathbf{a}), \tag{3.4b}$$

$$E_{\phi>}^s = jZ_0 A_0 \sum_{n=-\infty}^{\infty} a_n J'_{|n|}(k\mathbf{a}) H'_{|n|}(k\rho) e^{jn\phi} \qquad (\rho > \mathbf{a}), \tag{3.4c}$$

$$E_{\phi<}^s = jZ_0 A_0 \sum_{n=-\infty}^{\infty} a_n J'_{|n|}(k\rho) H'_{|n|}(k\mathbf{a}) e^{jn\phi} \qquad (\rho < \mathbf{a}), \tag{3.4d}$$

where H_n is the Hankel function of second kind and order n and $H'_n(x) = dH_n/dx$. The boundary conditions for the tangential electric and magnetic fields at the surface $\rho = \mathbf{a}$ are now enforced to obtain the n-series equations.

Since the total tangential electric field is zero on the metal, the scattered and the negative of the incident electric fields are equal there:

$$E_\phi^s(\mathbf{a},\phi) = -E_\phi^{\text{inc}}(\mathbf{a},\phi) \equiv A_0 \mathbb{E}(\phi) \qquad \phi \in I(\Gamma). \tag{3.5}$$

Substituting (3.4c) or (3.4d) ((3.4c) and (3.4d) guarantee the continuity of the tangential component of the scattered electric field across the interface $\rho=\mathrm{a}$) into (3.5), one obtains

$$(3.6)\qquad jZ_0\sum_{n=-\infty}^{\infty} a_n J'_{|n|}(k\mathrm{a})H'_{|n|}(k\mathrm{a})e^{jn\phi}=\mathbb{E}(\phi),\qquad \phi\in I(\Gamma).$$

The *dc* components of the fields can be extracted from this relation by introducing the functions $K_n(x)$ so that

$$(3.7\text{a})\qquad -j\pi J'_0(x)H'_0(x)=1+K_0(x)\qquad (n=0),$$

$$(3.7\text{b})\qquad j\pi x^2 J'_n(x)H'_n(x)=n[1+K_n(x)]\qquad (n>0)$$

where $K_n(0)\equiv 0$. As defined, the $K_n(x)\sim O(n^{-2})$ as $n\to\infty$ for any fixed x, and therefore, satisfy (2.54). Equation (3.6) thus becomes

(3.8)

$$\sum_{n=-\infty}^{\infty} |n|a_n\left[1+K_{|n|}(k\mathrm{a})\right]e^{jn\phi}=(k\mathrm{a})^2[1+K_0(k\mathrm{a})]\,a_0+\frac{(k\mathrm{a})^2\pi\mathbb{E}(\phi)}{Z_0},\qquad \phi\in I(\Gamma).$$

On the other hand, continuity of H_z across the apertures and the Wronskian relationship

$$(3.9)\qquad J'_{|n|}(k\mathrm{a})H_{|n|}(k\mathrm{a})-J_{|n|}(k\mathrm{a})H'_{|n|}(k\mathrm{a})=\frac{2j}{\pi k\mathrm{a}}$$

give

$$(3.10)\qquad \sum_{n=-\infty}^{\infty} a_n e^{jn\phi}=0,\qquad \phi\in I(L).$$

Defining the quantities

$$(3.11\text{a})\qquad \chi_m=K_{|m|}(k\mathrm{a}),$$

$$(3.11\text{b})\qquad \xi=(k\mathrm{a})^2[1+\chi_0(k\mathrm{a})],$$

$$(3.11\text{c})\qquad f(\phi)=\frac{(k\mathrm{a})^2\pi}{Z_0}\mathbb{E}(\phi),$$

(3.8) and (3.10) constitute the n-series problem:

$$(3.12\text{a})\qquad \sum_{m=-\infty}^{\infty} a_m e^{jm\phi}=0,\qquad \phi\in I(L),$$

$$(3.12\text{b})\qquad \sum_{m=-\infty}^{\infty} a_m|m|\tau_m e^{jm\phi}=\xi a_0+f(\phi),\qquad \phi\in I(\Gamma).$$

It is clearly seen that (3.12) coincides with the $\varepsilon_m=+1$ case of (2.51). Consequently, the unknown amplitudes $a_m(m=0,\pm 1,\cdots)$ are obtained from the solution system (2.62). The *currents induced* on the cylinder then follow immediately from (3.4a), (3.4b) and (3.9) as

$$(3.13)\qquad J_\phi(\mathrm{a},\phi)=H^s_{z<}(\mathrm{a},\phi)-H^s_{z>}(\mathrm{a},\phi)=\frac{2A_0}{j\pi k\mathrm{a}}\sum_{m=-\infty}^{\infty} a_m e^{jm\phi}.$$

The complementary problem, an E-polarized plane wave incident upon a circular cylinder, the metal and apertures now coinciding with L and Γ, respectively, has an analogous solution. Only the E_z, H_ρ and H_ϕ components of the field are excited, the tangential components being related as

$$H_\phi = \frac{-j}{\omega\mu}\partial_\rho E_z. \tag{3.14}$$

The Fourier expansions of the incident and scattered fields are now:

$$E_z^{\text{inc}} = A_0 \sum_{n=-\infty}^{\infty} \left[j^{|n|} J_{|n|}(k\rho) e^{-jn\phi^{\text{inc}}} \right] e^{jn\phi}, \tag{3.15a}$$

$$E_{z>}^{s} = A_0 \sum_{n=-\infty}^{\infty} c_n J_{|n|}(k\mathbf{a}) H_{|n|}(k\rho) e^{jn\phi} \qquad (\rho > \mathbf{a}), \tag{3.15b}$$

$$E_{z<}^{s} = A_0 \sum_{n=-\infty}^{\infty} c_n J_{|n|}(k\rho) H_{|n|}(k\mathbf{a}) e^{jn\phi} \qquad (\rho < \mathbf{a}), \tag{3.15c}$$

$$H_{\phi>}^{s} = -jY_0 A_0 \sum_{n=-\infty}^{\infty} c_n J_{|n|}(k\mathbf{a}) H_{|n|}'(k\rho) e^{jn\phi} \qquad (\rho > \mathbf{a}), \tag{3.15d}$$

$$H_{\phi<}^{s} = -jY_0 A_0 \sum_{n=-\infty}^{\infty} c_n J_{|n|}'(k\rho) H_{|n|}(k\mathbf{a}) e^{jn\phi} \qquad (\rho < \mathbf{a}), \tag{3.15e}$$

where $Y_0 = k/\omega\mu$ is the free space admittance. Continuity of H_ϕ across the apertures and the Wronskian relation (3.9) give

$$\sum_{n=-\infty}^{\infty} c_n e^{jn\phi} = 0, \qquad \phi \in I(\Gamma). \tag{3.16}$$

Furthermore, satisfaction of the boundary condition $E_z^s(\mathbf{a},\phi) = -E^{\text{inc}}(\mathbf{a},\phi)$ yields

$$\sum_{n=-\infty}^{\infty} c_n J_{|n|}(k\mathbf{a}) H_{|n|}(k\mathbf{a}) e^{jn\phi} = \tilde{\mathbb{E}}(\phi), \qquad \phi \in I(L), \tag{3.17a}$$

where

$$\tilde{\mathbb{E}}(\phi) = \sum_{n=-\infty}^{\infty} \tilde{\mathbb{E}}_n(\phi) e^{jn\phi} = \sum_{n=-\infty}^{\infty} \left[-j^{|n|} J_{|n|}(k\mathbf{a}) e^{-jn\phi^{\text{inc}}} \right] e^{jn\phi}. \tag{3.17b}$$

However, in contrast to the H-polarized case, the *dc* components of the field are properly extracted by introducing the functions $\tilde{K}_m(x)$ so that

$$[j\pi J_0(x) H_0(x)]^{-1} = \tilde{K}_0(x), \tag{3.18a}$$

$$[-j\pi J_m(x) H_m(x)]^{-1} = m(1 + \tilde{K}_m(x)), \qquad m > 0, \tag{3.18b}$$

where $\tilde{K}_m(0) = 0$. This choice is made to account for the logarithmic singularity of H_0 near $x = 0$. Furthermore, (2.54) is satisfied since $\tilde{K}_m(x) \sim O(m^{-2})$ as $m \to \infty$. Defining the coefficients

$$b_n = c_n J_{|n|}(k\mathbf{a}) H_{|n|}(k\mathbf{a}) - \tilde{\mathbb{E}}_n, \tag{3.19a}$$

$$\tilde{\tau}_n = (1 + \tilde{\chi}_n) \equiv 1 + \tilde{K}_{|n|}(k\mathbf{a}), \tag{3.19b}$$

$$\tilde{\xi} = \tilde{K}_0(k\mathbf{a}), \tag{3.19c}$$

the *n*-series system defined by (3.16) and (3.17) reduces to the form

$$\sum_{m=-\infty}^{\infty} b_m e^{jm\phi}=0, \qquad \phi\in I(L), \tag{3.20a}$$

$$\sum_{m=-\infty}^{\infty} b_m|m|\tilde{\tau}_m e^{jm\phi}=\tilde{\xi}b_0+\tilde{f}(\phi), \qquad \phi\in I(\Gamma), \tag{3.20b}$$

where

$$\tilde{f}(\phi)=\tilde{\xi}\tilde{\mathbb{E}}_0-\sum_{n\neq 0}|n|\tilde{\mathbb{E}}_n\tilde{\tau}_n e^{jn\phi} \tag{3.21a}$$

$$=\sum_{n=-\infty}^{\infty}\frac{\tilde{\mathbb{E}}_n e^{jn\phi}}{\left[j\pi J_{|n|}(k\mathbf{a})H_{|n|}(k\mathbf{a})\right]}. \tag{3.21b}$$

Therefore, since this system is of the same type as (3.12), its solution system also follows from (2.52), and hence has the same form as the one found in the original *H*-polarized problem. On the other hand, the currents on the cylinder are now defined as

$$J_z(\mathbf{a},\phi)=H^s_{\phi>}(\mathbf{a},\phi)-H^s_{\phi<}(\mathbf{a},\phi)=\frac{2Y_0A_0}{\pi k\mathbf{a}}\sum_{n=-\infty}^{\infty}c_n e^{jn\phi}. \tag{3.22}$$

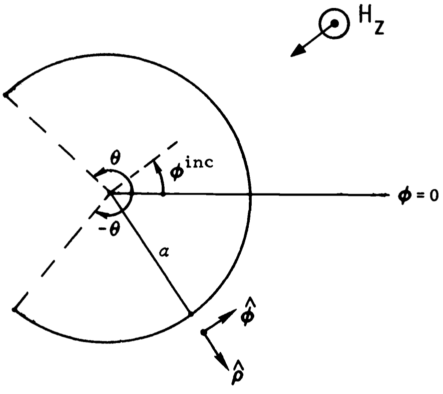


FIG. 1. *Configuration of the scattering of an H-polarized plane wave from a cylinder with an infinite axial slot.*

To illustrate the calculation of the induced currents, consider an *H*-polarized plane wave coupling to a circular cylinder with a single axial slot. The geometry of this problem is shown in Fig. 1. Equations (3.12) reduce in this case to the dual series equations:

$$\begin{aligned}&\sum_{m=-\infty}^{\infty} a_m e^{jm\phi}=0, && \phi\in(\theta,2\pi-\theta),\\ &\sum_{m=-\infty}^{\infty} a_m|m|\tau_m e^{jm\phi}=\xi a_0+f(\phi), && \phi\in(-\theta,\theta).\end{aligned} \tag{3.23}$$

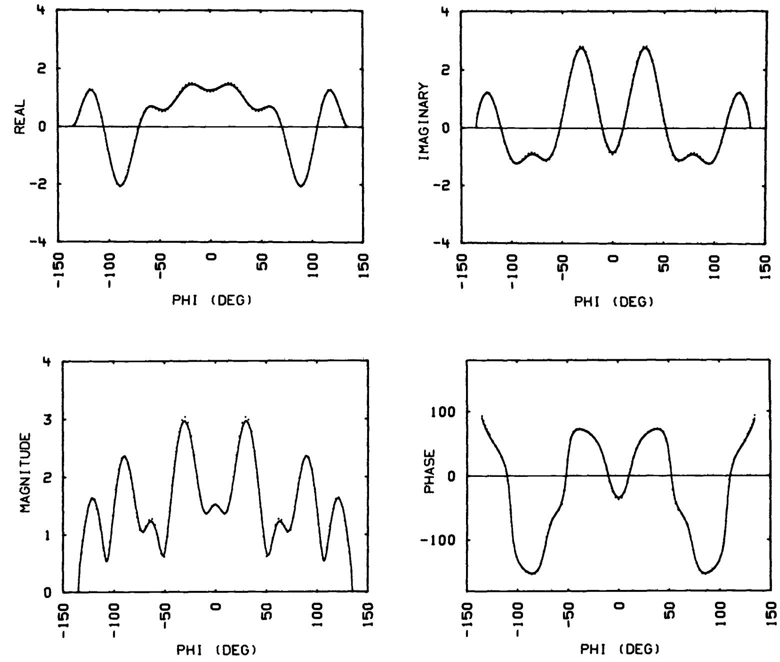


FIG. 2. *Currents calculated by the dual series* (----) *and the method of moments* (· · · ·) *for an H-polarized plane wave incident at* $\phi^{\text{inc}} = 180°$ *on a cylinder of radius* 1.0 λ *with an aperture angle* $\theta_{ap} = \pi - \theta = 45°$.

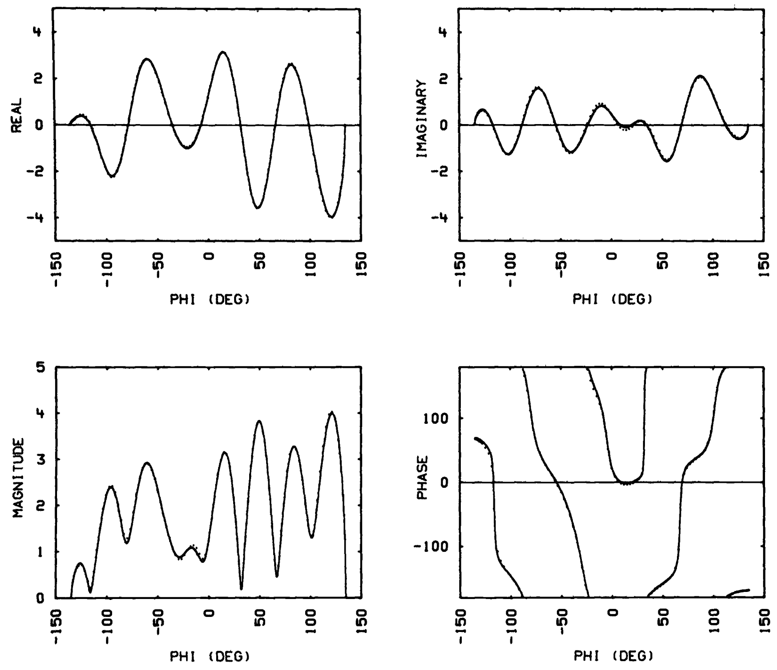


FIG. 3. *Currents calculated by the dual series* (----) *and the method of moments* (· · · ·) *for an H-polarized plane wave incident at* $\phi^{\text{inc}} = 135°$ *on a cylinder of radius* 1.0 λ *with an aperture angle* $\theta_{ap} = \pi - \theta = 45°$.

These dual series equations have the solution system

$$\text{(3.24)}\quad \begin{cases} x_m = ma_m = \xi V_m^0 a_0 + \sum\limits_{n=-\infty}^{\infty} f_n V_m^n - \sum\limits_{n=-\infty}^{\infty} \dfrac{|n|}{n} \chi_n x_n V_m^n + 2c_0 R_m & (m \neq 0), \\ 0 = \xi V_0^0 a_0 + \sum\limits_{n=-\infty}^{\infty} f_n V_0^n - \sum\limits_{n=-\infty}^{\infty} \dfrac{|n|}{n} \chi_n x_n V_m^n + 2c_0 R_0 & (m=0), \\ 0 = (1 + \xi W_0) a_0 + \sum\limits_{n=-\infty}^{\infty} f_n W^n - \sum\limits_{n=-\infty}^{\infty} \dfrac{|n|}{n} \chi_n x_n W^n + 2c_0 S & (\psi = \pi), \end{cases}$$

where $S_0^0 = S$ and

$$\text{(3.25)}\qquad f_n = -j^{|n|+1}(ka)^2 \pi J'_{|n|}(ka) e^{-jn\phi^{\text{inc}}}.$$

The coefficients V_m^n, R_m, W^n and S are given explicitly in [5]. They are combinations of Legendre functions and are readily computed. It has been found [5] that truncating f_n and χ_n in (3.24) for $|n|$ greater than some large value N and using Gauss elimination to solve the remaining finite system yields good numerical approximations for the coefficients c, a_0, $x_{\pm 1}, \cdots, x_{\pm N}$. The remaining coefficients, x_m, for $N < |m| \leqq M$ are given by the expression

$$\text{(3.26)}\qquad x_m = ma_m = \xi V_m^0 a_0 + \sum_{n=-N}^{N} f_n V_m^n - \sum_{n=-N}^{N} |n| a_n \chi_n V_m^n + 2c_0 R_m.$$

As N approaches ∞, this solution scheme becomes exact. The rate of convergence of the current sum (3.13) is then enhanced by handling the edge behavior analytically. In particular set

$$\text{(3.27)}\qquad \sum_{m=-\infty}^{\infty} a_m e^{jm\phi} = a_0 + \sum_{m \neq 0} \left(\frac{x_m - \tilde{x}_m}{m} \right) e^{jm\phi} + \sum_{m \neq 0} \frac{\tilde{x}_m}{m} e^{jm\phi}$$

where the term $\tilde{x}_m$ is a large m approximation of x_m. The first sum on the right-hand side of (3.27) is rapidly converging. The second sum is obtained analytically (see [5] for the details); it contains the singular component of the current near an edge of the aperture.

Currents generated with this dual series scheme (solid lines) and with a two-dimensional method of moments code (dotted lines) are shown in Figs. 2 and 3. In Fig. 2, the angle of incidence $\phi^{\text{inc}} = 180°$; in Fig. 3, $\phi^{\text{inc}} = 135°$. The radius of the cylinder in terms of wave length $(a/\lambda) = 1.0$ and the aperture angle $\theta_{ap} = \pi - \theta = 45°$ in both cases. Moreover, the truncation numbers were chosen to be large: $N = 25$ and $M = 190$, to guarantee the accuracy of the dual series results. Note that both figures demonstrate that the dual series solution readily models the singular behavior of the fields near the edges of the aperture. Furthermore, as discussed in [5], the dual series solution has revealed that the moment method solution will properly describe the current (especially in the shadow region) only if a nonuniform gridding that is finer near aperture edges is employed. The slight inaccuracy of the moment method solution present in both figures disappears when finer gridding is utilized.

4. Comments. The description of coupling to more complex structures such as slitted parabolic or elliptic cylinders leads to the more general n-series problem (2.65). As noted in §3, the structure is assumed to lie on a constant coordinate surface, and the incident and the interior and exterior scattered fields are expanded in the eigenmodes corresponding to that geometry. For instance, for a two-dimensional elliptic cylinder the fields would be expanded in terms of modified and periodic Mathieu functions. The n-series problem follows from enforcing the electromagnetic boundary conditions over the aperture and the metal.

The terminology "n-series problem" needs to be clarified since it is confusing to discover that in general one has a system of $2(n-1)$ equations for an n-series problem. For a single slit $n=2$ and a dual series equation system is obtained which agrees with the notation. On the other hand, for two slits $n=3$ and a system of four equations is obtained in general. However, assuming that the metal-aperture configuration is symmetric about the $\pi=0$ axis, only a triple series equation system need be treated. These symmetric problems are the only ones that have been treated in the past, for example, in [6]. The present approach is not restricted to problems of this type. Nonetheless, the terminology and the subsequent inconvenient notations were chosen so that they reduced to the standard ones encountered in dual and triple series problems.

Note that the Riemann–Hilbert results also explicitly contain, in addition to the correct edge behavior, the multipole behavior of the static solution of infinity. For instance, for a single slit case the dual series system leads to a solution (2.20) which has the limit $\lim_{|z|\to\infty} x(z) \sim c_0/z$. (In fact, since one also has from (2.5) that $\lim_{|z|\to\infty} x(z) \sim x_{-1}/z$, $c_0 \equiv x_{-1}$ in that case.) This indicates that at infinity, the static solution for the slitted cylinder behaves like a line charge or monopole. Similarly, for a cylinder with 2 slits (2.20) has the limit $\lim_{|z|\to\infty} x(z) \sim c_0 z^{-2} + c_1 z^{-1}$. Thus, the static solution contains a dipole as well as a monopole component at infinity.

N-series equations and their solution with Riemann–Hilbert techniques provide an effective approach to a large class of mixed boundary value problems. For instance, this generalized n-series approach generates analytic descriptions of the coupling of electromagnetic waves through apertures into open or enclosed regions. This was illustrated succinctly with the circular cylinder examples. The coupling to a circular cylinder with two axial slits and to a thin spherical shell with a circular aperture are currently under investigation with this method. The analytic solutions to such canonical problems are particularly useful because they are leading to the development of engineering "rules of thumb" for coupling to more general structures. Furthermore, they establish a standard to which large numerical coupling codes can be compared.

5. Appendix: the Riemann–Hilbert problem. Suppose that one is given a simple closed, smooth curve Γ dividing the complex plane into two open sets, the (bounded) interior S_+ and the exterior S_- and two Hölder continuous functions of position on that contour, $T(\gamma)$ and $F(\gamma)$, $T(\gamma)$ being nonvanishing. Let $x(z)$ be a sectionally analytic function, i.e., over the domains S_+ and S_- let $x(z)$ equal, respectively, the analytic functions $x_+(z)$ and $x_-(z)$. Then the Cauchy integral

$$x(z)=\frac{1}{2\pi i}\int_\Gamma \frac{F(\zeta)\,d\zeta}{\zeta-z} \tag{A.1}$$

solves the problem: Find a piecewise analytic function $x(z)$ vanishing at infinity that satisfies on Γ the prescribed transition condition

$$x_+(\gamma)-x_-(\gamma)=F(\gamma), \qquad \gamma\in\Gamma. \tag{A.2}$$

Note that on Γ, the function (A.1) is defined as a Cauchy principal value and satisfies a Hölder condition of the same type as F and the Plemelj–Sokhotskii conditions:

$$x_+(\gamma)=x(\gamma)+\frac{1}{2}F(\gamma), \tag{A.3a}$$

$$x_-(\gamma)=x(\gamma)-\frac{1}{2}F(\gamma). \tag{A.3b}$$

Moreover, the additional condition $x_-(\infty)=0$ can be modified. For instance, if $x(z)$ has a pole of order n at $z=\infty$, the solution of (A.2) is

$$x(z)=\frac{1}{2\pi i}\int_\Gamma \frac{F(\zeta)\,d\zeta}{\zeta-z}+P_n(z), \tag{A.4}$$

where $P_n(z)$ is a polynomial of order n in z, $P_0(z)$ being a constant.

The Riemann–Hilbert problem is a generalization of this problem. In particular, it is desired to find the sectionally analytic function $x(z)$ which satisfies on the contour Γ either the transition condition

$$x_+(\gamma)=T(\gamma)x_-(\gamma) \qquad \text{(homogeneous problem)}, \tag{A.5}$$

or

$$x_+(\gamma)=T(\gamma)x_-(\gamma)+F(\gamma) \qquad \text{(inhomogeneous problem)}. \tag{A.6}$$

A further extension of this problem to open curves and discontinuous coefficients is possible. Note that by generating a solution, $y(z)$, of the homogeneous problem (A.5):

$$y_+(\gamma)=T(\gamma)y_-(\gamma), \tag{A.7}$$

and defining the functions $\Phi=x/y$ and $\Psi=F/y_+$, the inhomogeneous problem (A.6) is reduced to the problem (A.2):

$$\Phi_+(\gamma)-\Phi_-(\gamma)=\Psi(\gamma). \tag{A.8}$$

Acknowledgments. The author wishes to thank J. B. Grant for his computational assistance and K. F. Casey for many enjoyable discussions on this topic. He especially wishes to acknowledge the many astute observations and valuable suggestions made by William A. Johnson during the course of this work. They greatly aided its completion and were very much appreciated.

REFERENCES

[1] H. A. BETHE, *Theory of diffraction by small holes*, Phys. Rev., 66 (1944), pp. 163–182.

[2] K. F. CASEY, *The natural frequencies of the axisymmetric modes of a hollow conducting sphere with a circular aperture*, National Radio Science Meeting, Los Angeles, 1981.

[3] ———, *Capacitive iris in a rectangular waveguide: an exact solution*, UCRL-88377, Lawrence Livermore National Laboratory, Livermore, CA, 1983.

[4] ———, *On the inductive iris in a rectangular waveguide*, UCRL-88308, Lawrence Livermore National Laboratory, Livermore, CA, 1983.

[5] W. A. JOHNSON AND R. W. ZIOLKOWSKI, *The coupling of an H-polarized plane wave to an infinite circular cylinder with an infinite axial slot: a dual series approach*, Radio Sci., 19 (1984), pp. 275–291.

[6] I. N. SNEDDON, *Mixed Boundary Valve Problems in Potential Theory*, North-Holland, Amsterdam, 1966.

[7] Z. S. Agranovich, V. A. Marchenko and V. P. Shestopalov, *The diffraction of electromagnetic waves from plane metallic lattices*, Zh. Tekhn. Fiz., 32 (1962), pp. 381–394; Sov. Phys.- Tech. Phys., 7 (1962), pp. 277–286.

[8] V. N. Koshparënok and V. P. Shestopalov, *Diffraction of a plane electromagnetic wave by a circular cylinder with a longitudinal slot*, Zh. Vychisl. Mat. i Mat. Fiz., 11 (1971), pp. 719–737; U.S.S.R. Comp. Math. and Math. Phys., 11 (1971), pp. 222–243

[9] F. D. Gakhov, *Boundary Value Problems*, Pergamon Press, New York, 1966.

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

012

POLYNOMIAL EXPANSIONS FOR SOLUTIONS OF THE SYSTEM $D^k_{X_1}U(X_1,\cdots,X_r)=D_{X_k}U(X_1,\cdots,X_r)$*

HANS KEMNITZ†

Abstract. This paper is an extension of results by P. C. Rosenbloom and D. V. Widder [Trans. Amer. Math. Soc., 92(1959), pp. 220–266], [Duke Math. J., 29(1962), pp. 497–503] concerning the expansion of a solution $u(x,t)$ of the heat equation, $D_x^2u(x,t)=D_tu(x,t)$, in a series of polynomial solutions.

It is found that a polynomial expansion

$$\sum_{n=0}^{\infty} a_nP_{r,n}(X_1,\cdots,X_r)$$

converges in an infinite strip $|X_r|<\sigma$, where the polynomials

$$P_{r,n}(X_1,\cdots,X_r)=n!\sum_{k_1+2k_2+\cdots+rk_r=n}\frac{X_1^{k_1}}{k_1!}\cdots\frac{X_r^{k_r}}{k_r!}$$

satisfy the system $D^k_{X_1}U(X_1,\cdots,X_r)=D_{X_k}U(X_1,\cdots,X_r)$.

This paper includes several applications of classical initial-value problems of parabolic differential equations. For example, it is found that there exists a solution of $(A_rD_x^r+\cdots+A_2D_x^2)u(x,t)=D_tu(x,t)$ which has a Maclaurin expansion in a strip $|t|<\sigma$ and which reduces to $f(x)$ for $t=0$ if and only if $f(x)$ is an entire function of special growth.

Introduction. In a recent paper [9], the author established criteria for the expansion of a solution of the parabolic equation

$$\text{(E1)}\qquad D_x^ru(x,t)=D_tu(x,t),$$

$r\geq 2$ a fixed positive number, in a series of polynomial solutions $v_{r,n}(x,t)$, where

$$v_{r,n}(x,t)=n!\sum_{k+rl=n}\frac{x^k}{k!}\frac{t^l}{l!}.$$

It was found that a polynomial expansion $\Sigma(a_n/n!)v_{r,n}$ exists in a strip $|t|<\sigma$, where σ is calculated by a Hadamard formula, i.e. $\limsup(re/n)|a_n|^{r/n}=\sigma^{-1}$. This is also a necessary and sufficient criterion for the representation of a solution of (E1) by a Maclaurin series. These results were extensions of the work of P. C. Rosenbloom and D. V. Widder [14], [17].

Our present goal is to establish that analogous criteria hold for a system of $r-1$ partial differential equations ($k=2,\cdots,r$):

$$\text{(S)}\qquad D^k_{X_1}U(X_1,\cdots,X_r)=D_{X_k}U(X_1,\cdots,X_r).$$

The starting point of this research was the parabolic equation

$$\text{(E2)}\qquad \sum_{\nu=2}^{r}A_\nu D_x^\nu u(x,t)=D_tu(x,t),\qquad A_r\neq 0,$$

which is the natural generalization of (E1). (S) and (E2) are related by the following fact: If $P(X_1,\cdots,X_r)$ is a polynomial solution of (S), then $p(x,t):=P(x,A_2t,\cdots,A_rt)$

*Received by the editors November 30, 1982, and in revised form September 1, 1983.

†Pirminweg 10, 7750 Konstanz, Federal Republic of Germany.

will be a polynomial solution of (E2). We define such polynomials as the coefficients of $z^n/n!$ in the expansion

$$\exp(X_1z+\cdots+X_rz^r)=:\sum_{n=0}^{\infty}P_{r,n}(X_1,\cdots,X_r)\frac{z^n}{n!}.$$

By use of Cauchy's rule for multiplying power series, we obtain the explicit expression

$$P_{r,n}(X_1,\cdots,X_r)=n!\sum_{k=0}^{[n/r]}\frac{X_r^k}{k!}\frac{P_{r-1,n-rk}(X_1,\cdots,X_{r-1})}{(n-rk)!}\tag{0.1}$$

where $[n/r]$ means the largest integer $\leq n/r$. The polynomials have several obvious properties

$$P_{1,n}(X_1)=X_1^n,\tag{0.2}$$

$$P_{r,n}(X_1,\cdots,X_r)=n!\sum_{k_1+2k_2+\cdots+rk_r=n}\frac{X_1^{k_1}\cdots X_r^{k_r}}{k_1!\cdots k_r!},\tag{0.3}$$

$$\lambda^nP_{r,n}(X_1,\cdots,X_r)=P_{r,n}(\lambda X_1,\cdots,\lambda^rX_r),\tag{0.4}$$

$$|P_{r,n}(X_1,\cdots,X_r)|\leq P_{r,n}(|X_1|,\cdots,|X_r|),\tag{0.5}$$

$$D_{X_k}P_{r,n}(X_1,\cdots,X_r)=\frac{n!}{(n-k)!}P_{r,n-k}(X_1,\cdots,X_r).\tag{0.6}$$

From (0.6) we see that for fixed r, $P_{r,n}$ is a solution of (S), and we also see that $v_{r,n}(x,t)=P_{r,n}(x,0,\cdots,0,t)$. In this paper we use $\mathbb{K}$ as the field of real or complex numbers, therefore D_zu means the partial derivative of u with respect to $z\in\mathbb{K}$, and $|\cdot|$ stands for the Euclidean norm in $\mathbb{K}$.

1. Strip of convergence. In this section we show that polynomial series

$$\sum_{n=0}^{\infty}\frac{a_n}{n!}P_{r,n}(X_1,\cdots,X_r)\tag{1.0}$$

converge in a strip $|X_r|<\sigma$, where σ is calculated by Hadamard's formula. For a discussion of series (1.0) it is essential to know the behavior of $P_{r,n}$ as $n\to\infty$. In [14] Rosenbloom and Widder created a method to estimate the heat polynomials: $v_{2,n}(x,t)$. For this purpose they used the Poisson representation

$$v_{2,n}(x,t)=\int_{-\infty}^{+\infty}k(x-y,t)y^n\,dy,$$

where $k(x,t)=\exp(-x^2/4t)/(4\pi t)^{1/2}$ is the fundamental solution of the heat equation. This was the beginning of extensive research into other partial differential equations of second order [2], [3], [5], [6], [7], [8], [10], [11], [16], [17]. In [9] the author presented a new method, based only on elementary calculus, for estimating the polynomials $v_{r,n}(x,t)$, in particular the heat polynomials. In this paper we follow the same methods.

LEMMA 1.1. *For* $0<\delta<+\infty$

$$\frac{P_{r,n}(|X_1|,\cdots,|X_r|)}{n!}\leq c_r\frac{(\delta+|X_r|)^{[n/r]}}{[n/r]!},$$

where $c_1\equiv1$, $c_{r+1}=c_r\cdot p_r^r\exp(\delta\cdot p_r^{r+1})$ *with* $p_r=(\delta+|X_r|)/\delta$.

Proof. We use an induction argument. For $r=1$ it follows easily from (0.2). Now, let $r\ge 2$ and $n=rm+s(0\le s\le r-1)$, then from (0.1) we obtain

$$\frac{1}{n!}P_{r,n}(|X_1|,\cdots,|X_r|)=\sum_{k=0}^{m}\frac{|X_r|^k}{k!}\frac{P_{r-1,(m-k)r+s}(|X_1|,\cdots,|X_{r-1}|)}{((m-k)r+s)!}$$

$$\le\sum_{k=0}^{m}\frac{|X_r|^k}{k!}c_{r-1}\frac{(|X_{r-1}|+\delta)^{m-k+\alpha}}{(m-k+\alpha)!},$$

where $\alpha:=[(m-k+s)/(r-1)]$. Since $(m-k)!\alpha!\le(m-k+\alpha)!$ we have

$$\le c_{r-1}\sum_{k=0}^{m}\frac{|X_r|^k}{k!}\frac{\delta^{m-k}}{(m-k)!}\left(\frac{|X_{r-1}|+\delta}{\delta}\right)^{m-k}\frac{(|X_{r-1}|+\delta)^\alpha}{\alpha!}.$$

Since $m-k\le(r-1)(\alpha+1)$ and $1\le(|X_{r-1}|+\delta)/\delta=p_{r-1}$

$$\le c_{r-1}p_{r-1}^{r-1}\sum_{k=0}^{m}\frac{|X_r|^k\delta^{m-k}}{k!(m-k)!}\cdot\frac{(\delta\cdot p_{r-1}^r)^\alpha}{\alpha!}.$$

Since $|z|^\alpha\le\alpha!\exp|z|$, the lemma is thus established. An appeal to Stirling's formula, $n!=n^ne^{-n}(2\pi n)^{1/2}.\,e^{\theta/12n}(0<\theta<1)$, yields

COROLLARY 1.2. *For* $0<\delta<+\infty$, $n=1,2,3,\cdots,|X_k|<M_k$ *for* $k=1,\cdots,r-1$

$$\frac{P_{r,n}(|X_1|,\cdots,|X_r|)}{n!}\le M_0n^{1/2}\left(\frac{(\delta+|X_r|)re}{n}\right)^{n/r},$$

where M_0 *is a constant depending on* δ, r *and* M_k.

LEMMA 1.3. *For* $n=rm+s$, $s\in\{0,\cdots,r-1\}$, $Y_1\ne0\ne Y_r$,

$$0\le Y_kY_r^{-k/r}<+\infty\quad for\ k=1,\cdots,r,$$

$$\frac{P_{r,n}(Y_1,\cdots,Y_r)}{n!}\ge\frac{|Y_r|^m}{m!}\frac{P_{r-1,s}(|Y_1|,\cdots,|Y_{r-1}|)}{s!}.$$

Proof. Note first that the right-hand side is a positive number. By use of relation (0.4), we have

$$\frac{|P_{r,n}(Y_1,\cdots,Y_r)|}{n!}=|Y_r|^{n/r}\frac{P_{r,n}(Y_1Y_r^{-1/r},\cdots,Y_rY_r^{-r/r})}{n!},$$

and from (0.1)

$$\frac{|P_{r,n}(Y_1Y_1^{-1/r},\cdots,Y_rY_r^{-r/r})|}{n!}\ge\frac{P_{r-1,s}(Y_1Y_r^{-1/r},\cdots,Y_{r-1}Y_r^{-(r-1)/r})}{m!s!}.$$

But this completes the proof.

COROLLARY 1.4. *For* $n=1,2,3,\cdots$.

$$Y_1\ne0\ne Y_r,\ 0\le Y_kY_r^{-k/r}<+\infty\quad for\ k=1,\cdots,r,$$

$$\frac{|P_{r,n}(Y_1,\cdots,Y_r)|}{n!}\ge m_0n^{-1/2}\left(\frac{|Y_r|re}{n}\right)^{n/r},$$

where m_0 *is a constant depend on* r *and* Y_k, *for* $1\le k\le r-1$.

Proof. This fact is easily proved by Stirling's formula.

Next, we will show that if series (1.0) converges at a point $Y := (Y_1,\cdots,Y_r)$, where Y satisfies the hypothesis of corollary (1.4), then it converges in a whole strip $Z_\sigma^r := \{Y\in\mathbb{K}^r : |Y_r|<\sigma\}$.

THEOREM 1.5. *If the series* (1.0) *converges at* $Y := (Y_1,\cdots,Y_r)$, *where* $Y_1\neq 0\neq Y_r$ *and* $0\leq Y_k Y_r^{-k/r}<+\infty$ *for* $k=1,\cdots,r$, *then*

i) *the series converges absolutely in the strip* $Z^r_{|Y_r|}$;

ii) *the series converges uniformly in any compact region of* $Z^r_{|Y_r|}$;

iii)

$$a_n = O\left(n^{1/2}\left(\frac{n}{re|Y_r|}\right)^{n/r}\right), \qquad n\to\infty.$$

Proof. Since the series (1.0) is assumed to converge at $(Y_1,\cdots,Y_r)$, the general term tends to zero as $n\to\infty$. It follows that by Corollary 1.4

$$\left|a_n m_0 n^{-1/2}\left(\frac{|Y_r|re}{n}\right)^{n/r}\right|\to 0, \qquad n\to\infty,$$

and hence

$$|a_n|\leq cn^{1/2}\left(\frac{n}{re|Y_r|}\right)^{n/r}, \qquad n\geq 1$$

for some constant c which depends on Y_k. By use of Corollary 1.2 we have for $|X_k|<M_k$

$$\left|\frac{a_n}{n!}P_{r,n}(X_1,\cdots,X_r)\right|\leq Cn\left(\frac{\delta+|X_r|}{|Y_r|}\right)^{n/r}, \qquad n\geq 1,$$

where C is a constant which depends on Y_k, δ and M_k. But the series

$$\sum_{n=1}^{\infty} n\left(\frac{\delta+|X_r|}{|Y_r|}\right)^{n/r}$$

converges for $|X_r|<|Y_r|-\delta$. Since δ may be taken arbitrarily small, an application of Weierstrass M-test completes the proof.

We are now in a position to establish the principal result of this section.

THEOREM 1.6. *If*

$$(*) \qquad \limsup_{n\to\infty}\frac{re}{n}|a_n|^{r/n}=\sigma^{-1}<+\infty$$

then the series

$$\sum_{n=0}^{\infty}\frac{a_n}{n!}P_{r,n}(X_1,\cdots,X_r)$$

has the following properties:

i) *if it converges in a strip* Z_τ^r, *then* $\tau\leq\sigma$;

ii) *it converges absolutely and uniformly compact by in* Z_σ^r.

Proof. If $0<\sigma_0<\sigma$, then the assumption (*) implies that, for $n \geq n(\sigma_0)$,

$$|a_n| \leq \left(\frac{n}{re\sigma_0}\right)^{n/r}.$$

With an application of Corollary 1.2 we obtain

$$\left|\frac{a_n}{n!}P_{r,n}(X_1,\cdots,X_r)\right| \leq M_0 n^{1/2}\left(\frac{\delta+|X_r|}{\sigma_0}\right)^{n/r}, \qquad n\geq 1.$$

But it follows immediately that

$$\sum_{n=1}^{\infty} n^{1/2}\left(\frac{\delta+|X_r|}{\sigma_0}\right)^{n/r}$$

converges for $|X_r|<\sigma_0-\delta$. Consequently, for δ arbitrarily near 0 and σ_0 arbitrarily near σ, conclusion ii) is established. Now, suppose that the series (1.0) converges in the whole strip Z^r_τ, where $\tau>\sigma$. Then in particular it would converge at $Y=(Y_1,\cdots,Y_r)$, with $Y_1\neq 0\neq Y_r$ and $0\leq Y_k Y_r^{-k/r}\leq +\infty$, where $\sigma<|Y_r|<\tau$. By Theorem 1.5, we obtain

$$\limsup_{n\to\infty} \frac{re}{n}|a_n|^{r/n} \leq |Y_r|^{-1}.$$

The desired contradiction is evident and i) is proved.

We can now say that σ is the radius of convergence of the series (1.0). But there exist series some of which converge for all Y, while others fail to converge when $Y_r\neq 0$. From [9, p. 645] we have an example of convergence outside the strip Z^r_σ. Let, for $n=rm+s$,

$$a_n=\begin{cases} 0, & s=0, \\ \dfrac{(rm+s)!}{((r-1)m)!}, & s\neq 0; \end{cases} \tag{1.7}$$

then the strip of convergence is bounded, $r(r/(r-1))^{r-1}=\sigma^{-1}$, but series $\Sigma(a_n/n!)P_{r,n}(0,\cdots,0,X_r)$ converges over the whole X_r-axis.

2. Related series expansions. In (0.6) we noted that the polynomials $P_{r,n}$ satisfy the equation

$$D_{X_k}P_{r,n}=\frac{n!}{(n-k)!}P_{r,n-k},$$

and therefore they are solutions of the system

$$D^k_{X_1}U=D_{X_k}U, \qquad k=2,\cdots,r. \tag{S}$$

From the §1, we know that the polynomial series $\Sigma(a_n/n!)P_{r,n}$ converges in a strip Z^r_σ. Since, every finite sum of the polynomials $P_{r,n}$ satisfies (S), we would expect an analogous result for an infinite series.

LEMMA 2.1. *If*

i) $U(X_1,\cdots,X_r):=\displaystyle\sum_{n=0}^{\infty}\frac{a_n}{n!}P_{r,n}(X_1,\cdots,X_r)$,

ii) $\limsup_{n\to\infty} \frac{re}{n} |a_n|^{r/n} = \sigma^{-1} < +\infty$,

then U satisfies (S) *and is analytic, as a function of r variables, in the whole strip Z_σ^r. The coefficients satisfy*

(iii) $a_{kn} = D_{X_k}^n U(0,\cdots,0)$.

Proof. By Theorem 1.6 the series in i) converges uniformly in any compact region of Z_σ^r. Hence $U(X_1,\cdots,X_r)$ is an analytic function in Z_σ^r. Setting all $X_\nu=0$, for $\nu\neq k$, we obtain from (0.3) that $P_{r,n}\equiv 0$ for $n\neq km$. But $P_{r,km}(0,\cdots,0,X_k,0,\cdots,0)=((km)!/m!)X_k^m$. Now we have

$$U(0,\cdots,0,X_k,0,\cdots,0)=\sum_{n=0}^{\infty}\frac{a_{kn}}{(kn)!}\frac{(kn)!}{n!}X_k^n$$

and iii) follows from Taylor's formula. For fixed integer p

$$\limsup_{n\to\infty}\frac{\mathrm{re}}{n}|a_{n+p}|^{r/n}=\sigma^{-1}$$

which follows directly from ii). By applying Theorem 1.6 the series $\Sigma(a_{n+p}/n!)P_{r,n}(X_1,\cdots,X_r)$ converges uniformly in any compact region of Z_σ^r. Consequently, we obtain in Z_σ^r:

$$\sum_n (a_{n+p}/n!)P_{r,n}=\sum_n D_{X_p}(a_n/n!)P_{r,n}=D_{X_p}\sum_n (a_n/n!)P_{r,n},$$

so that U is a solution of (S).

There is a close relation between polynomial series and analytic function. In the following, we will show that the expansion i) holds in some infinite strip $|X_r|<\sigma$ if and only if $U(X_1,\cdots,X_r)$ is analytic, as a function of r variables, at some point of the X_1-axis and satisfies the system (S).

THEOREM 2.2. *Under the conditions of Lemma* 2.1 *$U(X_1,\cdots,X_r)$ has the Maclaurin expansion*

$$U(X_1,\cdots,X_r)=\sum_{m_1=0}^{\infty}\cdots\sum_{m_r=0}^{\infty}a_{m_1+2m_2+\cdots+rm_r}\frac{X_1^{m_1}}{m_1!}\cdots\frac{X_r^{m_r}}{m_r!}.$$

Proof. From the hypothesis, it follows that the series $\Sigma_n(a_n/n!)P_{r,n}(|X_1|,\cdots,|X_r|)$ converges absolutely for $|X_r|<\sigma$. Hence, by (0.3)

$$U(X_1,\cdots,X_r)=\sum_{n=0}^{\infty}a_n\sum_{k_1+2k_2+\cdots rk_r=n}\frac{X_1^{k_1}}{k_1!}\cdots\frac{X_r^{k_r}}{k_r!}$$

$$=\sum_{k_1=0}^{\infty}\cdots\sum_{k_r=0}^{\infty}a_{k_1+2k_2+\cdots+rk_r}\frac{X_1^{k_1}}{k_1!}\cdots\frac{X_r^{k_r}}{k_r!},$$

and the result is established.

THEOREM 2.3. *If for $k=1,\cdots,r$ $\varepsilon_k>0$, $|X_k|<\varepsilon_k$*

i) $D_{X_1}^k U=D_{X_k}U,\ U(X_1,\cdots,X_r)=\sum_{m_1=0}^{\infty}\cdots\sum_{m_r=0}^{\infty}a_{m_1,\cdots,m_r}\frac{X_1^{m_1}}{m_1!}\cdots\frac{X_r^{m_r}}{m_r!}$, *then $U(X_1,\cdots,X_r)$ can be extended to an analytic function $\tilde{U}$ in the whole strip $Z^r_{\varepsilon_r}$,*

$$\text{ii)}\quad \tilde{U}(X_1,\cdots,X_r)=\sum_{n=0}^{\infty}(a_n/n!)P_{r,n}(X_1,\cdots,X_r),$$

where $a_n := a_{n,0,\cdots,0}$.

Proof. Since a power series may be differentiated term by term, we have for $|X_k|<\varepsilon_k$

$$D_{X_k}U=\sum_{m_1=0}^{\infty}\cdots\sum_{m_r=0}^{\infty}a_{m_1,\cdots,m_k+1,\cdots,m_r}\frac{X_1^{m_1}}{m_1!}\cdots\frac{X_r^{m_r}}{m_r!}.$$

Using $(D^k_{X_1}-D_{X_k})U(X_1,\cdots,X_r)\equiv 0$. For $|X_k|<\varepsilon_k$, we obtain the formula

$$\begin{aligned} a_{m_1,m_2,\cdots,m_r} &= a_{m_1+2m_2,0,m_3,\cdots,m_r} & \left(D^2_{X_1}=D_{X_2}\right),\\ &\ \vdots & \vdots\qquad\\ &= a_{m_1+2m_1+\cdots+rm_r,0,\cdots,0} & \left(D^k_{X_1}=D_{X_r}\right). \end{aligned}$$

Since the series in ii) converges for $|X_k|<\varepsilon_k$ (and hence absolutely for $0<Y_k<\varepsilon_k$) we have

$$\begin{aligned} U(Y_1,\cdots,Y_r) &= \sum_{m_1=0}^{\infty}\cdots\sum_{m_r=0}^{\infty}a_{m_1+2m,+\cdots+r_{mr}}\frac{Y_1^{m_1}}{m_1!}\cdots\frac{Y_r^{m_r}}{m_r!}\\ &= \sum_{n=0}^{\infty}a_n\sum_{m_1+2m_2+\cdots+rm_r=n}\frac{Y_1^{m_1}}{m_1!}\cdots\frac{Y_r^{m_r}}{m_r!}\\ &= \sum_{n=0}^{\infty}(a_n/n!)P_{r,n}(Y_1,\cdots,Y_r). \end{aligned}$$

But by Theorem 1.5 the latter series converges absolutely and uniformly compact in $Z^r_{Y_r}$. Since Y_r may be taken arbitrarily near ε_r, we obtain an analytic function $\tilde{U}(X_1,\cdots,X_r)$ which is the continuation of $U(X_1,\cdots,X_r)$. This completes the proof.

We know that an analytic function $U(X_1,\cdots,X_r)$, which satisfies the system (S), has an expansion of polynomials $P_{r,n}$. And conversely, a polynomial expansion of $P_{r,n}$ possesses a Maclaurin series which performs (S).

THEOREM 2.4. *If*

$$(*)\qquad \limsup_{n\to\infty}\frac{re}{n}|a_n|^{r/n}=\sigma^{-1}<+\infty$$

then the series

$$\sum_{m_1=0}^{\infty}\cdots\sum_{m_r=0}^{\infty}a_{m_1+2m_2+\cdots+rm_r}\frac{X_1^{m_1}}{m_1!}\cdots\frac{X_r^{m_r}}{m_r!}$$

converges in the strip Z^r_σ and, except perhaps when $X_1=0$, diverges for $|X_r|>\sigma$.

Proof. The convergence of the power series follows immediately from Lemma 2.1 and Theorem 2.2. Suppose now that the Maclaurin series converges at some point

$Y=(Y_1,\cdots,Y_r)$ with $Y_1\neq 0$ and $|Y_r|>\sigma$. Since the general term

$$a_{m_1+2m_2+\cdots+rm_r}\frac{Y_1^{m_1}}{m_1!}\cdots\frac{Y_r^{m_r}}{m_r!}$$

tends to zero, for $m_i\to\infty$, we have, in particular,

$$\left|a_{s+rm_r}\frac{Y_1^sY_r^{m_r}}{s!m_r!}\right|\to 0, \qquad m_r\to\infty,$$

and hence

$$\limsup_{n\to\infty}\frac{e}{n}|a_{s+rn}|^{1/n}<|Y_n|^{-1}.$$

Combining these r inequalities yields

$$\limsup_{n\to\infty}\frac{re}{n}|a_n|^{r/n}<|Y_r|^{-1}<\sigma^{-1}.$$

But this contradicts the condition of convergence for the related polynomial series.

3. Classical Cauchy problems. Before we start our discussion about initial-value problems, we intend to show some general aspects of transformations.

Setting $X_1=:x$, for $k\geq 2X_k=G_k(t)$, with $G_k\in C^1(I)$, I an open interval, $G_k'=g_k$, then we have for an expansion $U=\Sigma(a_n/n!)P_{r,n}$

$$D_x^kU(x,G_2(t),\cdots,G_r(t))=D_{X_1}^kU(X_1,\cdots,X_r),$$

$$D_tU(x,G_2(t),\cdots,G_r(t))=\sum_{\nu=2}^r g_\nu(t)D_{X_\nu}(X_1,\cdots,X_r).$$

If the series converges in Z_σ^r, then $u(x,t):=U(x,G_2(t),\cdots,G_r(t))$ satisfies the equation.

$$\sum_{\nu=2}^r g_\nu(t)D_x^\nu u(x,t)=D_tu(x,t)$$

for all $(x,t)\in\mathbb{K}\times I$ with $|G_r(t)|<\sigma$. It is also remarkable that there is no restriction of singularities which could result from $g_\nu(t)$. In the following we confine our research to equation

$$\text{(E2)}\qquad L_Xu(x,t):=\sum_{\nu=2}^r A_\nu D_x^\nu u(x,t)=D_tu(x,t), \qquad A_r\neq 0$$

at which we combine the results of [9] and [15]. Some other important aspects of (E2) are discussed in [4], [12], [13]. If we define $V_{r,n}(x,t)=P_{r,n}(x,A_2t,\cdots,A_rt)$, we obtain explicit expression

$$V_{r,n}(x,t)=n!\sum_{n_1+2n_2+\cdots+rn_r=n}\frac{A_2^{n_2}}{n_2!}\cdots\frac{A_r^{n_r}}{n_r!}\frac{x^{n_1}}{n_1!}t^{n_2+\cdots n_r}.$$

Now, we transfer the most important results from the first two sections.

THEOREM 3.1. *If*

$$(*)\qquad \limsup_{n\to\infty} \frac{re}{n}|a_n|^{r/n}\le|A_r\sigma|^{-1}<+\infty$$

then the series

$$\sum_{n=0}^{\infty}\frac{a_n}{n!}V_{r,n}(x,t)$$

i) *converges absolutely in* $Z_\sigma^2:=\{(x,t)\in\mathbb{K}^2\colon |t|<\sigma\}$ *and converges uniformly in any compact region of* Z_σ^2;
ii) *is an analytic function of x and t, and it satisfies* (E2)
iii) *has a Maclaurin representation in* Z_σ^2

$$\sum_{n,m=0}^{\infty} b_{m,n}\frac{x^m}{m!}\frac{t^n}{n!},$$

where $b_{m,n}$ *is given by*

$$n!\sum_{k_2+k_3+\cdots+k_r=n}\frac{A_2^{k_2}}{k_2!}\cdots\frac{A_r^{k_r}}{k_r!}a_{m+2k_2+\cdots+rk_r}.$$

Proof. i) and ii) are immediate consequences of Theorem 1.6 and Lemma 2.1. We only have to show the determination of the coefficients $b_{m,n}$. But this follows from Theorem 2.2 since

$$\sum_{n=0}^{\infty}(a_n/n!)V_{r,n}(x,)$$

$$=\sum_{n=0}^{\infty}a_n\sum_{k_1+\cdots+rk_r=n}\frac{A_2^{k_2}}{k_2!}\cdots\frac{A_r^{k_r}}{k_r!}\frac{x^{k_1}}{k_1!}t^{k_2+\cdots+k_r}$$

$$=\sum_{k=0}^{\infty}\sum_{m=0}^{\infty}\left(m!\sum_{k_2+\cdots+k_r=m}\frac{A_2^{k_2}}{k_2!}\cdots\frac{A_r^{k_r}}{k_r!}a_{k+2k_2+\cdots+rk_r}\right)\frac{x^k}{k!}\frac{t^m}{m!}.$$

Now we can apply our results to solve analytic Cauchy problems of (E2). For that we need to define the growth of an entire function [1, p.11]: $f(x)=\Sigma_m(a_m/m!)x^m$ has growth $\{\alpha,\beta\}$ if and only if $\limsup_{m\to\infty}(e/m)^{\alpha-1}|a_m|^{\alpha/m}<\alpha\beta$. The first classical Cauchy problem consist in finding a solution of

$$(\text{P1})\qquad L_xu(x,t)=D_tu(x,t),\ u(x,0)=f(x).$$

A solution is given by:

COROLLARY 3.2. *For* $\sigma>0$ *and* $f(x)=\Sigma_m(a_m/m!)x^m$

i) *If f has growth* $\{r/(r-1),((r-1)/r)(r\sigma)^{1/(1-r)}\}$, *then*

$$u(x,t):=\sum_{n=0}^{\infty}L_x^nf(x)\frac{t^n}{n!}$$

is defined in Z_σ^2 *and is a solution of* (P1) *for* $|t|<\sigma$.

ii) *If*

$$u(x,t)=\sum_{m,n=0}^{\infty} b_{m,n}\frac{x^m}{m!}\frac{t^n}{n!}$$

satisfies (P1) *for* $|t|<\sigma$, *then* $u(x,0)$ *has growth* $\{r/(r-1),((r-1)/r)(r\sigma)^{1/(1-r)}\}$.

Proof. Since $L_x^n=(A_rD_x^r+\cdots+A_2D_x^2)^n$, we obtain

$$L_x^n=n!\sum_{k_2+\cdots+k_r=n}\frac{A_2^{k_2}}{k_2!}\cdots\frac{A_r^{k_r}}{k_r!}D_x^{2k_2+\cdots+rk_r}.$$

We derive from this that

$$L_x^nf(x)=\sum_{m=0}^{\infty} n!\sum_{k_2+\cdots+k_r=n}\frac{A_2^{k_2}}{k_2!}\cdots\frac{A_r^{k_r}}{k_r!}a_{m+2k_2+\cdots+rk_r}\frac{x^m}{m!}.$$

Now, with an application of Theorem 3.1, i) follows from the equivalence of

$$\limsup_{n\to\infty}\frac{re}{n}|a_n|^{r/n}\leq|A_r\sigma|^{-1}$$

and

$$\limsup_{n\to\infty}\left(\frac{e}{n}\right)^{1/(r-1)}|a_n|^{r/n(r-1)}\leq|rA_r\sigma|^{1/(1-r)}.$$

We have seen, that a Maclaurin series has a representation of polynomials $V_{r,n}$ (Theorems 2.3, 3.1). Thus we obtain the growth of $u(x,0)$ from the last two equations. This completes the proof.

Corollary 3.2 shows us that a solution of (P1) exists over the whole x-axis. It is remarkable that a solution of the Cauchy problem

$$A_0w(x,t)+A_1D_xw(x,t)+L_xw(x,t)=D_tw(x,t),w(x,0)=\sum_{m=0}^{\infty}(a_m/m!)x^m$$

has the same property, because $w(x,t):=u(x+A_1t,t)\exp A_0t$ is the solution of the latter problem if $u(x,t)$ solves (P1) with $f(x)=\Sigma(a_m/m!)x^m$. An explicit expression is given by

$$w(x,t)=\exp A_0t\sum_{k=0}^{\infty}\frac{a_k}{k!}V_{r,k}(x+A_1t,t)$$

$$=\sum_{n=0}^{\infty}\sum_{k=0}^{\infty}\binom{n}{k}A_0^kL_x^{n-k}f(x+A_1t)\frac{t^n}{n!}.$$

Now, let us return to the equation (E2). The second classical Cauchy problem is described by

(P2) $\quad L_xu(x,t)=D_tu(x,t),\ D_x^su(0,t)=g_s(t),\quad$ for $s=0,\cdots,r-1,$

where $g_s(t)$ is an analytic function.

For the special case $(A_i=0, 2\leq i\leq r-1)$, the solution is given in [9, p.649] by

$$u(x,t)=\sum_{m=0}^{\infty}\sum_{s=0}^{r-1}g_s(t)\frac{x^{rm+s}}{(rm+s)!}.$$

This can also be written as

$$u(x,t)=\sum_{m=0}^{\infty} \mathrm{x}_m\Delta_t^m \mathrm{g}(t),$$

where $\mathrm{x}_m:=(x^{rm}/(rm)!,\cdots,x^{rm+r-1}/(rm+r-1)!)$, $\mathrm{g}:=(g_0,\cdots,g_{r-1})$ and Δ_t is a $(r\times r)$-matrix with $d_{ik}=\delta_{ik}D_t$, δ_{ik} is Kronecker's index.

In general we have to construct a suitable matrix. Let us assume that

$$u(x,t)=\sum_{m,n=0}^{\infty} b_{m,n}\frac{x^m}{m!}\frac{t^n}{n!}$$

is a solution of (P2). Then from Theorem 3.1iii

$$b_{m,n}=n!\sum_{n_2+\cdots+n_r=n}\frac{A_2^{n_2}}{n_2!}\cdots\frac{A_r^{n_r}}{n_r!}b_{m+2n_2+\cdots+rn_r,0},$$

we immediately obtain

$$\text{i)}\quad b_{m,n+1}=\sum_{\nu=2}^{r}A_\nu b_{m+\nu,n}.$$

Consider now an $(r\times r)$-matrix Δ_t, so that

$$u(x,t)=\sum_{m=0}^{\infty}\mathrm{x}_m\Delta_t^m\mathrm{g}(t).$$

Since the power series converges, and hence absolutely at a point(x_0,t_0), we have

$$\sum_{i=1}^{k}d_{s+1,i}^{(m)}g_{i-1}(t)=\sum_{n=0}^{\infty}b_{rm+s,n}\frac{t^n}{n!}.$$

Using $\Delta_t^{m+1}=\Delta_t\Delta_t^m$, we obtain

$$\Delta_t^{m+1}\mathrm{g}(t)=\Delta_t\sum_{n=0}^{\infty}\left(b_{rm,n}\frac{t^n}{n!},\cdots,b_{rm+r-1,n}\frac{t^n}{n!}\right).$$

From this we derive

$$\text{ii)}\quad b_{r(m+1)+s,n}\frac{t^n}{n!}=\sum_{k=1}^{r}d_{s+1,k}^{(1)}b_{rm+k-1,n}\frac{t^n}{n!}.$$

Combining i) with ii), we obtain a representation of the matrix Δ_t.

COROLLARY 3.3. *If*

$$\limsup_{n\to\infty}\frac{re}{n}|a_n|^{r/n}=|A_r\sigma|^{-1}<+\infty,$$

$$b_{s,n}=n!\sum_{n_2+\cdots+n_r=n}\frac{A_2^{n_2}}{n_2!}\cdots\frac{A_r^{n_r}}{n_r!}a_{s+2n_2+\cdots+rn_r},$$

$$g_s(t)=\sum_{n=0}^{\infty}(b_{s,n}/n!)t^n,$$

then

$$u(x,t)=\sum_{m=0}^{\infty}\mathbf{x}_m\Delta_t^m\mathbf{g}(t)=\sum_{n=0}^{\infty}\frac{a_n}{n!}V_{r,n}(x,t)$$

is the solution of (P2) *which exists in the whole strip* Z_σ^2. Δ_t *is given by* $(d_{ij}^{(1)}\in\Delta_t)$:

$$-d^{(1)}_{s+1,k}=\sum_{\nu=0}^{s-1}\frac{A_{r-s+\nu}}{A_r}d^{(1)}_{\nu+1,k}+\sum_{\nu=s}^{r-1}\frac{A_{\nu-s}}{A_r}$$

where $A_{-r+1}=\cdots=A_{-1}=A_1=0$, *and* $A_0=-D_t$.

Proof. From Theorem 3.1 we know

$$L_xu(x,t)=D_tu(x,t),$$
$$u(x,t)=\sum_{m,n=0}b_{m,n}\frac{x^m}{m!}\frac{t^n}{n!},$$

and hence $D_x^su(0,t)=g_s(t)$. We have only to show the determination of the coefficients $d^{(1)}_{s+1,k}$. But this follows directly from a comparison between

$$\text{i)}\quad A_rb_{r(m+1)+s,n}=b_{rm+s,n+1}-\sum_{\nu=2}^{r-1}A_\nu b_{rm+s+\nu,n}$$

and

$$\text{ii)}\quad b_{r(m+1)+s,n}\frac{t^n}{n!}=\sum_{k=1}^{r}d^{(1)}_{s+1,k}b_{rm+k-1,n}\frac{t^n}{n!}.$$

REFERENCES

[1] R. P. Boas, *Entire Functions*, Academic Press, New York, 1954.

[2] L. R. Bragg, *The radial heat polynomials and related functions*, Trans. Amer. Math. Soc., 119 (1965), pp. 270–290.

[3] F. M. Cholewinski and D. T. Haimo, *Expansions in terms of laguerre heat polynomials and of their Appell transforms*, J. Analyse Math., 24 (1971), pp. 285–322.

[4] J. B. Diaz and C. S. Means, *An initial value problem for a class of higher order partial differential equations related to the heat equation*, Ann. Matematica, 97 (1973), pp. 113–187.

[5] D. T. Haimo, *Expansions in terms of generalized heat polynomials and of their Appell transforms*, J. Math. Mech., 15 (1966), pp. 735–758.

[6] ———, *Series expansions of generalized temperature functions in n dimensions*, Canad J. Math., 18 (1966), pp. 794–802.

[7] ———, *Series representations of generalized temperature functions*, SIAM J. Appl. Math., 15 (1967), pp. 359–367.

[8] ———, *Series expansions for dual Laguerre temperatures*, Canad. J. Math., 24 (1972), pp. 1145–1153.

[9] H. Kemnitz, *Polynomial expansions for solutions of* $D_x^ru(x,t)=D_tu(x,t)$, $r=2,3,4,\cdots$, this Journal, 13 (1982), pp. 640–650.

[10] C. Y. Lo, *Polynomial expansions of solutions of* $u_{xx}+\varepsilon^2u_{tt}=u_t$, J. Reine Angew. Math., 252 (1972), pp. 88–103.

[11] ———, *Series expansions of solutions of* $u_{xx}+u_{yy}+\varepsilon^2u_{tt}=u_t$, this Journal, 3 (1972), pp. 461–473.

[12] H. M. Riedl, *Integraloperatoren und vollständige Familien von Lösungen bei parabolischen Gleichungen höherer Ordnung*, Dissertation, Konstanz 1977.

[13] E. L. Roetman, *Some observations about an odd order parabolic equation*, J. Differential Equations, 9 (1971), pp. 335–345.

[14] P. C. Rosenbloom and D. V. Widder, *Expansions in terms of heat polynomials and associated functions*, Trans. Amer. Math. Soc., 92 (1959), pp. 220–266.

[15] W. Watzlawek, *Über die Lösung des Cauchyschen Problems bei linearen partiellen Differentialgleichungen beliebiger Ordnung mittels Lie-Reihen*, Monatshefet für Mathematik, 70 (1966), pp. 366–376.

[16] D. V. Widder, *Series expansions of solutions of the heat equation in n dimensions*, Ann. Math. Pura Appl., (4) 55 (1961), pp. 389–409.

[17] ———, *Analytic solutions of the heat equation*, Duke Math. J., 29 (1962), pp. 497–503.

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

A NOTE ON SUMMABILITY AND ASYMPTOTICS*

K. SONI† AND R. P. SONI†

Abstract. The asymptotic expansions of the Fourier and the $H_\nu^{(1)}$-transform are discussed. The objective is to investigate the necessity and the sufficiency of the conditions under which such expansions have been obtained in recent years.

1. Introduction. The summability technique introduced by Olver [2] to obtain the asymptotic expansion of a class of integral transforms together with the error terms is well known. In certain cases, his basic ideas can be used without explicitly introducing a summability kernel. In particular, the Fourier transform expansion can be obtained from the Laplace transform expansion. Similarly, the Hankel and the Y_ν-transform expansions can be obtained from the corresponding K_ν-transform expansion. However, if we use this approach, and compare our conditions with those under which the same expansions have been obtained earlier, some interesting questions arise. For example, Olver [2] gives the Fourier transform expansion of functions $f(t)$ which belong to $C^p(0, \infty)$ under the condition that each of the integrals

$$I_k(x) = \int_1^\infty e^{ixt} f^{(k)}(t)\,dt, \tag{1.1}$$

$k = 0, 1, \cdots, p$, converge uniformly for all x sufficiently large. If we derive the same expansion from the corresponding Laplace transform expansion by using Abel's limit, we find that we do not require the convergence of the integrals $I_k(x)$, $k = 1, 2, \cdots, p-1$. It is enough to assume that $I_0(x)$ and $I_p(x)$ converge uniformly for all x sufficiently large. These conditions appear to be less restrictive than Olver's. On the other hand, if we want that similar expansion should be valid for every q, $1 \leqq q \leqq p$, we see that Olver's conditions are also necessary. Therefore, we ask the following question: Does the convergence of the integrals $I_0(x)$ and $I_p(x)$, $p > 1$, imply the convergence of the integrals $I_k(x)$ for $k = 1, 2, \cdots, p-1$? Here we should mention that we do not impose any conditions on the behavior of $f^{(k)}(t)$ as $t \to \infty$. In §2 we prove that if $I_0(x)$ and $I_p(x)$ converge at least for p distinct values of x different from zero, then $f^{(k)}(t) \to 0$ as $t \to \infty$ and consequently all $I_k(x)$, $k \leqq p$, converge for those x. Whether the conclusion holds when the number of such points is less than p, remains an open question.

In §3 we discuss similar questions in connection with the asymptotic expansion of the $H_\nu^{(1)}$-transform. Furthermore, due to the singularity of $H_\nu^{(1)}(t)$ near the origin, the remainder in the $H_\nu^{(1)}$-transform expansion cannot in general be expressed as an integral. We give an alternative form of the remainder which, we believe, is new. Then, by using a connection formula for the Bessel functions, we obtain the asymptotic expansion of the Hankel and the Y_ν-transform from that of the $H_\nu^{(1)}$-transform.

In §4, we state analogues of Watson's lemma for the Hankel and the Y_ν-transforms and compare our conditions with those given by Wong [7]. Finally we give some examples to emphasize the significance of certain conditions in the statement of these results.

* Received by the editors April 6, 1978, and in final revised form October 12, 1983.

† Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996.

2. Notation and main result on Fourier integrals. Unless otherwise indicated, we assume that $f(t)$ is complex valued and belongs to $C^p(0,\infty)$. Following the notation used in some recent papers on asymptotics, we write

$$f(t)=f_{n-1}(t)+R_n(t) \tag{2.1}$$

where

$$f_{n-1}(t)=\sum_{k=0}^{n-1} a_k t^{\lambda_k}, \qquad a_k\neq 0, \quad -1<\lambda_0<\lambda_1<\cdots<\lambda_{n-1}. \tag{2.2}$$

Furthermore, we assume that

$$\begin{aligned} R_n(t)&=O(t^{\lambda_n}), && t\to 0^+, \quad \lambda_{n-1}<\lambda_n,\\ R_n^{(p)}(t)&=O(t^{\lambda_n-p}), && t\to 0^+, \quad p<\lambda_n+1. \end{aligned} \tag{2.3}$$

We say that an integral converges if it converges either absolutely or as an improper Riemann integral.

THEOREM 1. *Let y be a fixed positive number and let $\lambda_{n-1}<p<\lambda_n+1$. If $I_k(y)$ are defined by* (1.1), *then*

(2.4)

$$\int_0^\infty e^{iyt}f(t)\,dt=\sum_{k=0}^{n-1} y^{-(\lambda_k+1)}e^{i(\lambda_k+1)\pi/2}\Gamma(\lambda_k+1)a_k+y^{-p}e^{ip\pi/2}\int_0^\infty e^{iyt}R_n^{(p)}(t)\,dt$$

if and only if the integrals $I_0(y)$ and $I_p(y)$ converge.

THEOREM 2. *If $I_0(y)$ and $I_p(y)$ converge for p distinct, positive values of y, then $I_k(y)$, $k=1,2,\cdots,p-1$, also converge for those y.*

Although we make no assumptions regarding the behavior of $f(t)$ as $t\to\infty$, the convergence of $I_0(y)$ and $I_p(y)$ for p distinct values of y implies that $f^{(k)}(t)\to 0$ as $t\to\infty$ for $k=0,1,\cdots,p-1$. The practical significance of our results is that if we know that $I_0(y)$ converges for at least p positive values of y and $I_p(y)$ converges uniformly for all y large enough, we can conclude that the expansion (2.4) is asymptotic. We should mention here that the condition $p>\lambda_{n-1}$ in Theorem 1 is not essential. It can be removed by using the complementary incomplete gamma functions as in an earlier result of one of the authors [3, Thm. 1]. Also, the lower limit in the integrals $I_k(x)$ is not important. It can be replaced by any positive number.

Proof of Theorem 1. Since $R_n^{(p)}(t)=f^{(p)}(t)-f_{n-1}^{(p)}(t)$, the necessity of the condition is clear. We will show that it is sufficient. Let

$$\psi(t)=\int_t^\infty e^{iyu}f^{(p)}(u)\,du, \tag{2.5}$$

$$\begin{aligned} f^{(p-1)}(t)-f^{(p-1)}(1)&=\int_1^t e^{-iyu}e^{iyu}f^{(p)}(u)\,du\\ &=\psi(1)e^{-iy}-\psi(t)e^{-iyt}-iy\int_1^t e^{-iyu}\psi(u)\,du. \end{aligned} \tag{2.6}$$

Since $\psi(t)\to 0$ as $t\to\infty$, it follows that $f^{(p-1)}(t)=o(t)$ as $t\to\infty$. Therefore

$$f^{(k)}(t)=o(t^{p-k}), \qquad t\to\infty, \quad k=0,1,\cdots,p-1. \tag{2.7}$$

Now let $s=x-iy$, $x>0$. By (2.3), $R_n^{(k)}(t)=O(t^{\lambda_n-k})$ as $t\to 0$ and by (2.7), $f^{(k)}(t)=o(e^{xt})$ for every $x>0$ as $t\to\infty$. Therefore, by a familiar technique in asymptotics which involves replacing $f(t)$ by $f_{n-1}(t)+R_n(t)$ and then integrating by parts p times (see, for example, [3]), we obtain

$$\int_0^\infty e^{-st}f(t)\,dt=\sum_{k=0}^{n-1} s^{-(\lambda_k+1)}\Gamma(\lambda_k+1)a_k+s^{-p}\int_0^\infty e^{-st}R_n^{(p)}(t)\,dt. \tag{2.8}$$

The equality (2.4) follows from (2.8). As $x\to 0^+$, $s=x-iy\to e^{-i\pi/2}y$ and the integrals in (2.8) approach the corresponding integrals in (2.4). The last assertion is justified by the convergence of the integrals $I_0(y)$, $I_p(y)$ and the well-known Abelian implication (see [2, Lemma 2]).

Proof of Theorem 2. Without loss of generality, we may assume that $f(t)$ belongs to $C^p(0,\infty)$ and as $t\to 0^+$, $f(t)=O(t^{\lambda_0})$ where $\lambda_0>p$. Assume that $I_0(y)$ and $I_p(y)$ converge for $y=y_j$, $j=1,2,\cdots,p$. By Theorem 1,

$$\int_0^\infty e^{iyt}f(t)\,dt=y^{-p}e^{ip\pi/2}\int_0^\infty e^{iyt}f^{(p)}(t)\,dt, \qquad y=y_j. \tag{2.9}$$

Let $0<a<\infty$. By using integration by parts p times,

$$\int_0^a e^{iyt}f(t)\,dt=-e^{iya}\Psi(y,a)+y^{-p}e^{ip\pi/2}\int_0^a e^{iyt}f^{(p)}(t)\,dt, \tag{2.10}$$

where

$$\Psi(y,a)=\sum_{k=1}^{p} y^{-k}e^{ik\pi/2}f^{(k-1)}(a). \tag{2.11}$$

By comparing (2.9) and (2.10) we conclude that $\Psi(y,a)\to 0$ as $a\to\infty$ for $y=y_j$, $j=1,2,\cdots,p$. This provides a system of p linear equations with nonsingular coefficient matrix. Therefore, $f^{(k-1)}(a)\to 0$ as $a\to\infty$, $k=1,2,\cdots,p$. Since,

$$\int_1^a e^{iyu}f^{(k-1)}(u)\,du=\left(e^{iya}f^{(k-1)}(a)-e^{iy}f^{(k-1)}(1)\right)/(iy)$$
$$-(1/iy)\int_1^a e^{iyu}f^{(k)}(u)\,du$$

and $I_0(y_j)$ converges for each y_j, it follows that each $I_k(y)$, $k=1,2,\cdots,p-1$, also converges for these values of y.

3. **$H_\nu^{(1)}$-transform.** We will need the modified Dirichlet test to obtain an expansion for the $H_\nu^{(1)}$-transform from that of the K_ν-transform. We state it here in a form needed in the proof of Theorem 3 (see [1, Thms. 17.2d, 17.3b]).

LEMMA 1. *Let* $c>0$. *If*

(i) $g(t)$ *is continuous in* $[\alpha,\infty)$ *and* $\int_\alpha^\infty g(t)\,dt$ *converges,*

(ii) $\phi(x,t)$, $\phi_t(x,t)$ *are continuous in* $[0,c]\times[\alpha,\infty]$,

(iii) $\phi(x,\alpha)$ *and* $\int_\alpha^\infty|\phi_t(x,t)|\,dt$ *are uniformly bounded in* $0\leqq x\leqq c$, *then*

$$G(x)=\int_\alpha^\infty \phi(x,t)g(t)\,dt \tag{3.1}$$

converges uniformly in $[0,c]$ *and*

$$\lim_{x\to 0^+}\int_{\alpha}^{\infty}\phi(x,t)g(t)\,dt=\int_{\alpha}^{\infty}\phi(0,t)g(t)\,dt. \tag{3.2}$$

In Theorems 3–6 below we assume that $f(t)$ satisfies the conditions stated at the beginning of §2, that is, $f(t)$ belongs to $C^p(0,\infty)$ and satisfies (2.1)–(2.3).

THEOREM 3. *Let* y *be a fixed positive number and let* $0\leq \operatorname{Re}\nu<\lambda_0+1$, $\operatorname{Re}\nu+2p<\lambda_n+1$. *Define the integrals* $\mathscr{I}_k(y)$ *by*

$$\mathscr{I}_k(y)=\int_1^{\infty}e^{iyt}t^{\nu+k+1/2}\mathscr{D}^k\big(t^{-\nu-1}f(t)\big)\,dt, \tag{3.3}$$

where

$$\mathscr{D}=t^{-1}\frac{d}{dt},\qquad \mathscr{D}^k=\mathscr{D}\mathscr{D}^{k-1}, \tag{3.4}$$

and the function $E(y)$ *by*

$$\begin{aligned}E(y)={}&(-y)^{-p}\int_0^1 H^{(1)}_{\nu+p}(yt)t^{\nu+p+1}\mathscr{D}^p\big(t^{-\nu-1}R_n(t)\big)\,dt\\&+(-y)^{-p}\int_1^{\infty}H^{(1)}_{\nu+p}(yt)t^{\nu+p+1}\mathscr{D}^p\big(t^{-\nu-1}f(t)\big)\,dt\\&-(-y)^{-p}\sum_{l=0}^{m-1}(-y)^{l+1}H^{(1)}_{\nu+p+l+1}(y)\big(\mathscr{D}^{p+l}t^{-\nu-1}f_{n-1}(t)\big)_{t=1}\\&-(-y)^{-p-m}\int_1^{\infty}H^{(1)}_{\nu+p+m}(yt)t^{\nu+p+m+1}\mathscr{D}^{p+m}\big(t^{-\nu-1}f_{n-1}(t)\big)\,dt,\end{aligned} \tag{3.5}$$

where $\lambda_{n-1}-p-\frac{1}{2}<m$. *Then*

$$\begin{aligned}&\int_0^{\infty}H^{(1)}_{\nu}(yt)f(t)\,dt\\&=\pi^{-1}\sum_{k=0}^{n-1}2^{\lambda_k}y^{-\lambda_k-1}e^{i(\lambda_k-\nu)\pi/2}\Gamma\Big(\frac{1}{2}(\lambda_k+\nu+1)\Big)\Gamma\Big(\frac{1}{2}(\lambda_k-\nu+1)\Big)a_k+E(y)\end{aligned} \tag{3.6}$$

if and only if $\mathscr{I}_0(y)$ *and* $\mathscr{I}_p(y)$ *converge.*

This theorem is analogous to Theorem 1 except for the comparatively complicated form of $E(y)$. In case $\lambda_{n-1}-1/2<p<(\lambda_n+1-\operatorname{Re}\nu)/2$, we can write

$$E(y)=(-y)^{-p}\int_0^{\infty}H^{(1)}_{\nu+p}(yt)t^{\nu+p+1}\mathscr{D}^p\big(t^{-\nu-1}R_n(t)\big)\,dt. \tag{3.7}$$

In general, however, this is not possible because for a given function f, no p may satisfy the condition necessary for the representation (3.7). The following theorem is an analogue of Theorem 2 for the differential operator $\mathscr{D}$.

THEOREM 4. *If* $\mathscr{I}_0(y)$ *and* $\mathscr{I}_p(y)$ *converge for* p *distinct positive values of* y, *then* $\mathscr{I}_k(y)$, $k=1,2,\cdots,p-1$, *also converge for those* y.

From (3.6), we obtain the following expansions for the Hankel and the Y_ν-transforms.

THEOREM 5. *If $f(t)$ is real and the integrals*

$$\int_1^\infty e^{iyt}t^{k-p-1/2}f^{(k)}(t)\,dt, \qquad k=1,2,\cdots,p, \tag{3.8}$$

and

$$\int_1^\infty e^{iyt}t^{-1/2}f(t)\,dt \tag{3.9}$$

converge, then for $\operatorname{Re}\nu+\lambda_0>-1, p>\lambda_{n-1}-1/2$,

$$\int_0^\infty J_\nu(yt)f(t)\,dt=\pi^{-1}\sum_{k=0}^{n-1}2^{\lambda_k}y^{-\lambda_k-1}\cos(\lambda_k-\nu)\pi/2 \cdot\Gamma\left(\frac{1}{2}(\lambda_k+\nu+1)\right)\Gamma\left(\frac{1}{2}(\lambda_k-\nu+1)\right)a_k +(-y)^{-p}\int_0^\infty J_{\nu+p}(yt)t^{\nu+p+1}\mathscr{D}^p\left(t^{-\nu-1}R_n(t)\right)dt, \tag{3.10}$$

and for $0\leqq\operatorname{Re}\nu<\lambda_0+1, \operatorname{Re}\nu+2p<\lambda_n+1$,

$$\int_0^\infty Y_\nu(yt)f(t)\,dt=\pi^{-1}\sum_{k=0}^{n-1}2^{\lambda_k}y^{-\lambda_k-1}\sin(\lambda_k-\nu)\pi/2 \cdot\Gamma\left(\frac{1}{2}(\lambda_k+\nu+1)\right)\Gamma\left(\frac{1}{2}(\lambda_k-\nu+1)\right)a_k+F(y) \tag{3.11}$$

where $F(y)$ is obtained from $E(y)$, defined in (3.5), *by replacing the Hankel functions $H^{(1)}_{\nu+\cdot}$ by the Neumann functions $Y_{\nu+\cdot}$.*

If f is complex valued, we can obtain the expansions (3.10) and (3.11) from (3.6) provided that the integrals (3.8) and (3.9) converge when y is replaced by $\pm y$. The convergence of the integrals (3.8) implies the convergence of the integral $\mathscr{I}_p(y)$, (see (3.3)). In the last theorem, we give a more precise relationship between them.

THEOREM 6. *If $y_j, j=1,2,\cdots,p$, is a set of p distinct positive numbers, the following three sets of conditions are equivalent.*

I. *The integrals* (3.8) *and* (3.9) *converge for $y=y_j, j=1,2,\cdots,p$.*

II. *The integrals $\mathscr{I}_0(y)$ and $\mathscr{I}_p(y)$ converge for $y=y_j, j=1,2,\cdots,p$.*

III. *The integrals*

$$\int_1^\infty e^{iyt}t^{-1/2}f^{(k)}(t)\,dt, \qquad k=0,1,\cdots,p, \tag{3.12}$$

converge for $y=y_j, j=1,2,\cdots,p$.

Remark. In Theorems 1, 3, and 5 we have given certain expansions which hold for one or more values of y under the given conditions. We make no claim that these expansions are asymptotic. In fact, even if the corresponding integrals converge for all y sufficiently large, the expansions may not be asymptotic unless some additional condition such as the uniform convergence of the integrals appearing in the remainder terms, is satisfied. The asymptotic nature of these expansions has been discussed by Olver [2], Soni [3], [4], Soni and Soni [5] and Wong [7]. (For other references, see [8]).

Proof of Theorem 3. Let

$$\eta(s)=(2s/\pi)^{1/2}e^sK_\nu(s) \tag{3.13}$$

where $K_\nu(s)$ is the modified Bessel function of the second kind. Thus $\eta(s)$ is analytic except perhaps at $s=0$ where it may have a branch point. By [6, p. 202],

$$\eta(s)\sim\sum_{n=0}^{\infty}(\nu,n)(2s)^{-n},\qquad |s|\to\infty,\qquad |\arg s|<\frac{3\pi}{2},\tag{3.14}$$

where

$$(\nu,n)=\Gamma(\nu+n+\tfrac{1}{2})\big(n!\Gamma(\nu-n+\tfrac{1}{2})\big)^{-1}.\tag{3.15}$$

Therefore, $\eta(-iy)\to 1$ as $y\to\infty$. Since, by [6, p. 78, (8)] and (3.13),

$$H_\nu^{(1)}(y)=(2/\pi)^{1/2}e^{-i(\pi\nu/2+\pi/4)}e^{iy}y^{-1/2}\eta(-iy),\tag{3.16}$$

$\eta(-iy)$ is nonzero for all y sufficiently large. Therefore, the integral

$$\mathscr{H}_k(y)=\int_1^\infty H^{(1)}_{\nu+k}(yt)t^{\nu+k+1}\mathscr{D}^k\big(t^{-\nu-1}f(t)\big)\,dt\tag{3.17}$$

converges if and only if the corresponding integral $\mathscr{I}_k(y)$ defined by (3.3), converges. Thus, if the convergence of $\mathscr{H}_0(y)$ and $\mathscr{H}_p(y)$ is a necessary condition, then so is the convergence of $\mathscr{I}_0(y)$ and $\mathscr{I}_p(y)$. To prove that this condition is also sufficient, we follow the technique used in the proof of Theorem 1. By the convergence of the integral $\mathscr{I}_p(y)$ it follows that $\mathscr{D}^k(t^{-\nu-1}f(t))=o(t^{2p-2k})$, $k=0,1,\cdots,p-1$. In particular, for every $x>0$,

$$\mathscr{D}^k\big(t^{-\nu-1}f(t)\big)=o(e^{xt}),\qquad t\to\infty.\tag{3.18}$$

Now, by the standard technique that was used to obtain (2.8) (see also [5, p. 166]), it follows that for $\operatorname{Re}s>0$,

(3.19)
$$\int_0^\infty K_\nu(st)f(t)\,dt=\sum_{k=0}^{n-1}2^{\lambda_k-1}s^{-\lambda_k-1}\Gamma\Big(\frac{1}{2}(\lambda_k+\nu+1)\Big)\Gamma\Big(\frac{1}{2}(\lambda_k-\nu+1)\Big)a_k+E_1(s)$$

where

$$E_1(s)=s^{-p}\int_0^\infty K_{\nu+p}(st)t^{\nu+p+1}\mathscr{D}^p\big(t^{-\nu-1}R_n(t)\big)\,dt.\tag{3.20}$$

For successive integration by parts, we use the relation

$$\int t^{\mu+1}K_\mu(t)\,dt=-t^{\mu+1}K_{\mu+1}(t).\tag{3.21}$$

We will prove that (3.6) follows from (3.19) when $\operatorname{Re}s\to 0^+$. If $s=x-iy$ and $x\to 0^+$, $x-iy\to e^{-i\pi/2}y$. We note that

$$K_\nu(-iyt)=(\pi/2)e^{i(\nu+1)\pi/2}H_\nu^{(1)}(yt).\tag{3.22}$$

Let

$$\phi(x,t)=e^{-xt}\eta\big((x-iy)t\big)\tag{3.23}$$

where $\eta(s)$ is defined by (3.13). If y is a fixed positive number, $\phi(x,t)$ satisfies the conditions of Lemma 1 in $[0,c) \times [1,\infty)$ for every $c>0$. Therefore, by the convergence of the integral $\mathscr{I}_0(y)$ and the relation (3.16),

(3.24)

$$\lim_{x\to 0^+}\int_1^\infty K_\nu((x-iy)t)f(t)\,dt = \lim_{x\to 0^+}(\pi/2)^{1/2}(x-iy)^{-1/2}\int_1^\infty \phi(0,t)e^{iyt}t^{-1/2}f(t)\,dt$$

$$=(\pi/2)e^{i(\nu+1)\pi/2}\int_1^\infty H_\nu^{(1)}(yt)f(t)\,dt.$$

Similarly, by Lemma 1 and the convergence of $\mathscr{I}_p(y)$,

$$\text{(3.25)}\qquad \lim_{x\to 0^+}\int_1^\infty K_{\nu+p}((x-iy)t)t^{\nu+p+1}\mathscr{D}^p\big(t^{-\nu-1}f(t)\big)\,dt$$

$$=(\pi/2)e^{i(\nu+p+1)\pi/2}\int_1^\infty H_{\nu+p}^{(1)}(yt)t^{\nu+p+1}\mathscr{D}^p\big(t^{-\nu-1}f(t)\big)\,dt.$$

The other limits follow in a straightforward manner. To find the limit of $E_1((x-iy)t)$ as $x\to 0^+$, we replace $R_n(t)$ by $f(t)-f_{n-1}(t)$ in the interval $(1,\infty)$ and then use (3.25) and the following expansion

$$\text{(3.26)}\qquad \int_1^\infty K_{\nu+p}(st)t^{\nu+p+1}\mathscr{D}^p\big(t^{-\nu-1}f_{n-1}(t)\big)\,dt$$

$$=\sum_{l=0}^{m-1} s^{-l-1}K_{\nu+p+l+1}(s)\mathscr{D}^{p+l}\big(t^{-\nu-1}f_{n-1}(t)\big)_{t=1}$$

$$+s^{-m}\int_1^\infty K_{\nu+p+m}(st)t^{\nu+p+m+1}\mathscr{D}^{p+m}\big(t^{-\nu-1}f_{n-1}(t)\big)\,dt.$$

Proof of Theorem 4. As in the proof of Theorem 2, assume that $f(t)$ belongs to $C^p(0,\infty)$ and as $t\to 0^+$, $f(t)=O(t^{\lambda_0})$ where $\lambda_0>\operatorname{Re}\nu+2p$. If $\mathscr{I}_0(y)$ and $\mathscr{I}_p(y)$ converge for $y=y_j$, $j=1,2,\cdots,p$, and we compare the $H_\nu^{(1)}$-transform expansion (3.6) with that obtained for the interval $(0,a)$, (see [5, eq. (2.4)]), we find that the function $\zeta(y,a)$ defined by

$$\text{(3.27)}\qquad \zeta(y,a)=\sum_{k=1}^{p}(-y)^{-k}a^{\nu+k}H_{\nu+k}^{(1)}(ya)\big(\mathscr{D}^{k-1}t^{-\nu-1}f(t)\big)_{t=a}$$

is such that $\zeta(y_j,a)\to 0$ as $a\to\infty$, $j=1,2,\cdots,p$. This implies that

$$\text{(3.28)}\qquad \lim_{t\to\infty} t^{\nu+k-1/2}\mathscr{D}^{k-1}\big(t^{-\nu-1}f(t)\big)=0,\qquad k=1,2,\cdots,p.$$

By the convergence of $\mathscr{I}_0(y)$ and (3.28), it follows that each of the integrals $\mathscr{I}_k(y)$, $k=1,2,\cdots,p-1$, converges.

Proof of Theorem 5. We note that

$$\text{(3.29)}\qquad \mathscr{D}^p\big(t^{-\nu-1}f(t)\big)=\sum_{k=0}^{p} c_{pk}t^{-2p+k-\nu-1}f^{(k)}(t),$$

for some constants c_{pk}. Therefore, the convergence of the integrals (3.8) and (3.9) implies the convergence of the integrals $\mathscr{I}_0(y)$ and $\mathscr{I}_p(y)$. Therefore the expansion (3.6) holds. Since f is real valued, the integrals (3.8) and (3.9) converge even when y is replaced by $(-y)$. Thus we obtain the expansion (3.6) with $H_\nu^{(1)}(yt)$ replaced by $H_\nu^{(2)}(yt)$. By using the relations

$$H_\nu^{(k)}(yt)=J_\nu(yt)+(-1)^{k-1}iY_\nu(yt), \qquad k=1,2,$$

we obtain the corresponding expansions for the Hankel and the Y_ν-transform. In the K_ν and consequently in the Y_ν-transform expansion, the conditions $0\leqq \operatorname{Re}\nu<\lambda_0+1$ and $\operatorname{Re}\nu+2p<\lambda_n+1$ are due to the behavior of these functions as $t\to 0^+$. In the case of the Hankel transform, these conditions can be replaced by the condition $\operatorname{Re}\nu+\lambda_0>-1$. The condition $p>\lambda_{n-1}-\frac{1}{2}$ is simply due to the representation of the remainder term as an integral.

Proof of Theorem 6. In the proof of Theorem 5 we have shown that I ⇒ II. Since III ⇒ I, we only need to prove that II → III. The convergence of the integrals $\mathscr{I}_0(y_j)$ and $\mathscr{I}_p(y_j)$ implies (3.28). Since $\mathscr{D}^k(t^{-\nu-1}f(t))$ can be written in the form (3.29), it follows that for $k=0,1,\cdots,p-1$,

$$f^{(k)}(t)=o(t^{1/2}), \qquad t\to\infty. \tag{3.30}$$

Since the integrals $\mathscr{I}_0(y_j)$ converge, by using integration by parts and (3.30) it follows that II ⇒ III.

4. Analogues of Watson's lemma and examples. By using the condition III of Theorem 6, we can state analogues of Watson's lemma for the Hankel and the Y_ν-transform as follows.

THEOREM 7. *Assume that*

(i) *$f(t)$ belongs to $C^\infty(0,\infty)$,*

(ii)

$$f(t)\sim\sum_{k=0}^{\infty}a_k t^{\lambda_k}, \qquad t\to 0^+, \tag{4.1}$$

where $\lambda_k<\lambda_{k+1}$, $k=0,1,\cdots,\lambda_k\to\infty$ as $k\to\infty$, and the asymptotic expansion for the successive derivatives of $f(t)$ as $t\to 0^+$ can be obtained from (4.1) *by term by term differentiation,*

(iii) *each of the integrals*

$$\int_1^\infty e^{iyt}t^{-1/2}f^{(k)}(t)\,dt, \qquad k=0,1,\cdots, \tag{4.2}$$

converges uniformly for all y sufficiently large.

Then the asymptotic expansions of the integrals

$$\mathscr{F}_1(y)=\int_0^\infty J_\nu(yt)f(t)\,dt, \qquad \operatorname{Re}\nu+\lambda_0>-1 \tag{4.3}$$

and

$$\mathscr{F}_2(y)=\int_0^\infty Y_\nu(yt)f(t)\,dt, \qquad 0\leqq\operatorname{Re}\nu<\lambda_0+1, \tag{4.4}$$

as $y\to\infty$ can be obtained by substituting the expansion (4.1) *for $f(t)$ and integrating term by term in the generalized sense.*

The analogue of the Watson's lemma for the Hankel transform was first given by Wong [7, Thm. 2]. In the statement of his result he gives the following condition in place of (iii):

$$t^{-1/2}f^{(k)}(t)\to 0 \quad \text{as}\, t\to\infty, \quad k=0,1,\cdots. \tag{4.5}$$

It is reasonable to assume that he also requires the uniform convergence of the integral $\mathscr{F}_1(y)$ for all y sufficiently large [7, §2]. However, his proof of the theorem prompts us to investigate whether any conditions besides those stated explicitly in his Theorem 2, are indeed necessary. In particular, we want to examine the extent to which the uniform convergence of $\mathscr{F}_1(y)$ for all y sufficiently large, is a necessary condition. $\mathscr{F}_1(y)$ can have an asymptotic expansion of the type given in [7] or in Theorem 7 above only if it is defined for all y sufficiently large and furthermore, is reasonably well behaved as $y\to\infty$. Therefore, we ask the following:

1. Does the convergence of the integral (4.3) for all y sufficiently large follow from (4.5) together with, perhaps, the convergence of (4.3) on a smaller set?

2. Does (4.5) together with the convergence of (4.3) for all y sufficiently large, imply that $\mathscr{F}_1(y)$ is reasonably well behaved as $y\to\infty$?

We give examples to show that the answer to both of these questions is in the negative. This indicates that some condition such as the uniform convergence or uniform boundedness of the integral (4.3) for all y sufficiently large, is necessary. Again, the condition (iii) implies (4.5) (see proof of Theorem 6). However, (4.5) does not guarantee the uniform convergence of the integrals (4.2) even when the integral (4.3) converges uniformly. As we have remarked earlier, without some such condition, we may have finite expansions of the type that we have discussed in the earlier sections but these need not be asymptotic.

It is possible to weaken the condition (iii) for the Hankel transform as well as for the Y_ν-transform by replacing the integrals (4.2) by some other integrals but it is questionable whether (iii) can be replaced by (4.5) alone (see [7, p. 804] and Example 3 below).

In the following examples, we assume that $\nu=-\frac{1}{2}$ so that $J_\nu(t)=(2/\pi)^{1/2}t^{-1/2}\cos t$ and $J_{\nu+1}(t)=(2/\pi)^{1/2}t^{-1/2}\sin t$. $I_0(y)$ and $I_1(y)$ are defined by

$$I_k(y)=\int_1^\infty J_{\nu+k}(yt)f^{(k)}(t)\,dk, \qquad k=0,1. \tag{4.6}$$

Example 1. Let

$$f(t)=t^{1/2}(\ln(t+2))^{-1}\sum_{n=1}^{\infty}e^{-n}\cos nt. \tag{4.7}$$

Clearly $I_0(y)$ converges for all $y>0$ except $y=n$, $n=1,2,\cdots$. Thus (4.5) and the convergence of the integral $I_0(y)$ for almost all y fail to guarantee the convergence for all y sufficiently large.

Example 2. Let

$$f(t)=t^{1/2}(\ln(t+2))^{-1}\sum_{n=1}^{\infty}e^{-n}\sin nt. \tag{4.8}$$

Again, $f(t)$ satisfies the condition (4.5), $I_0(y)$ converges for every $y>0$ but not uniformly. In fact, in the neighborhood of $y=n$, $n=1,2,\cdots$, $I_0(y)$ is unbounded and therefore, cannot have an asymptotic expansion in the usual sense as $y \to \infty$. Furthermore,

$$(4.9)\quad f'(t)=t^{1/2}(\ln(t+2))^{-1}\sum_{n=1}^{\infty} ne^{-n}\cos nt$$

$$+\left(\sum_{n=1}^{\infty} e^{-n}\sin nt\right)\left\{(2t^{1/2}\ln(t+2))^{-1}-t^{1/2}(t+2)^{-1}(\ln(t+2))^{-2}\right\}.$$

It is clear from (4.9) that $I_1(y)$ does not converge for $y=n$, $n=1,2,\cdots$. Therefore, even the convergence of $I_0(y)$ for all $y>0$ and the condition (4.5), fail to guarantee the convergence of $I_1(y)$ for all sufficiently large y.

Example 3. Let

$$(4.10)\qquad f(t)=t^{1/2}\sin t^2(\ln(t+2))^{-1}.$$

In this case $I_0(y)$ converges uniformly for $y \geqq 1$, $t^{-1/2}f(t)\to 0$ as $t\to\infty$ and $I_1(y)$ converges but not uniformly for all y sufficiently large. In fact, $I_1(y)$ is unbounded. (The proof of this is rather involved and is given in the appendix).

Appendix. Let

$$f(t)=t^{1/2}(\ln(t+2))^{-1}\sin t^2,\qquad t\geqq 0.$$

We prove that if $I_0(x)$ and $I_1(x)$ are defined by (4.6) and $\nu=-\frac{1}{2}$, then

(a) $I_0(x)$ converges uniformly in $[1,\infty)$.

(b) $I_1(x)$ converges for $x \geqq 1$ but fails to converge uniformly in $[X,\infty)$ for any $X \geqq 1$.

Proof of (a).

$$\sqrt{\pi/2}\,I_0(x)=\int_1^{\infty} x^{-1/2}\cos xt\,\sin t^2(\ln(t+2))^{-1}dt.$$

Let $\varepsilon>0$. Choose N_0 such that $0<(\ln N_0)^{-1}<\varepsilon/4$. For $N>M\geqq N_0$,

$$(A.1)\qquad \mathscr{I}(M,N,x)=x^{-1/2}\int_M^N \cos xt\,\sin t^2(\ln(t+2))^{-1}dt$$

$$=\frac{1}{2}x^{-1/2}\int_M^N \sin(t^2+xt)(\ln(t+2))^{-1}dt$$

$$+\frac{1}{2}x^{-1/2}\int_M^N \sin(t^2-xt)(\ln(t+2))^{-1}dt$$

$$=\mathscr{I}_1+\mathscr{I}_2.$$

By the second mean value theorem,

(A.2)

$$\mathscr{I}_1=2^{-1}x^{-1/2}(\ln(M+2))^{-1}(2M+x)^{-1}\int_M^{N'}(2t+x)\sin(t^2+xt)\,dt,\qquad M\leqq N'\leqq N.$$

Therefore, for $x\geqq 1$, $|\mathscr{I}_1|\leqq\varepsilon/8$.

Before we apply the second mean value theorem to $\mathscr{I}_2$, we must provide for the possibility that $2t-x$ may become zero for some t in the interval of integration. For this reason, we consider the following three cases:

(i) $\frac{1}{2}x+1 \leqq M$,

(ii) $\frac{1}{2}x-1 \leqq M \leqq \frac{1}{2}x+1$,

(iii) $M \leqq \frac{1}{2}x-1$.

Case (i). We can apply the second mean value theorem directly and obtain $|\mathscr{I}_2|<\varepsilon/8$.

Case (ii). Let $\alpha=\min(\frac{1}{2}x+1,N)$. Then

$$\mathscr{I}_2=\int_M^{\alpha}+\int_{\alpha}^{N} 2^{-1}x^{-1/2}(\ln(t+2))^{-1}\sin(t^2-xt)\,dt=\mathscr{I}_3+\mathscr{I}_4. \tag{A.3}$$

Clearly, $|\mathscr{I}_3|<\varepsilon/4$ for $x \geqq 1$. If $N<\frac{1}{2}x+1$, $\mathscr{I}_4=0$. Otherwise, as in Case (i), $|\mathscr{I}_4|<\varepsilon/8$. Therefore $|\mathscr{I}_2|<\varepsilon/2$.

Case (iii).

$$\begin{aligned}\mathscr{I}_2&=\int_M^{x/2-1}+\int_{x/2-1}^{N} 2^{-1}x^{-1/2}(\ln(t+2))^{-1}\sin(t^2-xt)\,dt\\&=\mathscr{I}_5+\mathscr{I}_6.\end{aligned} \tag{A.4}$$

Note that $(2t-x)$ is negative and increasing in $[M,\frac{1}{2}x-1]$ and, takes the maximum value -2 at $t=\frac{1}{2}x-1$. By two applications of the second mean value theorem,

(A.5)

$$\begin{aligned}\mathscr{I}_5&=2^{-1}x^{-1/2}(\ln(M+2))^{-1}\int_M^{\xi}\sin(t^2-xt)\,dt, \qquad M\leqq\xi\leqq\frac{1}{2}x-1\\&=2^{-1}x^{-1/2}(\ln(M+2))^{-1}(2\xi-x)^{-1}\int_{\xi'}^{\xi}(2t-x)\sin(t^2-xt)\,dt, \qquad M\leqq\xi'\leqq\xi.\end{aligned}$$

Therefore $|\mathscr{I}_5|<\varepsilon/8$ and as in Case (ii), $|\mathscr{I}_6|<\varepsilon/2$. This gives $|\mathscr{I}_2|<3\varepsilon/4$ in Case (iii).

By combining the estimates for $|\mathscr{I}_2|$ with the estimate for $|\mathscr{I}_1|$, we obtain, $|\mathscr{I}(M,N,x)|<\varepsilon$ for all $x\geqq 1$. This proves that $I_0(x)$ converges uniformly in $[1,\infty)$.

Proof of (b).

$$\begin{aligned}\sqrt{\pi/2}\,I_1(x)&=\frac{1}{2}x^{-1/2}\int_1^{\infty}\sin xt\,t^{-1}(\ln(t+2))^{-1}\sin t^2\,dt\\&\quad-x^{-1/2}\int_1^{\infty}\sin xt\,(\ln(t+2))^{-2}(t+2)^{-1}\sin t^2\,dt\\&\quad+2x^{-1/2}\int_1^{\infty}\sin xt\,(\ln(t+2))^{-1}t\cos t^2\,dt,\\&=\mathscr{I}_{11}+\mathscr{I}_{12}+\mathscr{I}_{13}.\end{aligned} \tag{A.6}$$

Clearly, $\mathscr{I}_{12}$ converges absolutely and uniformly in $x\geqq 1$. Also, by using the same technique as in (a) above, we can show that $\mathscr{I}_{11}$ converges uniformly in $x\geqq 1$. Therefore, we need to show that $\mathscr{I}_{13}$ converges, but not uniformly, in $[X,\infty)$ no matter how large X.

$$\text{(A.7)}\qquad \begin{aligned}\mathscr{I}_{13}&=x^{-1/2}\int_1^\infty t(\ln(t+2))^{-1}\sin(t^2+xt)\,dt\\ &\qquad -x^{-1/2}\int_1^\infty t(\ln(t+2))^{-1}\sin(t^2-xt)\,dt\\ &=\mathscr{I}_{14}-\mathscr{I}_{15}.\end{aligned}$$

By two applications of the second mean value theorem (as in $\mathscr{I}_5$), we can show that $\mathscr{I}_{14}$ converges uniformly in $x \geqq 1$. Again,

$$\text{(A.8)}\qquad \begin{aligned}\mathscr{I}_{15}&=\frac{1}{2}x^{-1/2}\int_1^\infty (\ln(t+2))^{-1}(2t-x)\sin(t^2-xt)\,dt\\ &\qquad +\frac{1}{2}x^{1/2}\int_1^\infty (\ln(t+2))^{-1}\sin(t^2-xt)\,dt\\ &=\frac{1}{2}\mathscr{I}_{16}+\frac{1}{2}\mathscr{I}_{17}.\end{aligned}$$

By the second mean value theorem, $\mathscr{I}_{16}$ converges uniformly in $x \geqq 1$. Also, for each fixed x, $\mathscr{I}_{17}$ converges (see $\mathscr{I}_2$ in (a) above). Therefore, $I_1(x)$ converges for each fixed $x \geqq 1$. To complete the proof, we must prove that $\mathscr{I}_{17}$ does not converge uniformly in $[X, \infty)$. For this purpose, consider the following integral:

$$\text{(A.9)}\qquad \begin{aligned}Q(M,x)&=\int_M^\infty x^{1/2}(\ln(t+2))^{-1}e^{i(t^2-xt)}\,dt\\ &=x^{1/2}\exp(-ix^2/4)\int_M^\infty (\ln(t+2))^{-1}e^{i(t-x/2)^2}\,dt\\ &=x^{1/2}\exp(-ix^2/4)\int_{M-x/2}^\infty \left(\ln\left(u+\frac{1}{2}x+2\right)\right)^{-1}e^{iu^2}\,du.\end{aligned}$$

For each M sufficiently large, consider those values of x for which $x=2M-2C$, $1 \leqq C \leqq C_1$ where C_1 is some fixed number greater than one. Then, using integration by parts,

$$\text{(A.10)}\qquad \begin{aligned}&x^{-1/2}\exp(ix^2/4)Q(M,x)\\ &\quad=(2C)^{-1}e^{iC^2}(\ln(M+2))^{-1}\\ &\qquad -i2^{-1}\int_C^\infty e^{iu^2}u^{-2}(\ln(u+M-C+2))^{-1}du\\ &\qquad -i2^{-1}\int_C^\infty e^{iu^2}u^{-1}(\ln(u+M-C+2))^{-2}(u+M-C+2)^{-1}du\\ &\quad=i(2C)^{-1}e^{iC^2}(\ln(M+2))^{-1}-i2^{-1}Q_1-i2^{-1}Q_2.\end{aligned}$$

By the second mean value theorem,

$$\text{(A.11)}\qquad |Q_2|<k(\ln(M+2))^{-2}(M+2)^{-1}$$

for some constant k. To obtain the behavior of Q_1, let

$$\phi(t) = -\int_t^\infty u^{-2} e^{iu^2}\,du. \tag{A.12}$$

Using integration by parts,

(A.13)

$$Q_1 = (\ln(M+2))^{-1}\phi(C) - \int_C^\infty \phi(u)(\ln(u+M-C+2))^{-2}(u+M-C+2)^{-1}du.$$

Since $\phi(t) = O(t^{-3})$, $t \to \infty$, the last integral in (A.13) is less than a constant multiple of $(\ln(M+2))^{-2}(M+2)^{-1}$. Therefore for $x = 2M - 2C$, as $M \to \infty$,

$$Q(M,x) \sim \frac{iM^{1/2}}{\sqrt{2}\,C\ln M} e^{-i(M-C)^2}\left\{e^{iC^2} - C\phi(C)\right\}.$$

Since $\mathscr{I}_{17}$ in (A.8) is the imaginary part of $Q(M,x)$, the conclusion follows.

REFERENCES

[1] W. Fulks, *Advanced Calculus*, 2nd ed., John Wiley, New York, 1969.

[2] F. W. J. Olver, *Error bounds for stationary phase approximations*, this Journal, 5 (1974), pp. 19–29.

[3] K. Soni, *On uniform asymptotic expansions of finite Laplace and Fourier integrals*, Proc. Roy. Soc. Edinburgh, 85A (1980), pp. 299–305.

[4] ______, *Asymptotic expansion of the Hankel transform with explicit remainder terms*, Quart. Appl. Math., 50 (1982), pp. 1–14.

[5] K. Soni and R. P. Soni, *A note on uniform asymptotic expansion of finite K_ν and related transforms with explicit remainder*, J. Math. Anal. Appl., 79 (1981), pp. 163–177.

[6] G. N. Watson, *Theory of Bessel Functions*, Cambridge, Univ. Press, London, 1958.

[7] R. Wong, *Error bounds for asymptotic expansion of Hankel transforms*, this Journal, 7 (1976), pp. 799–808.

[8] ______, *Error bounds for asymptotic expansion of integrals*, SIAM Rev., 22 (1980), pp. 401–435.

SIAM J. MATH. ANAL.
Vol. 16, No. 2, March 1985

N-WIDTH AND ENTROPY OF H_p-CLASSES IN $L_q(-1,1)$*

H. G. BURCHARD[†] AND K. HÖLLIG[‡]

Abstract. The n-width d_n, approximation numbers δ_n and entropy ε_n of the Hardy spaces H_p in $L_q(-1,1)$ are estimated. More precisely, denote by F^r the space of continuous functions which satisfy a Lipschitz condition of order r at ± 1. It is shown that

$$\exp(-2\alpha n^{1/2}) \ll \delta_n(H_p \cap F^r, L_\infty), d_n(H_p \cap F^r, L_\infty) \ll \exp(-\alpha n^{1/2}),$$
$$\exp(-2\beta n^{1/2}) \ll \delta_n(H_p, L_q), d_n(H_p, L_q) \ll \exp(-\beta n^{1/2}) \quad \text{for } p > q,$$
$$\exp(-2\gamma n^{1/3}) \ll \varepsilon_n(H_p \cap F^r, L_\infty) \ll \exp(-\gamma n^{1/3}),$$

where "$\ll$" indicates that the inequalities hold except for polynomial factors in n. The constants α, β, γ depend on p, q and r. For $p = \infty$, the factor 2 in the lower bound of the first inequality can be omitted.

AMS-MOS subject classification (1980). Primary 41A46, 30D55

Key words. Hardy spaces, n-width, entropy, upper and lower bounds, Wittaker series

1. Introduction. For analytic functions many of the standard approximation processes converge at an exponential rate. Using more sophisticated methods, it is still possible to obtain exponential convergence, even in the presence of singularities at the endpoints of an inverval of approximation.

In this paper we obtain precise upper and lower bounds for optimal convergence rates of approximation processes for the natural imbeddings of Hardy spaces into $L_q(-1,1)$ in the sense of n-width, approximation numbers (linear n-width) and also entropy. This makes it possible to assess the optimality of bounds previously obtained for special approximation operators.

As a model example, consider the class H_∞ of analytic functions f bounded in the unit disc. To obtain convergence in $L_\infty(-1,1)$ of approximation methods, some mild additional assumptions must be imposed about the behaviour of f at ± 1. For this, let F^r denote the class of functions in $L_\infty(-1,1)$ which satisfy a Lipschitz condition of order $r > 0$ at ± 1 (cf. (2.4)). In [6] A. Goncar has constructed piecewise polynomial approximation operators P_n of rank n such that

$$(1.1) \qquad \|f - P_n f\|_{L_\infty(-1,1)} \ll \exp(-\alpha n^{1/2})\left(\|f\|_{H_\infty} + \|f\|_{F^r}\right)$$

where $\alpha = \log(1+\sqrt{2})r^{1/2}$. Here, "$\ll$" indicates that the inequality holds except for a polynomial factor in n. R. De Vore and K. Scherer [4] showed that $\exp(-\sqrt{2}\,\alpha n^{1/2})$ is a lower bound for approximation by piecewise polynomial operators. In [9], [13]–[16] F. Stenger developed a theory for approximating analytic functions using Whittaker's cardinal series. In particular he obtained (1.1) with an improved value of the constant, $\alpha = (\pi/2)r^{1/2}$. For the approximation of functions f in $H_\infty \cap F^r$ we obtain the sharp lower bound: For some nonzero f in $H_\infty \cap F^r$

$$(1.2) \qquad \|f - P_n f\|_{L_\infty(-1,1)} \gg \exp\left(-\frac{\pi}{2}(rn)^{1/2}\right)\left(\|f\|_{H_\infty} + \|f\|_{F^r}\right),$$

* Received by the editors February 7, 1983. This research was sponsored by the U. S. Army under contract DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under grant MCS-7927062, Mod. 1.

[†] Department of Mathematics, Oklahoma State University, Stillwater, Oklahoma, 74078.

[‡] Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53705.

valid for an arbitrary rank n operator P_n (cf. Theorem 1). This establishes that approximation by Whittaker's cardinal series is optimal and $\exp(-(\pi/2)(rn)^{1/2})$ is the precise asymptotic order of the n-width of $H_\infty \cap F^r$ in $L_\infty(-1,1)$ up to a polynomial factor in n. We obtain results analogous to (1.1) and (1.2) for the n-width and approximation numbers of $H_p \cap F^r$ in $L_\infty(-1,1)$ (cf. Theorem 2) and of H_p in $L_q(-1,1)$, $p>q$ (cf. Theorem 3).

For entropy, however, the asymptotic behavior is different. For our model example we obtain (cf. Theorem 4)

$$\exp(-2\gamma n^{1/3}) \ll \varepsilon_n(H_\infty \cap F^r, L_\infty) \ll \exp(-\gamma n^{1/3})$$

where $\gamma = ((\pi^2/2) r \log 2)^{1/3}$. Thus, our results show that n-width tends to zero more rapidly then entropy. These estimates are in remarkable contrast to the results for Sobolev spaces where $\varepsilon_n \le d_n$ [7]. The slower decay of entropy seems to be typical for classes of analytic functions. The best known example appears to be the following. Set $\Delta = \{w \in \mathbf{C}: |w| < 1\}$. Then we have for $\rho < 1$

$$d_n(H_\infty, L_\infty(\rho\Delta)) >< \exp(-|\log\rho| n),$$
$$\varepsilon_n(H_\infty, L_\infty(\rho\Delta)) >< \exp(-\log 2 |\log\rho| n^{1/2}).$$

After stating our main results in §2 we prove in §3 auxiliary results regarding n-width and entropy. In §4 we introduce equivalent approximation problems on the real line and obtain basic approximation properties of weighted cardinal series. The proofs of Theorems 1–4 are given in §5.

2. Main results. Let T: $X \to Y$ be a bounded linear operator between Banach spaces X and Y. The *n-width* d_n, the *approximation numbers* (linear n-width) δ_n and the *entropy* ε_n of T are defined by

$$(2.1) \qquad d_n(T) = \inf_{\substack{V \subset Y \\ \dim V \le n}} \sup_{\|x\|_X \le 1} \operatorname{dist}_Y(Tx, V),$$

$$(2.2) \qquad \delta_n(T) = \inf_{\substack{P \in L(X,Y) \\ \operatorname{rank} P \le n}} \|T - P: X \to Y\|,$$

$$(2.3) \qquad \varepsilon_n(T) = \inf\left\{\varepsilon: \exists y_1, \cdots, y_{2^n} \in Y \text{ such that } TB(X) \subset \bigcup_{\nu=1}^{2^n} (y_\nu + \varepsilon B(Y))\right\}$$

where $B(X)$ denotes the closed unit ball of the B-space X. If T: $X \to Y$ is a continuous embedding, we write $a_n(X, Y)$ in place of $a_n(T)$. Here and in the sequel a_n stands for either one of the numbers d_n, δ_n or ε_n.

Let H_p, $1 \le p \le \infty$, denote the Hardy space [5], i.e. H_p is the class of analytic functions in the unit disc Δ for which

$$\|f\|_{H_p} = \sup_{0 \le s < 1} \left(\frac{1}{2\pi}\int_0^{2\pi} |f(se^{i\theta})|^p d\theta\right)^{1/p}, \qquad 1 \le p < \infty,$$
$$\|f\|_{H_\infty} = \sup_{|z|<1} |f(z)|$$

is finite. Given a conformal homeomorphism h of Δ onto a simply connected region $\Omega\subset\mathbf{C}$, one can define [6]

$$H_p(\Omega)=\{f\colon\Omega\to\mathbf{C}: f\circ h\in H_p\}.$$

Different conformal maps result in equivalent norms.

We denote by F^r the class of functions in $L_\infty(-1,1)$ which satisfy a Lipschitz condition of order $r>0$ at ± 1. Let $\lfloor r\rfloor$ ($\lceil r\rceil$) denote the least (largest) integer not less (greater) than r. F^r is the direct sum of $P_{2\lfloor r\rfloor-1}$, the space of polynomials of degree at most $2\lfloor r\rfloor-1$, and the space $\psi^r L_\infty(-1,1)$ of functions with zeros of order r at ± 1, $\psi(w)=1-w^2$. Let p_r be the projection of F^r onto $P_{2\lfloor r\rfloor-1}$ defined by the conditions

$$(f-p_rf)\psi^{-r}\in L_\infty(-1,1).$$

Then a norm on F^r can be defined by

$$\|f\|_{F^r}=\|p_rf\|_{L_\infty(-1,1)}+\|\psi^{-r}(f-p_rf)\|_{L_\infty(-1,1)}. \tag{2.4}$$

We study approximation of functions in $H_p\cap F^r$. To state our results we use the following notions of asymptotic equivalence. Let a_n, b_n, $n\in\mathbf{N}$ be two sequences of positive numbers. We write $a_n\lesssim b_n$ if there exists a positive constant c such that $a_n\leq cb_n$, and $a_n\ll b_n$ if there exists a positive constant j such that $a_n\lesssim n^jb_n$. The symbols $\gtrsim$, $\gg$ and $\simeq$, $><$ are defined similarly.

THEOREM 1. *For $r>0$ we have*

$$\delta_n(H_\infty\cap F^r,L_\infty(-1,1)),d_n(H_\infty\cap F^r,L_\infty(-1,1))><\exp\left(-\frac{\pi}{2}(rn)^{1/2}\right).$$

This result has already been mentioned in the introduction (cf. (1.1),(1.2)). The upper estimate is due to F. Stenger [13] and our lower bound shows the optimality of the order $\exp(-(\pi/2)(rn)^{1/2})$.

THEOREM 2. *For $r>0$ and $1\leq p\leq\infty$ we have*

$$\exp(-2\alpha n^{1/2})\ll\delta_n(H_p\cap F^r,L_\infty(-1,1)),d_n(H_p\cap F^r,L_\infty(-1,1))\ll\exp(-\alpha n^{1/2})$$

where $\alpha=\pi r/(2(r+1/p)^{1/2})$.

It is interesting to compare these rates with the estimates of F. Stenger [15], who considered the classes $H_p^*=\psi H_p$, where $\psi(w)=1-w^2$, and obtained

$$\exp\left(-\left(\sqrt{5}\,\pi+\varepsilon\right)n^{1/2}\right)\leq\delta_n^*(H_p^*,L_\infty(-1,1))$$
$$\leq\exp\left(-\left(\frac{\pi}{2(p')^{1/2}}-\varepsilon\right)n^{1/2}\right),\qquad n\geq n(\varepsilon).$$

Here δ_n^* is defined analogous to δ_n but restricting the class of rank n approximations P to methods based on point evaluation, i.e.

$$Pf=\sum_{j=1}^{n}f(x_j)s_j.$$

As we shall see in §5, H_p^* is similar to the (smaller) class $H_\infty\cap F^r$, $r=1-1/p=1/p'$. We obtain the bounds (valid also for δ_n^*, d_n)

$$\exp\left(-\pi\frac{1}{\sqrt{p'}}n^{1/2}\right)\ll\delta_n(H_p^*,L_\infty(-1,1))\ll\exp\left(-\frac{\pi}{2}\frac{1}{\sqrt{p'}}n^{1/2}\right). \tag{2.5}$$

In view of Theorem 1 we conjecture that the factor 2 in the lower bound of Theorem 2 can be omitted and $\exp(-\alpha n^{1/2})$ is the precise asymptotic rate of the n-width.

THEOREM 3. *For* $1 \leq q < p \leq \infty$ *we have*

$$\exp(-2\beta n^{1/2}) \ll \delta_n(H_p, L_q(-1,1)), d_n(H_p, L_q(-1,1)) \ll \exp(-\beta n^{1/2})$$

where $\beta = \frac{\pi}{2}(1/q - 1/p)^{1/2}$.

Remark. The proofs of Theorems 1–3 will show that the results are true for any s-number in the sense of Pietsch [11].

As mentioned in the introduction, Vitushkin's [19] estimates for entropy of classes of analytic functions show exponential decay of $\varepsilon_n(H_\infty, L_\infty(\rho\Delta))$ as $\exp(-cn^{1/2})$ for $0 < \rho < 1$. Notice that in this case the functions approximated are analytic in a neighborhood of the domain $\rho\Delta$ of approximation. In this paper we obtain estimates for entropy of imbeddings of analytic functions with singularities on the boundary $\{-1, 1\}$ of the interval of approximation, the rate being $\exp(-cn^{1/3})$. We attribute the curious exponent $1/3$ to the fact that singularities are allowed here. A typical result is as follows.

THEOREM 4. *For* $r > 0$ *and* $1 \leq p \leq \infty$ *we have*

$$\exp(-2\gamma n^{1/3}) \ll \varepsilon_n(H_p \cap F^r, L_\infty(-1,1)) \ll \exp(-\gamma n^{1/3})$$

where $\gamma = (\pi^2 r^2 \log 2/(2(r + 1/p)))^{1/3}$.

3. General properties of d_n, δ_n and ε_n. A. Pietsch has developed in [11] a general theory of "s-numbers" which includes n-width and approximation numbers as special cases. We list below some basic properties of d_n and δ_n which hold for entropy as well. Let a_n denote either one of the numbers d_n, δ_n or ε_n and let T: $X \to Y$ be a bounded linear operator, then we have

The numbers a_n form a monotone decreasing sequence, i.e.

$$(3.1) \qquad \|T\| = a_0(T) \geq a_1(T) \geq \cdots .$$

If T admits the factorization T: $Y \xrightarrow{R} Y' \xrightarrow{T'} X' \xrightarrow{S} X$ *we have*

$$(3.2) \qquad a_n(T) \leq \|R\| a_n(T') \|S\|.$$

The numbers a_n are additive, i.e.

$$(3.3) \qquad a_{n_0+n_1}(T_0 + T_1) \leq a_{n_0}(T_0) + a_{n_1}(T_1).$$

Properties (3.1)–(3.3) are direct consequences of the definitions (cf. [11]).

The following result is useful for obtaining lower bounds for n-width and approximation numbers,

LEMMA 1. [8]. *Let V be an $(n+1)$-dimensional subspace of X and let i: $V \to X$ be the canonical injection. Then we have*

$$d_n(i) = \delta_n(i) = 1.$$

We shall need some estimates for a_n in sequence spaces. By l_∞^m, l_∞ we denote $\mathbb{R}^m$, $\mathbb{R}^Z$ with supremum norm. In addition, we define the weighted space $l_{\infty,\rho}$ by

$$(3.4) \qquad l_{\infty,\rho} = \left\{ f \in l_\infty \colon \|f\|_{l_{\infty,\rho}} = \sup_{\nu \in \mathbb{Z}} \exp(\rho|\nu|)|f_\nu| < \infty \right\}.$$

LEMMA 2. *For* $m > n$ *we have*

$$\delta_n(l_\infty^m, l_\infty^m) = d_n(l_\infty^m, l_\infty^m) = 1$$

and for $N=2n-1, 2n$

$$d_N(l_{\infty,\rho}, l_\infty)=\delta_N(l_{\infty,\rho}, l_\infty)=\exp(-\rho n).$$

Proof. The first part of the Lemma is a consequence of Lemma 1.

Let P_N be the projection of l_∞ onto the span of the first N basis vectors $(\delta_{\nu,\mu})_{\mu\in\mathbf{Z}}$, $|\nu|\le N/2$. Then

$$\|i-P_N\colon l_{\infty,\rho}\to l_\infty\|=\exp(-\rho\lfloor N/2\rfloor)$$

is an upper bound for $\delta_N(i)$, $i\colon l_{\infty,\rho}\to l_\infty$ being the natural injection.

For the lower estimate consider the factorization of the identity

$$l_\infty^{N+1}\xrightarrow{I} l_{\infty,\rho}\hookrightarrow l_\infty\xrightarrow{P_{N+1}} l_\infty^{N+1}$$

where I is the canonical injection. Using (3.2) and Lemma 1 this yields

$$1=d_N(l_\infty^{N+1}, l_\infty^{N+1})\le\|I\|d_N(l_{\infty,\rho}, l_\infty)\|P_{N+1}\|\le\exp(\rho\lceil(N+1)/2\rceil)d_N(l_{\infty,\rho}, l_\infty).$$

LEMMA 3. *For* $\rho>0$ *we have*

$$e^{-2\rho}\exp(-(\rho n\log 2)^{1/2})\lesssim\varepsilon_n(l_{\infty,\rho}, l_\infty)\lesssim e^{\rho}\exp(-(\rho n\log 2)^{1/2}).$$

Proof. For $\varepsilon>0$ the unit ball B of $l_{\infty,\rho}$ contains the finite subset

$$A(\varepsilon)=\{a\in l_\infty\colon a_\nu\in\varepsilon\mathbf{Z}, |a_\nu|\le\exp(-|\nu|\rho)\}.$$

For the upper bound note that $A(2\varepsilon)$ is an l_∞ ε-net for B. It suffices therefore to show $\operatorname{card}A(2\varepsilon)\le 2^n$ when $\varepsilon=2e^{-\delta}$, $\delta=(\rho n\log 2)^{1/2}-\rho$. We estimate

$$\operatorname{card}A(2\varepsilon)=\prod_{\nu\in\mathbf{Z}}(2\lceil\exp(-|\nu|\rho+\delta)/4\rceil+1)\le\prod_{|\nu|<\delta/\rho}\exp(-|\nu|\rho+\delta)$$
$$\le\exp\left(-\rho\left(\frac{\delta}{\rho}-1\right)\frac{\delta}{\rho}+\delta\left(2\frac{\delta}{\rho}+1\right)\right)=\exp\left(\frac{\delta^2}{\rho}+2\delta\right)\le 2^n$$

as claimed.

For the lower bound fix $\tilde{\varepsilon}=\frac{1}{2}e^{-\delta}$, $\delta=(\rho n\log 2)^{1/2}+2\rho$. Then

$$\operatorname{card}A(2\tilde{\varepsilon})=\prod_{\nu\in\mathbf{Z}}(2\lceil\exp(-|\nu|\rho+\delta)\rceil+1)>\prod_{|\nu|<\delta/\rho}\exp(-|\nu|\rho+\delta)$$
$$\ge\exp\left(-\rho\left(\frac{\delta}{\rho}+1\right)\frac{\delta}{\rho}+\delta\left(2\frac{\delta}{\rho}-1\right)\right)=\exp\left(\frac{\delta^2}{\rho}-2\delta\right)\ge 2^n.$$

Given any l_∞ ε-net N for B with cardinality 2^n, since $\operatorname{card}A(2\tilde{\varepsilon})>2^n$, at least one of the l_∞-balls of radius ε with center in N must contain two distinct points of $A(2\tilde{\varepsilon})$, and this implies $\tilde{\varepsilon}\le\varepsilon$. This establishes the lower bound.

4. Approximation processes on the line. For the proofs of our theorems in §5 it will be convenient to consider equivalent approximation problems on the line. The conformal mapping

$$z=\log\frac{1+w}{1-w}$$

transforms the unit disk Δ one-to-one onto the parallel strip $\Omega=\{z\in\mathbf{C}\colon |\operatorname{Im} z|<\pi/2\}$ and at the same time maps the interval $(-1,1)$ onto $\mathbf{R}$. This substitution induces isometries from H_p onto $H_p(\Omega)$ and from $L_q(-1,1)$ onto the weighted space $\phi^{1/q}L_q(\mathbf{R})$, where

$$\phi(z)=\frac{dz}{dw}=1+\cosh z.$$

The norm on $\phi^{1/q}L_q(\mathbf{R})$ is given by

$$\|f\|_{\phi^{1/q}L_q(\mathbf{R})}=\|\phi^{-1/q}f\|_{L_q(\mathbf{R})}.$$

We also need other weighted function spaces. Let

$$\Omega_d=\frac{2d}{\pi}\Omega,\qquad 0<d\leq\pi,$$

i.e., $z\in\Omega_d$ iff $|\operatorname{Im} z|<d$. Then, for real λ, $f\in\phi^{\lambda}H_\infty(\Omega_d)$ iff $\phi^{-\lambda}f\in H_\infty(\Omega_d)$ and

$$\|f\|_{\phi^{\lambda}H_\infty(\Omega_d)}=\|\phi^{-\lambda}f\|_{H_\infty(\Omega_d)}.$$

Notice that ϕ maps Ω_π onto $\mathbf{C}\setminus\{x\in\mathbf{R}\colon x\leq 0\}$ and so ϕ^{λ} is holomorphic on Ω_π.

We note some simple properties of $\phi(z)$. If $z=x+iy\in\Omega_\pi$, then

$$c_y e^{|x|}\leq|\phi(z)|=\cosh x+\cos y\leq 2e^{|x|},\tag{4.1}$$

where $c_y=\frac{1}{2}$ for $|y|\leq\pi/2$ and $c_y=(1+\cos y)/2$ for $\pi/2<|y|<\pi$. If $w\in\Delta$, then $z=\log((1+w)/(1-w))\in\Omega_{\pi/2}$ and

$$1-|w|^2=2\cos y/|\phi(z)|.\tag{4.2}$$

We now establish equivalent approximation problems on the line. To do this, we first replace F^r by the simpler space $\psi^r L_\infty(-1,1)$.

LEMMA 4. *Let $r>0$, $k=2\lfloor r\rfloor$, $1\leq p, q\leq\infty$. Then for $a_n=\delta_n$, d_n*

$$a_{n+k}(H_p\cap F^r, L_q(-1,1))\lesssim a_n(H_p\cap\psi^r L_\infty(-1,1), L_q(-1,1))$$
$$\leq a_n(H_p\cap F^r, L_q(-1,1)).$$

For $a_n=\varepsilon_n$, the second inequality is still valid but the left-hand inequality is replaced by

$$\varepsilon_{m+n}(H_p\cap F^r, L_q(-1,1))-2^{-m/k}\lesssim\varepsilon_n(H_p\cap\psi^r L_\infty(-1,1), L_q(-1,1)).$$

Proof. Write the natural injection $i\colon H_p\cap F^r\to L_q(-1,1)$ in the form $i=(i-p_r)+p_r$, where p_r is the projection of F^r onto P_{k-1}, cf. (2.4). When $a_n=\delta_n$, d_n it follows from (3.3) and $\operatorname{rank} p_r=k$ that

$$a_{n+k}(i)\leq a_n(i-p_r).\tag{4.3}$$

This is the only step in the proof where entropy requires a different treatment, and we have

$$\varepsilon_{m+n}(i)\leq\varepsilon_n(i-p_r)+\varepsilon_m(p_r)\lesssim\varepsilon_n(i-p_r)+2^{-m/k}.\tag{4.4}$$

The second inequality comes from the factorization

$$p_r\colon H_p\cap F^r\to l_\infty^k\xrightarrow{j} l_\infty^k\to L_q(-1,1),$$

where j is the identity on l_∞^k and $\varepsilon_m(j)\simeq 2^{-m/k}$. We now proceed with $a_n=\delta_n$, d_n or ε_n. We may factor $i-p_r$ as

$$i-p_r\colon H_p\cap F^r\xrightarrow{T}H_p\cap\psi^rL_\infty(-1,1)\to L_q(-1,1),$$

with $Tf=(i-p_r)f$, and $\|T\|=1$ in view of (2.4) and the inequalities

$$\|f-p_rf\|_{H_p}\le\|f\|_{H_p}+\|p_rf\|_{H_p}\lesssim\|f\|_{H_p}+\|p_rf\|_{L_\infty(-1,1)}=\|f\|_{H_p\cap F^r}.$$

The left-hand inequalities of the Lemma now follow from (4.3), (4.4) and (3.2). The right-hand inequality is obvious from (2.4).

The substitution $z=\log((1+w)/(1-w))$ induces isometries $H_p\to H_p(\Omega)$, $L_q(-1,1)\to\phi^{1/q}L_q(\mathbb{R})$ and $\psi^rL_\infty(-1,1)\to 2^r\phi^{-r}L_\infty(\mathbb{R})$, noting that $\phi(z)=2/\psi(w)$. As a consequence we obtain the following equivalence.

LEMMA 5.

$$a_n\big(H_p\cap\psi^rL_\infty(-1,1),L_q(-1,1)\big)\simeq a_n\big(H_p(\Omega)\cap\phi^{-r}L_\infty(\mathbb{R}),\phi^{1/q}L_q(\mathbb{R})\big).$$

As one might expect when approximating functions on the line, the precise behavior of the functions near the boundary of Ω is not important for the rate of a_n. In fact, it turns out that what matters is merely the approximate rate of growth of $|f(z)|$ for $f\in H_p(\Omega)$. The next lemma is what we need for the reduction from $H_p(\Omega)$ to $\phi^\lambda H_\infty(\Omega_d)$.

LEMMA 6. *For $\varepsilon>0$ one has*

$$\varepsilon^{1/p}\|f\|_{\phi^{1/p}H_\infty(\Omega_{\pi/2-\varepsilon})}\lesssim\|f\|_{H_p(\Omega)}\lesssim\varepsilon^{-1/p}\|f\|_{\phi^{1/p-\varepsilon}H_\infty(\Omega)}.$$

Proof. The lower bound follows from an inequality of Hardy and Littlewood [5]. For $g\in H_p=H_p(\Delta)$

$$|g(w)|\le 2^{1/p}\big(1-|w|^2\big)^{-1/p}\|g\|_{H_p}.$$

For $f\in H_p(\Omega)$ let $g(w)=f(\log((1+w)/(1-w)))$. Then $\|f\|_{H_p(\Omega)}=\|g\|_{H_p}$ and hence by (4.2)

$$|f(z)|\le\cos^{-1/p}(y)|\phi(z)|^{1/p}\|f\|_{H_p(\Omega)}.$$

For the upper bound we observe that

$$\|\psi^{-1/p+\varepsilon}g\|_{H_p}\le\|\psi^{-1/p+\varepsilon}\|_{H_p}\|g\|_{H_\infty}$$

and

$$\|\psi^{-1/p+\varepsilon}\|_{H_p}\lesssim\left(\int_0^\pi\sin^{\varepsilon p-1}\theta\,d\theta\right)^{1/p}\cong\varepsilon^{-1/p}.$$

Remark. In analogy to similar characterizations of Hardy spaces on the upper half-plane [5], one can show that $H_p(\Omega)=\phi^{1/p}\mathcal{H}_p\neq\mathcal{H}_p$, where $f\in\mathcal{H}_p$ iff f is analytic in Ω and

$$\|f\|_{\mathcal{H}_p}=\sup_{|y|<\pi/2}\left\{\int_{-\infty}^{\infty}|f(x+iy)|^p\,dx\right\}^{1/p}<\infty.$$

The proof of this nonelementary result makes use of the factorization theorem for the Nevanlinna class N^+ and the fact that $(1-w^2)^{-1/p}$ is an outer function.

LEMMA 7. *Let* $\varepsilon>0$. *Then* $\|f\|_{L_q(\mathbb{R})} \lesssim \varepsilon^{-1/q}\|\phi^{\varepsilon} f\|_{L_\infty(\mathbb{R})}$.

This inequality is useful when proving upper bounds, as it implies $a_n(X,\phi^{1/q}L_q(\mathbb{R})) \lesssim \varepsilon^{-1/q}a_n(X,\phi^{1/q-\varepsilon}L_\infty(\mathbb{R}))$. The lower bounds require a different technique employing a regularization mapping $\phi^{1/q}L_q(\mathbb{R})$ into a weighted L_∞-space, cf. Lemma 9 below.

Next, we describe the approximation processes on the line to be used in the proofs of our main theorems in §5. Let $\rho>0$, $t>0$, $\nu\in\mathbb{Z}$ and define the functions

$$s_\nu(z)=s_{\nu t\rho}(z)=\phi^{-\rho}(z-\nu/t)\frac{\sin\pi(tz-\nu)}{\pi(tz-\nu)}.$$

Notice that s_ν is holomorphic in Ω_π and $s_\nu(\mu/t)=\delta_{\mu\nu}$. Let $S_{nt\rho}=\operatorname{span}\{s_{\nu t\rho}\colon |\nu|\leq n\}$ and define the interpolatory projections

$$P_{nt\rho}\colon C(\mathbb{R})\to S_{nt\rho},$$
$$P_{nt\rho}f=\sum_{|\nu|\leq n} f(\nu/t)s_{\nu t\rho}.$$

In addition to these finite-rank approximations we also need the series

$$P_{t\rho}f=\sum_{\nu\in\mathbb{Z}} f(\nu/t)s_{\nu t\rho}.$$

For $\rho=0$ this is the Whittaker cardinal series [18]. As mentioned in the introduction, the cardinal series was employed by Stenger [13] in obtaining his upper bounds. Lundin and Stenger [9] and Stenger [15] also used weighted cardinal series similar to ours. The next lemma implies that the series $P_{t\rho}f$ converges uniformly on $\mathbb{R}$ if $(f(\nu/t))_{\nu\in\mathbb{Z}}$ is a bounded sequence.

We now establish bounds on the condition number of the basis $\{s_\nu\}$. It is perhaps surprising that these crude estimates, where the coefficients grow as powers of n (the parameter t will turn out to be proportional to $n^{1/2}$), suffice for determining the order of a_n.

LEMMA 8. *For* $(a_\nu)_{\nu\in\mathbb{Z}}$ *in* $l_\infty(\mathbb{Z})$, $t\geq 1$

$$\|(a_\nu)\|_{l_\infty}\leq\Big\|\sum_{\nu\in\mathbb{Z}} a_\nu s_{\nu t\rho}\Big\|_{L_\infty(\mathbb{R})}\lesssim t\frac{e^{\rho/t}}{\rho}\|(a_\nu)\|_{l_\infty}. \tag{4.5}$$

Proof. For the upper bound we must estimate the Lebesgue function $L(x)=\sum_{\nu\in\mathbb{Z}}|s_{\nu t\rho}(x)|$. By (4.1)

$$\|L\|_{L_\infty(\mathbb{R})}\leq\sup_{x\in\mathbb{R}}\sum_{\nu\in\mathbb{Z}} e^{-\rho|x-\nu/t|}.$$

The value of the last sum evidently depends only on the residue of $xt \bmod 1$, hence let $0\leq xt\leq 1$. Then $|x-\nu/t|\geq(|\nu|-1)/t$ and for $t\geq 1$

$$\|L\|_{L_\infty(\mathbb{R})}\lesssim e^{\rho/t}\sum_{\nu\in\mathbb{N}} e^{-\rho\nu/t}\lesssim te^{\rho/t}/\rho.$$

A similar estimate can also be proved for $L_\infty(\mathbb{R})$ replaced by $L_q(\mathbb{R})$.

LEMMA 9. *For a positive integer* m, $\rho>0$ *and* $1\leq q\leq\infty$

$$(\rho+t)^{-1/q}n^{-1/q}\|(a_\nu)_{|\nu|\leq n}\|_{l_\infty^{2n+1}}\lesssim\Big\|\sum_{|\nu|\leq n} a_\nu s_{\nu t\rho}\Big\|_{L_q(\mathbb{R})}\lesssim\frac{n}{(\rho q)^{1/q}}\|(a_\nu)_{|\nu|\leq n}\|_{l_\infty^{2n+1}}. \tag{4.6}$$

Proof. The right-hand inequality follows from (4.1). To show the lower bound, for each $|\nu|\le n$ extend to $L_q(\mathbb{R})$ the linear functional on $S_{nt\rho}$ given by $\Sigma_{|\kappa|\le n} a_\kappa s_{\kappa t\rho} \to a_\nu$. A convenient extension is L_ν, where

$$L_\nu s = \sum_{|\mu|\le m} b_{\nu\mu}\frac{1}{2\varepsilon}\int_{\mu/t-\varepsilon}^{\mu/t+\varepsilon} s(x)\,dx, \qquad s\in L_q(\mathbb{R}),$$

and $B=(b_{\nu\mu})$ is the inverse of $A=(a_{\mu\kappa})$,

$$a_{\mu\kappa}=\frac{1}{2\varepsilon}\int_{\mu/t-\varepsilon}^{\mu/t+\varepsilon} s_{\kappa t\rho}(x)\,dx, \qquad \varepsilon>0, \qquad |\mu|,\ |\kappa|\le n.$$

As $s_{\kappa t\rho}(\mu/t)=\delta_{\kappa\mu}$, we have the bound

$$|a_{\mu\kappa}-\delta_{\mu\kappa}|\le\frac{1}{2\varepsilon}\int_{\mu/t-\varepsilon}^{\mu/t+\varepsilon}|s_{\kappa t\rho}(x)-s_{\kappa t\rho}(\mu/t)|\,dx$$

$$\le\varepsilon\sup_{x\in\mathbb{R}}\left|\frac{d}{dx}s_{\kappa t\rho}(x)\right|\le M\varepsilon(\rho+t)$$

where M is a constant. We choose $\varepsilon=(1/(8M))(\rho+t)^{-1}n^{-1}$ and obtain

$$\|A-I\|_\infty\le(2n+1)M\varepsilon(\rho+t)\le\frac{1}{2}.$$

Therefore

$$\|B\|_\infty\le\frac{1}{1-\|A-I\|_\infty}\le 2.$$

This gives for $1/q'+1/q=1$

$$\|L_\nu\|_{L_{q'}(\mathbb{R})}\lesssim\frac{1}{\varepsilon}\varepsilon^{1/q'}=\varepsilon^{-1/q}\cong(\rho+t)^{1/q}n^{1/q}.$$

Finally, in this section, we state a formula for the error of approximation $f-P_{nt\rho}f$ which follows from the calculus of residues and was extensively used by F. Stenger [13].

LEMMA 10. *If $f\in H_\infty(\Omega_d)$, $0<d<\pi$ and $x\in\Omega_d$, then*

$$(4.7)\qquad (f-P_{t\rho}f)(x)=\frac{\sin\pi tx}{2\pi i}\int_{\partial\Omega_d}\frac{\phi^{-\rho}(x-z)f(z)}{(z-x)\sin(\pi tz)}dz.$$

Proof. Denote by R_n the rectangle

$$\left\{x+iy\in\Omega_{d-\varepsilon}:\ |x|<\frac{n+1/2}{t}\right\}.$$

If we replace $P_{t\rho}$ by $P_{nt\rho}$ and $\partial\Omega_d$ by ∂R_n, then (4.7) follows from the residue theorem when $x\in R_n$. We now let $\varepsilon\to 0$ and $n\to\infty$ and apply the dominated convergence theorem. Here we are using the existence of nontangential limits in $L_\infty(\partial\Delta)$ for $f\in H_\infty=H_\infty(\Delta)$, as the lines $|\mathrm{Re}\,z|=\text{const}$, $z\in\Omega_d$, transform conformally to nontangential curves in Δ under the substitution $z=\log((1+w)/(1-w))$. We also make use of (4.1) and of

$$(4.8)\qquad |\sin(x+iy)|=\frac{1}{2}|2\cosh 2y-2\cos 2x|^{1/2}\ge\frac{1}{2}(e^{|y|}-1).$$

LEMMA 11. *Under the hypotheses of Lemma* 10, *when* $d=\pi/2$ *and* $t\geq 1$

$$(4.9)\qquad \|f-P_{tp}f\|_{L_\infty(\mathbb{R})}\lesssim\exp\left(-\frac{\pi^2}{2}t\right)\|f\|_{H_\infty(\Omega_d)}.$$

Proof. From (4.7) and (4.8)

$$\|f-P_{tp}f\|_{L_\infty(\mathbb{R})}\lesssim\left(\exp\left(\frac{\pi^2}{2}t\right)-1\right)^{-1}\|f\|_{H_\infty(\Omega_d)}\lesssim\exp\left(-\frac{\pi^2}{2}t\right)\|f\|_{H_\infty(\Omega_d)}.$$

For reference below we note

$$(4.10)\qquad \sup_{z\in\Omega_d}\left|\frac{\sin\pi(tz-\nu)}{\pi(tz-\nu)}\right|\lesssim\exp(\pi dt),$$

a consequence of the maximum modulus theorem and (4.8).

We note that approximation in $L_\infty(\mathbb{R})$ can be replaced by approximation in $C(\mathbb{R})$, the subspace of continuous functions in $L_\infty(\mathbb{R})$. More precisely, for a compact linear operator T: $X\to C(\mathbb{R})$ we have

$$(3.5)\qquad a_n(T)=a_n(jT)$$

where j: $C\to L_\infty$ is the injection. It is clear from (3.2) that $a_n(jT)\leq a_n(T)$ as $\|j\|=1$. To prove the reverse inequality note that the compactness of T implies that there exists a modulus of continuity $w(\delta,t)$ such that

$$\begin{aligned}&w(\delta,t)\to 0,\ \delta\to 0\\ &w(\delta,t)\leq w(\delta,t'),\qquad t\leq t',\end{aligned}$$

and for all $f\in T(B(X))$

$$\sup_{\substack{|s|,|s'|\leq t\\ |s-s'|\leq\delta}}|f(s')-f(s)|\leq w(\delta,t).$$

For $\varepsilon>0$ we choose a function $h>0$ such that $w(h(|t|),t)<\varepsilon$ and define a smoothing operator R_ε: $L_\infty\to C$ by

$$(R_\varepsilon f)(t)=\frac{1}{2h(|t|)}\int_{t-h(|t|)}^{t+h(|t|)}f(s)\,ds.$$

Then $\|R_\varepsilon\|=1$ and $a_n(jT)\geq a_n(R_\varepsilon jT)=a_n(R_\varepsilon T)\geq a_n(T)-\|T-R_\varepsilon T\|$ by (3.1), (3.2) and (3.3). From the definition of h we see that $\|T-R_\varepsilon T\|\leq\varepsilon$ which, since ε is arbitrary, finishes the proof.

5. Proofs. To simplify notation we shall use the abbreviations $X_{r,d}=\phi^r H_\infty(\Omega_d)$, $X_r=X_{r,\pi/2}$, $X=X_0$, $Y_r=\phi^r C(\mathbb{R})$, $Y=Y_0$.

Proof of Theorem 1. By Lemmas 4 and 5 it suffices to estimate $d_n(i)$, $\delta_n(i)$ with i: $X\cap Y_{-r}\to Y$. In view of the obvious inequality $d_n\leq\delta_n$ we shall bound δ_n from above and d_n from below. To prove the upper estimate we write $i=(i-P_{tr})+P_{tr}$ and by (3.1), (3.3) we have

$$\delta_n(i)\leq\|i-P_{tr}\|+\delta_n(P_{tr}).$$

By Lemma 11 we have

$$\|i-P_{tr}\| \lesssim \exp\left(-\frac{\pi^2}{2}t\right). \tag{5.1}$$

To estimate the second term we factor P_{tr} as

$$P_{tr}\colon Y_{-r} \xrightarrow{I} l_{\infty,r} \to l_\infty \xrightarrow{J} Y \tag{5.2}$$

where $(If)_\nu = f(\nu/t)$ and $Ja = \Sigma a_\nu s_{\nu tr}$. Clearly $\|I\|=1$ and since by Lemma 8 $\|J\| \lesssim t$ we obtain using (3.2) and Lemma 2

$$\delta_n(P_{tr}) \lesssim t \exp\left(-\frac{r}{2t}n\right). \tag{5.3}$$

Combining the estimate (5.1) with (5.2) and choosing $t=(rn)^{1/2}/\pi$ gives the upper bound.

To prove the lower estimate we consider the following factorization of the identity on l_∞^{2m+1},

$$l_\infty^{2m+1} \xrightarrow{J_m} X \cap Y_{-r} \xrightarrow{i} Y \xrightarrow{I_m} l_\infty^{2m+1},$$

where I_m and J_m are defined analogous to I and J. Using the estimates

$$\|s_\nu\|_X \lesssim \exp\left(\frac{\pi^2}{2}t\right),$$
$$\|s_\nu\|_{Y_{-r}} \lesssim \exp\left(\frac{r}{t}m\right), \qquad |\nu| \le m$$

for the norms of the basis functions s_ν we obtain, choosing $t=(2rm)^{1/2}/\pi$, $m=\lfloor n/2 \rfloor$

$$\|J_m\| \lesssim m\left(\exp\left(\frac{r}{t}m\right)+\exp\left(\frac{\pi^2}{2}t\right)\right) \lesssim n \exp\left(\frac{\pi}{\sqrt{2}}(rn)^{1/2}\right).$$

The lower bound follows now from (3.2) and Lemma 1.

We now formulate a general result which allows a unified treatment of the proofs of Theorems 2–4 and is of independent interest.

THEOREM 5. *Let a_n denote either one of the numbers d_n, δ_n or ε_n. For $\lambda \ge 0$, $\rho > 0$ and $t > 0$ we have*

$$a_n(l_{\infty,\rho/t}, l_\infty) \times \exp\left(-\frac{\pi^2}{2}\frac{\rho}{\lambda+\rho}t\right) \ll a_n(X_\lambda \cap Y_{-\rho}, Y) \tag{5.4}$$
$$\ll a_n(l_{\infty,\rho/t}, l_\infty) + \exp\left(-\frac{\pi^2}{2}\frac{\rho}{\lambda+\rho}t\right).$$

As a consequence we obtain

$$\exp\left(-\pi\left(\frac{\rho^2}{\lambda+\rho}n\right)^{1/2}\right) \ll d_n(X_\lambda \cap Y_{-\rho}, Y),\ \delta_n(X_\lambda \cap Y_{-\rho}, Y) \tag{5.5}$$
$$\ll \exp\left(-\frac{\pi}{2}\left(\frac{\rho^2}{\lambda+\rho}n\right)^{1/2}\right)$$

and

$$(5.6)\quad \exp\left(-(4\pi^2\log 2)^{1/3}\left(\frac{\rho^2}{\lambda+\rho}n\right)^{1/3}\right) \ll \varepsilon_n(X_\lambda\cap Y_{-\rho},Y)$$

$$\ll \exp\left(-\left(\frac{\pi^2}{2}\log 2\right)^{1/3}\left(\frac{\rho^2}{\lambda+\rho}n\right)^{1/3}\right).$$

Proof. To prove the upper estimate in (5.4) we write the embedding i: $X_\lambda\cap Y_{-\rho}\to Y$ in the form $i=(i-P_{t\sigma})+P_{t\sigma}$ where we choose $\sigma>\lambda,\rho$. As in the proof of Theorem 1 we estimate

$$a_n(i)\le\|i-P_{t\sigma}\|+a_n(P_{t\sigma})$$

and obtain for the second term (cf. (5.2))

$$a_n(P_{t\sigma})\le t a_n(l_{\infty,\rho/t},l_\infty).$$

Therefore we have to show

$$\|i-P_{t\sigma}\|\ll\exp\left(-\frac{\pi^2}{2}\frac{\rho}{\lambda+\rho}t\right).$$

To estimate $\sup_{x\in\mathbf{R}}|f(x)-P_{t\sigma}f(x)|$ we set $A=(\pi^2/2)t/(\lambda+\rho)$ and consider two cases.

(i) $|x|\le A$. Lemma 10 and the estimates (4.1), (4.8) imply that

$$|f(x)-P_{t\sigma}f(x)|\lesssim\exp\left(-\frac{\pi^2}{2}t\right)\int_{\mathbf{R}}\exp(-\sigma|x-z|+\lambda|z|)\,dz\,\|f\|_{X_\lambda}$$

$$\lesssim\exp\left(-\frac{\pi^2}{2}t+\lambda A\right)\|f\|_{X_\lambda}.$$

(ii) $|x|>A$. Since $|f(x)|\lesssim\exp(-\rho|x|)\|f\|_{Y_{-\rho}}$ it follows that

$$|P_{t\sigma}f(x)|\lesssim\sum_\nu\exp(-\rho|\nu/t|-\sigma|x-\nu/t|)\|f\|_{Y_{-\rho}}\lesssim t\exp(-\rho|x|)\|f\|_{Y_{-\rho}}.$$

This implies that for $|x|>A$

$$|f(x)-P_{t\sigma}f(x)|\lesssim\exp(-\rho A)\|f\|_{Y_{-\rho}}.$$

Combining the estimates (i) and (ii) completes the proof by our choice of A.

To prove the lower bound in (5.4) we consider for $\varepsilon>0$ and $A=[(\pi^2/2)t^2/(\lambda+\rho)]$ the following factorization of the embedding j: $l_{\infty,\rho/t+\varepsilon}\to l_\infty$

$$l_{\infty,\rho/t+\varepsilon}\xrightarrow{J_A}X_\lambda\cap Y_{-\rho}\xrightarrow{i}Y\xrightarrow{I_A}l_\infty$$

where $(I_Af)_\nu=f((\nu+b(\nu))/t)$ and $J_Aa=\sum_{\nu\in\mathbf{Z}}a_\nu s_{\nu+b(\nu),t,\sigma}$. Here $b(\nu)$ is defined as $b(\nu)=A\operatorname{sgn}\nu$ and σ is chosen larger than λ and ρ. Using the inequalities (4.1), (4.10) we obtain by a simple calculation, keeping in mind the choice of σ,

$$\|s_{\nu+b(\nu)}\|_{X_\lambda}\lesssim\exp\left(\frac{\pi^2}{2}t-\lambda A/t\right),$$

$$\|s_{\nu+b(\nu)}\|_{Y_{-\rho}}\lesssim\exp(\rho(A+|\nu|)/t).$$

Therefore

$$\begin{aligned}\|J_A a\|_{X_\lambda\cap Y_{-\rho}} &\leq \sum |a_\nu|\,\|s_{\nu+b(\nu)}\|_{X_\lambda\cap Y_{-\rho}}\\ &\lesssim \exp\left(\frac{\pi^2}{2}t-\lambda A/t\right)\sum |a_\nu| + \exp(\rho A/t)\sum \exp(\rho|\nu|/t)|a_\nu|\\ &\lesssim \left(\exp\left(\frac{\pi^2}{2}t-\lambda A/t\right)+\exp(\rho A/t)\right)\sum \exp(-\varepsilon|\nu|)\|a\|_{l_{\infty,\rho/t+\varepsilon}}\end{aligned}$$

and by our choice of A we get

$$\|J_A\| \lesssim \varepsilon^{-1}\exp\left(\frac{\pi^2}{2}\frac{\rho}{\lambda+\rho}t\right).$$

Since $\|I_A\|=1$ this implies

$$a_n(j)\lesssim \varepsilon^{-1}\exp\left(\frac{\pi^2}{2}\frac{\rho}{\lambda+\rho}t\right)a_n(i).$$

Taking into account the asymptotic behaviour of $a_n(j)=a_n(l_{\infty,\rho/t+\varepsilon}, l_\infty)$ the lower estimate follows by the appropriate choice of ε.

The inequalities (5.5) and (5.6) follow form (5.4) by substituting the bounds for $a_n(l_{\infty,\rho/t}, l_\infty)$ obtained in Lemmas 2 and 3 and choosing t appropriately. More precisely for $a_n=\delta_n$, d_n we choose $t=((2/\pi^2)(\lambda+\rho)n)^{1/2}$ and for $a_n=\varepsilon_n$ let

$$t=\left(\frac{4\log 2(\lambda+\rho)^2 n}{\pi^4\rho}\right)^{1/3}.$$

Using Theorem 5 we can now easily give the proofs of Theorems 2–4.

Proof of Theorem 2. In view of Lemmas 4 and 5 we have to estimate the n-width and approximation numbers of i: $H_p(\Omega)\cap Y_{-r}\to Y$.

For the upper estimate consider for $\varepsilon>0$ the following factorization of i

$$H_p(\Omega)\cap Y_{-r}\overset{j_1}{\to}X_{1/p,\pi/2-\varepsilon}\cap Y_{-r}\overset{T}{\to}X_{\lambda/p}\cap Y_{-\lambda r}\overset{j_2}{\to}Y\overset{T^{-1}}{\to}Y$$

where $\lambda=(\pi/2-\varepsilon)/(\pi/2)$ and T is defined by

$$(Tg)(z)=g(\lambda z).$$

Since by (4.1)

$$|\phi^{\lambda\rho}(z)|\simeq|\phi^\rho(\lambda z)| \quad \text{for } z, \lambda z\in\Omega_d,\quad d<\pi,$$

we have that

$$\begin{aligned}&\|T\colon X_{\rho,d}\to X_{\lambda\rho,d/\lambda}\|\lesssim 1, \qquad d, d/\lambda<\pi,\\ &\|T\colon Y_{-\rho}\to Y_{-\lambda\rho}\|\lesssim 1.\end{aligned}$$

Using (3.2), Lemma 6 and (5.5) we obtain from the above factorization

$$\delta_n(i)\leq\|j_1\|\|T\|\delta_n(j_2)\|T^{-1}\|\ll\varepsilon^{-1/\rho}\exp(-f(\varepsilon)n^{1/2})$$

where

$$f(\varepsilon)=\frac{\pi}{2}\left(\frac{(\lambda r)^2}{\lambda/p+\lambda r}\right)^{1/2}, \qquad \lambda=\frac{\pi/2-\varepsilon}{\pi/2}.$$

Since f is a smooth function of ε, we get by setting $\varepsilon=n^{-1/2}$

$$\delta_n(i)\ll\exp(-f(0)n^{1/2}-f'(\xi)), \qquad \xi\in[0,n^{-1/2}].$$

This proves the upper bound in view of

$$f(0)=\frac{\pi}{2}\left(\frac{r^2}{r+1/p}\right)^{1/2}.$$

To prove the lower bound we consider for $\varepsilon>0$ the embedding

$$j_1\colon X_{1/p-\varepsilon}\cap Y_{-r}\to Y,$$

which may be factored as

$$X_{1/p-\varepsilon}\cap Y_{-r}\xrightarrow{j_2}H_p(\Omega)\cap Y_{-r}\xrightarrow{i}Y.$$

By Lemma 6 we have $\|j_2\|\lesssim\varepsilon^{-1/p}$. Applying (3.2) and substituting the estimate (5.5) for $d_n(j_1)$, we get

$$\exp\left(-\pi\left(\frac{r^2}{1/p-\varepsilon+r}\right)^{1/2}n^{1/2}\right)\ll d_n(j_1)\lesssim\varepsilon^{-1/p}d_n(i).$$

As in the proof of the upper bound we write this inequality in the form

$$\varepsilon^{1/p}\exp(-g(\varepsilon)n^{1/2})\ll d_n(i).$$

Writing $g(\varepsilon)=g(0)+\varepsilon g'(\xi)$ and setting $\varepsilon=n^{-1/2}$ finishes the proof.

The proof of Theorem 4 is completely analogous and therefore omitted. We simply use the estimate (5.6) for ε_n instead of (5.5).

Proof of Theorem 3. Recall that by Lemmas 4 and 5 we may estimate $d_n(i)$, $\delta_n(i)$, where i: $H_p(\Omega)\to\phi^{1/q}L_q(\mathbb{R})$.

To prove the upper estimate we consider for $\varepsilon>0$, $\lambda=2(\pi/2-\varepsilon)/\pi$ and $\rho=1/q-(\lambda/2)(1/q+1/p)$ the following factorization of i:

$$i\colon H_p(\Omega)\xrightarrow{j_1}X_{1/p,\pi/2-\varepsilon}\xrightarrow{T}X_{\lambda/p}\xrightarrow{j_2}Y_{1/q-\rho}\xrightarrow{T^{-1}}Y_{(1/q-\rho)/\lambda}\xrightarrow{j_3}\phi^{1/q}L_q(\mathbb{R})$$

where T is defined by $(Tg)(z)=g(\lambda z)$. As we already pointed out in the proof of Theorem 2, $\|T\|$, $\|T^{-1}\|\lesssim 1$. By our choice of λ,ρ and since $p>q$, we have $\lambda/p<1/q-\rho$ and $(1/q-\rho)/\lambda<1/q$. Therefore the embeddings j_1, j_2, j_3, are well defined. By Lemmas 6, 7 $\|j_1\|\lesssim\varepsilon^{-1/p}$, $\|j_3\|\lesssim 1$. Using (3.2) we obtain

$$\delta_n(i)\lesssim\varepsilon^{-1/p}\delta_n(j_2).$$

Since $a_n(X_\sigma,Y_\tau)=a_n(X_{\sigma-\tau},Y)\lesssim a_n(X\cap Y_{\sigma-\tau},Y)$ this and (5.5) imply

$$\delta_n(i)\lesssim\varepsilon^{-1/p}\delta_n(X\cap Y_{\lambda/p-1/q+\rho},Y)\lesssim\varepsilon^{-1/p}\exp(-h(\varepsilon)n^{1/2}),$$

where $h(\varepsilon)=(\pi/2)(\lambda/p-1/q+\rho)$ with $\lambda=\lambda(\varepsilon)$, $\rho=\rho(\varepsilon)$ as defined above. Since $\lambda(0)=1$, $\rho(0)=0$ we may set $\varepsilon=n^{-1/2}$ and complete the proof as in the previous cases.

The lower estimate, however, cannot be obtained by this technique since there is no embedding $L_q(\mathbb{R}) \to Y_\lambda$. But we may work with L_q directly and proceed similarly as in the proof of Theorem 1. By Lemma 6 and isometries we obtain

$$d_n\big(H_p(\Omega), \phi^{1/q}L_q(\mathbb{R})\big) \gtrsim \varepsilon^{1/p} d_n\big(X_{1/p-\varepsilon}, \phi^{1/q}L_q(\mathbb{R})\big) = \varepsilon^{1/p} d_n\big(X_{1/p-1/q-\varepsilon}, L_q(\mathbb{R})\big).$$

Let $\rho = 1/q - 1/p + \varepsilon$. Then $\rho \geq \varepsilon$ and we can factor the identity

$$I\colon l_\infty^{2n+1} \overset{I_2}{\to} X_{-\rho} \overset{j}{\to} L_q(\mathbb{R}) \overset{I_1}{\to} l_\infty^{2n+1}.$$

Here, $I_2((a_\nu)) = \sum_{|\nu| \leq n} a_\nu s_{\nu t\rho}$. From Lemma 9 and the Hahn–Banach theorem there exist linear functionals L_ν on $L_q(\mathbb{R})$, $|\nu| \leq n$, such that

$$L_\nu\Bigg(\sum_{|\mu| \leq n} a_\mu s_{\mu t\rho}\Bigg) = a_\nu, \qquad \|L_\nu\| \lesssim (\rho + t)^{1/q} n^{1/q}$$

(actually, an explicit construction of L_ν is provided in the proof of the lemma). Let I_1 be defined by $I_1 f = (L_\nu f)_{|\nu| \leq n}$. Then

$$\|I_1\| \leq (\rho + t)^{1/q} n^{1/q}.$$

To estimate $\|I_2\|$ we have from (4.1), (4.9)

$$\|s_{\nu t\rho}\|_{\phi^{-\rho}H_\infty(\Omega)} \lesssim \sup_{x \in \mathbb{R}} e^{\rho|x| - \rho|x - \nu/t|} \exp\left(\frac{\pi^2}{2} t\right) = \exp\left(\rho|\nu|/t + \frac{\pi^2}{2} t\right).$$

Therefore

$$\|I_2\| \lesssim n \exp\left(\rho n/t + \frac{\pi^2}{2} t\right).$$

From Lemma 1 and (3.2) there follows now

$$1 = d_{2n}(I) \leq \|I_1\| d_{2n}(j) \|I_2\|,$$

hence

$$d_n(j) \gtrsim (\rho + t)^{-1/q} n^{-1-1/q} \varepsilon^{1/p} \exp\left(-\rho n/(2t) - \frac{\pi^2}{2} t\right).$$

The proof is completed by the choice of $t = \sqrt{\rho n}/\pi$ and then $\varepsilon = n^{-1/2}$.

Remark. Combining the ideas of the preceding proofs, we obtain the estimates (2.5). Indeed we have, cf. Lemma 5,

$$\delta_n\big(H_p^*, L_\infty(-1,1)\big) = 2\delta_n\big(\phi^{-1}H_p(\Omega), L_\infty(\mathbb{R})\big).$$

Let $\rho = 1 - 1/p = 1/p'$. From the obvious injection $X_{-\rho} \hookrightarrow X \cap Y_{-\rho}$, Lemma 6, (3.5), (5.5) and using T, λ as defined above

$$\delta_n\big(\phi^{-1}H_p(\Omega), L_\infty(\mathbb{R})\big) \lesssim \varepsilon^{-1/p} \delta_n\big(X_{-\rho,\pi/2-\varepsilon}, Y\big)$$

$$\lesssim \varepsilon^{-1/p} \delta_n\big(X_{-\rho}, Y\big) \lesssim \varepsilon^{-1/p} \delta_n\big(X \cap Y_{-\rho}, Y\big) \ll \exp\left(-\frac{\pi}{2}(\rho n)^{1/2}\right),$$

$\varepsilon = n^{-1/2}$, as in the proof of Theorem 2.

For the lower bound

$$\delta_n\big(\phi^{-1}H_p(\Omega), L_\infty(\mathbb{R})\big) \geq \varepsilon^{1/p}\delta_n\big(Y_{-\rho-\varepsilon}, Y\big) \gg \exp\big(-\pi(\rho n)^{1/2}\big),$$

where we proceed as in the proof of Theorem 3, but using Lemma 8 instead of Lemma 9.

The following slight extension of our results is now easily obtained. It concerns the question about the dependence of $a_n(H_p(D)\cap F^r, L_q(-1,1))$ on the domain D. A domain much smaller than Δ might be useful to consider when it is a matter of approximating functions with singularities very near interior points of the interval of approximation $(-1,1)$. Or, conversely, one might be interested in functions known to have singularities far from $(-1,1)$. Suitable domains generalizing Δ are given by

$$\Delta_d = \left\{ w \in \mathbb{C} : \left| \arg \frac{1+w}{1-w} \right| < d \right\}, \qquad 0 < d < \pi, \quad \Delta = \Delta_{\pi/2}.$$

These were considered by Stenger [13–16], who obtained upper bounds. The substitution $z = \log((1+w)/(1-w))$ maps Δ_d conformally and 1-to-1 onto Ω_d. Using the isometry $(Tg)(z) = g(\lambda z)$, $\lambda = 2d/\pi$, we find we can reduce the problem of bounding

$$a_n\big(H_p(\Delta_d)\cap F^r, L_q(-1,1)\big)$$

to that of bounding

$$a_n\big(H_p(\Omega)\cap Y_{-\lambda r}, \phi^{\lambda/q}L_q(\mathbb{R})\big)$$

which we have already solved. We state only two of the resulting estimates for illustration:

$$d_n(H_\infty(\Delta_d)\cap F^r, L_\infty(-1,1)) >< \exp\Big(-\frac{\pi}{2}(\lambda r n)^{1/2}\Big)$$

where $r > 0$, $\lambda = 2d/\pi$, and for $p > q/\lambda$

$$\exp(-2\beta n^{1/2}) \ll d_n\big(H_p(\Delta_d), L_q(-1,1)\big) \ll \exp(-\beta n^{1/2})$$

where $\beta = (\pi/2)(\lambda/q - 1/p)^{1/2}$ and similarly in the remaining cases.

REFERENCES

[1] L. V. Ahlfors, *Complex Analysis*, McGraw-Hill, New York, 1966.

[2] B. D. Bojanov, *Best quadrature formula for a certain class of analytic functions*, Zastosowania Matematijki Appl. Mat. XIV (1974), pp. 441–447. (In Russian.)

[3] H. G. Burchard and K. Höllig, *N-width and entropy of H_p-classes in $L_q(-1,1)$*, abstract 81T-B3, Abstracts, Amer. Math. Soc., 2 (1981), pp. 240–241.

[4] R. DeVore and K. Scherer, *Variable knot, variable degree spline approximation to x^β*, In Proc. Conference on Quantitative Approximation, Bonn, New York, 1979.

[5] P. L. Duren, *Theory of H^p-spaces*, Academic Press, New York, 1970.

[6] A. A. Goncar, *Piecewise polynomial approximation*, Mat. Zametki, 11 (1972), pp. 129–134. (In Russian.)

[7] K. Höllig, *Diameters of classes of smooth functions*, In Proc. Conference on Quantitative Approximation, Bonn, 1979, pp. 163–175.

[8] G. G. Lorentz, *Approximation of Functions*, Holt, Rinehart and Winston, New York, 1966.

[9] L. Lundin and F. Stenger, *Cardinal-type approximation of a function and its derivatives*, this Journal, 10 (1979), pp. 139–160.

[10] C. A. Micchelli and S. D. Fisher, *N-width of sets of analytic functions*, Duke Math. J., 48 (1980), pp. 789–801.

[11] A. PIETSCH, *S-numbers of operators in Banach spaces*, Studia Mathematatica, 51 (1974), pp. 201–223.

[12] K. SCHERER, *On optimal global error bounds obtained by scaled local error estimates*, Numer. Math., 36 (1981), pp. 151–176.

[13] F. STENGER, *Approximations via Whittaker's cardinal function*, J. Approx. Theory, 17 (1976), pp. 222–240.

[14] ______, *Upper and lower estimates on the rate of convergence of approximations in H_p*, Bull. AMS (1977), pp. 145–148.

[15] ______, *Optimal convergence of minimum-norm approximations in H_p*, Numer Math., 29 (1978), pp. 345–362.

[16] ______, *Numerical methods based on Whittaker cardinal, or sinc functions*, SIAM Rev., 23 (1981), pp. 165–224.

[17] E. T. WHITTAKER, *On the functions which are represented by the expansions of the interpolation theory*, Proc. Roy. Soc. Edinburgh, 35 (1915), pp. 181–194.

[18] J. M. WHITTAKER, *On the cardinal function of interpolation theory*, Proc. Edinburgh Math. Soc. Ser 1, 2 (1927), pp. 41–46.

[19] A. G. VITUSHKIN, *Theory of Transmission and Processing of Information*, Pergamon, New York, 1961.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

001

SYSTEMS OF DIFFERENTIAL EQUATIONS THAT ARE COMPETITIVE OR COOPERATIVE II: CONVERGENCE ALMOST EVERYWHERE*

MORRIS W. HIRSCH†

Abstract. A vector field in n-space determines a competitive (or cooperative) system of differential equations provided all of the off-diagonal terms of its Jacobian matrix are nonpositive (or nonnegative). The main results in this article are the following. A cooperative system cannot have nonconstant attracting periodic solutions. In a cooperative system whose Jacobian matrices are irreducible the forward orbit converges for almost every point having compact forward orbit closure. In a cooperative system in 2 dimensions, every solution is eventually monotone. Applications are made to generalizations of positive feedback loops.

Introduction. This paper studies the limiting behavior of solutions of systems

$$\frac{dx^i}{dt}=F^i(x^1,\cdots,x^n) \qquad (i=1,\cdots,n) \tag{1}$$

which are either cooperative:

$$\frac{\partial F^i}{\partial x^j}\geqq 0 \quad \text{for } i\neq j,$$

or competitive:

$$\frac{\partial F^i}{\partial x^j}\leqq 0 \quad \text{for } i\neq j.$$

The main results are as follows:

If (1) is cooperative, there are no attracting nonconstant periodic solutions (Theorem 2.4).

If (1) is cooperative and irreducible and its flow is $\{\phi_t\}$, then $\phi_t(x)$ approaches the equilibrium set for almost every point x whose forward orbit has compact closure (Theorem 4.1).

If $n=2$ every solution to (1) is eventually monotone (Theorem 2.7).

The author's earlier paper [6] investigated compact limit sets of (1) from a geometrical and topological point of view. The present paper is concerned more directly with dynamical behavior. While it is formally independent of [6], it uses some of the same techniques.

Most of the results concern cooperative systems which are irreducible in the sense that the matrices $[(\partial F^i/\partial x^j)(p)]$ are irreducible. This has the important consequence that the flow $\{\phi_t\}$ corresponding to (1) has positive derivatives for $t>0$, i.e. the matrices $D\phi_t(p)$ have only positive entries. By Kamke's theorem such a flow is strongly

* Received by the editors April 26, 1983.

† Department of Mathematics, University of California, Berkeley, California 94720. This research was supported in part by the National Science Foundation under grant MCS 80-02858.

monotone, that is,

$$\phi_t(x) < \phi_t(y) \quad \text{if } x \leqq y, \quad x \neq y, \quad t > 0.$$

The class of irreducible cooperative vector fields is contained in the wider class of vector fields whose flows have eventually positive derivatives: $D\phi_t > 0$ for all $t > t_0$, for some $t_0 \geqq 0$. A flow of this type is eventually strongly monotone, which is as useful as being strongly monotone for most purposes. Moreover suitably small perturbations (in the weak C^1 topology) of such a vector field have positive derivatives in a given compact set, and thus are eventually strongly monotone in such a set. This has the great advantage of allowing the use of perturbation methods. In particular we make crucial use of the Closing Lemma of C. Pugh in the proof of Theorem 4.1.

Section 1 contains results on the monotonicity of various kinds of flows. Section 2 studies equilibria and closed orbits of eventually monotone flows. The key result is Theorem 2.2 which gives a useful sufficient condition for a solution to converge.

In §3 the Closing Lemma is exploited to prove basic results (3.7), (3.8), (3.9) about the relative position of ω-limit sets. These are applied in §4 to derive conditions under which almost all bounded forward trajectories converge (Theorems 4.1 and 4.4). Theorem 4.6 implies that compact attractors contain equilibria. Theorem 4.7 shows that invariant functions are usually constant.

Section 5 shows how the theorems proved in earlier sections can be applied to systems in the nonnegative orthant $\mathbb{R}^n_+$. The positive feedback loop (Selgrade [13], [14]) is used as an example. The Closing Lemma is discussed in the Appendix.

Many of the results in this paper can be extended to strongly monotone semiflows in ordered Banach spaces, including those defined by solutions to certain parabolic evolution equations. For these results see the author's paper [19].

The following terminology will be used throughout the paper: $\mathbb{R}$ is the field of real numbers; real n-space is $\mathbb{R}^n$, the vector space of n-tuples $x = (x^1, \cdots, x^n)$ of real numbers. F: $W \to \mathbb{R}^n$ is a C^1 (continuously differentiable) vector field on a nonempty open set $W \subset \mathbb{R}^n$. For any $x \in W$ we denote the maximally defined solution of the vector differential equation

$$\frac{d\xi}{dt} = F(\xi), \qquad \xi(0) = x$$

by $t \to \phi_t(x)$, $t \in I(x) \subset \mathbb{R}$. For each $t \in \mathbb{R}$ the set of $x \in w$ for which $\phi_t(x)$ is defined is a (possibly empty) open set $W(t) \subset W$ and ϕ_t: $W(t) \to W(-t)$ is a C^1 diffeomorphism. The collection of maps $\{\phi_t\}_{t \in \mathbb{R}}$ is called the *flow* of F, or of the differential equation $dx/dt = F(x)$. For $x \in W$ we also write $x(t)$ for $\phi_t(x)$.

The forward trajectory of $x \in W$ is the parametrized curve $t \to \phi_t(x)$ ($t \geqq 0$, $t \in I(x)$). Its image is the *forward orbit* of x, $O_+(x)$. The backward trajectory of x and the backward orbit $O_-(x)$ are analogously defined.

A subset $X \subset W$ is *positively* (respectively, *negatively*) *invariant* if $O_+(x) \subset X$ (resp. $O_-(x) \subset X$) for all $x \subset X$. It is *invariant* if it is both positively and negatively invariant.

The *ω-limit set* $\omega(x)$ of a point $x \in W$, or of a solution $x(t)$, is the set of $p \in W$ such that $x(t_k) \to p$ for some sequence $t_k \to \infty$. The *α-limit set* $\alpha(x)$ is similarly defined with $t_k \to -\infty$.

If $O_+(x)$ has compact closure in W then $\omega(x)$ is a nonempty compact connected invariant set. An analogous statement is true of $\alpha(x)$.

We say a forward trajectory $x(t)$ *approaches* a subset $S\subset W$ if $O_+(x)$ has compact closure in W and $\omega(x)\subset S$. This implies

$$\lim_{t\to\infty}\inf_{y\in S}|x(t)-y|=0.$$

If $V\subset\mathbb{R}^n$ is open, h: $V\to\mathbb{R}^n$ is C^1 and $p\in V$, then $Dh(p)$ denotes the $n\times n$ matrix $[(\partial h^i/\partial x^j)(p)]$. In particular there are the matrices $DF(p)$ and $D\phi_t(p)$.

Let $x, y\in\mathbb{R}^n$. We write:

$$\begin{aligned} x<y &\quad \text{if } x^i<y^i \quad \text{for all } i,\\ x\leqq y &\quad \text{if } x^i\leqq y^i \quad \text{for all } i. \end{aligned}$$

The notations $y>x$, $y\geqq x$ have the obvious meaning. We call x positive if $x>0$ (the zero vector). Similar notation applies to matrices.

The *closed positive orthant* is the set

$$\mathbb{R}^n_+=\{x\in\mathbb{R}^n\colon x\geqq 0\}.$$

If X, $Y\subset\mathbb{R}^n$ are any subsets we write $X<Y$ if $x<y$ for all $x\in X$, $y\in Y$. We analogously define $X>Y$, $X\leqq Y$, etc.

An $n\times n$ matrix $A=[A_{ij}]$ is *irreducible* if whenever the set $\{1,\cdots,n\}$ is expressed as the union of two disjoint proper subsets S, S', then for every $i\in S$ there exists j, $k\in S'$ such that $A_{ij}\neq 0$, $A_{ki}\neq 0$. This means the linear map A: $\mathbb{R}^n\to\mathbb{R}^n$ does not map into itself any nonzeo proper linear subspace spanned by a subset of the standard basis. Equivalently: the directed graph with vertices $1,\cdots,n$ and directed edges (i,j) for $A_{ij}\neq 0$, is connected by directed paths.

If $x\in\mathbb{R}^n$ then $|x|$ is the Euclidean norm $(\Sigma x_i^2)^{1/2}$. If A is a real $n\times n$ matrix then $\|A\|$ is the operator norm

$$\max\{|Ax|\colon x\in\mathbb{R}^n \text{ and } |x|=1\}.$$

If K, L are sets then

$$K\setminus L=\{x\in K\colon x\notin L\}.$$

1. Monotone flows. For reference we quote a corollary of a result of Kamke [8].

KAMKE'S THEOREM. *Let $V\subset\mathbb{R}^n$ be an open set and G: $\mathbb{R}\times V\to\mathbb{R}^n$ a continuous map such that $G^i(t,x^1,\cdots,x^m)$ is a nondecreasing in x^k, for all $k\neq i$. Let*

$$\xi,\eta\colon [a,b]\to V$$

be solutions of

$$\frac{dx}{dt}=G(t,x)$$

such that $\xi(a)<\eta(a)$ (resp. $\xi(a)\leqq\eta(a)$). Then $\xi(t)<\eta(t)$ (resp. $\xi(t)\leqq\eta(t)$) for all $t\in[a,b]$.

For a proof see Coppel [2] or W. Walter [16].

Notice that the assumption on G is satisfied if $\partial G^i/\partial x^k\geqq 0$ whenever $k\neq i$ and V is convex. In fact V need only be *p-convex*: whenever $x, y\in V$ and $x\leqq y$ then V contains the entire line segment joining x and y.

Now consider the system

$$(1)\qquad \frac{dx^i}{dt}=F^i(x), \qquad i=1,\cdots,n$$

defined by the C^1 vector field F: $W\to\mathbb{R}^n$. The flow $\{\phi_t\}$ of the system is called:

cooperative if the off-diagonal entries of $DF(z)$ are $\geqq 0$ for all $z\in w$;

competitive if the off-diagonal entries of $DF(t)$ are $\leqq 0$ for all $z\in w$;

irreducible if $DF(t)$ is irreducible for all $z\in W$.

These adjectives will also be applied to the system (1) or the vector field F, with the same meaning.

It is a well-known consequence of Kamke's theorem that when W is convex, the flow of a cooperative system (1) preserves the ordering $\leqq$ in $\mathbb{R}^n$. In 1.4 and 1.5 below we extend this result.

We say $\{\phi_t\}$ has *nonnegative* (resp. *positive*) derivatives if $D\phi_t\geqq 0$ (resp. >0) for all $t>0$, $z\in W$.

THEOREM 1.1. *Let F be a cooperative vector field. Then*:

(a) $\{\phi_t\}$ *has nonnegative derivatives.*

(b) *If F is also irreducible then $\{\phi_t\}$ has positive derivatives.*

Proof. Fix $t\in W$. For all $t\in J(z)$ define matrices

$$A(t)=DF(\phi_t(z)), \qquad M(t)=D\phi_t(z).$$

Then $M(t)$ satisfies the variational equation

$$(2)\qquad \frac{dM}{dt}=A(t)M$$

with initial condition

$$(3)\qquad M(0)=I,$$

where I is the $n\times n$ identity matrix. The right-hand side of (2) is a matrix function $G(t,M)$ whose entries are

$$G_{ik}(t,M_{11},\cdots,M_{nn})=\sum_{j=1}^{n}A_{ij}(t)M_{jk}.$$

It is easily verified that (because F is cooperative)

$$\frac{\partial G_{ij}}{\partial M_{rs}}\geqq 0 \quad \text{if } (i,k)\neq(r,s).$$

It follows from Kamke's theorem that the solution $M(t)$ to (2), (3) satisfies $M_{ik}(t)\geqq 0$ for all i,k and all $t\geqq 0$ because the constant map $N(t)=0$ (the $n\times n$ zero matrix) is also a solution to (2), and $M(0)\geqq N(0)$. Thus $M(t)\geqq 0$ for all $t\geqq 0$.

Now assume F irreducible. If $t_0>0$ and $M(t_0)>0$ then Kamke's theorem implies $M(t)>0$ for all $t>t_0$. Suppose it is not the case that $M(t)>0$ for all $t>0$. Then there exists $t_1>0$ such that $M_{ij}(t_1)=0$ for some i,j. It follows that for every $t\in[0,t_1]$, $M_{ij}(t)=0$ for some i,j (depending on t). One of the sets

$$[0,t_1]\cap M_{ij}^{-1}(0)$$

must have an interior point. Therefore there exist i,j and an interval $[a,b]\subset\mathbb{R}_+$ such that $M_{ij}(t)=0$ for all $t\in[a,b]$. But the following lemma contradicts this.

LEMMA. *Suppose i, $k\in\{1,\cdots,n\}$ and $t_0>0$ are such that $M_{ik}(t_0)=0$. Then $(d/dt)M_{ik}(t_0)>0$.*

To prove the lemma define

$$S=\{r\colon M_{rk}(t_0)=0\}.$$

Then S is nonempty (since $i\in S$) and $S\neq\{1,\cdots,m\}$ (since the matrix $M(t_0)$ is nonsingular for all t). Now the matrix $A(t_0)$ is irreducible. Therefore there exists $j\in\{1,\cdots,\}\backslash S$ with $A_{ij}(t_0)\neq 0$. Clearly $j\neq i$, so $A_{ij}(t_0)>0$ since F is cooperative. We now have

$$\frac{d}{dt}M_{ik}(t_0)=\sum_{r=1}^{n}A_{ir}(t_0)M_{rk}(t_0).$$

Since $M_{ik}(t_0)=0$ we can write the sum as

$$\sum_{\substack{r=1\\ r\neq i}}^{n}A_{ir}(t_0)M_{0k}(t_0).$$

In this expression each summand is the product of nonnegative numbers, and the terms with $r=j$ are positive. Thus all $M_{ik}(t_i)>0$. This proves the lemma; the proof of the theorem is now complete. QED.

Perturbations of a cooperative irreducible vector field need not preserve the property that its flow have positive derivatives. But a slightly weaker, equally useful property is preserved: that of having *eventually positive derivatives.* A flow $\{\phi_t\}$ has this property in a set V provided there exists $t_0>0$ such that $D\phi_t(z)>0$ for all $t\geqq t_0$, $z\in V$.

THEOREM 1.2. *Assume $K\subset W$ is a compact set in which the flow $\{\phi_t\}$ has eventually positive derivatives. Then there exists $\delta>0$ with the following property. Let $\{\psi_t\}$ denote the flow of a C^1 vector field G such that*

$$(4)\qquad |F(x)-G(x)|+\|DF(x)-DG(x)\|<\delta\quad \textit{for all } x\in K.$$

Then there exists $t_*>0$ such that if $t\geqq t_*$ and $\psi_s(z)\in K$ for all $s\in[0,t]$ then $D\psi_t(z)>0$. In particular: if K is positively invariant under $\{\psi_t\}$ then $\{\psi_t\}$ has eventually positive derivatives in K.

Proof. Fix $t_0>0$ so that $D\phi_t(z)>0$ for all $t\geqq t_0$, $z\in K$. Fix $\delta>0$ so small that (4) implies

$$D\psi_t(z)>0\quad\text{if } t_0\leqq t\leqq 2t_0\quad\text{if } z\in K.$$

Now fix $t\geqq 2t_0$. Write

$$t=r+kt_0$$

where

$$t_0\leqq r<2t_0\quad\text{and}\quad k\text{ is an integer}\geqq 1.$$

Put $jt_0=s_j$ for $j=0,\cdots,k$. Let z be such that $\psi_s(z)\in K$ for all $s\in[0,t]$. Put $\psi_{s_j}(z)=z_j$. By the chain rule

$$(5)\qquad D\psi_t(z)=D\psi_r(z_k)D\psi_{t_0}(z_{k-1})\cdots D\psi_{t_0}(t_0).$$

Now $D\psi_{t_0}(z_i)>0$ for $i=0,\cdots,k-1$ because $z_i\in K$. Also $D\psi_r(t_k)>0$ because $z_k\in K$ and $t_0\leqq r<2t_0$. QED.

Let $W'\subset\mathbb{R}^n$ be a subset. A map f: $W'\to\mathbb{R}^n$ is *monotone* in W' (resp. *strongly monotone*) provided $x\leqq y$ implies $f(x)\leqq f(y)$ (resp. $f(x)<f(y)$ when $x\neq y$).

LEMMA 1.3. *Suppose f: $W'\to\mathbb{R}^n$ is C^1. If $Df(z)\geqq 0$ (resp. $Df(z)>0$) for all $z\in W'$ and W' is p-convex then f is monotone (resp. strongly monotone) in W'.*

Proof. Let a, $b\in W$ with $a\leqq b$. For $s\in[0,1]$ put $a_s=a+s(b-a)$. The lemma follows from the formula

$$f(b)-f(a)=\int_0^1 Df(a_s)(b-a)\,ds. \qquad \text{QED.}$$

The flow $\{\phi_t\}$ is *eventually (strongly) monotone* if there exists $t_0\geqq 0$ such that ϕ_t is (strongly) monotone for all $t>t_0$. If this holds with $t_0=0$ then the flow is called *(strongly) monotone.*

THEOREM 1.4. *Let $V\subset W$ be p-convex. If the flow $\{\phi_t\}$ has (eventually) positive derivatives in V then it is (eventually) strongly monotone in V. If the flow has (eventually) nonnegative derivatives in V then it is (eventually) monotone in V.*

Proof. This follows from Lemma 1.3. QED.

THEOREM 1.5. *Let $V\subset W$ be p-convex. If the vector field F is cooperative (resp., cooperative and irreducible) then its flow $\{\phi_t\}$ is monotone (resp. strongly monotone) in V.*

Proof. Apply Theorems 1.1 and 1.4. QED.

Notice also that every open set W is locally convex, hence the flow of a cooperative vector field in W monotone in some neighborhood of any point.

The following theorem is a converse to Theorem 1.4.

THEOREM 1.6. *Suppose $\{\phi_t\}$ is (eventually) monotone in an open set $V\subset W$. Then $\{\phi_t\}$ has (eventually) nonnegative derivatives in V.*

Proof. Fix $t>0$ such that ϕ_t is monotone in V. Fix $x\in V$ and $v\in\mathbb{R}^n_+$. Then

$$D\phi_t(x)v=\lim_{h\to 0}h^{-1}(\phi_t(x+hv)-\phi_t(x)).$$

If $h>0$ is sufficiently small we have $x+hv\in v$ and $\phi_t(x+hv)\geqq\phi_t(x)$. Thus $D\phi_t(x)v$ is a limit of vectors in $\mathbb{R}^n_+$, so $D\phi_t(x)v\geqq 0$. This shows $D\phi_t(x)$ maps $\mathbb{R}^n_+$ into itself. QED.

For convenience we present a summary of some of the implications between various kinds of systems. The following abbreviations are used:

C:	cooperative,
CI:	cooperative and irreducible,
ND:	nonnegative derivatives,
PD:	positive derivatives,
EPD:	eventually positive derivatives,
M:	monotone,
SM:	strongly monotone,
ESM:	eventually strongly monotone,
$\Rightarrow$:	implications,
$\overset{x}{\Rightarrow}$:	implication is valid in p-convex sets.

THEOREM 1.7.

$$\begin{tikzcd}
 & M & \\
C \arrow[r, Rightarrow] & ND \arrow[r, Rightarrow, "x"] & M \\
CI \arrow[u, Rightarrow] \arrow[r, Rightarrow] & PD \arrow[u, Rightarrow] \arrow[r, Rightarrow, "x"] \arrow[d, Rightarrow] & SM \arrow[u, Rightarrow] \arrow[d, Rightarrow] \\
 & EPD \arrow[r, Rightarrow, "x"] & ESM
\end{tikzcd}$$

2. Equilibria and close orbits. Let $\{\phi_t\}$ denote the flow of a C^1 vector field F defined on the open set $W \subset \mathbb{R}^n$.

A point $p \in W$ is an *equilibrium* if $F(p)=0$, or equivalently, if $\phi_t(p)=p$ for all $t \in \mathbb{R}$.

Let p be an equilibrium. Then $D\phi_t(p)=\exp(tDF(p))$. The spectrum (set of complex eigenvalues) of $D\phi_t(p)$ is related to that of $DF(p)$ by

$$\operatorname{Spec} D\phi_t(p) = \{e^{t\lambda}\colon \lambda \in \operatorname{Spec} DF(p)\}.$$

The equilibrium p is called:

simple if $0 \notin \operatorname{Spec} DF(p)$; or equivalently, if $1 \notin \operatorname{Spec} D\phi_t(p)$ for some $t \in \mathbb{R}$;

hyperbolic if $\operatorname{Re}\lambda \neq 0$ for all $\lambda \in \operatorname{Spec} DF(p)$; or equivalently, if $|\mu| \neq 1$ for all $\mu \in \operatorname{Spec} D\phi_t(p)$ and some (hence all) $t \neq 0$;

a *sink* if $\operatorname{Re}\lambda < 0$ for all $\lambda \in \operatorname{Spec} DF(p)$; or equivalently, if $|\mu| < 1$ for all $\mu \in \operatorname{Spec} D\phi_t(p)$ and some (hence all) $t > 0$.

It is well known that a sink p is asymptotically stable; that is, every neighborhood of p contains a positively invariant neighborhood N of p such that

(1) $\phi_t(x)$ converges to p uniformly in $x \in N$ as $t \to \infty$.

A weaker notion is that of a *trap*: an equilibrium p such that there is some open set N, not necessarily containing p, such that (1) holds.

It is not known how to characterize traps in terms of the vector field without integrating it. It is easy to see, however, if p is a trap then $\operatorname{Re}\lambda \leqq 0$ for all λ in $\operatorname{Spec} DF(p)$. It follows that a simple trap is a sink, and $\operatorname{Div} F \leqq 0$ at a trap.

An *ω-colimit point* of points u, v in W is a point $p \in W$ such that

$$p = \lim_{k \to \infty} u(t_k) = \lim_{k \to \infty} v(t_k)$$

for some sequence $t_k \to \infty$.

LEMMA 2.1. *Let $\{\phi_t\}$ be eventually monotone in an open set $W_0 \subset W$. Let $p \in W$ be an ω-colimit point of points x, y in W_0 where $x < y$. Then p is an equilibrium.*

Proof. Let $T > 0$ be so large that $x(t) < y(t)$ for all $t \geqq T$. Let $t_k \to \infty$, p, $y(t_k) \to p$. Choose $k_0 \in \mathbb{Z}_+$ so large that $t_k \geqq T$ for all $k \geqq k_0$. Put $t_{k_0} = s$. The set

$$U = \phi_s^{-1}\{z \in W_0\colon x(s) < z < y(s)\}$$

is a nonempty open set. If $u \in U$ then

$$x(s) < u(s) < y(s).$$

Therefore for $r \geqq T$

$$\phi_r(x(s)) < \phi_r(u(s)) < \phi_r(y(s)),$$

i.e.

$$x(r+s) < u(r+s) < y(r+s).$$

It follows that for all k such that $t_k \geqq T+s$,

$$x(t_k) < u(t_k) < y(t_k).$$

This implies $u(t_k) \to p$ as $k \to \infty$, uniformly in $u \in U$.

As a consequence,

$$\lim_{k \to \infty} \operatorname{diam} \phi_{t_k}(U) = 0$$

where for any $X \subset \mathbb{R}^n$,

$$\operatorname{diam} X = \sup\{|a-b| \colon a, b \in X\}.$$

Now each ϕ_{t_k} maps solution curves to solution curves, preserving parameterization up to an additive constant. It follows that there exists $\tau > 0$ with the following property. For every $t \in [0, \tau]$ and every $\varepsilon > 0$ there exists z in the ε-neighborhood of p such that

$$|z(t) - z| < \varepsilon \quad \text{for all } t \in [0, \tau].$$

Letting $\varepsilon \to 0$ we get

$$|p(t) - p| = 0 \quad \text{for all } t \in [0, \tau]$$

which implies p is an equilibrium. QED.

The following result is basic to the rest of the paper.

THEOREM 2.2. *Assume the flow* $\{\phi_t\}$ *is eventually monotone in an open set* W_0: *let* $t_1 \geqq 0$ *be such that* $\phi_t | W_0$ *is monotone for all* $t > t_1$. *Let* $x(t)$ *be a solution defined for all* $t \geqq 0$. *Suppose* $T > 0$ *is such that* $x(0) \in W_0$ *and* $x(t) \in W_0$, *and either* $x(0) < x(T)$ *or* $x(0) > x(T)$. *If* $p \in W$ *is a limit point of* $\{x(kT) \colon k \in \mathbb{Z}_+\}$ *then the following are true*:

(a) *p is a trap.*

(b) $p = \lim_{t \to \infty} x(t)$.

(c) *Assume* $p \in W_0$. *If* $x(T) > x(0)$ (*resp.* $x(T) < x(0)$) *then* $p > x(t)$ (*resp.* $p < x(t)$) *for all* $t > t_1$.

Proof. We assume $x(0) < x(T)$, the other case being similar.

For all $t > t_1$ we have

$$x(t) < x(t+T).$$

Letting $t = T, 2T, 3T, \cdots$, we find that for all sufficiently large integers $k > 0$:

$$x(kT) < x((k+1)T). \tag{2}$$

It follows that

$$p = \lim_{k \to \infty} x(kT), \qquad (k \in \mathbb{Z}_+). \tag{3}$$

Now apply Lemma 2.1 with $x = x(0)$, $y = x(T)$: clearly p is an ω-colimit point of x, y so p is an equilibrium. For all $s \in [0, T]$, $k \in \mathbb{Z}_+$, we have:

$$\phi_{kT+s}(x) = \phi_s(\phi_{kT}x) \to \phi_s(p) = p$$

as $k \to \infty$. This shows

$$\phi_t(x) \to p \quad \text{as } t \to \infty.$$

Similarly $\phi_t(y) \to p$ as $t \to \infty$. Finally, p is a trap because if $x < z < y$, $z \in w_0$ then

$$\phi_t(x) < \phi_t(z) < \phi_t(y)$$

which shows $\phi_t(z) \to p$ uniformly for z is the nonempty open set

$$\{z \in W_0 \colon x < z < y\}.$$

To prove (c) assume $p \in W_0$ and set

$$C = \{y \in W_0 \colon y < p\}.$$

Since $\phi_t(p) = p$ it follows that $\phi_t(C) \subset C$ for all $t > t_1$. By (2) and (3), $x \in c$. Therefore $x(t) \in C$ for all $t > t_1$, proving (c). QED.

THEOREM 2.3. *Let $\{\phi_t\}$ be eventually monotone.*

(a) *If $y \in \omega(x)$ and $y < x$ or $y > x$ then y is a trap and* $\lim_{t \to \infty} x(t) = y$.

(b) *There cannot exist u, v in $\omega(x)$ with $u < v$.*

(c) *If the flow is eventually strongly monotone there cannot exist u, v in $\omega(x)$ with $u \leqq v$, $u \neq v$.*

Proof. (a) Suppose $y > x$, the other case being similar. There exist arbitrarily large $T > 0$ such that $\phi_T(x) > x$. The conclusion now follows from Theorem 2.2(a) and 2.2(b).

(b) There exists x' in the forward orbit of x so close to u that $x' < v$. Part (a) implies $\lim_{t \to \infty} x'(t) = v$. Therefore $\omega(x) = \{v\}$, contradicting $u \neq v$.

(c) Apply part (b) to $u(t_0)$, $v(t_0)$ where $t_0 > 0$ is so large that $u(t_0) < u(t_0)$. QED.

COROLLARY. *Let $\{\phi_t\}$ be eventually strongly monotone and suppose E is totally ordered. If $x \in W$ is such that $\omega(x)$ is a nonempty compact subset of E then $x(t)$ converges to an equilibrium as $t \to \infty$.*

Proof. Suppose $\phi(x)$ contains two equilibria u and v. Then $u < v$ by strong monotonicity, contradicting Theorem 2.3(b). QED.

By a *closed orbit* we mean the image of a nonconstant periodic solution.

THEOREM 2.4. *An eventually monotone flow cannot have an attracting closed orbit.*

Proof. Let γ be a closed orbit of an eventually monotone flow. Let $y \in \gamma$. In every neighborhood of y there exists a point $x > y$. Since y is not an equilibrium it follows from 2.3(a) that $y \notin \omega(x)$. Since γ is invariant, $\gamma \cap \omega(x) = \varnothing$. This shows that γ is not an attractor. QED.

More generally, a similar argument shows that if a minimal set M is an attractor then M is a single point. We consider attractors again in Theorem 4.6.

The following convergence criterion can be considered an infinitesimal analogue of Theorem 2.2. For cooperative systems it is well known.

THEOREM 2.5. *Let $\{\phi_t\}$ have eventually nonnegative derivatives. Suppose $x \in W$ is such that $F(x) \geqq 0$ (or $F(x) \leqq 0$). Then all coordinates $x^i(t)$ are eventually nondecreasing (or eventually nonincreasing). If $\omega(X) \neq \varnothing$ then $x(t)$ converges to an equilibrium p. If $F(x) > 0$ (or $F(x) < 0$) then p is a trap. If $\{\phi_t\}$ has eventually positive derivatives then p is a trap.*

Proof. Suppose $F(x) \leqq 0$, the other case being similar. Let $t_0 \geqq 0$ be large enough so that $D\phi_t(x) \geqq 0$ for all $t \geqq t_0$. Now

$$F(x(t)) = D\phi_t(x)F(x).$$

Therefore for all $t \geqq t_0$ it follows that $F(x(t)) \leqq 0$. In other words $(d/dt)x^i(t) \leqq 0$ for all $i = 1, \cdots, n, t \geqq t_0$. Thus $x(t)$ is eventually nonincreasing.

If $p \in \omega(x)$ then $x(t)$ must converge to p. Each of the last three hypotheses implies the existence of $t_1 > 0$ such that $F(x(t)) < 0$ for all $t \geqq t_1$. Put $y = x(t_1)$. Then $y(T) < y$ for sufficiently small $T > 0$. By Theorem 2.3(a) $v(t)$ converges to a trap as $t \to \infty$. QED.

The following lemma says that solutions lying in certain two-dimensional affine subspaces are eventually monotone.

LEMMA 2.6. *Let $y(t)$ be a trajectory of a flow having eventually nonnegative derivatives. Suppose $y(t)$ defined for all $t \geqq 0$. Suppose there exist $t_0 \geqq 0$, and distinct j, $k \in \{1, \cdots, n\}$, such that $y_i(t)$ is constant for all $t \geqq t_0$, $i \neq j$, k. Then $y_j(t)$ and $y_k(t)$ are monotone for sufficiently large t. If $\omega(y) \neq \varnothing$ then $y(t)$ converges.*

Proof. We may assume $F(y(t)) \neq 0$ for all $t \geqq t_0$. If $F(y(t_1) \geqq 0$ or $F(y(t_1)) \leqq 0$ for some $t \geqq t_0$ the conclusion follows from Theorem 2.5. In the contrary case, the vector $(F_j(y(t)), F_k(y(t))) \in \mathbb{R}^2$ is confined, for all $t \geqq t_0$, to the second or fourth open quadrant. This implies eventual monotonicity of $y_j(t)$ and $y_k(t)$. QED.

THEOREM 2.7. *Let F be a vector field in an open subset of the plane. Assume that F is cooperative, or competitive, or that the flow has eventually nonnegative derivatives. Let $x(t)$ be a solution defined for $-\infty < t \leqq 0$, or for $0 \leqq t < \infty$. Then each $x^i(t)$ is monotone for $|t|$ sufficiently large.*

Proof. It suffices to consider the case where the flow has eventually nonnegative derivatives: this holds if F is cooperative; and if F is competitive it suffices to prove the lemma for $-F$ (which is cooperative).

If $x(t)$ is constant there is nothing more to prove. Assume $x(t)$ is not constant. Suppose $x(t)$ is defined for $-\infty < t \leqq 0$. Let $t_0 > 0$ be such that ϕ_t has nonnegative derivatives for $t \geqq t_0$.

Case 1. Assume there exists $t_1 \leqq 0$ such that $F(x(t_1))$ is in open quadrant II or IV (i.e. it is neither $\leqq 0$ nor $\geqq 0$). Then Theorem 2.6 implies that for all $t \leqq t_1 - t_0$, $F(x(t))$ is in quadrant II or IV. Since $F(x(t)) \neq 0$, $F(x(t))$ cannot pass directly from II to IV. Therefore $F(x(t))$ is in the same quadrant for all $t \leqq t_1 - t_0$. Therefore $x^1(t)$, $x^2(t)$ are monotone for $t \leqq t_1 - t_0$.

Case 2. $F(t)$ is never in II or IV. Then $F(t)$ must stay in I or III for all $t < 0$ and again $x^i(t)$ is monotone, $i = 1, 2$.

When $x(t)$ is defined for all $t > 0$ a similar argument applies. QED.

COROLLARY 2.8. *Let $\{\phi_t\}$ be a cooperative or competitive flow in $\mathbb{R}^2$ for which the nonnegative quadrant $\mathbb{R}^2_+$ is positively invariant. Then every bounded trajectory* $[0, \infty) \to$ *$\mathbb{R}^2_+$ converges.*

Versions of the last result have been proved many times: see for example Albrecht et al. [1], Grossberg [4], Hirsch–Smale [7], Kolmogorov [9], Rescigno–Richardson [12].

3. ω-Limits. Throughout this section we assume given a C^1 vector field F on a p-convex open set $W \subset \mathbb{R}^n$, whose flow $\{\phi_t\}$ has eventually positive derivatives. It follows from Theorem 1.5 that $\{\phi_t\}$ is eventually strongly monotone.

The following notation is used:

$E \subset W$ is the set of equilibria;

x, y are distinct points of W and $x \leqq y$.

The main results of this section are Theorems 3.7, 3.8 and 3.9.

LEMMA 3.1. *Let p be an ω-colimit point of x, y (see* §2*). Then $p \in E$. If $x(t)$, $y(t)$ converge to p as $t \to \infty$, then p is a trap.*

Proof. Follows from Lemma 2.1. QED.

Throughout the rest of this section we assume that $\omega(x)$ *and* $\omega(y)$ *are compact and nonempty.*

LEMMA 3.2. *If* $p\in\omega(x)\backslash E$ *then* $p<q$ *for some* $q\in\omega(y)$.

Proof. It is easy to see that there exists $q\in\omega(y)$ such that for some sequence $t_k\rightarrow\infty$,

$$p=\lim_{k\rightarrow\infty}x(t_k),\qquad q=\lim_{k\rightarrow\infty}y(t_k).$$

Monotonicity implies $p\leqq q$. If $p=q$ then p is an ω-colimit point of x,y. Then $p\in E$ by Lemma 3.1, a contradiction. Therefore $p\neq q$.

Let $T>0$ be so large that ϕ_t is strongly monotone for all $g\geqq T$. Since $\omega(x)$ and $\omega(y)$ are compact and negatively invariant, we can define

$$p_0=\phi_{-T}(p)\in\omega(x),$$
$$q_0=\phi_{-T}(q)\in\omega(y).$$

Clearly $p_0\neq q_0$, and

$$p_0=\lim_{k\rightarrow\infty}\phi_{t_k-T}(x),$$
$$q_0=\lim_{k\rightarrow\infty}\phi_{t_k-T}(y).$$

By the argument above it follows that $p_0\leqq q_0$. Therefore

$$p=\phi_T(p_0)<\phi_T(q_0)=q. \qquad \text{QED.}$$

COROLLARY 3.3. $\omega(X)\cap\omega(y)\subset E$.

Proof. Suppose $p\in\omega(x)\cap\omega(y)\backslash E$. Then by Lemma 3.2 there exists $q\in\omega(y)$ with $q>p$; and also $p\in\omega(y)$. But this is impossible by Theorem 2.3(b). QED.

LEMMA 3.4. *Let* K,L *be compact invariant sets such that* $K\geqq L$. *Then either* $K>L$ *or else there is an equilibrium* b *such that* $K\cap L=\{b\}$ *and* $a<b<c$ *for all* $a\in K\backslash\{b\}$, $c\in L\backslash\{b\}$.

Proof. Clearly $K\cap L$ is compact and invariant and $K\geqq K\cap L\geqq L$. Suppose b, $d\in K\cap L$. Then $b\leqq d$ and $d\leqq b$, so $b=d$. Therefore $K\cap L=\{b\}$. Since $\{b\}$ is invariant $b\in E$.

Choose $T>0$ so large that ϕ_T is strongly monotone. Let $x\in K$, $y\in L$ be distinct. Then $\phi_{-T}(x)\geqq\phi_{-T}(y)$ because $K\geqq L$, so $x>y$. This implies $K>L$ if $K\cap L=\varnothing$; and also if $K\cap L=\{b\}$ then $K\backslash b>b>L\backslash b$. QED.

LEMMA 3.5. *Let* $K\subset\omega(x)$, $M\subset\omega(y)$ *be nonempty subsets with* $K<M$. *If one of the sets* K,M *is compact and positively invariant then* $\omega(x)\leqq\omega(y)$.

Proof. First assume K is compact and positively invariant. Set

$$V=\{z\in W\colon z>K\}.$$

Then $V\cap\omega(y)\neq\varnothing$ and monotonicity implies $y(t)\in V$ for all sufficiently large $t>0$. This implies $\omega(y)\geqq K$. I claim $\omega(y)>K$. If not, by Lemma 3.4 there exists an equilibrium b such that $K\leqq b\leqq\omega(y)$ and $K\cap\omega(y)=\{b\}$. Now $\{b\}=\omega(y)$, otherwise $c>b$ for some $c\in\omega(y)$, contradicting Theorem 2.3(a). But then $b\in K\cap M$, contradicting $K<M$. Therefore $\omega(y)>K$.

Put $M_0=\omega(y)$. Then M_0 is compact and positively invariant and nonempty. By what has already been proved it follows that $\omega(x)<M_0$, i.e. $\omega(x)<\omega(y)$. In a similar way one shows that if M is compact and positively invariant then $\omega(x)<\omega(y)$. QED.

LEMMA 3.6. *Let $p\in\omega(x)$, $q\in\omega(y)$, $p<q$. If p or q is a periodic point then $\omega(x)<\omega(y)$.*

Proof. If p is an equilibrium, apply Lemma 3.5 with $K=\{p\}$, $M=\{q\}$.

A similar argument is used if q is an equilibrium.

Suppose p belongs to a closed orbit γ and q is not an equilibrium. By Lemma 3.4 γ is disjoint from $\omega(y)$. By the Closing Lemma (Pugh [10], or see Appendix) there is a C^1 vector field G whose flow $\{\psi_t\}$ has a closed orbit β through q. Moreover G can be chosen to coincide with F outside any given neighborhood of the orbit closure of Q, and to C^1-approximate F as closely as desired. Now the orbit closure of q is in the compact set $\omega(y)\subset W\backslash\gamma$. Therefore we can choose G so close to F that $\{\psi_t\}$ is eventually strongly monotone, by Theorems 1.3 and 1.5; and we can choose G to coincide with F in a neighborhood of γ. Therefore γ is a closed orbit of $\{\psi_t\}$.

We have closed orbits γ,β of $\{\psi_t\}$ and points $p\in\gamma$, $q\in\beta$ with $p<q$. It follows from Corollary 2.9 that $\gamma<\beta$. In particular, $\gamma<q$.

We now consider the original flow $\{\phi_t\}$. From Lemma 3.5 (with $K=\gamma$ and $M=\{q\}$) we conclude that $\omega(x)<\omega(y)$. QED.

THEOREM 3.7. *Suppose there exist $p\in\omega(x)$, $q\in\omega(y)$ with $p<q$. Then $\omega(x)<\omega(y)$.*

Proof. If p is an equilibrium then the theorem follows from Theorem 3.5. From now on assume p is not an equilibrium.

By Corollary 3.3 the orbit of p is disjoint from $\omega(y)$. We apply the Closing Lemma of Pugh [10] (see Appendix) to obtain a vector field G whose flow $\{\psi_t\}$ has a closed orbit γ through p. As in the proof of Lemma 3.6, we can choose $\{\psi_t\}$ so close to $\{\phi_t\}$ as to be eventually strongly monotone. Moreover there exists $T>0$ with the following property. For any $\varepsilon>0$ we can choose G so that $G=F$ outside the ε-neighborhood N_ε of the solution arc

$$\{\phi_t(p):\ |t|\leqq T\}.$$

For small ε, N_ε is disjoint from $\omega(y;\{\phi_t\})$. Therefore for sufficiently small ε, $\omega(y;\{\phi_t\})$ is also an ω-limit set for the flow $\{\psi_t\}$, of some point y_0. It is easy to see that we can choose $y_0>p$. It now follows from Lemma 3.6 that

$$\gamma=\omega(p;\{\psi_t\})<\omega(y_0;\{\psi_t\}).$$

Thus

$$p<\omega(y_0;\{\psi_t\})=\omega(y,\{\phi_t\}).$$

Now apply Lemma 3.5 to $\{\phi_t\}$, with $K=p, M=\omega(y;\{\phi_t\})$. QED.

THEOREM 3.8. *Exactly one of the following conditions holds*:

(a) $\omega(x)<\omega(y)$.

(b) $\omega(x)=\omega(y)\subset E$.

Proof. For any $p\in\omega(x)$ there exists $q\in\omega(y)$ such that $x(t_k)\to p$, $y(t_k)\to q$ for some sequence $t_k\to\infty$. Then $p\leqq q$. If $p\neq q$ then $p<q$ by eventual strong monotonicity; and then $\omega(x)<\omega(y)$ by Theorem 3.7. If $p=q$ then p is an ω-colimit point of x,y and so $p\in E$ by Lemma 2.1.

Suppose (a) is false. Then the results just proved show that for all $p\in\omega(x)$, whenever q is chosen as above then $p=q\in E$. Thus $\omega(x)\subset\omega(y)\cap E$. A similar argument shows $\omega(y)\subset\omega(x)\cap E$, and thus (b) holds. QED.

4. Convergence almost everywhere. We continue the assumption of §3: $W\subset\mathbb{R}^n$ is a p-convex open set and $\{\phi_t\}$ is a flow in W having eventually positive derivatives.

Let $W^c\subset W$ be the set of points whose forward orbit has compact closure in W.

THEOREM 4.1. *There is a set $Q \subset W^c$ having Lebesgue measure zero, such that $x(t)$ approaches the equilibrium set E as $t \to \infty$, for all $x \in W^c \backslash Q$.*

The proof of Theorem 4.1 uses Lemma 4.2 and 4.3 below. Put

$$Q = \{x \in W^c \colon \omega(x) \not\subset E\}.$$

It is easy to prove that W^c and Q are Borel sets. We must prove that Q has measure zero.

Fix a vector $v > 0$ in $\mathbb{R}^n$. Let $E^{n-1} \subset \mathbb{R}^n$ be the hyperplane orthogonal to v and let $\pi \colon \mathbb{R}^n \to E^{n-1}$ be orthogonal projection. To prove that Q has measure zero it suffices by Fubini's theorem to prove:

(1) For all $w \in E^{n-1}$, $Q \cap \pi^{-1}(w)$ is countable.

We need two lemmas. Let $\pi_i \colon \mathbb{R}^n \to \mathbb{R}$ be the ith coordinate projection. Let a and $b > a$ be two points of W. Since W is a p-convex, W contains the line segment L from a to b. For each $i \in \{1, \cdots, n\}$ define

$$S_i(L) = \{x \in Q \cap L \colon \pi_i(\omega(x)) \text{ has more than one element}\}.$$

LEMMA 4.2. *$S_i(L)$ is countable for $i = 1, \cdots, n$.*

Proof. For any $z \in S_i(L)$ let $K_z \subset \mathbb{R}$ be the interval spanned by $\pi_i(\omega(z))$. Let x, $y \in S_i(L)$ be distinct. Since $a < b$ we may assume $x < y$. Since x, $y \in Q$, Theorem 3.8 implies $\omega(x) < \omega(y)$. Therefore $K_x \cap K_y = \varnothing$. Thus the intervals K_x, $x \in S_i(L)$, are pairwise disjoint. Each interval contains a rational number, so this family of intervals is countable. Hence $S_i(L)$ is countable. QED.

LEMMA 4.3. *$Q \cap L$ is countable.*

Proof. If $x \in Q \cap L$ then $\omega(x)$ is not a single point. Therefore $x \in S_i(L)$ for some $i \in \{1, \cdots, n\}$, so $Q \cap L = \bigcup_{i=1}^n S_i(L)$. It follows from Lemma 4.2 that $Q \cap L$ is countable. QED.

Proof of 4.1. Lemma 4.3 implies (1) since $W \cap \pi^{-1}(w)$ is the union of a countable family of line segments of the type L. This completes the proof of Theorem 4.1.

By imposing some slightly generic behavior on the equilibria we obtain stronger conclusions.

THEOREM 4.4. (a) *Assume E is countable. Then $x(t)$ converges to a trap as $t \to \infty$, for almost all $x \in W^c$.*

(b) *Assume all equilibria are simple. Then $x(t)$ converges to a sink as $t \to \infty$, for almost all $x \in W^c$.*

Proof. Since a simple trap is a sink, (b) follows from (a). Recall the definition

$$Q = \{x \in W^c \colon \omega(x) \not\subset E\},$$

and define

$$N = \{x \in W^c \backslash Q \colon \omega(x) \text{ is not a trap}\}.$$

If $x \in W^c \backslash Q$ then $\omega(x)$ is a compact connected nonempty subset of E. Therefore, since E is countable, $\omega(x)$ is a single equilibrium, which we denote by $e(x)$. We obtain a map $e \colon W^c \backslash Q \to E$.

Let L be a line segment in W parallel to a positive vector. The map

$$e \colon L \cap N \to E$$

is injective: If x,y are distinct points of $L\cap N$ we may assume $x<y$; then $e(x)$ is not a trap so $e(x)\neq e(y)$ by Lemma 3.1.

This proves that $L\cap N$ is countable. In Lemma 4.3 we proved that $L\cap Q$ is countable. Therefore an argument using Fubini's theorem, similar to the proof of Theorem 4.1, shows that $Q\cup N$ has measure zero, which proves Theorem 4.4. QED.

The *stable set* of an equilibrium p is

$$S(p)=\{x\in W\colon \omega(x)=\{p\}\}.$$

COROLLARY 4.5. *Assume E is countable. If $p\in E$ is not a trap, then $S(p)$ has measure zero.*

Proof. Use Lemma 4.3(a). QED.

Even without assuming countability of E one can sometimes prove the existence of traps.

THEOREM 4.6. *Let $K\subset W$ be a compact attractor. Then K contains a trap. If all equilibria in K are simple then K contains a sink.*

Proof. By definition K is a compact nonempty positively invariant set having a neighborhood U such that

$$\varnothing \neq \omega(x)\subset K \quad \text{for all } x\in U.$$

It follows easily from Theorem 4.1 that K contains an equilibrium. Let $p\in K\cap E$ be maximal for the vector ordering $\leqq$. By Theorem 4.1 there exists $y\in U$ such that $y>p$ and $\omega(y)\cap K\cap E\neq\varnothing$. Let $q\in\omega(y)\cap K\cap E$. Then $q\geqq p$ since $y>p$ and the flow is eventually monotone. Therefore $q=p$ by maximality. In this way we can find $y_2>y_1>p$ such that $y_i(t)\to p$ as $t\to\infty$, $i=1,2$. This implies p is a trap (hence a sink if it is simple), since $y(t)\to p$ uniformly in $y\in\{x\in W\colon y_2>x>y_2\}$. QED.

We turn to invariant functions. As an example, consider the cooperative irreducible system

$$\frac{dx}{dt}=-2x^3+y+z,$$
$$\frac{dy}{dt}=x^3-2y+2z,$$
$$\frac{dz}{dt}=x^3+y-3z.$$

The function $x+y+z$ is invariant, as is seen by adding up the right-hand sides of the equations. The equilibrium set is the curve $z=(3/4)x^4$, $y=(5/4)x^3$. Thus the equilibria are quite degenerate. This is not an accident: Theorem 4.3 shows that when E is nondegenerate there are very likely to be traps, whereas a continuous invariant function must be constant on the domain of attraction of a trap. The following result strengthens this conclusion.

THEOREM 4.7. *Suppose the equilibrium set is countable. Let $A\subset W$ be a connected open set such that almost every point of A has compact forward orbit closure. Then every continuous invariant function f is constant on A.*

Proof. Let the traps be p_1, $p_2,\cdots$; let the domain of attraction of p_i be D By Theorem 4.4(a) the set $\bigcup_i A\cap D_i$ is dense in A.

Clearly f is constant on each $\overline{D}_i$. Therefore $f(A)$ is countable. Since A is connected, $f(A)$ is connected and countable, hence $f(A)$ is a single point, and the theorem is proved. QED.

5. Systems in $\mathbb{R}^n_+$. In this section we consider a C^1 vector field F: $\mathbb{R}^n \to \mathbb{R}$ and the corresponding system

$$\frac{dx^i}{dt} = F^i(x^1, \cdots, x^n), \qquad i = 1, \cdots, n \tag{1}$$

such that

(2) the system is cooperative and irreducible;

(3) $F(0) \geqq 0$;

(4) for any $x \in \mathbb{R}^n_+$ there exists $y > x$ with $F^i(y) < 0$, $i = 1, \cdots, n$.

An example is a "positive feedback loop" of n species, of the form

$$\begin{aligned} \frac{dx^1}{dt} &= f(x^n) - A_1 x^1, \\ \frac{dx^j}{dt} &= x^{j-1} - A_j x^j \quad \text{for } j = 2, \cdots, n \end{aligned} \tag{5}$$

where f: $\mathbb{R} \to \mathbb{R}$ is C^1 and the following conditions obtain:

(5a) $A_j > 0$, $\quad j = 1, \cdots, n$;

(5b) $f(s) > 0$ for all $s > 0$ and $f'(s) > 0$ for all $s \geqq 0$;

(5c) $f(s_k)/s_k < A_1 \cdots A_n$ for some sequences $s_k \to \infty$.

Systems of this kind have been studied by Selgrade [13], [14] and Griffiths [3]. See also Walter [15].

THEOREM 5.1. *Let system* (1) *satisfy* (2), (3), (4). *Then*:

(a) $\mathbb{R}^n_+$ *is positively invariant*;

(b) *every forward orbit closure in* $\mathbb{R}^n_+$ *is compact*;

(c) *the forward trajectory of almost every point of* $\mathbb{R}^n_+$ *approaches the set of equilibria.*

The next result shows that small perturbations of (1) enjoy similar properties. Let $\mathscr{V}(\mathbb{R}^n)$ denote the space of C^1 vector fields on $\mathbb{R}^n$ with the weak topology (see Appendix).

THEOREM 5.2. *Let $K \subset \mathbb{R}^n_+$ be a compact set. Then there exists a weak C^1 neighborhood $\mathscr{N} \subset \mathscr{V}(\mathbb{R}^n)$ of F such that if $G \in \mathscr{N}$ and $\mathbb{R}^n_+$ is positively invariant for the flow of G, then*:

(a) *the forward orbit closure of any point of K is compact*;

(b) *the forward trajectory of almost every point of K approaches the set of equilibria.*

Proof of Theorem 5.1. By (1), (2) and Theorem 1.5 the flow $\{\phi_t\}$ is strongly monotone. Evidently (3) implies (a), and (3) and (4) imply (b); given $y > x$ as in (4), for all $g \geqq 0$ we have

$$0 \leqq \phi_t(0) \leqq \phi_t(x) < \phi_t(y) \leqq y$$

showing that the forward orbit of x is bounded. Finally, (c) follows from Theorem 4.1.

Proof of Theorem 5.2. Fix $q \in \operatorname{int} \mathbb{R}^n_+$ such that

$$q > K \quad \text{and} \quad F(g) < 0,$$

using property (4). Set

$$\Gamma = \{x \in \mathbb{R}^n_+ \colon 0 \leqq x \leqq q\}.$$

Because F is cooperative and irreducible and $F(x) < 0 < F(q)$, it follows that $F(x)$ points *into* $\operatorname{int} \Gamma$ for any $x \in \partial F \backslash \{0\}$. If G is sufficiently near F then $G(x)$ points into Γ for all $x \subset \partial \Gamma$ such that $x^i = q^i$ for some $i \in \{1, \cdots, n\}$. If we assume that $\mathbb{R}^n_+$ is positively invariant for the flow $\{\psi_t\}$ of such a G, then $\psi_t(\Gamma)\subset\Gamma$ under $\{\psi_t\}$ for all $t>0$. Conclusion (a) is now obvious.

If, in addition to the properties above, G is sufficiently near F, then $\{\psi_t\}$ has eventually positive derivatives in Γ by Theorem 1.2. Conclusion (b) follows from Theorem 4.1. QED.

THEOREM 5.3. *In the feedback loop* (5) *assume that* 0 *is a regular value of the function*

$$s\rightarrow f(s)-A_1\cdots A_n s.$$

Then the forward trajectory of almost every point of $\mathbb{R}^n_+$ *converges to a sink.*

Proof. The hypothesis is exactly the condition that the vector field in (5) have simple equilibria. The theorem now follows from Theorems 5.1 and 4.4(b).

THEOREM 5.4. *In the system* (5) *we have*:

(a) *The forward trajectory of almost every point in* $\mathbb{R}^n_+$ *converges to an equilibrium.*

(b) *If* $f(s)$ *is an analytic for* $s>0$ *then every invariant function on* $\mathbb{R}^n_+$ *is constant.*

Proof. Part (a) follows from the Corollary to Theorem 2.3 because the set E of equilibria of (5) is totally ordered. Part (b) follows from Theorem 5.1 and (5a, b) once we show that E is countable. But in the contrary case one proves easily that $f(s)\equiv A_1\cdots A_n s$, violating (5c). QED.

Appendix. The Closing Lemma.

Let M be a smooth manifold without boundary. Denote by $\mathscr{V}(M)$ the set of C^1 vector fields on M in the weak C^1 topology.

THEOREM (Closing Lemma). *Let* $F\in\mathscr{V}(M)$ *have flow* $\{\phi_t\}$. *Let* $p\in M$ *be a nonwandering point of* $\{\phi_t\}$ *which is not an equilibrium, and which belongs to some compact invariant set. Let* $\mathscr{N}\subset\mathscr{V}(M)$ *be a neighborhood of* F. *Then there exists* $T>0$ *with the following property. For every neighborhood* $U\subset M$ *of the solution curve* $\{\phi_t(p)\colon -T\leqq t\leqq T\}$ *there exists* $G\in\mathscr{N}$ *such that* $G=F$ *in* $M\setminus U$ *and the flow of* G *has a closed orbit through* p.

COROLLARY. *For every* $F,p,\mathscr{N}$ *as above and every neighborhood* $N\in M$ *of the closure of the orbit of* p *there exists* $G\in\mathscr{N}$ *having a closed orbit through* p, *such that* $G=F$ *in* $M\setminus N$.

In this paper we need only the case where M is an open set in $\mathbb{R}^n$. From now on we assume M is of this kind.

A basic neighborhood of $F\in\mathscr{V}(M)$ is then a set of the form

$$\mathscr{N}=\mathscr{N}(F;K,\varepsilon)$$

where $K\subset M$ is compact, $\varepsilon>0$, and $g\in\mathscr{N}$ if and only if

$$|G(x)-F(x)|+\|DG(x)-DF(x)\|<\varepsilon$$

for all $x\in K$.

A point $p\in M$ is *nonwandering* for a flow $\{\phi_t\}$ provided that for every neighborhood $V\subset M$ of p and every real number $T>0$ there exists $t>T$ such that $\phi_t(U)\cap U$ is not empty. In particular, $\alpha-$ and ω-limit points are nonwandering.

The Closing Lemma was proved in Pugh [10]. The proof has a gap, but a complete proof has been given by Robinson [11]; see also Robinson [17] and a forthcoming paper by Pugh and Robinson [18]. The gap has also been filled by D. Hart [5].

The gap concerns the topology of $\mathscr{V}(M)$. What is actually proved in [10] is that the vector field G can be chosen so that for a given $\varepsilon>0$, compact $K\subset M$, and $s>0$ the flow $\{\psi_t\}$ of G satisfies

$$|F(x)-G(x)|+|\psi_t(x)-\phi_t(z)|+\|D\psi_t(x)-D\phi_t(x)\|<\varepsilon$$

for all $x\in K$, $t\in[-s,s]$. This result in fact suffices for the applications of the Closing Lemma in this paper.

REFERENCES

[1] F. ALBRECHT, H. GATZKE, A. HADDAD AND N. WAX, *The dynamics of two interacting populations*, J. Math. Anal. Appl., 46 (1974), pp. 658–670.

[2] W. A. COPPEL, *Stability and Asymptotic Behavior of Differential Equations*, D. C. Heath, Boston, 1969.

[3] J. S. GRIFFITH, *Mathematics of cellular control processes*, II: *Positive feedback to one gene*, J. Theor. Biology, 29 (1968), pp. 209–216.

[4] S. GROSSBERG, *Competition, decision and consensus*, J. Math. Anal. Appl., 66 (1978), pp. 470–493.

[5] D. HART, *On the smoothness of generators for flows and foliations*, Ph. D. Thesis, Univ. California, Berkeley, 1980.

[6] M. W. HIRSCH, *Systems of differential equations which are competitive or cooperative*. I: *Limit sets*, this Journal, 13 (1982), pp. 167–179.

[7] M. W. HIRSCH AND S. SMALE, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.

[8] E. KAMKE, *Zue Theorie der Systeme gewohnlicher differentialgleichungen*, II, Acta Math., 58 (1932), pp. 57–85.

[9] A. N. KOLMOGOROV, *Sulla teoria di Volterra della lotta per l'esistenza*, Giorn. Ist. Ital. Attuari, 7 (1936), pp. 74–80.

[10] C. PUGH, *An improved Closing Lemma and general density theorem*, Amer. J. Math., 89(1967), pp. 1010–1021.

[11] C. ROBINSON, *The C^1 Closing Lemma*, Northwestern Univ., Evanston, IL, 1970 (typewritten notes).

[12] A. RESCIGNO AND I. RICHARDSON, *The struggle for life*, I: *Two species*, Bull. Math. Biophys., 2 (1967), pp. 377–388.

[13] J. SELGRADE, *Mathematical analysis of a cellular control process with positive feedback*, SIAM J. Appl. Math., 36 (1979), pp. 219–229.

[14] ______, *Asymptotic behavior of solutions to single loop positive feedback systems*, J. Differential Eq., 38 (1980), pp. 80–103.

[15] C. WALTER, *Stability properties and periodic behavior of controlled biochemical systems*, Nonlinear Problems in the Physical Sciences, Lecture Notes in Mathematics 322, Springer-Verlag, New York, 1973.

[16] ______, *Differential and Integral Inequalities*, Springer-Verlag, New York, 1970.

[17] C. ROBINSON, *Introduction to the Closing Lemma*, Lecture Notes in Mathematics 668, Springer-Verlag, New York, pp. 223–230.

[18] C. PUGH AND C. ROBINSON, *The C^1 Closing Lemma, including Hamiltonians, ergodic theory and dynamical systems*, Ergodic Theory and Dynam. Sys., 3 (1983), pp. 261–313.

[19] M. W. HIRSCH, *Differential equations and convergence almost everywhere in strongly monotone semiflows*, Contemporary Math., 17 (1983), pp. 267–285.

[20] K. P. HADELER AND D. GLAS, *Quasimonotone systems and convergence to equilibrium in a population genetic model*, J. Math. Anal. Appl., to appear.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

GROWTH PROPERTIES OF STATIONARY KLEIN–GORDON EQUATIONS*

TAKAŜI KUSANO[†] AND CHARLES A. SWANSON[‡]

Abstract. Elliptic partial differential equations of the stationary Klein–Gordon type are considered in exterior domains in Euclidean n-space. Sufficient conditions are found for the existence of infinitely many positive solutions, and explicit asymptotic behavior as $|x|\to\infty$ is given for both bounded and unbounded positive solutions. Variable signed and unbounded nonlinearities are allowed in the differential equation with respect to any of the variables.

1. Introduction. Our purpose is to prove existence of positive solutions of the semilinear differential equation

$$\Delta u-q(|x|)u+f(x,u)=0, \qquad x\in\Omega, \tag{1.1}$$

in an exterior domain $\Omega\subset R^N$, $N\geq 2$, where q is nonnegative and locally Hölder continuous in $R_+=(0,\infty)$ and f is locally Hölder continuous in $\Omega\times R_+$. Explicit asymptotic behavior of both bounded and unbounded positive solutions of (1.1) will be obtained as $|x|\to\infty$. Unlike earlier work [2], [3], $f(x,u)$ is not required to be one-signed or monotone in u, and q is not required to be bounded in R_+. Furthermore, $f(x,u)$ is allowed to change sign with respect to x for fixed u. An example of (1.1), for which explicit results are obtained in §3, is

$$\Delta u-\rho^2|x|^{2r}u+p(x)(u^m-bu^n)=0, \qquad x\in\Omega, \tag{1.2}$$

where b,m,n and ρ are positive constants, $m<n$, $2r$ is a positive integer, and p is locally Hölder continuous in Ω.

Solitary waves in nonlinear quantum field theory arise as solutions of the Klein–Gordon (wave) equation [1], [6]

$$\psi_{tt}=\Delta\psi-q_0(|x|)\psi+f_0(x,|\psi|)\psi.$$

In particular, standing waves $\psi(x,t)=e^{i\omega t}u(x)$ (where ω denotes a real constant) exist corresponding to any positive solution $u(x)$ of a stationary equation of type (1.1) with exponential decay as $|x|\to\infty$.

2. Positive solutions of semilinear ODE. Let p_0, p_1 and p_2 be positive continuous functions in a positive interval $[t_0,\infty)$, and define

$$z_1(t)=p_0(t), \qquad z_2(t)=p_0(t)P_1(t) \tag{2.1}$$

for $t\geq t_0$, where

$$P_1(t)=\int_{t_0}^{t}p_1(s)\,ds.$$

* Received by the editors October 11, 1983.

[†] Department of Mathematics, Hiroshima University, Hiroshima 730, Japan. The work of the author was supported in part by Grant-in-Aid for Scientific Research 58460004, Ministry of Education, Japan.

[‡] Department of Mathematics, University of British Columbia, Vancouver, Canada V6T 1Y4. The work of this author was supported by the Natural Sciences and Engineering Research Council of Canada.

It is assumed throughout that $\lim_{t\to\infty} P_1(t) = +\infty$, $z_1(t)$ is bounded above, and $z_2(t)$ is bounded away from zero in $[t_0, \infty)$. Clearly z_1 and z_2 are linearly independent, asymptotically ordered, positive solutions of the linear differential equation

$$(2.2)\qquad Lz \equiv \frac{1}{p_2(t)}\frac{d}{dt}\left[\frac{1}{p_1(t)}\frac{d}{dt}\left(\frac{z}{p_0(t)}\right)\right]=0, \qquad t\geq t_0.$$

We use the notation

$$M=\sup_{t\geq t_0} z_1(t), \qquad \mu=\inf_{t\geq t_0} z_2(t).$$

Sufficient conditions will be established for the semilinear differential equations

$$(2.3_\pm)\qquad Ly\pm h(t,y)=0, \qquad t\geq t_0,$$

to have positive solutions $y_1(t)$, $y_2(t)$ which have the same asymptotic behavior as $z_1(t)$, $z_2(t)$ as $t\to\infty$. The hypotheses on $h(t,y)$ will be selected from the list below.

(h_1) There exists a positive constant c such that $h(t,u)$ is continuous, nonnegative and nondecreasing in u for $0<u\leq c$ and for all $t\geq t_0$.

(h_2) There exists a positive constant c such that $h(t,u)$ is continuous, nonnegative and nonincreasing in u for $0<u\leq c$ and for all $t\geq t_0$.

(h_3) There exists a positive constant c such that $h(t,u)$ is continuous, nonnegative and nondecreasing in u for $u\geq c$ and for all $t\geq t_0$.

(h_4) There exists a positive constant c such that $h(t,u)$ is continuous, nonnegative and nonincreasing in u for $u\geq c$ and for all $t\geq t_0$.

(h_5) $\int_{t_0}^{\infty} P_1(t)p_2(t)h(t,kp_0(t))\,dt<\infty$ for all $k\in(0,c/M]$.

(h_6) $\int_{t_0}^{\infty} p_2(t)h(t,kz_2(t))\,dt<\infty$ for all $k\geq c/\mu$.

THEOREM 2.1. *Conditions* $\{(\mathrm{h}_1),(\mathrm{h}_5)\}$ *or* $\{(\mathrm{h}_2),(\mathrm{h}_5)\}$ *imply the existence of infinitely many positive solutions* $y(t)$ *of equations* $(2.3_\pm)$ *such that*

$$(2.4)\qquad \lim_{t\to\infty}\frac{y(t)}{z_1(t)}=\text{constant}>0.$$

Proof. The proof will first be given for (2.3_+) under hypotheses $(\mathrm{h}_1),(\mathrm{h}_5)$. For arbitrary $k\in(0,c/M]$, choose $T=T(k)\geq t_0$ such that

$$(2.5)\qquad \int_T^{\infty} P_1(t)p_2(t)h(t,kp_0(t))\,dt\leq\frac{k}{2}.$$

The integrand is continuous and nonnegative for $t\geq T$ by (h_1), since $kp_0(t)\leq (c/M)\sup_{t\geq T}p_0(t)\leq c$ for $t\geq T$. Let $\mathscr{C}$ be the space of all continuous functions in $[T,\infty)$ with the compact open topology, and define

$$(2.6)\qquad \mathscr{Y}=\left\{y\in\mathscr{C}\colon \frac{k}{2}p_0(t)\leq y(t)\leq kp_0(t) \quad \text{for } t\geq T\right\},$$

a closed convex subset of $\mathscr{C}$. Let $\mathscr{M}$ be the mapping on $\mathscr{Y}$ defined by

$$(2.7)\qquad (\mathscr{M}y)(t)=p_0(t)\left[k-\int_t^{\infty}\left(\int_t^s p_1(\sigma)\,d\sigma\right)p_2(s)h(s,y(s))\,ds\right], \qquad t\geq T.$$

The Schauder–Tykhonov fixed-point theorem shows that $\mathscr{M}$ has a fixed point $y\in\mathscr{Y}$, i.e. $(\mathscr{M}y)(t)=y(t)$ for all $t\geq T$, since one easily verifies that $\mathscr{M}$ is a continuous mapping from $\mathscr{Y}$ into itself such that $\mathscr{M}\mathscr{Y}$ is relatively compact (using Ascoli's theorem). Differentiation of the integral equation $\mathscr{M}y=y$ twice completes the proof that $y(t)$ is a positive solution of (2.3$_+$), and clearly $\lim_{t\to\infty}y(t)/p_0(t)=k$ from (2.7).

In the case of (2.3$_-$), the mapping (2.7) is replaced by

$$(2.8)\qquad (\mathscr{M}y)(t)=p_0(t)\left[\frac{k}{2}+\int_t^\infty\left(\int_t^s p_1(\sigma)\,d\sigma\right)p_2(s)h(s,y(s))\,ds\right],\qquad t\geq T,$$

and virtually the same argument yields the existence of a positive solution of (2.3$_-$) with $\lim_{t\to\infty}y(t)/p_0(t)=k/2$.

If (h$_2$), (h$_5$) hold, condition (2.5) is replaced by

$$(2.9)\qquad \int_T^\infty P_1(t)p_2(t)h(t,kp_0(t))\,dt\leq k$$

for some $T\geq t_0$, and (2.6) is replaced by

$$(2.10)\qquad \mathscr{Y}=\{y\in\mathscr{C}: kp_0(t)\leq y(t)\leq 2kp_0(t)\text{ for }t\geq T\}.$$

Positive solutions $y(t)$ of (2.3$_+$) or (2.3$_-$) such that $\lim_{t\to\infty}y(t)/p_0(t)=2k$ or $\lim_{t\to\infty}y(t)/p_0(t)=k$, respectively, are then obtained in the same way as fixed points of the mappings

$$(2.11)\qquad (\mathscr{M}y)(t)=p_0(t)\left[2k-\int_t^\infty\left(\int_t^s p_1(\sigma)\,d\sigma\right)p_2(s)h(s,y(s))\,ds\right],\qquad t\geq T,$$

$$(2.12)\qquad (\mathscr{M}y)(t)=p_0(t)\left[k+\int_t^\infty\left(\int_t^s p_1(\sigma)\,d\sigma\right)p_2(s)h(s,y(s))\,ds\right],\qquad t\geq T.$$

THEOREM 2.2. *Conditions* $\{(\mathrm{h}_3),(\mathrm{h}_6)\}$ *or* $\{(\mathrm{h}_4),(\mathrm{h}_6)\}$ *imply the existence of infinitely many positive solutions* $y(t)$ *of* (2.3$_\pm$) *such that*

$$(2.13)\qquad \lim_{t\to\infty}\frac{y(t)}{z_2(t)}=\text{constant}>0.$$

Proof. We indicate the proof in the case of hypotheses (h$_3$), (h$_6$); the proof is similar in the other case and will be omitted. For arbitrary $k\geq c/\mu$, (h$_6$) shows that $T=T(k)\geq t_0$ can be chosen so that

$$\int_T^\infty p_2(t)h(t,kz_2(t))\,dt\leq\frac{k}{2}.$$

The integrand is continuous and nonnegative for $t\geq T$ by (h$_3$), since $kz_2(t)\geq(c/\mu)\inf_{t\geq T}z_2(t)\geq c$ for $t\geq T$. The set (2.6) used in Theorem 2.1 is now replaced by

$$\mathscr{Y}=\left\{y\in\mathscr{C}: \frac{k}{2}z_2(t)\leq y(t)\leq kz_2(t)\text{ for }t\geq T\right\}.$$

A parallel argument to that used in Theorem 2.1 establishes fixed points of the mappings defined by

$$(\mathscr{M}y)(t)=p_0(t)\left[\frac{k}{2}P_1(t)+\int_T^t p_1(s)\int_s^\infty p_2(\sigma)h(\sigma,y(\sigma))\,d\sigma\,ds\right],\quad t\geq T,$$

$$(\mathscr{M}y)(t)=p_0(t)\left[kP_1(t)-\int_T^t p_1(s)\int_s^\infty p_2(\sigma)h(\sigma,y(\sigma))\,d\sigma\,ds\right],\quad t\geq T,$$

yielding positive solutions of (2.3_+), (2.3_-), respectively, satisfying (2.13). □

3. Stationary Klein–Gordon equations. Existence and growth properties of positive solutions of (1.1) in an exterior domain $\Omega\subset R^N$ will be established under hypothesis (H_0) below and under other hypotheses listed in the sequel.

(H_0) q is nonnegative and locally Hölder continuous in $R_+=(0,\infty)$, and f is locally Hölder continuous in $\Omega\times R_+$.

The radial component of the linear part of (1.1) (i.e. f replaced by 0) is

$$Lz=t^{1-N}\frac{d}{dt}\left(t^{N-1}\frac{dz}{dt}\right)-q(t)z=0. \tag{3.1}$$

Since (3.1) is nonoscillatory in (t_0,∞) for some $t_0>0$, the Trench disconjugacy theory [7] shows that (3.1) has linearly independent, eventually positive, asymptotically ordered solutions $z_1(t)$, $z_2(t)$, and furthermore that (3.1) has the factorized form (2.2), where

$$p_0=z_1,\qquad p_1=\left(\frac{z_2}{z_1}\right)',\qquad p_2=\frac{1}{p_0p_1}. \tag{3.2}$$

Then z_1 and z_2 are given by (2.1), and it is easily seen that $z_1(t)$ is bounded and $z_2(t)$ is bounded away from zero in (t_0,∞).

We use the notation

$$\Omega_t=\{x\in\Omega:|x|>t\},\qquad t>0.$$

Hypotheses for the main theorems are to be selected from the following list:

(H_1) There exist a locally Hölder continuous function $\Phi:R_+\times R_+\to R$ and a positive constant c such that $\Phi(t,u)$ is nonnegative and nondecreasing in u (for fixed t) for all $u\in(0,c]$, $t\geq t_0$, and such that $|f(x,u)|\leq\Phi(|x|,u)$ for all $x\in\Omega_{t_0}$, $0<u\leq c$.

(H_2) The same as (H_1), except that "nondecreasing" is replaced by "nonincreasing."

(H_3) There exist a locally Hölder continuous function $\Phi:R_+\times R_+\to R$ and a positive constant c such that $\Phi(t,u)$ is nonnegative and nondecreasing in u (for fixed t) for all $u\geq c$, $t\geq t_0$, and such that

$$|f(x,u)|\leq\Phi(|x|,u)\quad\text{for all } x\in\Omega_{t_0},\quad u\geq c.$$

(H_4) The same as (H_3), except that "nondecreasing" is replaced by "nonincreasing."

$$\int_{t_0}^\infty p_2(t)\frac{z_2(t)}{z_1(t)}\Phi(t,kz_1(t))\,dt<\infty\quad\text{for all } k\in(0,c/M]. \tag{H_5}$$

$$\int_{t_0}^\infty p_2(t)\Phi(t,kz_2(t))\,dt<\infty\quad\text{for all } k\geq c/\mu. \tag{H_6}$$

THEOREM 3.1. *Either* $(H_0),(H_1),(H_5)$ *or* $(H_0),(H_2),(H_5)$ *imply the existence of infinitely many positive solutions* $u(x)$ *of* (1.1) *in some domain* Ω_τ *such that* $u(x)/z_1(|x|)$ *is bounded and bounded away from zero in* Ω_τ.

THEOREM 3.2. *Either* $(H_0),(H_3),(H_6)$ *or* $(H_0),(H_4),(H_6)$ *imply the existence of infinitely many positive solutions* $u(x)$ *of* (1.1) *in some domain* Ω_τ *such that* $u(x)/z_2(|x|)$ *is bounded and bounded away from zero in* Ω_τ.

Proof of Theorem 3.1. Consider the ordinary differential equations

$$Ly+\Phi(t,y)=0, \qquad t\geq t_0, \tag{3.3}$$

$$Ly_0-\Phi(t,y_0)=0, \qquad t\geq t_0, \tag{3.4}$$

where the linear operator L is given by (3.1), or, equivalently, (2.2). Equations (3.3) and (3.4) have the form $(2.3_\pm)$, and Theorem 2.1 is applicable. In view of (3.2), $P_1(t)$ can be taken to be $z_2(t)/z_1(t)$, and hence condition (h_5) reduces to (H_5). Then Theorem 2.1 implies that (3.3) and (3.4) have positive solutions $y(t)$ and $y_0(t)$, respectively, in an interval $[T,\infty)$ such that

$$\lim_{t\to\infty}\frac{y(t)}{z_1(t)}=k, \qquad \lim_{t\to\infty}\frac{y_0(t)}{z_1(t)}=k_0 \tag{3.5}$$

for any positive constants $k\in(0,c/M]$, $k_0\in(0,c/M]$. With any choice $k_0<k$ it follows that there exists a number $\tau\geq T$ such that $0<y_0(t)<y(t)$ for all $t\geq\tau$. Define

$$v(x)=y(|x|), \qquad w(x)=y_0(|x|) \quad \text{for } x\in\Omega_\tau.$$

Then v,w satisfy the partial differential equations

$$\Delta v-q(|x|)v+\Phi(|x|,v)=0,$$
$$\Delta w-q(|x|)w-\Phi(|x|,w)=0, \qquad x\in\Omega_\tau,$$

and so (H_1) or (H_2) shows that they satisfy the partial differential inequalities

$$\Delta v-q(|x|)v+f(x,v)\leq 0,$$
$$\Delta w-q(|x|)w+f(x,w)\geq 0, \qquad x\in\Omega_\tau.$$

Therefore v is a supersolution of (1.1), and w is a subsolution of (1.1) in Ω_τ such that $0<w(x)<v(x)$ throughout Ω_τ. Furthermore, the second partial derivatives of v and w are locally Hölder continuous by the regularity hypothesis (H_0) for (3.3) and (3.4). It follows from a theorem of Noussair and Swanson [4, p. 125] that (1.1) has a positive solution $u\in C^{2+\alpha}_{\mathrm{loc}}(\Omega_\tau)$ $(0<\alpha<1)$ satisfying $w(x)\leq u(x)\leq v(x)$ for all $x\in\Omega_\tau$. The boundedness properties in Theorem 3.1 are immediate consequences of (3.5). □

Proof of Theorem 3.2. Since (H_6) is equivalent to (h_6), Theorem 2.2 shows that (3.3) and (3.4) have positive solutions $y(t)$ and $y_0(t)$, respectively, in some interval $[T,\infty)$ such that

$$\lim_{t\to\infty}\frac{y(t)}{z_2(t)}=k, \qquad \lim_{t\to\infty}\frac{y_0(t)}{z_2(t)}=k_0$$

for constants k and k_0 which can be chosen to satisfy $0<k_0<k$. The proof is then completed in the same way as that of Theorem 3.1. □

Example 3.3. In the case of the prototype (1.2), (3.1) becomes

$$t^{1-N}\frac{d}{dt}\left(t^{N-1}\frac{dz}{dt}\right)-\rho^2t^{2r}z=0. \tag{3.6}$$

A fundamental set $\{z_1(t),z_2(t)\}$ of eventually positive, asymptotically ordered solutions of (3.6) has the asymptotic behavior [5, p. 285]

$$z_1(t)\sim t^{-\lambda}\exp\left(-\frac{\rho t^{r+1}}{r+1}\right),\qquad z_2(t)\sim t^{-\lambda}\exp\left(\frac{\rho t^{r+1}}{r+1}\right)$$

as $t\to\infty$, where $\lambda=(N+r-1)/2$. Hence

$$p_2(t)\sim\frac{t^{\lambda-r}}{2\rho}\exp\left(-\frac{\rho t^{r+1}}{r+1}\right)\quad\text{as } t\to\infty,$$

$$\frac{p_2(t)z_2(t)}{z_1(t)}\sim\frac{t^{\lambda-r}}{2\rho}\exp\left(\frac{\rho t^{r+1}}{r+1}\right)\quad\text{as } t\to\infty.$$

Define $\Phi(t,u)=P(t)g(u)$, where

$$P(t)=\max_{|x|=t}|p(x)|,\qquad g(u)=u^m-bu^n.$$

Then $f(x,u)=p(x)(u^m-bu^n)$ satisfies

$$|f(x,u)|\le\Phi(|x|,u)$$

for all $x\in\Omega$ and $0<u\le(1/b)^{1/(n-m)}$. Also $g'(u)\ge0$ if $0<u\le(m/nb)^{1/(n-m)}$. It follows that (H_1) is satisfied with the choice $c=(m/nb)^{1/(n-m)}$. Hypothesis (H_5) reduces to

$$\int_{t_0}^{\infty}P(t)t^{-(m-1)\lambda-r}\exp\left(-\frac{(m-1)\rho t^{r+1}}{r+1}\right)dt<\infty, \tag{3.7}$$

which is evidently satisfied, for instance, if $m>1$ and $P(t)$ is bounded by a polynomial in t. If (3.7) holds, Theorem 3.1 shows that (1.2) has a bounded positive solution $u(x)$ in Ω_τ, for some $\tau>0$, such that

$$c_1|x|^{-\lambda}\exp\left(-\frac{\rho|x|^{r+1}}{r+1}\right)\le u(x)\le c_2|x|^{-\lambda}\exp\left(-\frac{\rho|x|^{r+1}}{r+1}\right),\qquad x\in\Omega_\tau,$$

for some positive constants c_1 and c_2.

To obtain unbounded positive solutions, define $\Phi(t,u)=P(t)\tilde g(u)$, where now $\tilde g(u)=bu^n-u^m$ and $P(t)$ is as before. Then $f(x,u)=p(x)(u^m-bu^n)$ satisfies

$$|f(x,u)|\le\Phi(|x|,u)$$

for all $x\in\Omega$ and $u\ge(1/b)^{1/(n-m)}$, and hypothesis (H_3) is satisfied with the choice $c=(1/b)^{1/(n-m)}$. Hypothesis (H_6) reduces to

$$\int_{t_0}^{\infty}P(t)t^{-(n-1)\lambda-r}\exp\left(\frac{(n-1)\rho t^{r+1}}{r+1}\right)dt<\infty. \tag{3.8}$$

Theorem 3.2 shows that (3.8) is sufficient for (1.2) to have an unbounded positive solution $u(x)$ in Ω_T such that

$$c_1|x|^{-\lambda}\exp\left(\frac{\rho|x|^{r+1}}{r+1}\right)\leq u(x)\leq c_2|x|^{-\lambda}\exp\left(\frac{\rho|x|^{r+1}}{r+1}\right), \qquad x\in\Omega_T,$$

for some positive constants c_1 and c_2.

REFERENCES

[1] Melvyn S. Berger, *On the existence and structure of stationary states for a nonlinear Klein–Gordon equation*, J. Functional Anal., 9 (1972), pp. 249–261.

[2] Kurt Kreith and Charles A. Swanson, *Asymptotic solutions of semilinear elliptic equations*, J. Math. Anal. Appl., 98 (1984), pp. 148–157.

[3] Takaŝi Kusano and Charles A. Swanson, *Asymptotic properties of semilinear elliptic equations*, Funkcial. Ekvac., 26 (1983), pp. 115–129.

[4] Ezzat S. Noussair and Charles A. Swanson, *Positive solutions of quasilinear elliptic equations in exterior domains*, J. Math. Anal. Appl., 75 (1980), pp. 121–133.

[5] L. Sirovich, *Techniques of Asymptotic Analysis*, Springer-Verlag, New York-Heidelberg-Berlin, 1971.

[6] Walter A. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149–162.

[7] William F. Trench, *Canonical forms and principal systems for general disconjugate equations*, Trans. Amer. Math. Soc., 189 (1974), pp. 319–327.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

ELEMENTARY WAVE SOLUTIONS OF THE EQUATIONS DESCRIBING THE MOTION OF AN ELASTIC STRING*

MICHAEL SHEARER†

Abstract. The equations of planar motion of an elastic string form a 4×4 system of first order conservation laws. Two of the characteristic fields correspond to genuinely nonlinear longitudinal shocks and rarefaction waves, involving changes in the tension in the string, but not the slope. The other two fields correspond to contact discontinuities, across which the slope of the string jumps, reflecting the absence of any resistance to bending.

Here, the tension T is related to the local elongation $\xi>1$ in such a way as to ensure strict hyperbolicity: $T'(\xi)>T(\xi)/\xi\geq0$. The other main assumption is chosen to reflect properties of typical materials such as nylon and rubber. That is, $T''(\xi)$ is negative for $\xi<\xi_I$ and positive for $\xi>\xi_I$ for some $\xi_I>1$. The principal result of the paper is that the Riemann problem has a unique solution among combinations of centered waves, with a natural entropy condition placed on shocks. That is, any initial jump discontinuity in the tension, slope and velocity of the string can be resolved into combinations of longitudinal waves and contact discontinuities. This is illustrated for the plucked string problem, whose solution (valid until a wave first hits an end of the string) necessarily involves longitudinal waves.

1. Introduction.

Consider an elastic string whose configuration in $\mathbb{R}^3$ at time t is specified by a function $r(\cdot,t):[0,1]\to\mathbb{R}^3$. The interval $[0,1]$ is the *reference configuration*, each point of which identifies a material point in the string. Let $n(x,t)$ denote the tension in the string, taken to act tangentially to the string. In the absence of external forces, the equations of motion for the string are ([1])

$$\left[n(x,t)\frac{r_x(x,t)}{|r_x(x,t)|}\right]_x=r_{tt},\qquad 0<x<1,\quad t>0,\tag{1.1}$$

where we have additionally assumed the mass density $\rho(x)$ is constant (and have incorporated it into n). It will be further assumed that the tension $n(x,t)$ depends explicitly only upon the local extension $|r_x(x,t)|$

$$n(x,t)=T(|r_x(x,t)|),\tag{1.2}$$

for some smooth function $T:(0,\infty)\to\mathbb{R}$ satisfying

$$T'(\xi)>0\quad\text{for all }\xi>0.\tag{1.3}$$

Assumption (1.2) ignores other possible material properties of the string such as viscoelastic and memory effects; the absence of an explicit dependence of T upon x corresponds to a uniformity assumption on the string. (See [1], [2] for more general constitutive assumptions on n.) Inequality (1.3) states that the tension increases with an increase in the local extension, and guarantees that system (1.1) is hyperbolic providing the tension is not negative.

In §2, we analyze the structure of elementary wave solutions of equation (1.1), namely shocks (involving jumps in $|r_x|$), rarefaction waves and contact discontinuities. Since (1.1) is readily written as a first order system of hyperbolic conservation laws (see

* Received by the editors May 10, 1983, and in revised form September 9, 1983.

† Department of Mathematics, Duke University, Durham, North Carolina 27706.

(1.7) below), the specification of elementary waves is straightforward, and has been given elsewhere ([2], [4]). The principal purpose of this paper is to characterize solutions of the Riemann problem, which involves resolving an initial discontinuity in (r_x, r_t) into a combination of elementary waves. The waves may be characterized as longitudinal or transverse. The longitudinal waves are genuinely nonlinear providing $T''(|r_x|)\neq 0$, whereas the transverse waves are linearly degenerate. To solve the Riemann problem, we make some basic assumptions. First, only plane motion of the string will be considered. This reduces system (1.1) to a pair of coupled wave equations, and we may take

$$r(x,t)=(r_1(x,t),r_2(x,t))\in\mathbb{R}^2.$$

Since there are no external forces, the added degree of freedom of three-dimensional motion simply allows the string to bend in any direction. As remarked in §2, this corresponds to a rather uninteresting extra degree of freedom in the contact discontinuities, which involve jumps essentially only in the slope of the string.

The second assumption concerns the stress-strain law specifying T as a function of the local extension $|r_x|$. It will be convenient to let the reference configuration correspond to an equilibrium and unstretched configuration of the string, so that

$$(1.4)\qquad T(1)=0.$$

The following material properties appear to be typical (see [2]), and dictate our assumptions on T.

1. Either (a) $T''(\xi)<0$ for all ξ, or (b) there exists $\xi_I>1$ such that

$$(1.5)\qquad (\xi-\xi_I)T''(\xi)>0 \quad \text{for all } \xi>0.$$

Case (a) would simplify our analysis, and may be included in (1.5) for our purposes by simply taking ξ_I very large. A piecewise affine stress-strain law will not be considered.

2. There may or may not be values of ξ at which

$$T'(\xi)=\frac{T(\xi)}{\xi}.$$

This is important because the characteristic speeds for (1.1) are roots of $T'(\xi)$ and $T(\xi)/\xi$. To solve the Riemann problem, in §4, it will be assumed that (1.5) holds and

$$(1.6)\qquad T'(\xi)>\frac{T(\xi)}{\xi} \quad \text{for all } \xi.$$

Under (1.4), (1.6), system (1.1) is *strictly hyperbolic* providing $|r_x|>1$, since the characteristic speeds $[T'(|r_x|)]^{1/2}$ and $[T(|r_x|)/|r_x|]^{1/2}$ are distinct. If (1.6) is violated, the structure of the individual elementary waves is unchanged, but the solution of the Riemann problem is substantially more complicated. The paper of Keyfitz and Kranzer [4] solves the Riemann problem when the initial values lie in some neighborhood of a hypersurface $\xi=\xi_{NS}$ where strict hyperbolicity breaks down i.e., $T'(\xi_{NS})=T(\xi_{NS})/\xi_{NS}$. The implications of such a failure of (1.6) are discussed briefly in §6; the solution of the Riemann problem in this case, and under condition (1.5), will be presented in a future paper.

The graph of a typical T satisfying (1.3)–(1.6) is illustrated in Fig. 1. In §4, the Riemann problem is solved for any initial jump in (r_x, r_t) for which the tension is everywhere positive.

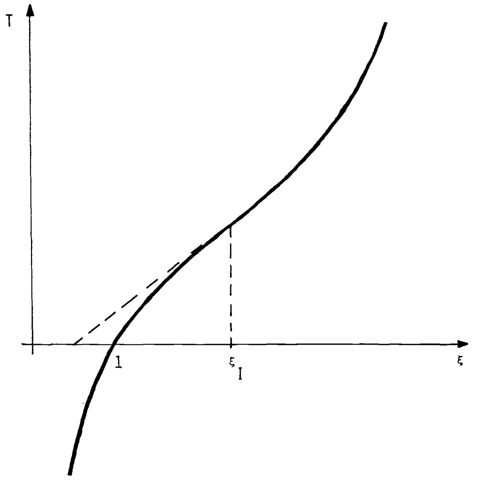


FIG. 1. *The graph of T satisfying* (1.3)–(1.6).

Under (1.3)–(1.6), system (1.7) below is a 4×4 system of first order strictly hyperbolic conservation laws (where $\xi > 1$). Two of the characteristic families are linearly degenerate, while the remaining two families lose genuine nonlinearity precisely due to the presence of an inflection point in the stress-strain curve, Fig. 1. Systems for which genuine nonlinearity breaks down on hypersurfaces were studied by Wendroff [6] and Liu [5], with Riemann problems being solved using a construction of rarefaction-shocks. A similar construction of longitudinal rarefaction-shocks is used here (see the definition of $P_-(U_0)$ in §3). It is the presence of two linearly degenerate fields that places the current work outside the scope of the general results of [5]. In particular, the coordinate system that arises here in solving the Riemann problem resembles that of plane polar coordinates, whereas that of [5] is a curvilinear version of Cartesian coordinates.

In §5, the plucked string problem is solved for small time (i.e., before waves hit the ends of the string) as an example of a Riemann problem. A feature of the solution is that it necessarily involves longitudinal waves, which are specifically excluded by equations modelling small transverse vibrations of an elastic string.

In what follows, we write $r_x = (p,q)$, $\xi = |r_x|$, and $r_t = (u,v)$. Equation (1.1), for planar motion, then becomes the first order 4×4 system

$$\frac{\partial}{\partial t}\begin{pmatrix} p \\ q \\ u \\ v \end{pmatrix} = \frac{\partial}{\partial x}F(p,q,u,v) \tag{1.7}$$

where

$$F(p,q,u,v) = \begin{pmatrix} u \\ v \\ pT(\xi)/\xi \\ qT(\xi)/\xi \end{pmatrix}. \tag{1.8}$$

The characteristic values for system (1.7) are eigenvalues of the Jacobian matrix $F'(U)$, where $U = (p,q,u,v)$, namely

$$\lambda_{\pm}(U)=\lambda_{\pm}(\xi)=\pm[T'(\xi)]^{1/2}, \qquad \mu_{\pm}(U)=\mu_{\pm}(\xi)=\pm\left[\frac{T(\xi)}{\xi}\right]^{1/2}.$$

2. Elementary waves. In this section, we restrict attention to single elementary waves separating regions in (x,t)-space where r_x and r_t are constant. For this purpose, we consider system (1.7) for $-\infty<x<\infty$, $t>0$, and all elementary waves will be centered at the origin. The standing assumptions are (1.2)–(1.5), and we shall consider only planar motion, described by (1.7), (1.8), except for a remark concerning three-dimensional motion.

Let $U_0=(p_0,q_0,u_0,v_0)\in\mathbb{R}^4$ be a given state of the string. U_0 may be joined to a state $U=(p,q,u,v)$ by a shock $x=st$ with speed s if U_0, U, s satisfy the Rankine–Hugoniot conditions

$$(2.1) \qquad -s(U-U_0)=F(U)-F(U_0)$$

where F is given by (1.8). We distinguish two types of solutions of equations (2.1).

Genuinely nonlinear shocks.

$$(2.2) \qquad \begin{gathered} \xi\neq\xi_0, \quad pq_0=qp_0, \quad s=\pm[(T(\xi)-T(\xi_0))/(\xi-\xi_0)]^{1/2}, \\ \begin{pmatrix} u \\ v \end{pmatrix}=\begin{pmatrix} u_0 \\ v_0 \end{pmatrix}\mp\frac{[(T(\xi)-T(\xi_0))(\xi-\xi_0)]^{1/2}}{\xi_0}\operatorname{sgn}(\xi-\xi_0)\begin{pmatrix} p_0 \\ q_0 \end{pmatrix}. \end{gathered}$$

Note that the slope q/p of the string is unchanged across such a shock wave, while the tension $T(\xi)$ undergoes a jump. Only the tangential component of the velocity experiences a jump.

Linear waves or contact discontinuities.

$$(2.3) \qquad \begin{gathered} \xi=\xi_0, \quad s=\pm[T(\xi)/\xi]^{1/2} \quad \text{if } (p,q)=\xi(\cos\theta,\sin\theta), \\ \begin{pmatrix} u \\ v \end{pmatrix}=\begin{pmatrix} u_0 \\ v_0 \end{pmatrix}\pm\frac{[\xi T(\xi)]^{1/2}}{\xi_0}\begin{pmatrix} p_0 \\ q_0 \end{pmatrix}\mp[\xi T(\xi)]^{1/2}\begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}. \end{gathered}$$

Across linear waves, the slope $\tan\theta$ of the string jumps, while the tension is continuous. In the (p,q)-plane, (2.3) describes a circle centered at the origin with radius ξ_0; in the (u,v)-plane, (2.3) describes a circle centered at $(u_0,v_0)\pm[\xi T(\xi)]^{1/2}(p_0,q_0)/\xi_0$ with radius $[\xi T(\xi)]^{1/2}$. Note that (2.2), (2.3) are together equivalent to the jump condition (2.1).

Next we describe the rarefaction waves. These are solutions of (1.7) of the form $U=U(x/t)$, with $x/t=\lambda_{\pm}(U(x/t))$. For such a solution, we must have

$$(2.4) \qquad U'(\eta)=w_{\pm}(U(\eta))$$

where $w_{\pm}(U)$ is the right eigenvector of $-F'(U)$ corresponding to the eigenvalue $\lambda_{\pm}(U)$, normalized by $w_{\pm}\cdot\operatorname{grad}\lambda_{\pm}=1$. For a rarefaction wave to separate constant states U_0 on the left and U on the right, (2.4) leads to

$$(2.5) \qquad \begin{gathered} \xi\neq\xi_0, \qquad qp_0=pq_0, \\ \begin{pmatrix} u \\ v \end{pmatrix}=\begin{pmatrix} u_0 \\ v_0 \end{pmatrix}-\frac{1}{\xi_0}\int_{\xi_0}^{\xi}\lambda_{\pm}(\nu)\,d\nu\begin{pmatrix} p_0 \\ q_0 \end{pmatrix}, \end{gathered}$$

with $\lambda_{\pm}(\xi)=\eta$. In particular, $\lambda_{\pm}(\nu)$ must be increasing as ν varies from ξ_0 to ξ:

$$\begin{aligned}&\pm(\xi-\xi_0)<0 \quad \text{and} \quad \xi<\xi_I \quad \text{if } \xi_0<\xi_I,\\ &\pm(\xi-\xi_0)>0 \quad \text{and} \quad \xi>\xi_I \quad \text{if } \xi_0>\xi_I.\end{aligned} \tag{2.6}$$

In order for solutions of the Riemann problem, discussed in §4, to be unique, the following additional "admissibility" condition will be imposed on genuinely nonlinear shocks: If the shock $x=st$ separates states U_0 on the left and U_1 on the right, then the shock is admissible if

$$s\left[(T(\xi)-T(\xi_0))/(\xi-\xi_0)-(T(\xi_1)-T(\xi_0))/(\xi_1-\xi_0)\right]\geq 0 \tag{2.7}$$

for all ξ between ξ_0 and ξ_1.

Remarks. 1. Condition (2.7) is a condition on the slopes of chords in the graph of T, analogous to the chord condition appropriate for a single second order quasilinear wave equation [6]. That it is natural here arises from the observation that shock waves represent purely longitudinal motion of the string. In particular, by choosing a coordinate system (for $r(x,t)$) that moves with the string near the shock, system (1.1) reduces to a single quasilinear wave equation.

2. Let

$$\eta=r_t\cdot\frac{r_t}{2}+\int_1^{|r_x|}T(\nu)\,d\nu, \qquad Q=-r_x\cdot\frac{r_t}{|r_x|}.$$

The energy η and energy flux Q play the role of an entropy/entropy-flux pair, in the sense that smooth solutions of (1.1) satisfy the conservation law

$$\eta_t+Q_x=0, \tag{2.8}$$

while admissible genuinely nonlinear shocks satisfy

$$\eta_t+Q_x\leq 0 \tag{2.9}$$

in the sense of distributions. Linear waves however satisfy (2.8) in the sense of distributions. Shock wave solutions of (1.7) satisfying (2.9) are not necessarily admissible in the sense of (2.7). However, a short calculation shows that they do satisfy the following viscosity criterion. Replace $T(\xi)$ by $\hat{T}_\varepsilon(\xi,\xi_t)=T(\xi)+\varepsilon\xi_t$. The term $\varepsilon\xi_t$ reflects a simple rate dependence of the tension (see e.g. [2]). Each admissible shock wave solution of (1.7) is the pointwise limit as $\xi\to 0+$ of a travelling wave solution of the modified system

$$r_{tt}=\left[\hat{T}_\varepsilon\frac{r_x}{|r_x|}\right]_x.$$

3. Genuinely nonlinear shocks and rarefaction waves will be collectively referred to as *longitudinal waves*. For fixed U_0, the locus of points U to which U_0 may be joined by a longitudinal wave forms straight line segments with slope q/p in both the (p,q) plane (where it passes through (p_0,q_0)) and in the (u,v) plane (where it passes through (u_0,v_0)).

4. For three-dimensional motion, both the longitudinal waves and linear waves described above occur, but now the linear waves have the extra degree of freedom that the discontinuity in r_x may occur in any direction (while keeping $|r_x|$ continuous). The corresponding jump in velocity r_t is computed from the wave speed $\pm[T(\xi)/\xi]^{1/2}$ and the jump in r_x, by noting that $r(x,t)$ remains continuous. This degeneracy in the linear

waves arises from the fact that the string approximation neglects any resistance to bending. Three-dimensional motion will not be pursued further here, although the solution of the Riemann problem in the next section may easily be suitably generalized.

For each $U_0 \in \mathbb{R}^4$, define the *shock curves* $S_{\pm}(U_0)$ to be the locus of points $U \in \mathbb{R}^4$ satisfying (2.2) and (2.7) for some s, with $s>0$ on $S_+(U_0)$ and $s<0$ on $S_-(U_0)$. Similarly, let $C_{\pm}(U_0)$ be the *linear wave curves*, of points $U \in \mathbb{R}^4$ satisfying (2.3), and let $R_{\pm}(U_0)$ be the *rarefaction curves* of points $U \in \mathbb{R}^4$ satisfying (2.5), (2.6). Note that $C_{\pm}(U_0)$ are defined only for $\xi_0 > 1$.

3. The structure of the wave curves. Let $U_0 = (p_0, q_0, u_0, v_0)$ be fixed. The curve $\gamma_-(U_0)$ in $\mathbb{R}^4$, to be constructed in this section, will consist of points U_1 to which U_0 may be joined by successively slower centered longitudinal waves moving to the left, with U_0 on the left of this family of waves, and U_1 on the right. Similarly, points U on a curve $\gamma_+(U_0)$ will represent a family of successively faster centered longitudinal waves moving to the right separating U_0 on the left from U on the right. Correspondingly, we define curves $\tilde{\gamma}_{\pm}(U_0)$ by

$$(3.1) \qquad U \in \tilde{\gamma}_{\pm}(U_0) \quad \text{if and only if} \quad U_0 \in \gamma_{\pm}(U) \quad \text{(respectively)}.$$

The structure of $\gamma_{\pm}(U_0)$ is complicated by the fact that genuine nonlinearity breaks down at the inflection point $\xi = \xi_I$ of $T(\xi)$. The way to overcome this for a single wave equation is described by Wendroff [6], a procedure generalized by Liu [5]. As remarked in §2, the longitudinal waves may be regarded as weak solutions of a single wave equation (governing longitudinal motion), and indeed our construction of $\gamma_{\pm}(U_0)$ closely follows that in [6], except that we also keep track of $C_{\pm}(U)$ as U varies on $\gamma_{\pm}(U_0)$.

If $0 < \xi \neq \xi_I$, define $\xi^* \neq \xi$ by

$$(3.2) \qquad T'(\xi) = (T(\xi) - T(\xi^*))/(\xi - \xi^*),$$

with the proviso that $\xi^* = \infty$ if no finite ξ^* can be found to satisfy (3.2). Under assumptions (1.2)–(1.5), ξ^* is uniquely defined, and is necessarily finite for $\xi > \xi_I$. Similarly, let $\hat{\xi} \neq \xi$ satisfy

$$(3.3) \qquad T'(\hat{\xi}) = (T(\xi) - T(\hat{\xi}))/(\xi - \hat{\xi})$$

unless no finite $\hat{\xi}$ can be found to satisfy (3.3), in which case set $\hat{\xi} = \infty$. Again $\hat{\xi}$ is uniquely determined, and is necessarily finite for $\xi > \xi_I$. The significance of equations (3.2), (3.3) is that they express equality between the shock speed of certain longitudinal shock waves, and a corresponding characteristic speed $\lambda_{\pm}$, thus defining limits upon the validity of the admissibility criterion (2.7). Note that

$$(3.4) \qquad \begin{array}{l} \operatorname{sgn}(\xi - \xi_I) = \operatorname{sgn}(\xi_I - \xi^*) = \operatorname{sgn}(\xi_I - \hat{\xi}) = \operatorname{sgn}(\hat{\xi} - \xi^*), \\ \hat{\xi}^* = \xi. \end{array}$$

In the construction of $\gamma_{\pm}(U_0)$, the two cases $\xi_0 > \xi_I$ and $\xi_0 < \xi_I$ have to be considered separately.

A. $\xi_0 < \xi_I$. Let $\mathscr{S}_-(U_0)$ be the section of $S_-(U_0)$ with $\xi < \xi_0$, and let $\mathscr{S}_-^*(U_0)$ be the section of $S_-(U_0)$ with $\xi > \xi_0^*$. Note that the corresponding shock speed s coincides at $\xi = \xi_0^*$ with the characteristic speed λ_-. Now $R_-(U_0)$ corresponds to values of ξ with $\xi_0 < \xi \leq \xi_I$. For $\xi_I < \xi_1 < \xi_0^*$, we have $\xi_0 < \hat{\xi}_1 < \xi_I$, and $\hat{\xi}_1$ defines a point $\hat{U}_1$ on $R_-(U_0)$. Further $\hat{U}_1$ may be joined to a point U_1 (for which $\xi = \xi_1$) by a shock, with speed

$-[T'(\hat{\xi}_1)]^{1/2}$, the slowest speed in the rarefaction wave joining U to $\hat{U}_1$ (see Fig. 3). Let $P_-(U_0)$ denote the curve of such points U_1, parameterized by $\xi_1 \in (\xi_I, \xi_0^*)$. Note that $P_-(U_0)$ joins $R_-(U_0)$ to $\mathscr{S}_-^*(U_0)$. Now set

$$\gamma_-(U_0) = \mathscr{S}_-(U_0) \cup R_-(U_0) \cup P_-(U_0) \cup \mathscr{S}_-^*(U_0). \tag{3.5}$$

A similar analysis leads to

$$\gamma_+(U_0) = R_+(U_0) \cup S_+(U_0) \cup R_+(\hat{U}_0), \tag{3.6}$$

where $\hat{U}_0 \in S_+(U_0)$ is the end point, where $\xi = \hat{\xi}_0$.

B. $\xi_0 > \xi_I$. The formulae (3.5), (3.6) again define $\gamma_\pm(U_0)$.

To discuss the behavior of $C_-(U)$ as U varies along $\gamma_-(U_0)$, and of $C_+(U)$ for $U \in \tilde{\gamma}_+(U_0)$, it will be convenient to use the same labels for the projections of all quantities U, γ_+, etc. onto either the (p,q)-plane or the (u,v)-plane.

PROPOSITION 1. *Suppose* (1.2)–(1.6) *hold, and let* $U_0 \in \mathbb{R}^4$ *be fixed.*

i) *The family* $\{C_-(U) : U \in \gamma_-(U_0)\}$ *of circles in the* (u,v)*-plane is nested, and fills the entire plane.*

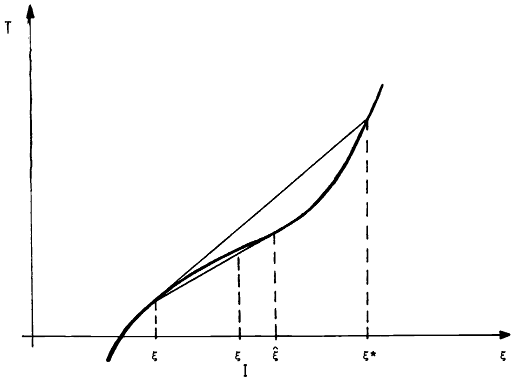


FIG. 2. *Definition of* $\hat{\xi}, \xi^*$.

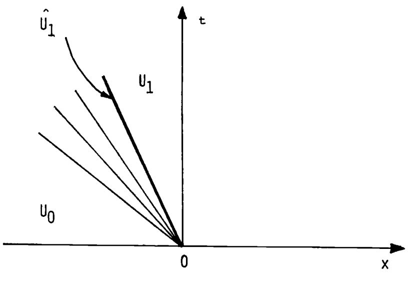


FIG. 3. $U_1 \in P_-(U_0)$.

ii) *The same is true of the family* $\{C_+(U): U\in\tilde{\gamma}_+(U_0)\}$.

Proof. Let $U\in\gamma_-(U_0)$, with $\xi>1$. Then $C_-(U)$ is a circle in the (u,v)-plane with center $\bar{U}$ given by

$$(3.7)\qquad \begin{pmatrix}\bar{u}\\ \bar{v}\end{pmatrix}=\begin{pmatrix}u\\ v\end{pmatrix}-\frac{[\xi T(\xi)]^{1/2}}{\xi}\begin{pmatrix}p\\ q\end{pmatrix}.$$

Let $d(\xi)=[(u_0-\bar{u})^2+(v_0-\bar{v})^2]^{1/2}$ be the distance from $\bar{U}$ to U_0 in the (u,v)-plane. For $\xi<\xi_0$, U lies between $\bar{U}$ and U_0 in the (u,v) plane, on $\gamma_-(U_0)$, whereas for $\xi>\xi_0$, U_0 lies between U and $\bar{U}$. Since the radius $[\xi T(\xi)]^{1/2}$ of $C_-(U_0)$ in the (u,v)-plane is zero when $\xi=1$, and increases with ξ, to prove part (i) of Proposition 1, it is sufficient to observe the following fact, which follows from (1.6) and a case-by-case calculation.

LEMMA. *If* (1.2)–(1.6) *hold, then* $d'(\xi)>0$ *for all* $\xi>1$.

The proof of part (ii) involves a similar calculation.

4. The Riemann problem.

The Riemann problem is the initial value problem

$$(4.1)\qquad U_t=F(U)_x,\qquad -\infty<x<\infty,\quad t>0,$$

$$(4.2)\qquad U(x,0)=\begin{cases}U_L & \text{if } x<0,\\ U_R & \text{if } x>0,\end{cases}$$

where F is given in (1.8), and $U_L\neq U_R$ are given points in $\mathbb{R}^4$.

THEOREM 1. *Suppose* T *satisfies* (1.3)–(1.6). *Let* $U_L, U_R\in\mathbb{R}^4$ *have* $\xi_L\geq 1$ *and* $\xi_R\geq 1$. *Then the Riemann problem* (4.1), (4.2) *has a unique solution among those functions* $U=U(x/t)$ *that are piecewise* C^1, *whose jump discontinuities are admissible shock waves or contact discontinuities, and for which the tension* T *satisfies* $T\geq 0$ *everywhere.*

Proof. For a fixed U_0, let $U_1\in C_-(U_0)$. The corresponding linear wave moves to the left with speed $c_0=[T(\xi_0)/\xi_0]^{1/2}$. Suppose U_0 is on the left of this wave. If U_1 could be joined to another state U by a combination of longitudinal waves moving to the left, with U_1 on the left, then we have $U\in\gamma_-(U_1)$. But every wave in such a combination necessarily has speed greater than c_0, by (1.6). Thus, in any centered combination of linear and longitudinal waves moving in the same direction, the longitudinal waves travel faster than the linear waves. This property depends crucially upon assumption (1.6) (see also §6.).

To solve the Riemann problem, we need to show that U_L may be joined to a state $U_1\in\gamma_-(U_L)$ by a combination of longitudinal waves moving to the left, and U_R may be joined to a state $U_3\in\tilde{\gamma}_+(U_R)$ by a combination of longitudinal waves moving to the right, in such a way that U_1 is joined to U_3 by a combination of linear waves, one moving to the left and one moving to the right. The intermediate state U_2 between these two linear waves must satisfy

$$(4.3)\qquad U_2\in C_-(U_1)\cap C_+(U_3).$$

Let θ_2 be the angle the string makes with the horizontal when in state U_2 (i.e. $\tan\theta_2=q_2/p_2$). Since ξ is constant on each of $C_-(U_1)$, $C_+(U_3)$, (4.3) implies $\xi_1=\xi_2=\xi_3=\xi$, say. Now ξ specifies a single point on $\gamma_-(U_L)$ and a single point on $\tilde{\gamma}_+(U_R)$, while for fixed ξ,θ_2 specifies a single point on $C_-(U_1)$ and a single point on $C_+(U_3)$. We need only guarantee that these two points are coincident, to satisfy (4.3). This is the case precisely when $C_-(U_1)$ and $C_+(U_3)$ are tangent in the (u,v)-plane and not coincident. (Note that they both have radius $[\xi T(\xi)]^{1/2}$ as circles in the (u,v)-plane.) The situation is illustrated in Fig. 4.

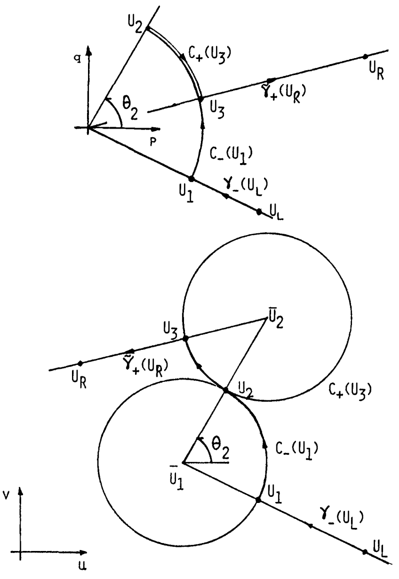


FIG. 4. *Solution of the Riemann problem. Arrows indicate increasing x/t.*

Let $\xi = \xi_1 = \xi_3$, corresponding to points $U_1 \in \gamma_-(U_2)$, $U_3 \in \tilde{\gamma}_+(U_R)$. Since only those solutions of (4.1), (4.2) for which $T \geq 0$ everywhere are being considered, we restrict ξ by $\xi \geq 1$. Consider $\xi = 1$. If $U_1 = U_3$ in the (u,v)-plane, we may join U_1 to U_3 by a degenerate linear wave (for which the tension is zero), which is stationary in the (x,t)-plane. Now, for $\xi > 1$, the curves $C_-(U_1)$, $C_+(U_3)$ are nested in the (u,v)-plane, so they necessarily overlap for $\xi > 1$, since they shrink to the point $U_1 = U_3$ as $\xi \to 1+$. In particular, there is no value of $\xi > 1$ for which $C_-(U_1)$, $C_+(U_3)$ have tangential intersection. Thus, the only solution of the Riemann problem in this context is the degenerate one described above, with U_L joined to a point U_1 by a combination of longitudinal waves moving to the left, with $\xi_1 = 1$, U_1 joined to U_3 by a linear wave that is stationary in the (x,t)-plane, and across which the tension is zero, and U_3 joined to U_R by a combination of longitudinal waves moving to the right.

The only other possibility is that $U_1 \neq U_3$ in the (u,v)-plane when $\xi = 1$. By Proposition 1, there exists a unique value ξ_2 of ξ for which $C_-(U_1)$ and $C_+(U_3)$ are tangent and not coincident. This completes the proof.

Remark. We have excluded the possibility $T < 0$ in some parts of the solution for two reasons. Firstly it could be that $U_1 = U_3$ for some $\xi < 1$, which leads to nonuniqueness of solutions, since necessarily in this case $U_1 \neq U_3$ in the (u,v)-plane when $\xi = 1$. Secondly, if the tension is negative anywhere in the string, the string will tend to bend, since there is no resistance to bending. Thus any such solution may be regarded as unphysical. Note that the equations (4.1) are no longer hyperbolic if $T < 0$, since the characteristic values $\pm[T(\xi)/\xi]^{1/2}$ are then imaginary.

Finally, we note that since $\xi_1=\xi_2=\xi_3=\xi$, the solution of the Riemann problem may be reduced to the solution of a single nonlinear equation for ξ. This equation simply expresses the fact that $C_-(U_1)$ and $C_+(U_3)$ are tangent in the (u,v)-plane precisely when the distance between their centers $\overline{U}_1$, $\overline{U}_3$ is twice the radius:

$$(\bar{u}_1-\bar{u}_3)^2+(\bar{v}_1-\bar{v}_3)^2=4\xi T(\xi). \tag{4.4}$$

If θ_L and θ_R are the angles the string initially makes with the horizontal on the left and right respectively, we have

$$\begin{aligned}\begin{pmatrix}\bar{u}_1\\ \bar{v}_1\end{pmatrix}&=\begin{pmatrix}u_1\\ v_1\end{pmatrix}-[\xi T(\xi)]^{1/2}\begin{pmatrix}\cos\theta_L\\ \sin\theta_L\end{pmatrix},\\ \begin{pmatrix}\bar{u}_3\\ \bar{v}_3\end{pmatrix}&=\begin{pmatrix}u_3\\ v_3\end{pmatrix}+[\xi T(\xi)]^{1/2}\begin{pmatrix}\cos\theta_R\\ \sin\theta_R\end{pmatrix},\end{aligned} \tag{4.5}$$

where (u_k,v_k) $(k=1,3)$ are given explicitly by ξ, and the construction of $\gamma_-(U_L)$, $\gamma_+(U_L)$ described in §3. Let $g(\xi)=(\bar{u}_1-\bar{u}_3)^2+(\bar{v}_1-\bar{v}_3)^2-4\xi T(\xi)$. Setting $\xi=1$, we have (from (1.4))

$$g(1)\geq 0 \tag{4.6}$$

with equality only in the degenerate case, for which $\xi=1$ solves the Riemann problem. Proposition 1 implies that for large enough $\xi=\bar{\xi}$,

$$g(\bar{\xi})<0. \tag{4.7}$$

Consequently, to solve the Riemann problem computationally, one would first determine $\bar{\xi}$ satisfying (4.7) (the first guesses being ξ_L,ξ_R), and then solve

$$g(\xi)=0 \tag{4.8}$$

using interval division, or a more efficient algorithm readily available in scientific programming libraries. Having determined ξ, the intermediate angle θ_2 is simply the angle that the line joining $(\bar{u}_1,\bar{v}_1)$ to $(\bar{u}_3,\bar{v}_3)$ makes with the horizontal.

5. The plucked string. As an example of solutions of the Riemann problem, solutions of the initial boundary value problem for the plucked string are described in this section. The solution is valid until a longitudinal wave first hits an end of the string, so the boundary conditions play no role, except to restrict the initial conditions. It is not hard to analyze the reflection of longitudinal waves from the ends, but the subsequent interactions with linear waves are complicated (especially with the rather general form of T considered in this paper). Accordingly, we do not attempt to continue the solution past the first time at which a wave hits an end.

The initial boundary value problem specifies the function $r(x,t)=(r_1,r_2)(x,t)$ as follows (see Fig. 5):

$$r_{tt}=(r_xT(\xi)/\xi)_x,\quad \xi=|r_x|,\qquad 0<x<1,\quad t>0,$$

$$r_1(0,t)=r_2(0,t)=0,\quad r_1(1,t)=L,\quad r_2(1,t)=0,\quad t>0,$$

$$r_1(x,0)=\begin{cases}ax/x_0 & \text{if } 0\leq x\leq x_0,\\ (a(1-x)+L(x-x_0))/(1-x_0) & \text{if } x_0\leq x\leq 1,\end{cases}$$

$$r_2(x,0)=\begin{cases} bx/x_0 & \text{if } 0\le x\le x_0,\\ b(1-x)/(1-x_0) & \text{if } x_0\le x\le 1,\end{cases}$$

$$r_t(x,0)=0, \qquad 0\le x\le 1;$$

where the constants x_0, L, a and b are arbitrary apart from the constraints

$$\begin{gathered} 0<x_0<1, \quad L>0, \quad 0<a<L, \quad b\ge 0,\\ a^2+b^2>x_0^2, \qquad (L-a)^2+b^2\ge(1-x_0)^2. \end{gathered} \tag{5.1}$$

Condition (5.1) simply guarantees that the tension is initially everywhere positive. Since L may be less than one, no such assumption is made concerning the horizontal string in equilibrium: $r(x,0)=(Lx,0)$.

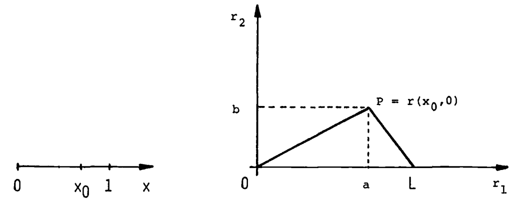


FIG. 5. *Initial configuration of the plucked string.*

In terms of system (1.7), the initial boundary value problem becomes (where $U=(p,q,u,v)$):

$$U_t=F(U)_x, \qquad 0<x<1, \quad t>0, \tag{5.2}$$

$$(u,v)(0,t)=(u,v)(1,t)=0, \quad t>0, \tag{5.3}$$

$$(p,q)(x,0)=\begin{cases} (a,b)/x_0 & \text{if } 0\le x<x_0,\\ (L-a,-b)/(1-x_0) & \text{if } x_0<x\le 1,\end{cases} \tag{5.4}$$

$$(u,v)(x,0)=(0,0) \qquad 0\le x\le 1. \tag{5.5}$$

Equation (5.2) with initial conditions (5.4), (5.5) constitute a Riemann problem centered at $x=x_0$, $t=0$.

To describe the solutions of the Riemann problem in terms of the analysis of sections three and four, we assume T satisfies conditions (1.3)–(1.6). Note that (in the terminology of §4) U_L and U_R are both at the origin in the (u,v)-plane.

Set $\xi_L=[a^2+b^2]^{1/2}/x_0$, $\xi_R=[(L-a)^2+b^2]^{1/2}/(1-x_0)$. The solution of (5.2)–(5.5) for small $t\ge 0$ will involve intermediate states $U_1\in\gamma_-(U_L)$, $U_2\in C_-(U_1)\cap C_-(U_3)$ and $U_3\in\tilde{\gamma}_+(U_R)$. Without loss of generality, we consider the situation when

$$\xi_L\le\xi_R. \tag{5.6}$$

Now $C_-(U_L)$, $C_+(U_R)$ intersect at the origin in the (u,v)-plane. Therefore, if $\xi_1=\xi_3=\xi_R$, then $C_-(U_1)$ and $C_+(U_3)$ overlap. Thus, since $\xi_1=\xi_2=\xi_3=\xi$ for the solution, where ξ satisfies (4.8), we have $g(\xi_R)<0$, so that $g(\xi)=0$ implies

$$\xi<\xi_R. \tag{5.7}$$

Unfortunately, there are no simple conditions on the parameters a, b, x_0, L which serve to further usefully determine ξ and hence the qualitative structure of the solution of the Riemann problem. In Table 1, we simply consider various relationships between ξ_L, ξ_R, ξ_L, and ξ, indicating the form the solution takes. Note that whether $\xi > \xi_L$, $\xi < \xi_L$ or $\xi = \xi_L$ depends upon whether $g(\xi_L) > 0$, $g(\xi_L) < 0$ or $g(\xi_L) = 0$ respectively. Apart from degenerate cases (e.g., $b = 0$), there will always be a linear wave moving left and a linear wave moving right; the indications Left, Right in Table 1 refer to the structure of the longitudinal waves moving to the left or right respectively. A composite wave means a composite of one or more elementary waves. The precise structure of the composite waves may be determined by examining the sign of $g(\xi)$ for $\xi = \xi_L$, ξ_I, ξ_R^*, the points at which the structure of $\gamma_-(U_L)$ and $\tilde{\gamma}_+(U_R)$ changes. The sign of $g(\xi)$ at these points has no simple relationship with the parameters a, b etc. in the initial conditions (5.4).

Note that in every case, the solution necessarily involves longitudinal waves. In particular, suppose the tension is initially small (i.e. less than $T(\xi_I)$) on both sides of the initial discontinuity. The solution involves a shock in each direction if the initial jump in tension is small enough, since for $\xi_L = \xi_R$, we necessarily have $\xi < \xi_L$. On the other hand, if the initial jump in tension is sufficiently large, there is a shock moving toward the side with larger tension, and a rarefaction in the other direction. The cross-over between these two types of solutions occurs precisely when the initial conditions are such that $\xi = \xi_L$, i.e. $g(\xi_L) = 0$.

TABLE 1

Possible wave configurations for the plucked string, when $\xi_L \leq \xi_R$.

	$\xi_L \leq \xi_R \leq \xi_I$	$\xi_L < \xi_I \leq \xi_R$	$\xi_I \leq \xi_L \leq \xi_R$
$\xi_L < \xi < \xi_R$	Left: rarefaction Right: shock	composite composite	shock rarefaction
$\xi < \xi_L$	Left: shock Right: shock	shock composite	composite composite
$\xi = \xi_L$	Left: no wave Right: shock	no wave composite	no wave rarefaction

6. Remarks. To solve the Riemann problem, we have depended heavily upon assumption (1.6), which in particular establishes a fixed order for the characteristic speeds. However, (1.6) is not satisfied by all materials, and in this section, we make some remarks concerning the alternative. That is, suppose (1.3)–(1.5) hold, but (1.6) is replaced by

$$(6.1) \qquad T'(\xi_I) < \frac{T(\xi_I)}{\xi_I}.$$

Now, the definition of the curves $\gamma_\pm, \tilde{\gamma}_\pm$, corresponding to combinations of longitudinal waves does not depend on (1.6) being valid. However, the geometric construction, given in §4, of solutions of the Riemann problem necessarily breaks down under (6.1) for many choices of U_L, U_R. This is because a linear wave may now travel faster than a neighboring longitudinal wave in (x, t)-space. In particular, solutions of the Riemann problem will often involve rarefaction waves with embedded linear waves. Physically, the tension changes continuously through such a combination wave, but the slope undergoes a single jump, precisely when $T'(\xi)=T(\xi)/\xi$. Broadly speaking (and we leave the details to a further paper), the study of such combinations of waves may proceed along the geometric lines described in §§3 and 4. The simple criterion (4.4) for solutions of the Riemann problem only holds provided each longitudinal wave in the supposed solution so obtained travels faster than the linear wave moving in the same direction. In all other situations, the solution must involve combinations of rarefaction waves and linear waves as described above. Some progress with this problem has been made by Keyfitz and Kranzer [4], where T is assumed to be convex or concave, and there is just one point ξ where $T'(\xi)=T(\xi)/\xi$.

Throughout this paper, it has been mathematically convenient to consider $T(\xi)$ defined for all $\xi>0$. That is, the string may undergo local stretching of an arbitrarily large magnitude without breaking. More realistically, suppose T satisfies (1.3)–(1.6), but is defined only on an interval $(0,\xi_{\max}]$, where $\xi_{\max}<\xi_I$ corresponds to the tension $T(\xi_{\max})<\infty$ at which the string will break. The curves $\gamma_-(U_L)$, $\tilde{\gamma}_+(U_R)$ are now terminated at $\xi=\xi_{\max}$, and the Riemann problem can only be solved for U_L, U_R in a restricted region in $\mathbb{R}^4$. Specifically, let U_L be fixed, with $\xi_L<\xi_{\max}$, and $U_-\in\gamma_-(U_L)$ have $\xi=\xi_{\max}$. By Proposition 1, U_L may be joined to a point $U=(p,q,u,v)$ by a combination of longitudinal and linear waves moving to the left if and only if $\xi=(p^2+q^2)^{1/2}<\xi_{\max}$ and (u,v) lies within the projection of $C_-(U_-)$ onto the (u,v)-plane. Similarly, for each U_R with $\xi_R<\xi_{\max}$, we have $U_+\in\tilde{\gamma}_+(U_R)$, with $\xi=\xi_{\max}$. The Riemann problem can be solved if and only if, in the (u,v)-plane

$$C_-(U_-)\cap C_+(U_+)\neq\varnothing \tag{6.2}$$

and the intersection is nontangential. The simplest interpretation of this criterion is that if the initial jump in velocity is too large, then the string will break. One way to see this is to extend $T(\xi)$ for $\xi>\xi_{\max}$, maintaining (1.3)–(1.6). If (6.2) fails, then the solution of the Riemann problem necessarily involves values of $\xi>\xi_{\max}$.

REFERENCES

[1] S. S. Antman, *The equations for large vibrations of strings*, Amer. Math. Monthly, 87 (1980), pp. 359–370.

[2] N. Cristescu, *Dynamic Plasticity*, North-Holland, Amsterdam, 1967.

[3] C. M. Dafermos, *The entropy rate admissibility criterion for solutions of hyperbolic conservation laws*, J. Differential Equations, 14 (1973), pp. 202–212.

[4] B. L. Keyfitz and H. C. Kranzer, *A system of hyperbolic conservation laws arising in elasticity theory*, Arch. Rat. Mech. Anal., 72 (1980), pp. 219–241.

[5] T. P. Liu, *The Riemann problem for general systems of conservation laws*, J. Differential Equations, 18 (1975), pp. 218–234.

[6] B. Wendroff, *The Riemann problem for materials with non-convex equations of State*, I: *isentropic flow*, J. Math. Anal. Appl., 38 (1972), pp. 454–466.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

THE NUMBER OF PEAKS OF POSITIVE SOLUTIONS OF SEMILINEAR PARABOLIC EQUATIONS*

WEI-MING NI† AND PAUL SACKS‡

Abstract. We consider positive, radially symmetric solutions $u(x,t)$ of a class of semilinear parabolic initial-boundary value problems in a ball of $\mathbb{R}^N$. It is shown that there exists a time $T<\infty$ such that $r \to u_r(r,t) \leqq 0$ for $t \geqq T$, $r=|x|$.

Introduction. We consider positive solutions of semilinear parabolic problems with radial symmetry,

$$(0.1)\qquad \begin{array}{ll} u_t-\Delta u=f(t,r,u,u_r), & x\in\Omega,\quad t>0,\\ u(x,t)=0, & x\in\partial\Omega,\\ u(x,0)=u_0(x), & x\in\Omega, \end{array}$$

where $\Omega=\{x\in\mathbb{R}^N:|x|<R\}$, $r=|x|$, $u_r=(x\cdot\nabla u)/r$ and $u_0(x)=u_0(r)\geqq 0$. The main goal of this article is to show that under certain conditions on f there exists a time $T^*>0$ depending on f and u_0 such that $u_r(r,t)<0$ for $t\geqq T^*$ and $r\in(0,R]$. That is to say that the maximum of $x\to u(x,t)$ eventually occurs at $x=0$ and there are no other critical points.

Such a result may be obtained by fairly straightforward arguments provided that the asymptotic behavior of the solution u is sufficiently well understood. To illustrate, if $f=f(u)$ and $u(x,t)$ is uniformly bounded, then one sometimes knows the asymptotic profile of u ([G-V], [H]). This means that there exists a function $\alpha(t)$ such that $\lim_{t\to\infty}\alpha(t)u(x,t)=w(x)$, where $w(x)>0$ is typically a solution of some elliptic equation and satisfies $w_r<0$ for $r>0$. For example, if $f\equiv 0$ we take $\alpha(t)=e^{\lambda_1 t}$ and $w(x)=\psi_1(x)$, λ_1 and ψ_1 being the first eigenvalue and eigenfunction for $-\Delta$ in Ω with zero boundary conditions. One then deduces that $\alpha(t)u(\cdot,t)$ converges to w in $C^2(\Omega)$ from which it follows that $u_r(r,t)<0$ for $r>0$ and $t\geqq T^*$, if T^* is sufficiently large. Using this argument, however, one does not obtain any estimate on the time T^*. The method we use does yield a fairly explicit bound for T^* in many cases. This quantitative aspect is essential for the application which originally motivated this work ([N-S]).

For the more general class of equations 0.1, as well as for unbounded solutions when $f=f(u)$, detailed knowledge of the asymptotic shape is of course more difficult to come by. Also our theorem may sometimes be applied in cases where the solution u does not exist for all $t>0$. See Remark (iv) after the statement of Theorem 1.

Generally speaking, for the application of Theorem 1 below, we will require some knowledge of the behavior of $\|u(\cdot,t)\|_{L^\infty(\Omega)}$ only. This may be fairly easy to obtain using comparison techniques; see Corollary 2.

We mention also that our result is related to some recent work of Matano [M] who considers some one-dimensional problems of the type 0.1. He defines the "lap number", $l(t)$, which is roughly speaking a count of the number of local extrema of $x\to u(x,t)$.

* Received by the editors December 6, 1983.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

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

Under various conditions on f and the boundary conditions he shows that $l(t)$ is nonincreasing in time. Under a somewhat different set of conditions we are showing that actually $l(t)\to 1$ in finite time.

We now state the main results. Denote by $\gamma(t;\, a)$ the solution of

$$(0.2)\qquad \gamma'=f(t,0,\gamma,0),\qquad \gamma(0)=a$$

insofar as it exists.

We will use the following hypotheses on f.

(H)(i) f is C^∞ in all variables, and for each $t>0$ the relation

$$f(t,r,u,u_r)=F(t,x,u,\nabla u)$$

defines F as a real analytic function for $x\in\Omega$, $u>0$, and $\nabla u\in\mathbb{R}^N$.

(ii) $f(t,r,0,0)\geqq 0$ for $t\geqq 0$, $r\in[0,R]$.

(iii) $f_r(t,r,u,0)=0$ for $t\geqq 0$, $r\in[0,R]$, $u>0$.

THEOREM 1. *Suppose f satisfies hypotheses* (H) *and u is a global classical solution of* 0.1 *with $u_0(x)=u_0(|x|)\geqq 0$.*

Suppose that for every $a>0$ there exists $T=T(a)\in(0,\infty]$ such that

$$(0.3)\qquad \overline{\lim_{t\to T}}\,\|u(\cdot,t)\|_{L^\infty(\Omega)}<\underline{\lim_{t\to T}}\,\gamma(t;\,a).$$

Then there exists $T^*<\infty$ such that

$$(0.4)\qquad u_r(r,t)<0$$

for $t\geqq T^*$ and $r\in(0,R]$.

Remarks. (i). The qualitative assumption on the smoothness of f may be significantly weakened by the use of approximation arguments; see the beginning of §4. For example if $f=f(u)$ then we may replace (H) by $f(0)\geqq 0$ and f is locally Lipschitz on $[0,\infty)$.

The analyticity requirement on F is met, for example, if

$$f(t,r,u,u_r)=g(t,r^2,u,ru_r)$$

and g is analytic in its last three arguments.

(ii) Condition (H) is most stringent in its restriction on the r dependence of f. This is somewhat natural, since even for time-independent solutions of (0.1) one cannot expect (0.4) to hold without significant hypotheses on the r dependence ([G-N-N]). A typical case when (H)(iii) holds is for f in the form

$$f(t,r,u,u_r)=f_1(t,r,u)u_r+f_2(t,u).$$

(iii) In §4 we will comment further on the way that T^* depends on f and u_0.

(iv) The condition that u be a global solution of 0.1 may sometimes be dispensed with, namely if one can show that T^* is less than the existence time for u. For example if $f(u)=u^p$, $p>1$, one can show that this is the case for initial values which are sufficiently close to large multiples of $\psi_1(x)$, the first eigenfunction of $-\Delta$ in Ω with zero boundary conditions.

(v) The conclusion (0.4) may be deduced for nonclassical solutions of (0.1) provided that they may be suitably approximated by classical solutions; see [N-S] for example.

As a corollary we mention some specific cases when (0.3) can be verified. The interested reader can supply more such examples.

COROLLARY 2. *Let f satisfy* (H) *and also any one of the following.*

(i) $0<A\leqq f(t,0,u,0)$ *and* $f(t,r,u,u_r)\leqq B<\infty$;

(ii) $A\leqq f(t,0,u,0)/u$ *and* $f(t,r,u,u_r)/u\leqq B$, $B<A+\lambda_1$;

(iii) $f(t,0,u,0)\geqq g(u)>0$ *where* $\int_a^\infty ds/g(s)<\infty$ *for every* $a>0$.

If u is a global classical solution of (0.1) *with* $u_0(x)=u_0(|x|)\geqq 0$, *then there exists* $T^*<\infty$ *such that*

$$u_r(r,t)<0$$

for $t\geqq T^*$ *and* $r\in(0,R]$.

The following technical result is the main ingredient in the proof of Theorem 1. It says roughly that if for some $t_0>0$ there is a local minimum of $r\to u(r,t_0)$ at r_0, then there is a "sufficiently smooth" curve $\eta(t)$ defined on $[0,t_0]$ such that $r\to u(r,t)$ has a local minimum at $\eta(t)$. The fact that there is a level curve of u_r which has this smoothness property is perhaps of some independent interest.

PROPOSITION 3. *Let f satisfy* (H) *and let $u(x,t)$ be a global classical solution of* (0.1) *with* $u_0(x)=u_0(|x|)\geqq 0$. *Assume also that* $u_0(|x|)>0$ *for* $|x|<R$, $u_{0r}(R)<0$ *and* u_0 *is analytic in* Ω.

Suppose $u_r(r^*,t^*)>0$ *for some* $r^*\in[0,R)$, $t^*>0$.

Then there exists a continuous function η: $[0,t^*]\to[0,R)$ *such that*

(i) $u_r(\eta(t),t)=0$, $0\leqq t\leqq t^*$;

(ii) $u_{rr}(\eta(t),t)\geqq 0$, $0\leqq t\leqq t^*$;

(iii) *if* $\eta(t_0)=0$ *for some* $t_0>0$ *then* $\eta(t)\equiv 0$ *for* $t\in[t_0,t^*]$;

(iv) η *is differentiable from the left except possibly at finitely many points.*

In §1 we give the short proof of Theorem 1 using Proposition 3, and also the proof of Corollary 2. The arguments in the proof of Corollary 2 are more or less standard, but we include them for the convenience of the reader. Proposition 3 is proved in §§2 and 3.

Finally in §4 we discuss several ways in which Theorem 1 may be generalized or modified. The first, which has already been mentioned, is the weakening of the smoothness assumptions on f. Secondly, in the case when $\Omega\subset\mathbb{R}$ we may dispense with the symmetry assumption on both u_0 and f. We obtain a result like Theorem 1, except that 0.4 is replaced by the conclusion that the solution has eventually no interior local minimum. Lastly we show that Theorem 1 remains valid if the Dirichlet boundary condition is replaced by a nonlinear Neumann type condition $\partial u/\partial n+\psi(u)=0$, $x\in\partial\Omega$.

1.

LEMMA 1.1. *Let f satisfy* (H) *and let u be a global classical solution of* (0.1) *with* $u_0(x)=u_0(|x|)\geqq 0$, $u_0\not\equiv 0$.

Then for each $t>0$ *we have*

(i) $u(x,t)=u(|x|,\mathrm{t})$,

(ii) $u(x,t)>0$ *for* $x\in\Omega$,

(iii) $u_r(R,t)<0$,

(iv) $r\to u(r,t)$ *is analytic on* $[0,R)$.

Proof. Property (i) is an immediate consequence of the uniqueness of classical solutions of 0.1 (see e.g. [L-S-U]) while (ii) and (iii) are the strong maximum principle and Hopf boundary point lemma respectively, [P-W]. The analyticity result may be found, for example in Friedman [F]. □

Proof of Theorem 1. If $u_0\equiv 0$, the conclusion is obvious. Otherwise, using Lemma 1.1 and replacing $u(x,t)$ by $u(x,t+\tau)$ if necessary, $\tau>0$, we may assume that $u_0>0$ in Ω, $u_{0r}(R)<0$ and u_0 is analytic in Ω.

Let $a>0$ be defined by

$$a=\min\{u_0(r):u_{0r}(r)=0\}$$

and pick T^* so that

$$\gamma(T^*;\, a)>\|u(\cdot,T^*)\|_{L^\infty(\Omega)}. \tag{1.1}$$

If $u_r(r^*,T^*)>0$ for some $r^*\in(0,R)$, we may find a function $\eta(t)$ as in Proposition 3. Set $h(t)=u(\eta(t),t)$. By $h'(t)$ we will mean left-hand derivative; clearly $h'(t)$ exists wherever $\eta'(t)$, the left-hand derivative of η exists.

As long as $\eta(t)>0$ and η' exists we have

$$\begin{aligned}h'(t)&=u_{rr}(\eta(t),t)+u_r(\eta(t),t)\left(\eta'(t)+\frac{N-1}{\eta(t)}\right)+f(t,r,u(\eta(t),t),\,u_r(\eta(t),t))\\ &\geqq f(t,0,h(t),0)\end{aligned}$$

using the hypotheses on f and Proposition 3. If $\eta(t_0)=0$ for some t_0, then by Proposition 3(iii), $h(t)=u(0,t)$ for $t\geqq t_0$, so that

$$h'(t)=\Delta u(0,t)+f(t,0,u(0,t),\,u_r(0,t))\geqq f(t,0,h(t),0)$$

again.

Therefore

$$h'(t)\geqq f(t,0,h(t),0)$$

except possibly at finitely many points, $h(t)$ is continuous on $[0,T^*]$ and $h(0)\geqq a$.

It follows that

$$h(T^*)\geqq\gamma(T^*;\, a)$$

(note that piecewise left differentiability suffices here) and so

$$\|u(\cdot,T^*)\|_{L^\infty(\Omega)}<h(T^*) \tag{1.2}$$

in view of (1.1). But (1.2) is clearly impossible; hence we must have

$$u_r(r,T^*)\leqq 0$$

for $r\in[0,R]$. Applying the strong maximum principle to u_r gives the conclusion (0.4). □

Proof of Corollary 2. (i) For any $a>0$ we have $\lim_{t\to\infty}\gamma(t;\, a)=\infty$. On the other hand u is a positive subsolution of $v_t-\Delta v=B$, all of whose solutions are uniformly bounded ([L-S-U]). Hence u itself is uniformly bounded which implies (0.3).

(ii) Let $w(x,t)=e^{-Bt}u(x,t)$ so that $w_t-\Delta w\leqq 0$. By standard results it follows that $\|w(\cdot,t)\|_{L^\infty(\Omega)}=O(e^{-\lambda_1 t})$ as $t\to\infty$, and therefore $\|u(\cdot,t)\|_{L^\infty(\Omega)}=o(e^{-At})$ as $t\to\infty$, since $B<A+\lambda_1$. But $\gamma(t;\, a)\geqq ae^{-At}$, hence (0.3) holds.

(iii) In this case we have $\lim_{t\to T}\gamma(t;\, a)=\infty$ for some $T=T(a)<\infty$. If $u(x,t)$ exists for all $t>0$ then (0.3) must hold.

2. For simplicity we will carry out the proof of Theorem 3 for the case $f(t,r,u,u_r)=f(u)$ only. The modifications necessary to handle the general case are straightforward.

This section contains two lemmas which will be used for the proof of Proposition 3. The first is a slight modification of the Hopf boundary point lemma. We use the following notation.

$$B_\varepsilon(r_0,t_0)=\{(r,t):(r-r_0)^2+(t-t_0)^2<\varepsilon^2\},$$
$$H_\varepsilon(r_0,t_0)=\{(r,t):(r-r_0)^2+(t-t_0)^2<\varepsilon^2,\ t<t_0\},$$
$$I_T=(0,T)\times(0,R).$$

Let L be the differential operator defined by

$$Lw=w_{rr}+a(r,t)w_r+b(r,t)w$$

where a and b are bounded smooth functions.

LEMMA 2.1. *Assume* $w\in C^2(H_\varepsilon(r_0,t_0))\cap C^1(\overline{H_\varepsilon(r_0,t_0)})$ *for some* $\varepsilon>0$ *and* w *satisfies*

$$\begin{aligned} &w_t=Lw && \text{in } H_\varepsilon(r_0,t_0),\\ &w(r_0-\varepsilon,t_0)=0, &&\\ &w>0 && \text{in } \overline{H_\varepsilon(r_0,t_0)}\setminus\{(r_0-\varepsilon,t_0)\}. \end{aligned}$$

Then $w_r(r_0-\varepsilon,t_0)>0$.

Similarly, if

$$\begin{aligned} &w_t=Lw && \text{in } H_\varepsilon(r_0,t_0),\\ &w(r_0+\varepsilon,t_0)=0, &&\\ &w>0 && \text{in } \overline{H_\varepsilon(r_0,t_0)}\setminus\{(r_0+\varepsilon,t_0)\} \end{aligned}$$

then

$$w_r(r_0+\varepsilon,t_0)>0.$$

Remark. In comparison with the usual statement of the Hopf boundary point lemma [P-W] we are just pointing out that no assumption on w need be made for $t>t_0$. This is seen simply by examining the proof.

Now let $v(r,t)=u_r(r,t)$. Note that v satisfies

$$v_t=v_{rr}+\frac{N-1}{r}v_r-\frac{N-1}{r^2}v+f'(u(r,t))v, \tag{2.1}$$

$$v(0,t)=0,\quad v(R,t)<0,\quad t>0. \tag{2.2}$$

LEMMA 2.2. *Assume the hypotheses of Proposition* 3, *and suppose* $v(r_0,t_0)=u_r(r_0,t_0)>0$ (<0). *Then there exists a function* $\rho\in C[0,t_0]$, $0<\rho(t)<R$ *such that* $\rho(t_0)=r_0$ *and* $v(\rho(t),t)>0$ (<0) *on* $[0,t_0]$.

Proof. First suppose $v(r_0,t_0)>0$. Choose $\alpha>0$ so that $v>0$ on $B_\alpha(r_0,t_0)$ and denote by K the maximal connected component of $\{v>0\}\cap I_{t_0}$ which intersects $B_\alpha(r_0,t_0)$. Choose $\lambda>\|f'(u)\|_{L^\infty(I_{t_0})}$ and set $w=e^{-\lambda t}v$. Then

$$w_t<w_{rr}+\frac{N-1}{r}w_r \tag{2.3}$$

for $(r,t)\in K$, $w>0$ in K. Hence w achieves a positive maximum at some point $(\bar r,\bar t)\in\overline{K}$. We cannot have $\bar r=0$ or $\bar r=R$ by (2.2), and (2.3) rules out $\bar t=t_0$ or a point of K. Hence

$\bar{K}\cap\{t=0\}$ is not empty and there is a point $(r_1,0)\in\bar{K}$ with $v(r_1,0)>0$. Since $v(r,t)>0$ also for (r,t) near $(r_1,0)$, it follows that there exists a smooth path $\Gamma\subset K\cup\{(r_0,t_0)\}\cup\{(r_1,0)\}$ connecting (r_0,t_0) and $(r_1,0)$. We need to show that there is also a path with the same properties which may be parameterized by t.

Let $\delta=\min_\Gamma v$, so $\delta>0$. There exists $\varepsilon>0$ such that for any $p\in\Gamma$, $v\geqq\delta/2$ on $B_\varepsilon(p)\cap I_{t_0}$. It is easy to check that we may choose a finite number of points $p_k\in\Gamma$ so that $\Gamma\subset\bigcup_{k=1}^n B_\varepsilon(p_k)$. With no loss of generality $p_1=(r_0,t_0)$, $p_n=(r_1,0)$ and $B_\varepsilon(p_k)\cap B_\varepsilon(p_{k+1})\neq\varnothing$ for each k.

Let Γ' be the polygonal path obtained by connecting successive centers p_k by line segments. Clearly $\Gamma'\subset K\cup\{p_1\}\cup\{p_n\}$. By removing finitely many segments if necessary we may assume that Γ' is simple.

Now we may regard Γ' as being parameterized by a normalized arc length $\Gamma'=\{(r(s),t(s)):0\leqq s\leqq 1\}$, with $(r(0),t(0))=p_1$, $(r(1),t(1))=p_n$. We have $\Gamma'=\bigcup_{k=1}^l\Gamma_k$ where each Γ_k is a line segment. Consider all segments Γ_k along which $\partial t/\partial s>0$ and let (r_2,t_2) be the endpoint with largest t coordinate; we may suppose that $t_2<t_0$. The horizontal line $\{t=t_2\}$ must intersect Γ' at exactly one point (r_3,t_2) corresponding to a smaller value of the parameter s. Applying the maximum principle we see that $v>0$ on the horizontal segment from (r_3,t_2) to (r_2,t_2). We replace the part of Γ' connecting these two points with this horizontal segment, so that Γ' still has the required properties.

Repeating this procedure finitely many times we obtain a polygonal path Γ'' from p_1 to p_n along which the t coordinate is nonincreasing. Since v will still be bounded away from zero on Γ'', it is possible to make a small perturbation to get yet another path Γ''' connecting p_1 to p_n along which the t coordinate is strictly decreasing. This path is the graph of a continuous function $\rho(t)$ which has all the required properties.

For the case $v(r_0,t_0)<0$ the argument is similar, reversing the direction of inequalities when necessary. In this case it is possible that the minimum of w of $\bar{K}$ ($K=$ maximal connected component of $\{v<0\}$ containing a neighborhood of (r_0,t_0)) could occur at a point $(\bar{r},\bar{t})$ with $\bar{r}=R$. However by Lemma 1.1(iii) and the assumption that $u_{0r}(R)<0$ it follows that $v(r,t)\leqq-\delta$ for $t\in[0,t_0]$ and $r\in[R-\varepsilon,R]$ for some $\varepsilon,\delta>0$. Thus $\bar{K}\cap\{t=0\}$ must still contain a point $(r_1,0)$ with $r_1<R$ and $v(r_1,0)<0$; the rest of the argument is carried out as before. □

3. In this section we prove Proposition 3. To begin with it is necessary to construct $\eta(t)$.

Fix some $T>t^*$ and set $K=$ maximal connected component of $I_T\cap\{v>0\}$ which contains (r^*,t^*); recall $v=u_r$. Define

$$A(\tau)=K\cap\{t=\tau\},$$
$$\eta(\tau)=\inf\{r:r\in A(\tau)\}.$$

By Lemma 2.2 $A(\tau)\neq\varnothing$ for $\tau\in[0,t^*]$, hence η: $[0,t^*]\to[0,R)$. From the construction it is clear that properties (i) and (ii) of Proposition 3 hold. Furthermore, for any $t\in[0,t^*]$ there exists $\delta>0$, depending on t, such that $v(r,t)>0$ for $r\in(\eta(t),\eta(t)+\delta)$; this is a consequence of the analyticity of $v=u_r$.

LEMMA 3.1. *η is continuous on* $[0,t_0]$.

Proof. Fix $t_0>0$. We first show that

$$\varlimsup_{t\to t_0}\eta(t)\leqq\eta(t_0). \tag{3.1}$$

If (3.1) fails then there exists $t_n \to t_0$ such that $\eta(t_n) \to r_1 > \eta(t_0)$. Pick $r_2 \in A(t_0)$ such that $r_2 < r_1$. Since $v(r_2, t_0) > 0$, we have $v > 0$ on $B_\delta(r_2, t_0)$ for sufficiently small δ; in particular $B_\delta(r_2, t_0) \subset K$ for small enough δ. But then for sufficiently large n we must have $\eta(t_n) < r_2$, a contradiction.

Next we claim that

$$(3.2) \qquad \varliminf_{t \downarrow t_0} \eta(t) \geqq \eta(t_0).$$

Suppose (3.2) fails; then there exists $t_n \downarrow t_0$ such that $\eta(t_n) \to r_1 < \eta(t_0)$. On the interval $[r_1, \eta(t_0)] v(\cdot, t_0) \not\equiv 0$ by the analyticity of v, hence either $v > 0$ or $v < 0$ somewhere on this segment. In either case there exists $r_2 \in (r_1, \eta(t_0))$ and $\delta > 0$ such that $B_\delta(r_2, t_0)$ does not intersect K. Next pick n so that $t_n < t_0 + \delta$, and choose $r_3 \in (\eta(t_n), r_2)$ so that $r_3 \in A(t_n)$.

By Lemma 2.2 there is a continuous function $\rho(t)$, $\rho(t_n) = r_3$ with $v(\rho(t), t) > 0$ for $t \in [0, t_n]$. In particular $(\rho(t), t) \in K$ for $t \in [0, t_n]$. But then $\rho(t_0) < \eta(t_0)$ since $(\rho(t), t) \notin B_\delta(r_2, t_0)$ for any t. This contradicts the definition of $\eta(t_0)$.

Finally we must show that

$$(3.3) \qquad \varliminf_{t \uparrow t_0} \eta(t) \geqq \eta(t_0).$$

If not, then there exists $t_n \uparrow t_0$ such that $\eta(t_n) \to r_1 < \eta(t_0)$. As in the previous case there exists $r_2 \in (r_1, \eta(t_0))$ and $\delta > 0$ such that $B_\delta(r_2, t_0)$ does not intersect K. By Lemma 2.2 we can find a function $\rho \in C[0, t_0]$, $\rho(t_0) = r_2$ with $(\rho(t), t) \notin K$ for all t.

Next pick $n > 0$, $r_3 \in A(t_n)$, $r_3 < \rho(t_n)$ and $r_4 \in A(t_0)$. Since the points (r_3, t_n) and (r_4, t_0) both belong to K there must be a path in K which connects them. This path cannot cross the graph of $\rho(t)$, hence it must cross the segment $t = t_0$ at some value of $r < \eta(t_0)$. But this contradicts the definition of $\eta(t_0)$.

Straightforward modification of the above argument gives the continuity at $t_0 = 0$. □

Next, we prove that property (iii) of Proposition 3 holds. By Lemma 3.1 it is enough to check that if $\eta(t_0) = 0$ then $\eta(t_0 + h) = 0$ for sufficiently small $h > 0$.

If $\eta(t_0) = 0$, then $v(r, t_0) > 0$ for $t \in (0, r_0]$, for some $r_0 > 0$. Thus $v(r_0, t) > 0$ for $t \in [t_0, t_1]$, some $t_1 > t_0$. Using the equation for $e^{-\lambda t} v$ as in the proof of Lemma 2.2, it follows that we must have $v(r, t) > 0$ for $r \in (0, r_0]$, $t \in [t_0, t_1]$. This rectangle must belong to K, hence, $\eta(t) = 0$ for $t \in [t_0, t_1]$.

It remains to show the left differentiability of η. Define

$$\begin{aligned}
&J = \{t_0 \in (0, t^*] : \eta(t_0) > 0\}, \\
&J_1 = \{t_0 \in J : \eta(t) < \eta(t_0) \text{ for } t \in [t_0 - \delta, t_0) \text{ some } \delta > 0\}, \\
&J_2 = \{t_0 \in J : \eta(t) > \eta(t_0) \text{ for } t \in (t_0 - \delta, t_0) \text{ some } \delta > 0\}, \\
&J_3 = J \backslash (J_1 \cup J_2).
\end{aligned}$$

We will show

(3.4) (i) $\eta(t)$ is C^∞ at each point of J_1,

(3.5) (ii) $\eta(t)$ is C^∞ except for finitely many points of J_2,

(3.6) (iii) $\eta(t)$ is left differentiable at each point of J_3.

Taken together with property (iii) of Proposition 3, these facts establish that η is left differentiable, except possibly at finitely many points.

LEMMA 3.2. *The statement* (3.4) *holds.*

Proof. Fix $t_0 \in J_1$. By the implicit function theorem it is enough to show that $v_r(\eta(t_0), t_0) \neq 0$. This in turn follows from Lemma 2.1 provided we show that there exists $\varepsilon > 0$ such that

$$\overline{H_\varepsilon(\eta(t_0)+\varepsilon, t_0)} \subset \{v>0\} \cup \{(\eta(t_0), t_0)\}. \tag{3.7}$$

Actually, it is enough to show that

$$H_\varepsilon(\eta(t_0)+\varepsilon, t_0) \subset \{v \geqq 0\} \tag{3.8}$$

because then $v > 0$ in the interior of H_ε; hence (3.7) holds with ε replaced by $\varepsilon/2$.

Suppose then that (3.8) fails for every $\varepsilon > 0$; then there exists $\varepsilon_n \to 0$ and points $(\rho_n, \tau_n) \in H_{\varepsilon_n}(\eta(t_0)+\varepsilon_n, t_0)$ such that $v(\rho_n, \tau_n) < 0$ and $\rho_n > \eta(t_0) > \eta(\tau_n)$ for all n.

Pick $r_1 \in A(\tau_1)$, $r_1 < \eta(t_0)$ such that $v(\cdot, \tau_1) > 0$ on $(\eta(\tau_1), r_1)$, and pick $r_2 \in A(t_0)$. There is a path $\Gamma \subset K$ connecting (r_1, τ_1) and (r_2, t_0). Let $\delta = \inf_\Gamma v$ so $\delta > 0$.

Define a function $\sigma(t)$ in the following way. Pick n_0 so that $v(r,t) \leqq \delta/2$ in $H_{\varepsilon_{n_0}}(\eta(t_0)+\varepsilon_{n_0}, t_0)$. Let $\sigma(t) = \eta(t)$ for $t \geqq t_0$. For $t \in [\tau_{n_0}, t_0]$ $\sigma(t)$ is the function whose graph is the segment from $(\eta(t_0), t_0)$ to (ρ_{n_0}, τ_{n_0}). For $0 \leqq t \leqq \tau_{n_0}$ $\sigma(t)$ is the continuous function given by Lemma 2.2 with $\sigma(\tau_{n_0}) = \rho_{n_0}$. Then $v(\sigma(t), t) \leqq \delta/2$ for all t, $v(\sigma(t), t) \leqq 0$ for $t \geqq t_0$ and $v(\sigma(t), t) < 0$ for $t \leqq \tau_{n_0}$.

We claim that $\sigma(\tau_1) > r_1$; if not then $\sigma(\tau_1) < \eta(\tau_1)$ since $v(r, \tau_1) > 0$ for $r \in (\eta(\tau_1), r_1)$. Since also $\sigma(\tau_{n_0}) = \rho_{n_0} > \eta(\tau_{n_0})$ we must have $\sigma(\bar\tau) = \eta(\bar\tau)$ for some $\bar\tau \in (\tau_1, \tau_{n_0})$, hence $v(\sigma(\bar\tau), \bar\tau) = 0$, a contradiction.

Finally, let Γ have the parameterization $(r(s), t(s))$ $0 \leqq s \leqq 1$ with $(r(0), t(0)) = (r_1, \tau_1)$ and $(r(1), t(1)) = (r_2, t_0)$. Since $\sigma(t(0)) > r(0)$ and $\sigma(t(1)) < r(1)$, it is simple to see, using the continuity of σ that we must have $\sigma(t(s)) = r(s)$ for some $s \in (0,1)$. But then $v(r(s), t(s)) = v(\sigma(r(s)), t(s)) \leqq \delta/2$, contradicting the fact that $v \geqq \delta$ on Γ. □

Before proceeding to the proof of (3.5) we make two observations.

(i) If $\eta(t_0) > 0$ there exists $\delta > 0$ such that $v(r, t_0) < 0$ for $r \in (\eta(t_0) - \delta, \eta(t_0))$. Otherwise, by analyticity we would have $v \geqq 0$ for $r \in (\eta(t_0) - \delta, \eta(t_0) + \delta)$, $v(r,t) > 0$ for $r = \eta(t_0) \pm \delta$, $t \in [t_0, t_0 + \delta)$, some $\delta > 0$. Applying the maximum principle gives $v > 0$ in the rectangle $(\eta(t_0) - \delta, \eta(t_0) + \delta) \times (t_0, t_0 + \delta)$ which contradicts the fact that $\lim_{t \to 0} \eta(t) = \eta(t_0)$.

(ii) The set $\{v > 0\} \cap I_T$ has only finitely many maximal connected components. To see this we note that by Lemma 2.2 each component must intersect $\{t = 0\}$, and then use the analyticity of u_0.

LEMMA 3.3. *The statement* (3.5) *holds.*

Proof. Using arguments as in Lemma 3.2, it is enough to show that there is a finite set $\hat J \subset J_2$ such that for $t_0 \in J_2 \backslash \hat J$ there exists $\varepsilon > 0$ such that

$$H_\varepsilon(\eta(t_0) - \varepsilon, t_0) \subset \{v \leqq 0\}. \tag{3.9}$$

If this is not the case, then there must exist $t_0, \hat t_0 \in J_2$, $t_0 < \hat t_0$, a sequence $\varepsilon_n \downarrow 0$ and points $(\rho_n, \tau_n) \in H_{\varepsilon_n}(\eta(t_0) - \varepsilon_n, t_0)$, $(\hat\rho_n, \hat\tau_n) \in H_{\varepsilon_n}(\eta(\hat t_0) - \varepsilon_n, \hat t_0)$, with (ρ_n, τ_n), $(\hat\rho_n, \hat\tau_n)$ all belonging to a component $K_1 \neq K$ of $\{v > 0\} \cap I_T$. By remark (i) above we may pick $r_1 < \eta(t_0)$ so that $v(\cdot, t_0) < 0$ on $[r_1, \eta(t_0))$. Let $\rho(t)$ be the continuous function given by Lemma 2.2 with $\rho(t_0) = r_1$, and pick n_0 such that $r_1 < \rho_{n_0} < \eta(\tau_{n_0})$. There must be a path $\Gamma \subset K_1$ connecting (ρ_{n_0}, τ_{n_0}) to $(\hat\rho_{n_0}, \hat\tau_{n_0})$. If Γ has the parameterization $(r(s), t(s))$

$0 \leqq s \leqq 1$ and $t(s_0)=t_0$, then $r(s_0) \notin [r_1, \eta(t_0)]$. But then Γ must meet either $\rho(t)$ or $\eta(t)$ for $t > t_0$, a contradiction either way. □

LEMMA 3.4. *The statement* (3.6) *holds.*

Proof. Suppose $t_0 \in J_3$ and η is not differentiable from the left at t_0. First we must have $v_r(\eta(t_0), t_0)=0$ by the implicit function theorem. Also there must exist a sequence $s_n \uparrow t_0$ and $\varepsilon \neq 0$ such that

$$\frac{\eta(s_n)-\eta(t_0)}{s_n - t_0} = \varepsilon \tag{3.10}$$

since otherwise η has left derivative 0 at t_0.

Since $t_0 \in J_3$ there is also a sequence $t_n \uparrow t_0$ such that $\eta(t_n)=\eta(t_0)$. Thus $v(\eta(t_0), t_n) = v(\eta(t_n), t_n)=0$ which implies

$$\frac{\partial^k v}{\partial t^k}(\eta(t_0), t_0)=0 \tag{3.11}$$

for every integer $k \geqq 0$, by successive application of Rolle's theorem. We claim that

$$\frac{\partial^k v}{\partial r^k}(\eta(t_0), t_0)=0$$

for all integers $k \geqq 0$. This will contradict the analyticity of v and therefore complete the proof.

We will actually prove the stronger statement that every partial derivative of v vanishes at $(\eta(t_0), t_0)$. The proof is by induction on the order of the derivative. We have already established that

$$D^\alpha v(\eta(t_0), t_0)=0 \quad \text{for } |\alpha|=0,1$$

where D^α denotes a partial derivative of order $|\alpha|$.

Consider first the case $|\alpha|=2$. From (2.1) we see that $v_{rr}(\eta(t_0), t_0)=0$ while $v_{tt}(\eta(t_0), t)=0$ by (3.11). To obtain the vanishing of v_{rt}, we write Taylor's theorem with remainder for v restricted to the segment from $(\eta(t_0), t_0)$ to $(\eta(s_n), s_n)$. This gives

$$\begin{aligned} v(\eta(s_n), s_n) &= v(\eta(t_0), t_0) + v_t(\eta(t_0), t_0)(s_n - t_0) + v_r(\eta(t_0), t_0)(\eta(s_n)-\eta(t_0)) \\ &\quad + \frac{v_{rr}}{2}(p_n)(\eta(s_n)-\eta(t_0))^2 + v_{rt}(p_n)(\eta(s_n)-\eta(t_0))(s_n - t_0) \\ &\quad + \frac{v_{tt}(p_n)}{2}(s_n - t_0)^2, \end{aligned}$$

where p_n denotes some point on this segment. Upon rearrangement we obtain

$$v_{rt}(p_n) = -\frac{1}{2}\left[v_{rr}(p_n)\varepsilon + \frac{v_{tt}(p_n)}{\varepsilon} \right]$$

and letting $n \to \infty$ gives $v_{rt}(\eta(t_0), t_0)=0$.

For the general induction step, assume $D^\alpha v(\eta(t_0), t_0)=0$ for $|\alpha|=k$. We have

$$\frac{\partial^{k+1} v}{\partial t^{k+1}}(\eta(t_0), t_0)$$

by (3.11), and applying D^{β} to (2.1) for any multi-index β of order $k-1$ yields $k-1$ equations expressing a partial derivative of order $k+1$ in terms of derivatives of order k and lower. In this way we obtain the vanishing of all derivatives of order $k+1$ except possibly

$$\frac{\partial}{\partial r}\frac{\partial^k v}{\partial t^k}(\eta(t_0),t_0).$$

This last one is handled by the use of Taylor's theorem and (3.10) as in the case $k=1$.

This completes the proof. □

4. (I). It is of interest to remove the smoothness assumptions which have been made on the nonlinearity f. While Proposition 3 uses this hypothesis in an essential way, the conclusion of Theorem 1 can be shown to be valid for more general f's provided we can make suitable approximations. For simplicity we consider the case $f(t,r,u,u_r)=f(u)$.

It is first of all necessary to determine more precisely how the time T^* in Theorem 1 depends on the data of the problem. Given $v\in C^1(\overline{\Omega})$, $v=v(|x|)$, we define

$$a(v)=\min\{v(r):v_r(r)=0\}.$$

An examination of the proof of Theorem 1 shows the following. If f satisfies (H), then for any $t_0>0$, u satisfies $u_r(r,t)\leqq 0$ for $t\geqq T$ if

$$\|u(\cdot,T+t_0)\|_{L^\infty(\Omega)}<\gamma(T;\,a(u(\cdot,t_0))). \tag{4.1}$$

THEOREM 2. *In the statement of Theorem* 1, *the hypotheses* (H) *may be replaced by*

$f=f(u)$,

$f(0)\geqq 0$,

f locally Lipschitz on $[0,\infty)$.

Proof. Replacing $u(x,t)$ by $u(x,t+\tau)$ if necessary, $\tau>0$, we may assume $u_0>0$ in Ω, $u_{0r}(R)<0$ and $u_0\in C^1(\overline{\Omega})$.

Choose $R'<R$ such that $u_{0r}(r)<0$ for $r\in[R',R]$ and set

$$\alpha=\min\{u_0(r):r\leqq R'\};$$

note $\alpha>0$.

Pick T so that

$$\|u(\cdot,T+t_0)\|_{L^\infty(\Omega)}<\gamma\left(T;\,\frac{\alpha}{2}\right) \tag{4.2}$$

for all sufficiently small $t_0>0$. Let

$$M=\max\left(\|u\|_{L^\infty((0,T+\varepsilon)\times\Omega)},\ \|\gamma\|_{L^\infty(0,T+\varepsilon)}\right)+1$$

where $\varepsilon>0$ is so small that M is finite, and choose a sequence of functions $\{f_n\}$ uniformly bounded on $[0,\infty)$, $f_n(0)\geqq 0$, f_n analytic on $(0,M)$ and $f_n\to f$ uniformly on $[0,M]$.

Define $u_n(x,t)$ to be the solution of

$$\begin{aligned} &v_t-\Delta v=f_n(v), && x\in\Omega,\quad t>0,\\ &v(x,t)=0, && x\in\partial\Omega,\\ &v(x,0)=u_0(x), && x\in\Omega. \end{aligned} \tag{4.3}$$

By standard arguments [L-S-U] each u_n exists for all $t \geqq 0$ and $u_n \to u$ in $C^1(\overline{\Omega} \times [0, T])$. By [G-N-N, Thm. 5.2] $u_{nr}(r,t) < 0$ for $r \in [R', R]$, all n and all $t > 0$. Pick $t_0 \in (0, \varepsilon)$ such that $u(r, t_0) > 3\alpha/4$ for $r \in [0, R']$. For sufficiently large n we must have $u_n(r, t_0) > \alpha/2$ for $r \in [0, R']$. In particular we may assume that

$$a(u_n(\cdot, t_0)) > \frac{\alpha}{2} \tag{4.4}$$

for all n.

If $\gamma_n(t;\, a)$ denotes the solution of

$$\gamma' = f_n(\gamma), \qquad \gamma(0) = a,$$

we have $\gamma_n \to \gamma$ uniformly for $t \in [0, T + \varepsilon]$ and any fixed $a > 0$.

From all of the above remarks it follows that

$$\|u_n(\cdot, T + t_0)\|_{L^\infty(\Omega)} < \gamma_n(T;\, a(u_n(\cdot, t_0)))$$

for n sufficiently large.

Thus

$$u_{nr}(r, t) \leqq 0$$

for $t \geqq T$ and large enough n from which (0.4) follows. □

4. (II). In the case that $\Omega \subset \mathbb{R}$ we may dispense with the symmetry requirement on u_0, and consider f in the form $f = f(t, x, u, u_x)$, with $f(t, x, u, 0) \geqq 0$, $f_x(t, x, u, 0) = 0$, f infinitely differentiable in all variables, and analytic in (x, u, u_x) for fixed t.

Instead of (0.4) we obtain the conclusion that there exists $T^* < \infty$ such that for $t \geqq T^*$

$$x \to u(x, t)$$

has no local minimum in $(-R, R)$.

To prove this it is only necessary to define $\eta(t)$ from Proposition 3 in a slightly different way. Note that Lemma 2.2 is still valid in this situation.

If $x \to u(x, t^*)$ has a local minimum at x^*, then there exists x_1 and x_2, $-R < x_1 < x^* < x_2 < R$, such that

$$u_x(x_1, t^*) < 0, \qquad u_x(x_2, t^*) > 0. \tag{4.5}$$

We set $K =$ maximal connected component of $\{u_x > 0\}$ which contains (x_2, t^*). If we define

$$A(\tau) = K \cap \{t = \tau\},$$
$$\eta(\tau) = \inf\{r : r \in A(\tau)\},$$

then by (4.5), Lemma 2.2, and the Hopf boundary point lemma, $\eta(\tau)$ is defined and satisfies $-R < \eta(\tau) < R$ on $[0, t^*]$. The remainder of the proof proceeds as in §3.

As an example we can handle Burgers' equation

$$u_t + uu_x = u_{xx}.$$

In this case (0.3) can be verified since $\lim_{t \to \infty} \gamma(t;\, a) = a$ and $\|u(\cdot, t)\|_{L^\infty(\Omega)} \to 0$ as $t \to \infty$ (see [S, Chap. 21]).

If the domain Ω is an annulus in $\mathbb{R}^n$ and $u_0(x)=u_0(|x|)$, then (0.1) is a one-dimensional problem of the type just discussed, so we obtain the same conclusion.

4. (III). Finally, we can replace the zero boundary condition by the second boundary condition

$$\frac{\partial u}{\partial n}+\psi(u)=0, \qquad x\in\partial\Omega$$

where $\partial/\partial n$ is differentiation in the outward normal direction. Since $\partial u/\partial n=\partial u/\partial r$ the rotational symmetry of the problem is preserved. We assume first of all that $u\psi(u)\geqq 0$, so that $u(x,t)\geqq 0$ if $u_0(x)\geqq 0$. Under these circumstances Theorem 1 still follows from Proposition 3, and Proposition 3 is valid as long as we can be sure that the curve $\eta(t)$ does not run into the boundary $|x|=R$. For example if we require $u\psi(u)>0$ for $u>0$, then this latter possibility is ruled out.

Note that we do not expect to obtain the conclusion of Theorem 1 for the homogeneous Neumann condition $\psi(u)\equiv 0$. For example if $\Omega=(-\pi,\pi)$ and $f\equiv 0$, then

$$u(x,t)=2+e^{-4t}\cos 2x$$

is a positive solution of (0.1) for which (0.4) does not hold for any $t>0$.

Acknowledgment. We would like to thank N. Alikakos for some helpful conversations.

REFERENCES

[F] A. Friedman, *On the regularity of the solutions of nonlinear elliptic and parabolic systems of partial differential equations*, J. Math. Mech., 7 (1958), pp. 43–59.

[G-N-N] B. Gidas, W. Ni and L. Nirenberg, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[G-V] A. Gmira and L. Veron, *Asymptotic behavior of the solution of a semilinear parabolic equation*, Monat. für Math., 94 (1982), pp. 299–312.

[H] C. Holland, *Limiting behavior of a class of nonlinear reaction diffusion equations*, Quart. Appl. Math., Oct. 1982, pp. 293–296.

[L-S-U] O. A. Ladyzhenskaya, V. A. Solonnikov and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[M] H. Matano, *Nonincrease of the lap number for a one-dimensional semilinear parabolic equation*, J. Fac. Sci. Univ., Tokyo IA, 29 (1982), pp. 401–441.

[N-S] W. Ni and P. Sacks, *Singular behavior in nonlinear parabolic equations*, Trans. Amer. Math. Soc., to appear.

[P-W] M. Protter and H. Weinberger, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[S] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

MODULATIONAL STABILITY OF GROUND STATES OF NONLINEAR SCHRÖDINGER EQUATIONS*

MICHAEL I. WEINSTEIN†

Abstract. The modulational stability of ground state solitary wave solutions of nonlinear Schrödinger equations relative to perturbations in the equation and initial data is studied. In the "subcritical case" ground states are shown by variational methods to be stable modulo time-dependent adjustments (modulations) of free parameters. These parameters satisfy the *modulation equations*, a coupled system of nonlinear ODE's governing the amplitude, phase, position and speed of the dominant solitary wave part of the solution.

1. Introduction. The initial-value problem (IVP) for the nonlinear Schrödinger equation (NLS)

$$(1.1)^1 \qquad 2i\phi_t+\Delta\phi+|\phi|^{2\sigma}\phi=0, \qquad 0<\sigma<\frac{2}{N-2},$$

$$(1.2) \qquad \phi(x,0)=\phi_0(x), \qquad x\in\mathbb{R}^N$$

arises in the mathematical description of a diverse set of physical phenomena. Some of these are

(a) the propagation of a narrow electromagnetic beam through a medium with an index of refraction dependent on the field intensity [1], [7], [18], [27],

(b) electromagnetic (Langmuir) waves in a plasma [31], [32], and

(c) the motion of a vortex filament for the Euler equations of fluid mechanics [15].

NLS has nonlinear bound states which are "localized" finite energy solutions. These are believed to describe wave phenomena that are observed in the above physical contexts. Such solutions of (1.1) can be found in the form

$$(1.3)^2 \qquad \psi^0(x,t)=u(x)e^{it/2}.$$

Substitution of (1.3) into (1.1) implies

$$(1.4) \qquad \Delta u-u+|u|^{2\sigma}u=0, \qquad 0<\sigma<\frac{2}{N-2}.$$

The existence of infinitely many H^1 solutions of (1.4) follows from work of Strauss [24] (see also [4]). Among them is a real, positive, and radial solution which we call the *ground state* and denote by $R(x)$. To describe a physical phenomenon, a nonlinear bound state should be stable. In this paper we study the stability of the ground state relative to small perturbations in the equation and initial data.

*Received by the editors April 14, 1983, and in final revised form November 14, 1983.

† Department of Mathematics, Stanford University, Stanford, California 94305. Present address: Department of Mathematics, Princeton University, Princeton, New Jersey 08544. This research was supported in part by the National Science Foundation, the Office of Naval Research, the Air Force Office of Scientific Research and the Army Research Office.

[1] For $N=1$ or 2 we allow $0<\sigma<\infty$.

[2] If we seek solutions of the form $\psi^0=ue^{iEt/2}$, E real, then the rescaling $W(x)=E^{1/2\sigma}u(Ex)$ leads to (1.4).

We now discuss the sense in which we expect the ground state to be stable and then state our results in this direction. Let a, η_0, $\xi=(\xi_1,\cdots,\xi_N)$ and $\theta_0=(\theta_{0,1},\cdots,\theta_{0,N})$ be real constants. By the scaling properties of (1.1) or by direct verification it can be seen that the functions

$$(1.5)\qquad \psi(x,t;a,\theta_0,\xi,\eta_0)=a^{-1/\sigma}R\left(a^{-1}[\theta-\theta_0]\right)\exp\left[i(\xi\cdot(\theta-\theta_0)+(\eta-\eta_0))\right]$$

form a $(2N+2)$-parameter family of solutions of (1.1) for $0<\sigma<2/(N-2)$ provided the following relations hold:

$$(1.6)\qquad \begin{aligned} &(\mathrm{a})\quad \frac{\partial\theta_i}{\partial t}=-\xi_i, &&(\mathrm{c})\quad \frac{\partial n}{\partial t}=\frac{a^{-2}+\xi\cdot\xi}{2},\\ &(\mathrm{b})\quad \frac{\partial\theta_i}{\partial x_j}=\delta_{ij}, &&(\mathrm{d})\quad \frac{\partial\eta}{\partial x_j}=0,\qquad 1\leqq i,j\leqq N. \end{aligned}$$

In the "critical" case $\sigma=2/N$, there is at least one more scale invariance of (1.1). In particular, for $b\in\mathbb{R}$ the functions [20]

$$(1.7)\qquad \begin{aligned} &\psi(x,t;a,b,\theta_0,\xi,\eta_0)\\ &\quad=(a+bt)^{-N/2}R\left((a+bt)^{-1}[\theta-\theta_0]\right)\\ &\quad\quad\cdot\exp\left[i\left(\xi\cdot(\theta-\theta_0)+\frac{b}{2}(a+bt)^{-1}|\theta-\theta_0|^2+\eta-\eta_0\right)\right] \end{aligned}$$

form a $(2N+3)$-parameter family of solutions of (1.1) provided (1.6a, b, d) are satisfied together with the following extension of (1.6c):

$$(1.6\mathrm{c}')\qquad \frac{\partial\eta}{\partial t}=\frac{(a+bt)^{-2}+\xi\cdot\xi}{2}.$$

Note that (1.7) reduces to (1.5) when $b=0$. We will refer to the functions (1.5) when $\sigma\neq 2/N$ and (1.7) when $\sigma=2/N$ as the *ground state traveling wave family* or simply *ground state family* of NLS.

To study the stability of the ground state family we consider the perturbed IVP

$$(1.8)\qquad 2i\phi_t^\varepsilon+\Delta\phi^\varepsilon+|\phi^\varepsilon|^{2\sigma}\phi^\varepsilon=\varepsilon F(|\phi^\varepsilon|)\phi^\varepsilon,\qquad \phi^\varepsilon(x,0)=R(x)+\varepsilon S(x).$$

In general, the solution ϕ^ε of (1.8) will not evolve in the simple form

$$(1.9)\qquad \phi^\varepsilon(x,t)=\left[R(x)+\varepsilon w_1+\varepsilon^2 w_2+\cdots\right]e^{it/2}$$

with $\varepsilon w_1+\varepsilon^2 w_2+\cdots$ genuinely small for large times (say of order $1/\varepsilon$). The possibility of the *linearized perturbation*, εw_1, becoming nonnegligible for large time can be displayed as follows. Differentiation with respect to the free parameters of ψ in (1.5) or (1.7) generates solutions of the linearized (about the ground state) equation with polynomial growth in time. Since in general εw_1 will contain these solutions of the linearized problem or "secular modes", εw_1 will not remain small for large times.

When $\sigma<2/N$ there are $2N+2$ secular modes associated with the $2N+2$ parameter family of solutions (1.5). When $\sigma=2/N$, 2 more secular modes arise giving $2N+4$. One of these new modes is associated with the new parameter b in (1.7). The other, however, has not been associated with a classical symmetry of equation (1.1). This is further discussed in [20].

In §2 we show that these secular modes are the only source of linear instability of the ground state. In particular, we show that by constraining the evolution of the linearized perturbation w_1 to the space H^1 with the secular modes removed, w_1 is controllable in H^1 (Theorem 2.12). We call this space M. All theorems of §2 are stated and proofs carried out in an arbitrary spatial dimension N, although the technical point about $N(L_+)$ in part b of Proposition 2.8 has been completely proved, only in dimension $N=1$ for all σ and in dimension $N=3$ for $0<\sigma\leqq 1$ (see appendix A). Since the stability analysis of the ground state concerns the case $\sigma\leqq 2/N$, our results are completely rigorous in dimensions $N=1$ and 3, and are lacking only in the above mentioned technical point in other dimensions.

A natural remedy to the growth of w_1 in (1.9) is the use of the ground state *family* with slowly varying parameters: $a(\varepsilon t)$, $\xi(\varepsilon t)$ etc. as the leading order Ansatz. The idea is to choose the slow functions a, ξ etc. to constrain the evolution of w_1 to M, thereby ensuring that $\varepsilon w_1(t)$ is genuinely small for times of order $1/\varepsilon$. More precisely, in §3 we prove our main result which we state now as (later as Theorem 1′)

THEOREM 1. *Let $\sigma<2/N$. Expand the solution $\phi^\varepsilon(x,t)$ of* (1.8) *as*

$$\phi^\varepsilon(x,t)=\left(\lambda^{1/\sigma}R(\lambda(\theta-\theta_0))+\varepsilon w_1+\varepsilon^2 w_2+\cdots\right)e^{i[\xi\cdot(\theta-\theta_0)+\eta-\eta_0]} \tag{1.10}$$

with

$$w_1=w_1\left(\lambda(\theta-\theta_0),\int_0^t\lambda^2\,ds\right).$$

For a class of perturbations F, if the $2N+2$ parameters λ,ξ,θ_0 and η_0 evolve as functions of $T\equiv\varepsilon t$, according to the coupled system of $2N+2$ ordinary differential equations (3.5) (*the modulation equations*), *then*

(i) *$w_1\in M$ for $t>0$, and*

(ii) *for any $T_0>0$*

$$\sup_{0\leqq t\leqq T_0/\varepsilon}\|\varepsilon w_1(t)\|_{H^1}=\alpha(\varepsilon), \tag{1.11}$$

where $\alpha(\varepsilon)\downarrow 0$ as $\varepsilon\to 0$.

The modulation equations have been derived previously for various one-dimensional formal perturbation theories [16], [17], [19], [22]. An aim of this paper is to present some justification for these perturbation techniques. Results on the nonlinear stability of ground states were obtained by Cazenave [5] for a logarithmic NLS and by Cazenave and Lions [6] for (1.1) with $\sigma<2/N$. In these works, perturbations of the initial data alone are considered. They prove H^1 stability of the ground state modulo "adjustments in the free phase and centering parameters." These adjustments are incorporated in the norm and are not explicitly constructed. (This idea has been used for other equations as well [2], [3], [10].) Thus, their work shows that initial data near a solitary wave evolves nearly as a solitary wave with unspecified, possibly different position and phase for times $t>0$. The modulation theory presented and analyzed here provides an approximate and constructive answer to the questions (a) *where the solitary wave is located* and (b) *what its phase is* for $t>0$ (see Example 3.1). When small perturbations of the equation (e.g. dissipation) occur, the phase and centering parameters are, in general, not sufficient to describe the dominant part of the solution for $t>0$. Modulation theory gives the additional explicit information on the *amplitude* and *speed* of the solitary wave needed to track the solution closely (see Example 3.2).

In the mathematical theory of NLS, the case $\sigma=2/N$ has been understood to play a distinguished role. It is also of physical interest since the case $\sigma=1$, $N=2$ arises in modelling self-focusing optical beams [18]. The cases $\sigma<2/N$, $\sigma=2/N$ and $\sigma>2/N$ are called, respectively, *subcritical*, *critical* and *supercritical* cases. A consequence of work of Ginibre and Velo [13] is that if $\sigma<2/N$, (1.1)–(1.2) has global solutions in $C([0,\infty);H^1(\mathbb{R}^N))$ for all $\phi_0\in H^1(\mathbb{R}^N)$. In the case $\sigma\geqq 2/N$ Glassey [14] has displayed a class of initial data for which the solution of the IVP "blows up" in finite time in H^1, i.e. there is a finite time T, such that $\int|\nabla_x\phi(x,t)|^2dx\to\infty$ as $t\to T$ (see also [28]).

Numerical observations in certain critical cases (Zakharov–Synakh [33] for $(\sigma,N)=(1,2)$ and later Sulem–Sulem–Patera [26] for $(\sigma,N)=(1,2)$ and $(2,1)$) indicate that when $\sigma=2/N$, the ground state plays an important role in the structure of solutions developing singularities. This phenomenon is discussed in detail in the articles [20], [26].

Related analytical results were obtained in [29] where it was proved using a particular variational characterization of the ground state (exploited further in §2) that when $\sigma=2/N$, a sharp sufficient condition for global existence in (1.1)–(1.2) is

$$\int|\phi_0(x)|^2dx<\int R^2(x)\,dx. \tag{1.12}$$

The linearized stability results of §2 for the case $\sigma=2/N$ also give information on the special role of the ground state. In this case the techniques of [6] do not apply.

Notation. All integrals are understood to be taken over $\mathbb{R}^N$, the N-dimensional Euclidean space, unless otherwise indicated.

1) $W=(u,v)^t=\binom{u}{v}\in\mathbb{R}^2$,

2) $\|f\|_p=\left(\int|f(x)|^pdx\right)^{1/p}$, $\|W\|_p^p=\|u\|_p^p+\|v\|_p^p$,

$L^p(\mathbb{R}^N)=L^p=\{f\colon\|f\|_p<\infty\}$,

3) $\|f\|_{H^s}^2=\int|\hat{f}(\xi)|^2\left(1+|\xi|^2\right)^s d\xi$,

$H^s(\mathbb{R}^N)=H^s=\{f\colon\|f\|_{H^s}<\infty\}$,

4) $N(A)=$ null space of an operator A,

5) for functions $f,g\in\mathbb{R}^N$, $(f,g)=\sum_{i=1}^N\int f_ig_i\,dx$.

2. The linearized NLS operator. We first consider the stability of the gound state $\psi^0(x,t)=R(x)e^{it/2}$ of (1.1) by seeking a solution to the perturbed equation (1.8) of the form (1.9). Small perturbations of initial data may be incorporated in the equation through a redefinition of ϕ^ε (see Example 3.1, (3.7)). Substitution of (1.9) into (1.8) and linearization yields the following IVP for the linearized perturbation w:

$$2iw_t+\Delta w-w+(\sigma+1)R^{2\sigma}w+\sigma R^{2\sigma}\overline{w}=F(R)R,\qquad w(x,0)=0. \tag{2.1}$$

We will also study the homogeneous IVP, (2.1) with $F\equiv0$ and $w(x,0)=w_0(x)$, which we denote by (2.1H). IVP (2.1H) has a useful conserved quantity given in the following:

THEOREM 2.1 *Let $w(x,t)$ be an H^1-solution of* (2.1H). *Then*

$$\int\left(|\nabla w|^2+|w|^2-(\sigma+1)R^{2\sigma}|w|^2-\frac{\sigma}{2}R^{2\sigma}(w^2+\overline{w}^2)\right)dx \tag{2.2}$$

is independent of time.

Proof. We multiply (2.1) by $\bar{w}_t$, take the real part and then integrate by parts. This gives (d/dt) (2.2) = 0. □

It is expedient to work with the real and imaginary parts of w. We set $w \equiv u + iv$ and make the following definitions:

$$(2.3) \quad \text{(a)} \quad L_+ \equiv -\Delta + 1 - (2\sigma+1)R^{2\sigma}, \qquad \text{(c)} \quad L = \begin{pmatrix} 0 & L_- \\ -L_+ & 0 \end{pmatrix},$$

$$\text{(b)} \quad L_- = -\Delta + 1 - R^{2\sigma}, \qquad \text{(d)} \quad W = \begin{pmatrix} u \\ v \end{pmatrix}.$$

Now (2.1) can be written as the real system:

$$(2.4) \qquad 2W_t = LW + G, \qquad W|_{t=0} = 0,$$

where G is a 2-vector with components $\operatorname{im} F \cdot R$ and $-\operatorname{re} F \cdot R$. We denote the homogeneous IVP ($G \equiv 0$ and $W|_{t=0} = W_0$) by (2.4H). Let Ω_t denote the propagator or solution operator for (2.4H). Thus $W(t) \equiv \Omega_t W_0$. Theorem 2.1 can now be expressed as:

COROLLARY 2.2. *Let $W \in H^1 \times H^1$ be a solution of* (2.4H). *Then, for $t > 0$,*

$$(2.5) \qquad Q(W) \equiv Q(u,v) \equiv (L_+u,u) + (L_-v,v) = Q(u_0,v_0).$$

We would like to use $Q^{1/2}$ as a norm, measuring the size of the perturbation W. However, Q is not positive definite on $H^1 \times H^1$. We now introduce a subspace on which we will see that $Q^{1/2}$ is equivalent to the $H^1 \times H^1$ norm.

DEFINITION 2.3. For $\sigma \leqq 2/N$ we set

$$(2.6) \qquad M \equiv H^1 \times H^1 \cap \left[N_g(L^*)\right]^{\perp},$$

where $B^{\perp} = \{a = (a_1,a_2) | \int a_1b_1 + a_2b_2\,dx = 0 \text{ for all } b = (b_1,b_2) \in B\}$ and $N_g(A) =$ generalized null space of $A = \bigcup_{j=1}^{\infty} N(A^j)$.

In appendix B we derive and display explicitly the elements of $N_g(L)$ and $N_g(L^*)$ (Theorems B.2–3). This explicit information and Definition 2.3 imply the following orthogonality relations which we require in the coming analysis:

PROPOSITION 2.4. *Let $(f,g)^t \in M$. Then* (i) *For $\sigma < 2/N$, the following $2N+2$ orthogonality relations hold*:

$$(2.7) \quad \text{(a)} \quad (f,R) = 0, \qquad \text{(c)} \quad \left(g, R_{x_j}\right) = 0,$$

$$\text{(b)} \quad (f, x_jR) = 0 \qquad \text{(d)} \quad \left(g, \frac{1}{\sigma}R + x \cdot \nabla R\right) = 0, \qquad 1 \leqq j \leqq N.$$

(ii) *For $\sigma = 2/N$, we have the $2N+4$ orthogonality conditions*: (2.7) *and the two new relations*

$$(2.8) \quad \text{(a)} \quad \left(f, |x|^2R\right) = 0,$$

$$\text{(b)} \quad (g,\rho) = 0, \quad \textit{where } L_+\rho = -|x|^2R.$$

The following result shows that under suitable restrictions the linearized energy controls a classical norm.

THEOREM 2.5. *Let $\sigma \leqq 2/N$ and $(f,g)^t \in M$. There exist positive constants K and K', independent of f and g, such that*

$$(2.9) \qquad K\left(\|f\|_{H^1}^2 + \|g\|_{H^1}^2\right) \leqq Q(f,g) \leqq K'\left(\|f\|_{H^1}^2 + \|g\|_{H^1}^2\right).$$

Thus for $\sigma \leqq 2/N$, M is a closed linear subspace on which the functional $Q^{1/2}$ defines a norm equivalent to the $H^1 \times H^1$ norm.

The upper bound in (2.9) is simple and holds for any f and g in H^1. The difficult part is the lower estimate. We base the proof on several propositions. The first is a particular characterization of the ground state R, introduced in [29, Thm. B]. For $u \in H^1$ we define the functional

$$(2.10) \qquad J^{\sigma,N}(u) = \frac{\|\nabla u\|_2^{\sigma N} \|u\|_2^{2+\sigma(2-N)}}{\|u\|_{2\sigma+2}^{2\sigma+2}}, \qquad 0 < \sigma < \frac{2}{N-2}.$$

PROPOSITION 2.6. *For $0 < \sigma < 2/(N-2)$*

$$(2.11) \qquad \alpha \equiv \inf_{u \in H^1(\mathbb{R}^N)} J^{\sigma,N}(u)$$

is attained at a function R with the following properties:

(1) *$R > 0$ and $R = R(|x|)$,*

(2) *$R \in H^1(\mathbb{R}^N) \cap C^\infty(\mathbb{R}^N)$,*

(3) *R is a solution of (1.4).*

In [29] the functional $J^{\sigma,N}$ was minimized to obtain the optimal constant of a classical interpolation estimate of Nirenberg [23] and Gagliardo [11], [12]. There, the optimal constant, which is expressible in terms of α, is used in an a priori estimate derived from the conserved quantities of (1.1) to obtain the condition (1.12) for global existence when $\sigma = 2/N$.

We now use $J^{\sigma,N}$ to prove the following result which is at the heart of Theorem 2.5:

PROPOSITION 2.7. *Let $\sigma \leqq 2/N$. Then*

$$(2.12) \qquad \inf_{(f,R)=0} (L_+ f, f) = 0.$$

Proof I.[3] First, note that $L_+ \nabla R = 0$ and $(\nabla R, R) = 0$. Therefore the infimum in (2.12) is nonpositive. Since $J^{\sigma,N}$ attains its minimum at R,

$$(2.13) \qquad \left.\frac{d^2}{d\varepsilon^2}\right|_{\varepsilon=0} J^{\sigma,N}(R + \varepsilon\eta) \geqq 0$$

for all $\eta \in C_0^\infty(\mathbb{R}^N)$. Now (2.13) can be written as

$$(2.14) \qquad (T\eta, \eta) \geqq 0, \quad \text{where}$$

$$(2.15) \qquad (2+\sigma(2-N))^{-1} Tz \equiv L_+ z + (2-N) a_{\sigma N}(R,z) R - b_{\sigma N}(R,z)\Delta R + (\sigma N - 2) c_{\sigma N}(\Delta R, z)\Delta R.$$

Here, $a_{\sigma N}$ $b_{\sigma N}$ and $c_{\sigma N}$ are constants, dependent on σ and N, that are positive for $0 < \sigma < 2/(N-2)$. Thus, (2.14) and (2.15) imply

$$(2.16) \qquad (L_+ f, f) \geqq (2 - \sigma N) c_{\sigma N} (\Delta R, f)^2$$

for any f with $(f, R) = 0$. Since the right-hand side of (2.16) is nonnegative for $\sigma \leqq 2/N$ the result follows. □

[3] A second, more general, proof is given in Appendix E (see also [30]).

In what follows we will need the next result, part (b) of which has been proved completely only in dimension $N=1$ for all $\sigma>0$ and in dimension $N=3$ for $0<\sigma\leqq 1$ (see appendix A for the proofs). This suffices for our modulational stability analysis of the ground state for $\sigma\leqq 2/N$ to be complete in dimensions $N=1$ and 3.

PROPOSITION 2.8.

(a) L_- *is a nonnegative self-adjoint operator in* L^2 *with null space* $N(L_-)=\operatorname{span}\{R\}$.

(b) L_+ *is a self-adjoint operator in* L^2 *with null space* $N(L_+)=\operatorname{span}\{R_{x_i}; 1\leqq i\leqq N\}$.

The next two results show that by imposing the additional constraints defining M, $Q^{1/2}$ controls the H^1 norm.

PROPOSITION 2.9. *Let* $\sigma\leqq 2/N$. *There exists a positive constant* $C_{\sigma N}^+$ *such that for any* f *satisfying orthogonality conditions* (2.7a, b) (*and* (2.8a) *when* $\sigma=2/N$)

$$(L_+f,f)\geqq C_{\sigma N}^+(f,f). \tag{2.17}$$

PROPOSITION 2.10. *Let* $\sigma\leqq 2/N$. *There exists a positive constant* $C_{\sigma N}^-$ *such that for any* g *satisfying orthogonality relations* (2.7c, d) (*and* (2.8b) *when* $\sigma=2/N$)

$$(L_-g,g)\geqq C_{\sigma N}^-(g,g). \tag{2.18}$$

Proof of Proposition 2.9. We first consider the case $\sigma<2/N$. Let $\tau\equiv\inf_{\|f\|_2=1}(L_+f,f)$, where f is constrained by (2.7a, b). We will prove $\tau>0$ by showing that the assumption $\tau=0$ leads to a contradiction. This will suffice by Proposition 2.7. We first show that $\tau=0$ implies the minimum is attained in the admissible class. We then can consider an associated Lagrange multiplier problem to deduce $\tau>0$.

Let $\{f_\nu\}$ be a minimizing sequence i.e. $\|f_\nu\|_2=1$, $(L_+f_\nu,f_\nu)\downarrow 0$ and f_ν satisfies (2.7a, b). Then for any $\eta>0$ we can choose f_ν so that

$$0<\int(\nabla f_\nu)^2dx+\int f_\nu^2dx\leqq(2\sigma+1)\int R^{2\sigma}f_\nu^2dx+\eta. \tag{2.19}$$

Since $\|f_\nu\|_2$ is finite, (2.19) implies $\|f_\nu\|_{H^1}$ are uniformly bounded. Thus a subsequence f_ν exists that converges weakly to some H^1 function f_*. By weak convergence f_* satisfies (2.7a, b). We also have $\int R^{2\sigma}f_\nu^2dx\to\int R^{2\sigma}f_*^2dx$ by Hölder's inequality, interpolation, and the uniform decay of R. Thus $f_*\not\equiv 0$, by (2.19) since η is arbitrary.

We now show that the minimum is attained at f_* and $\|f_*\|_2=1$. By Fatou's lemma $\|f_*\|_2\leqq 1$. Suppose $\|f_*\|_2<1$. Then, define $g_*=f_*/\|f_*\|_2$ which is admissible. Let $\zeta\in L^2$, $\|\zeta\|_2=1$. By weak convergence of f_ν to f_*, $(\zeta,\nabla f_*)=\liminf_{\nu\to\infty}(\zeta,\nabla f_\nu)\leqq\liminf_{\nu\to\infty}\|\nabla f_\nu\|_2$. Maximizing over all such ζ, we obtain

$$\|\nabla f_*\|_2\leqq\liminf_{\nu\to\infty}\|\nabla f_\nu\|_2.$$

Since $(R^{2\sigma}f_\nu,f_\nu)\to(R^{2\sigma}f_*,f_*)$, we have

$$(L_+f_*,f_*)\leqq\liminf_{\nu\to\infty}(L_+f_\nu,f_\nu)=0.$$

Hence, $(L_+g_*,g_*)\leqq 0$. By Proposition 2.7, $(L_+g_*,g_*)=0$. Thus we can take $\|f_*\|_2=1$ and the minimum to be attained there.

Since the minimum is attained at an admissible function $f_* \not\equiv 0$, there exists $(f_*, \lambda, \beta, \gamma)$ among the critical points of the Lagrange multiplier problem

$$
(2.20) \qquad \begin{aligned}
&\text{(a)} \quad (L_+ - \lambda)f = \beta R + \gamma \cdot xR, \qquad \beta \in \mathbb{R}, \quad \gamma \in \mathbb{R}^N,\\
&\text{(b)} \quad \|f\|_2 = 1,\\
&\text{(c)} \quad f \text{ satisfies } (2.7\text{a,b}).
\end{aligned}
$$

By (2.20) $\lambda = (L_+ f, f)$, so $\lambda = \tau = 0$ is a critical value. Therefore, we need to conclude that

$$
(2.21) \qquad L_+ f_* = \beta R + \gamma \cdot xR
$$

has no nontrivial solutions (f_*, β, γ) satisfying the side constraints. Taking the inner product of (2.21) with ∇R, integrating by parts, and using that $R_{x_i} \in N(L_+)$, we find $\gamma \equiv 0$. Therefore $f_* = -1/2\beta((1/\sigma)R + x \cdot \nabla R) + \theta \cdot \nabla R$, $\theta \in \mathbb{R}^N$. Here we use that $L_+((1/\sigma)R + x \cdot \nabla R) = -2R$ (by inspection) and part (b) of Proposition 2.8. Now $\theta = 0$ by (2.7b). Also, since $(f_*, R) = -1/2\beta(1/\sigma - N/2)\, \|R\|_2^2$, (2.7a) is violated when $\sigma < 2/N$. Thus $f_* \equiv 0$, a contradiction. We now conclude that $\tau > 0$. This settles the case $\sigma < 2/N$.

In the case $\sigma = 2/N$, we show that $\tau' \equiv \inf_{\|f_2\| = 1}(L_+ f, f)$ where f is constrained by (2.7a, b) and (2.8a) is strictly positive. The proof of part (i) adapts and we conclude that if $\tau' = 0$, then the minimum is attained at an admissible function f_*. We are thus led to the Lagrange multiplier problem

$$
(2.22) \qquad \begin{aligned}
&\text{(a)} \quad L_+ f_* = \beta R + \gamma \cdot xR + \delta |x|^2 R,\\
&\text{(b)} \quad \|f_*\|_2 = 1,\\
&\text{(c)} \quad f_* \text{ satisfies (2.17) and (2.19)}.
\end{aligned}
$$

We now argue as before: (2.22a) implies that $f_* = -1/2\beta((N/2)R + x \cdot \nabla R) - \delta\rho + \theta \cdot \nabla R$ since $L_+((N/2)R + x \cdot \nabla R) = -2R$, $L_+ \nabla R = 0$ and $L_+ \rho = -|x|^2 R$ (see (B.15)). Since $((N/2)R + x \cdot \nabla R, |x|^2 R) = -\int |x|^2 R^2 dx \neq 0$, $(\rho, R) \neq 0$ (see (B.16)), and $(\nabla R, xR) \neq 0$, the constraints (2.7a, b) and (2.8a) imply $f_* \equiv 0$. Hence $\tau' > 0$. This completes the proof of Proposition 2.9. □

Proof of Proposition 2.10. We first consider the case $\sigma < 2/N$. Let $\mu = \inf_{\|g\|_2} = 1(L_- g, g)$, where g is constrained by (2.7a, b). By a proof similar to that in part (i) of Proposition 2.9, if $\mu = 0$, then the minimum is attained at an admissible function $g_* \not\equiv 0$. Since L_- is nonnegative (Proposition 2.8a), $g_* = R/\|R\|_2$. But g_* does not satisfy (2.7d) since $(R, (1/\sigma)R + x \cdot \nabla R) = (1/\sigma - N/2)\|R\|_2^2$, which does not vanish for $\sigma < 2/N$. Thus $g_* \equiv 0$, a contradiction. We conclude $\mu > 0$ in the case $\sigma < 2/N$.

For $\sigma = 2/N$, let $\mu' = \inf_{\|g\|_2 = 1}(L_- g, g)$, where g is constrained by (2.7c, d) and (2.8b). Although $g_* = R/\|R\|_2$ now satisfies (2.7d), by (B.17) it violates (2.8b) since $(R, \rho) = -\int |x|^2 R^2 dx \neq 0$. Hence, $\mu' > 0$. □

Proof of Proposition 2.5. For $\sigma \leqq 2/N$, $(f, g)^t \in M$ implies by Proposition 2.4 and Propositions 2.9–2.10

$$
Q(f,g) \equiv (L_+ f, f) + (L_- g, g) \geqq C^+(f,f) + C^-(g,g) \geqq K_1\left(\|f\|_2^2 + \|g\|_2^2\right).
$$

The lower estimate of (2.9) now follows easily. □

That the space M is "natural" for the linearized evolution (2.4H) is clear from

PROPOSITION 2.11. *Let $\sigma \leqq 2/N$. Then, Ω_t maps M into itself.*

The proof of this proposition is a simple but lengthy computation which we give in appendix C. The idea is to compute the evolution of the $2N+2$ modes defining M, for $\sigma<2/N$ ($2N+4$ modes, for $\sigma=2/N$). These satisfy a simple linear system of ODE's in time ((C.3) or (C.7)). Thus if we assume that W_0 has a vanishing component in M, then so will $W(t)$.

We conclude this section with the following result on the "H^1 control" of Ω_t on M. It is the main analytical tool of §3.

THEOREM 2.12. *Let* $\sigma \leqq 2/N$, *and consider* (2.4) *with* $W_0 \in M$ *and* $G \in M$. *Then,* $W(t) \in M$ *for* $t>0$ *and*

$$K\|W(t)\|_{H^1}^2 \leqq Q(W(t)) \leqq K'\|W(t)\|_{H^1}^2. \tag{2.23}$$

In addition under the above hypothesis, for (2.4H) *we have*

$$K\|W(t)\|_{H^1}^2 \leqq Q(W_0) \leqq K'\|W_0\|_{H^1}^2. \tag{2.24}$$

Proof. Since Ω_t: $M \to M$ (Proposition 2.11) estimate (2.9) holds with $(f,g)^t = W(t)$ for all t (Theorem 2.5). This is precisely (2.23). For (2.4H) $Q(W(t))$ is conserved (Corollary 2.2). Thus (2.24) follows. □

3. Modulational stability. In this section we will prove Theorem 1. Our first aim is to derive the modulation equations referred to in the statement of the theorem. Let $\tau \equiv \int_0^t \lambda^2(\varepsilon s)\,ds$, $\Theta \equiv \lambda(\theta-\theta_0)$. We now let the $2N+2$ parameters vary slowly, i.e. $a^{-1} \equiv \lambda = \lambda(T)$, $\xi_i = \xi_i(T)$, $\theta_{0,i} = \theta_{0,i}(T)$, and $\eta_0 = \eta_0(T)$, $1 \leqq i \leqq N$ where $T \equiv \varepsilon t$. We will determine the slow time evolution of the functions λ, ξ, θ_0, and η_0 that will ensure (1.11). Substitution of (1.10) into (1.8), use of (1.6), and the balance of terms of order ε yields the IVP

$$W_\tau = LW + G, \quad W_0 = 0, \quad W = (u,v)^t. \tag{3.1}$$

We display the source term, $G = G(\varepsilon t) = (f^\varepsilon(\varepsilon t), g^\varepsilon(\varepsilon t))^t$ explicitly:

$$\begin{aligned} &\text{(a)} \quad f^\varepsilon = A^\varepsilon + \operatorname{im} F(R)R, \\ &\text{(b)} \quad g^\varepsilon = B^\varepsilon - \operatorname{re} F(R)R; \end{aligned} \tag{3.2}$$

$$\begin{aligned} &\text{(a)} \quad A^\varepsilon = -2\lambda^{1/\sigma-3}\dot{\lambda}\left(\frac{1}{\sigma}R + \Theta\cdot\nabla R\right) + 2\lambda^{1/\sigma-1}\dot{\theta}_0\cdot\nabla R, \\ &\text{(b)} \quad B^\varepsilon = -2\lambda^{1/\sigma-3}\dot{\xi}\cdot\Theta R + 2(\xi\cdot\dot{\theta}_0 + \dot{\eta}_0)\lambda^{1/\sigma-2}R. \end{aligned} \tag{3.3}$$

Our aim is to control the large τ (or t) behavior of W in a suitable norm. In §2 the natural norm with which to study (3.1) was found to be the "linearized energy" Q (see 2.5). By Theorem 2.12, if $G^\varepsilon \in M$ for $\tau>0$, then

$$K\|W(\tau)\|_{H^1}^2 \leqq Q(W(\tau)). \tag{3.4}$$

By Proposition 2.4, $G^\varepsilon \in M$ for $\tau>0$ implies that f^ε and g^ε satisfy the $2N+2$-orthogonality conditions (2.7). These relations are called the *modulation equations*. After some simplification we find that the conditions (2.7) reduce to the following system of

$2N+2$ ordinary differential equations in time for the $2N+2$ parameters λ, ξ, θ_0, and η_0:

$$(3.5)\qquad \begin{aligned} &(\mathrm{a}) && 2\left(\frac{1}{\sigma}-\frac{N}{2}\right)\|R\|_2^2\dot{\lambda}=\lambda^{-1/\sigma+3}(f^\varepsilon,R), \\ &(\mathrm{b}) && \|R\|_2^2\dot{\theta}_{0,j}=\lambda^{-1/\sigma+1}(f^\varepsilon,\Theta_j R), \qquad 1\leqq j\leqq N, \\ &(\mathrm{c}) && \|R\|_2^2\dot{\xi}_j=\lambda^{-1/\sigma+3}\left(g^\varepsilon,R_{\Theta_j}\right), \qquad 1\leqq j\leqq N, \\ &(\mathrm{d}) && 2\left(\frac{1}{\sigma}-\frac{N}{2}\right)\|R\|_2^2(\xi\cdot\dot{\theta}_0+\dot{\eta}_0)=\lambda^{-1/\sigma+2}\left(g^\varepsilon,\frac{1}{\sigma}R+\Theta\cdot\nabla R\right). \end{aligned}$$

Note that in the critical case $\sigma=2/N$, (3.5) becomes singular. This is a manifestation of the need to incorporate the two additional generalized eigenmodes (recall there are $2N+4$ when $\sigma=2/N$) with which one can constrain the evolution to M. Since we have not found a $(2N+4)$-parameter family of solutions (we only have the $(2N+3)$-parameter family (1.7)) we have not carried this out for $\sigma=2/N$.

We now restate Theorem 1 more precisely and give the proof.

THEOREM 1$'$. *Consider the* IVP (1.8), *where* ϕ^ε *the solution is expanded as in* (1.10). *Suppose the* $2N+2$ *"slow" functions* $\lambda(T)$, $\xi(T)$, $\theta_0(T)$, *and* $\eta_0(T)$ *solve* (3.5) *and are such that* $|f^\varepsilon(\varepsilon t)|$ *and* $|g^\varepsilon(\varepsilon t)|$ *of* (3.2) *are uniformly bounded* (*independently of* ε) *on any t-interval* $0\leqq t\leqq T_0/\varepsilon$. *Then*, (1.11) *holds*.

Proof of Theorem 1$'$. By (3.1), we have

$$W^\varepsilon(\tau)=\Omega_\tau\int_0^\tau \Omega_{-s}G(\varepsilon s)\,ds,$$

where $\Omega_\tau=\exp[\frac{1}{2}L\tau]$ is a unitary group acting in the space M with norm $Q^{1/2}$ (Theorem 2.5). The modulation equations (3.5) imply $G(\varepsilon\tau)\in M$ for all τ corresponding to the t-interval $[0, T_0/\varepsilon]$. Thus, by Theorem 2.12, $W^\varepsilon(\tau)\in M$ and

$$(3.6)\qquad K\|\varepsilon W^\varepsilon(\tau)\|_{H^1}^2\leqq Q\left[\varepsilon\int_0^\tau \Omega_{-s}G(\varepsilon s)\,ds\right].$$

We now want to study the behavior of the right-hand side of (3.6) for fixed time intervals of order $1/\varepsilon$ as $\varepsilon\to 0$. Heuristically, since G is "almost constant" the mean ergodic theorem [9] should imply that $\varepsilon\int_0^\tau \Omega_{-s}G(\varepsilon s)ds$ tends to the projection of G onto the null space of L in the space M. But the null space of L in M is empty since the only possible members, $(0,R)^t$ and $(0,\nabla R)^t$, do not lie in M. Thus we should have $Q(\varepsilon\int_0^\tau \Omega_{-s}G(\varepsilon s)ds)\downarrow 0$ for $0\leqq\tau\leqq 0(1/\varepsilon)$ as $\varepsilon\to 0$. This argument is made rigorous by invoking Lemma D.1. The assertion (1.11) now follows from this and (3.6). □

We now study (1.8) for two specific perturbations. We apply Theorem 1$'$, the hypotheses of which are easily verified.

Example 3.1. Initial data near the ground state.

Consider, for $0<\sigma<2/N$, (1.1) with

$$\phi^\varepsilon(x,0)=\psi\left(x,0;\lambda^{in},0,x_0,\eta_0^{in}\right)+\varepsilon S(x), \qquad S\in H^1.$$

Expand $\phi^\varepsilon(x,t)$ as in (1.10) with w_1, the linearized perturbation, defined by

$$(3.7)\qquad w_1=(u+\mathrm{re}\,S)+i(v+\mathrm{im}\,S).$$

Then, $W=(u,v)^t$ satisfies (3.1) with

$$(3.8)\quad \begin{aligned} &(a)\ \ f^\varepsilon = A^\varepsilon + L_-\operatorname{im} S,\\ &(b)\ \ g^\varepsilon = B^\varepsilon - L_+\operatorname{re} S \end{aligned}$$

and A^ε, B^ε defined in (3.3). The modulation equations (3.5) reduce to

$$(3.9)\quad \begin{aligned} &(a)\ \ \dot\lambda = 0 \Rightarrow \lambda(\varepsilon t) \equiv \lambda^{in},\\ &(b)\ \ \dot\xi_i = 0 \Rightarrow \xi_i \equiv 0; \end{aligned}$$

$$(3.10)\quad \begin{aligned} &(a)\ \ \|R\|_2^2\dot\theta_{0,i} = -2\int \operatorname{im} S(x) R_{\Theta_i}\big(\lambda^{in}(x-\theta_0)\big)\,dx, \qquad 1\leqq i\leqq N,\\ &(b)\ \ (1/\sigma - N/2)\|R\|_2^2\dot\eta_0 = \int \operatorname{re} S(x) R\big(\lambda^{in}(x-\theta_0)\big)\,dx. \end{aligned}$$

Therefore, to leading order, perturbations of ground state initial values induce modulations of the parameters θ_0 and η_0 alone. This is consistent with the nonlinear stability results in [6].

Example 3.2. Simple dissipation.
Consider the IVP (1.8) with

$$(3.11)\quad \begin{aligned} &(a)\ \ F^\varepsilon(z) = -i\kappa,\ \kappa>0\\ &(b)\ \ \phi^\varepsilon(x,0) = \psi\big(x,0;\lambda^{in},0,x_0,\eta_0^{in}\big). \end{aligned}$$

This has been considered in one spatial dimension in [17], [19]. Following the procedure outlined earlier, we obtained the following modulation equations:

$$(3.12)\quad \begin{array}{lll} \dot\lambda = -\kappa\dfrac{\sigma}{2-\sigma N}\lambda, & \lambda(0)=\lambda^{in}, & \\ \dot\theta_{0,i}=0, & \theta_{0,i}(0)=x_{0,i}, & 1\leqq i\leqq N,\\ \dot\xi_i = 0, & \xi_i(0)=0, & 1\leqq i\leqq N,\\ \dot\eta_0 = -\xi\cdot\dot\theta_0, & \eta_0(0)=\eta_0^{in}. & \end{array}$$

These equations are easily solved:

$$(3.13)\quad \begin{array}{ll} \lambda(\varepsilon t) = \lambda^{in}\exp\left[-\kappa\dfrac{\sigma}{2-\sigma N}\varepsilon t\right], & \xi_i(\varepsilon t)\equiv 0,\\ \theta_{0,i}(\varepsilon t)\equiv x_{0,i}, & \eta_0(\varepsilon t)\equiv \eta_0^{in}. \end{array}$$

The Ansatz (1.10) and (3.13) imply that the effect of small simple dissipation is a slow exponential decrease in the amplitude of the ground state profile, accompanied by a simultaneous broadening of the profile due to the changing spatial scale, $\lambda(\varepsilon t)(x-x_0)$.

Appendix A. Partial proof of Proposition 2.8.

Proof of Proposition 2.8a. Since $L_-R=0$ and $R\in L^2$, we have $0\in\sigma(L_-)$. Since $R>0$, it is the ground state. Thus L_- is nonnegative. Finally, $N(L_-)=\operatorname{span}\{R\}$ since the ground state is nondegenerate.

Proof of Proposition 2.8b *for* $N=1$. Since $R''-R+R^{2\sigma+1}=0$ $\big(R(x)=(\sigma+1)^{1/2\sigma}\operatorname{sech}^{1/\sigma}(\sigma x)>0\big)$, $R'\in N(L_+)$. Define $w=R'/R''(0)$. Then $w(0)=0$ and $w'(0)=1$. Let v be the solution of $L_+v=0$ with $v(0)=1$ and $v'(0)=0$. We then have the

Wronskian relation $w(x)v'(x)-w'(x)v(x)=-1$. Therefore, $R'(x)v'(x)-R''(x)v(x)=-R''(0)=\sigma(\sigma+1)^{1/2\sigma}$. We then have that $(v/R')'=\sigma(\sigma+1)^{1/2\sigma}/(R')^2$. To prove that $v\notin N(L_+)$ and therefore that $N(L_+)=\operatorname{span}\{R'\}$ it will suffice to show that $|v|\to\infty$ as $x\to\infty$. For $\varepsilon>0$ and $x>0$,

$$v(x)=\frac{v(\varepsilon)}{R'(\varepsilon)}R'(x)+\int_\varepsilon^x\frac{ds}{[R'(s)]^2}\sigma(\sigma+1)^{1/2\sigma}R'(x).$$

Since $R'(x)=O(e^{-x})$ as $x\to\infty$, i.e., $|R'(x)|\leqq Ce^{-x}$, and $R'<0$ for $x>0$,

$$v(x)\leqq v(\varepsilon)\frac{R'(x)}{R'(\varepsilon)}+\int_\varepsilon^x\frac{ds}{C^2e^{-2s}}R'(x)\to-\infty\quad\text{as } x\to\infty.$$

Proof of Proposition 2.8b *for* $N=3$, $0<\sigma\leqq1$ *and partial proof in all other* $(N>1)$ *cases*. Since the equation $\Delta R-R+R^{2\sigma+1}=0$ is translation invariant in space, $\nabla R=R'(|x|)x/|x|\in N(L^+)$. Now L_+ is an operator with a potential dependent only on $r=|x|$. Hence an L^2 solution of $L_+v=0$ can be expanded in a series with functions of the form $f(r)Y(\theta)$, where $f\in L^2(0,\infty;r^{N-1}dr)$ and $Y\in L^2(S^{N-1})$. These functions must satisfy

$$\text{(A.1)}\qquad A_kf\equiv\left(-\frac{d^2}{dr^2}-\frac{N-1}{r}\frac{d}{dr}+1-(2\sigma+1)R^{2\sigma}+\frac{\lambda_k}{r^2}\right)f=0,$$

$$\text{(A.2)}\qquad -\Delta_{S^{N-1}}Y=\lambda_kY,$$

where $\lambda_k=k(N-2+k)$.

The null functions ∇R correspond to $k=1$. Therefore to prove the theorem it will suffice to show that there is no solution of (A.1) satisfying the correct boundary conditions at $r=0$ and $r=\infty$ for $k\in\{0,2,3,4,\cdots\}$.

We handle the case $k\geqq2$ as follows. The function R' is an eigenfunction of A_1 corresponding to the eigenvalue zero. Since R' has no interior zeros it is the ground state of A_1. Hence A_1 is a nonnegative operator. Setting $k\equiv1+\delta$, $\delta>0$, we find $A_k=A_1+\delta(N+\delta)r^{-2}$ which is a positive operator. Hence $f\in L^2$ and $A_kf=0$ for $k\geqq2$ implies $f\equiv0$.

The case $k=0$ has not yet been completely resolved. Coffman, in proving the uniqueness of the positive decaying solution of $R''+(2/r)R'-R+R^3=0$ $(\sigma=1,N=3)$, first shows that $|(\partial u/\partial r)(r,\alpha_0)|\to\infty$ as $r\to\infty$, where $u(r,\alpha)$ is the solution of the IVP

$$u''+\frac{2}{r}u'-u+u^3=0,\qquad u(0,\alpha)=\alpha$$

and α_0 is the initial value generating the ground state $u(r,\alpha_0)=R(r)$ [8, Lemma 4.2]. This lemma holds as well, with minor changes in the proof, for the range $0<\sigma\leqq1$, $N=3$.[4] This eliminates the $k=0$ mode in the case $0<\sigma\leqq1$, which includes the subcritical and critical cases in dimension three. (More general results on the uniqueness of the ground state of (1.4) were obtained by McLeod and Serrin [21].) For general σ and N, since the problem is reduced to consideration of the ODE (A.1) with $\lambda_k=0$, we can say that $\dim N(L_+)$ is either N or $N+1$.

[4] C. V. Coffman, personal communication.

Appendix B. Generalized null spaces. In this appendix we display explicitly the generalized null spaces of L and L^*, denoted by $N_g(L)$ and $N_g(L^*)$. The results of this section depend on Proposition 2.8b which has been completely proved in dimension one for all σ, and in dimension three for $0<\sigma\leqq 1$.

We begin with the following useful observation which follows by direct verification, using (1.4).

PROPOSITION B.1.

$$\text{(B.1)}\quad \begin{aligned} &\text{(a)}\ L_-R=0, && \text{(c)}\ L_+\nabla R=0,\\ &\text{(b)}\ L_-xR=-2\nabla R, && \text{(d)}\ L_+\left(\frac{1}{\sigma}R+x\cdot\nabla R\right)=-2R. \end{aligned}$$

We are interested in the null spaces of L, L^2, $L^3,\cdots$. For example, the equation $LW=0$, where $W=(u,v)^t$ reduces to $L_-v=0$ and $L_+u=0$. By Proposition 2.9, $(0,R)^t$ and $(\nabla R,0)^t$ span $N(L)$. The equation $L^2W=0$ implies $L_-L_+u=0$ and $L_+L_-v=0$. By Proposition B.1 we find that $N(L^2)-N(L)$ is spanned by $((1/\sigma)R+x\cdot\nabla R,0)^t$ and $(0,xR)^t$.

Continuing, we consider $L^3W=0$ which implies

$$\text{(B.2)}\quad \begin{aligned} &\text{(a)}\ L_-L_+L_-v=0,\\ &\text{(b)}\ L_+L_-L_+u=0. \end{aligned}$$

We seek to construct functions in $N(L^3)-N(L^2)$. Equation (B.2a) implies $L_+L_-v=cR$, c constant. By Proposition B.1,

$$\text{(B.3)}\qquad L_-v=-\frac{c}{2}\left(\frac{1}{\sigma}R+x\cdot\nabla R\right).$$

Now (B.3) will have an H^1 solution if the following solvability condition holds:

$$\text{(B.4)}\qquad 0=\left(R,\frac{1}{\sigma}R+x\cdot\nabla R\right)=\left(\frac{1}{\sigma}-\frac{N}{2}\right)\|R\|_2^2.$$

Thus for $\sigma\neq 2/N$, (B.3) has no solutions giving rise to an element of $N(L^3)-N(L^2)$.

Consider now (B.2b). This implies

$$\text{(B.5)}\qquad L_-L_+u=c\cdot\nabla R,\qquad c\in\mathbb{R}^N.$$

By (B.1b)

$$\text{(B.6)}\qquad L_+u=-\frac{c}{2}\cdot xR.$$

Since $(\nabla R,xR)\neq 0$, (B.6) has no solution generating an element of $N(L^3)-N(L^2)$.

It follows that for $\sigma\neq 2/N$, $N(L^3)=N(L^2)$. Similarly, for $\sigma\neq 2/N$ $N(L^k)=N(L^2)$, $k\geqq 3$. The same procedure can be followed to deduce $N_g(L^*)$. We summarize these observations.

THEOREM B.2. *Let $\sigma \neq 2/N$. $N_g(L) = \cup_{j=1}^{2} N(L^j)$ and $N_g(L^*) = \cup_{j=1}^{2}[N(L^*)^j]$ are spanned by the following two $2N+2$-dimensional biorthogonal sets*:

$$\text{(B.7)} \quad \begin{aligned} &\text{(a)} \quad a_1 = \alpha_1^{-1} = (0, -R)^t, \\ &\text{(b)} \quad a_{2,j} = \alpha_2^{-1}(-R_{x_j}, 0)^t, \qquad 1 \leqq j \leqq N, \\ &\text{(c)} \quad a_{3,j} = \alpha_2^{-1}(0, x_j R)^t, \qquad 1 \leqq j \leqq N, \\ &\text{(d)} \quad a_4 = \alpha_1^{-1}\left(\frac{1}{\sigma}R + x \cdot \nabla R, 0\right)^t, \quad \textit{where} \\ &\qquad \alpha_1 = \left(\frac{N}{2} - \frac{1}{\sigma}\right)\|R\|_2^2, \qquad \alpha_2 = \frac{1}{2}\|R\|_2^2. \end{aligned}$$

$$\text{(B.8)} \quad \begin{aligned} &\text{(a)} \quad b_1 = \left(0, \frac{1}{\sigma}R + x \cdot \nabla R\right)^t, \\ &\text{(b)} \quad b_{2,j} = (x_j R, 0)^t, \qquad 1 \leqq j \leqq N, \\ &\text{(c)} \quad b_{3,j} = (0, -R_{x_j})^t, \qquad 1 \leqq j \leqq N, \\ &\text{(d)} \quad b_4 = (-R, 0)^t, \end{aligned}$$

where

$$\text{(B.9)} \qquad (a_i, b_k) = \delta_{ik} \quad \textit{and} \quad (a_{i,m}, b_{k,l}) = \delta_{im}\delta_{mk}\delta_{kl}.$$

In the critical case, $\sigma = 2/N$, (B.4) implies that (B.3) has a solution generating an element of $N(L^3) - N(L^2)$. In fact (B.3) can be solved explicitly since

$$\text{(B.10)} \qquad L_-|x|^2 R = -4\left(\frac{N}{2}R + x \cdot \nabla R\right).$$

Consider now the equation $L^4 W = 0$, when $\sigma = 2/N$. This implies

$$\text{(B.11)} \quad \begin{aligned} &\text{(a)} \quad L_-L_+L_-L_+u = 0, \\ &\text{(b)} \quad L_+L_-L_+L_-v = 0. \end{aligned}$$

We seek a solution generating an element of $N(L^4) - N(L^3)$. (B.11a) and (B.1a) imply

$$\text{(B.12)} \qquad L_+L_-L_+u = cR, \qquad c \text{ constant}.$$

(B.1d) implies

$$\text{(B.13)} \qquad L_-L_+u = -\frac{c}{2}\left(\frac{N}{2}R + x \cdot \nabla R\right).$$

(B.10) implies

$$\text{(B.14)} \qquad L_+u = \frac{c}{8}|x|^2 R.$$

(B.14) has a solution since the inhomogeneous term $|x|^2 R$ satisfies the solvability condition $(|x|^2 R, \nabla R) = 0$. We define a particular solution of (B.12) $u = -(c/8)\rho$, where ρ is the unique radial and even solution of

$$\text{(B.15)} \qquad L_+\rho = -|x|^2 R.$$

Thus $(\rho, 0)^t \in N(L^4) - N(L^3)$.

It is easily seen that (B.11b) has no solution giving rise to an element of $N(L^4)-N(L^3)$. Similarly it can be checked that $N(L^k)=N(L^4)$ $k\geqq 5$, when $\sigma=2/N$. $N_g(L^*)$ is similarly deducible.

We summarize the structure of the generalized null spaces when $\sigma=2/N$ in

THEOREM B.3. *Let* $\sigma=2/N$. $N_g(L)=\cup_{j=1}^{4}N(L^j)$ *and* $N_g(L^*)=\cup_{j=1}^{4}N[(L^*)^j]$ *are spanned by the following* $2N+4$*-dimensional biorthogonal sets*:

$$\text{(B.16)}\quad \begin{aligned}
&\text{(a)}\quad n_1=\beta_1^{-1}(0,-R)^t,\\
&\text{(b)}\quad n_{2,j}=\beta_2^{-1}(-R_{x_j},0)^t,\qquad 1\leqq j\leqq N,\\
&\text{(c)}\quad n_{3,j}=\beta_2^{-1}(0,x_jR)^t,\qquad 1\leqq j\leqq N,\\
&\text{(d)}\quad n_4=\beta_1^{-1}\left(\frac{N}{2}R+x\cdot\nabla R,0\right)^t,\\
&\text{(e)}\quad n_5=\beta_1^{-1}\left(0,|x|^2R\right)^t+\gamma_1(0,R)^t,\\
&\text{(f)}\quad n_6=\beta_1^{-1}(\rho,0)^t,\ \textit{where}
\end{aligned}$$

$$\text{(B.17)}\quad \begin{aligned}
&\beta_1=-(R,\rho)=\frac{1}{2}\left(|x|^2R,R\right),\qquad \beta_2=\frac{1}{2}\|R\|_2^2,\\
&\gamma_1=\alpha_1^{-2}\left(|x|^2R,\rho\right),\qquad \gamma_2=(2\beta_1)^{-1}\left(|x|^2R,\rho\right).
\end{aligned}$$

$$\text{(B.18)}\quad \begin{aligned}
&\text{(a)}\quad m_1=(0,\rho)^t,\\
&\text{(b)}\quad m_{2,j}=(x_jR,0)^t,\qquad 1\leqq j\leqq N,\\
&\text{(c)}\quad m_{3,j}=\left(0,-R_{x_j}\right)^t,\qquad 1\leqq j\leqq N,\\
&\text{(d)}\quad m_4=\left(-\frac{1}{2}|x|^2R,0\right)^t+\gamma_2(-R,0)^t,\\
&\text{(e)}\quad m_5=\left(0,\frac{N}{2}R+x\cdot\nabla R\right)^t,\\
&\text{(f)}\quad m_6=(-R,0)^t.
\end{aligned}$$

Appendix C. The secular evolution. We set $S\equiv N_g(L)$ for $\sigma\leqq 2/N$. Recalling Definition 2.3 of M, we have by the biorthonormality of $N_g(L^*)$ and $N_g(L)$ the following:

PROPOSITION C.1. *For* $\sigma\leqq 2/N$, $H^1\times H^1\cong M\oplus S$.

Proposition 2.11 stated that M is mapped to itself by Ω_t. The following result describes the evolution in the complementing space S.

THEOREM C.2. *Consider* (2.4) *with* $W_0\equiv(u_0,v_0)^t\in S$ *and* $G\equiv(f,g)^t\in S$. *Then*, $W(t)=(u(t),v(t))^t\in S$ *and has the following form.*

(a) *For* $\sigma<2/N$

$$\text{(C.1)}\qquad W(t)=\sum_{j=1}^{4}\mu_j(t)\cdot a_j,\quad \textit{where}$$

$$\text{(C.2)}\qquad \mu_j(t)=\left(b_j,W(t)\right),\qquad 1\leqq j\leqq 4.$$

The functions μ_j *satisfy the system of* ODE's

$$\text{(C.3)}\qquad \begin{aligned} &\text{(a)}\ \ 2\dot{\mu}_1(t)=-2\mu_4(t)+c_1, && \\ &\text{(b)}\ \ 2\dot{\mu}_{2,k}(t)=2\mu_{3,k}(t)+c_{2,k}, && 1\leqq k\leqq N, \\ &\text{(c)}\ \ 2\dot{\mu}_{3,k}(t)=c_{3,k}, && 1\leqq k\leqq N, \\ &\text{(d)}\ \ 2\dot{\mu}_4(t)=c_4, && \textit{where} \end{aligned}$$

$$\text{(C.4)}\qquad c_j=(b_j,G).$$

(b) *For* $\sigma=2/N$

$$\text{(C.5)}\qquad W(t)=\sum_{j=1}^{6}\nu_j(t)n_j,\quad \textit{where}$$

$$\text{(C.6)}\qquad \nu_j(t)=(m_j,W(t)),\qquad 1\leqq j\leqq 6.$$

The functions $\nu_j(t)$ *satisfy the system of* ODE's

$$\text{(C.7)}\qquad \begin{aligned} &\text{(a)}\ \ 2\dot{\nu}_1(t)=-2\nu_4(t)+\beta_1\gamma_1\nu_6(t)+d_1, \\ &\text{(b)}\ \ 2\dot{\nu}_2(t)=2\nu_3(t)+d_2, \\ &\text{(c)}\ \ 2\dot{\nu}_3(t)=d_3, \\ &\text{(d)}\ \ 2\dot{\nu}_4(t)=-4\nu_5(t)+d_4, \\ &\text{(e)}\ \ 2\dot{\nu}_5(t)=\nu_6(t)+d_5, \\ &\text{(f)}\ \ 2\dot{\nu}_6(t)=d_6,\ \ \textit{where} \end{aligned}$$

$$\text{(C.8)}\qquad d_j=(m_j,G).$$

Proof. We substitute the expansion (C.1) ((C.5) when $\sigma=2/\text{N}$) into (2.4) and equate coefficients of like modes. This yields system (C.3) ((C.7) when $\sigma=2/N$). □

Appendix D. A mean ergodic theorem for slowly varying functions. To prove Theorem 1′ we use the following lemma:

LEMMA D.1. *Let* L *be a skew-adjoint operator on a separable Hilbert space* H, *and* $\exp(Ls)$ *be a corresponding unitary group of transformations. Let* $J^\varepsilon=J(\varepsilon s)$ *be a continuous* H*-valued function with* H*-norm bounded, independently of* ε, *for* $0\leqq s\leqq T/\varepsilon$, *where* $T>0$ *is fixed. Furthermore, assume*

$$J(\varepsilon s)\in N^{\perp}(L)\quad \textit{for}\ 0\leqq s\leqq T/\varepsilon,\quad \varepsilon>0.$$

Then,

$$\text{(D.1)}\qquad \lim_{\varepsilon\to 0}\frac{\varepsilon}{T}\left\|\int_0^{T/\varepsilon}e^{Ls}J(\varepsilon s)\,ds\right\|_H=0.$$

We prove (D.1) in the form

$$\text{(D.2)}\qquad \lim_{\varepsilon\to 0}\frac{1}{T}\left\|\int_0^{T}e^{Ls/\varepsilon}f(s)\,ds\right\|_H=0$$

if $f(s)\in N^{\perp}(L)$ for $0\leqq s\leqq T$.

Let $\{\phi_j\}$ be a countable orthonormal basis of H. Let $\delta>0$ be arbitrary. Choose $n=n(\delta,T)$ so that

$$\sup_{0\leqq t\leqq T}\|f(t)-f^n(t)\|_H<\delta/2, \tag{D.3}$$

where $f^n(t)=\sum_{j=1}^n c_j(t)\phi_j$ and $c_j(t)=(\phi_j,f(t))$. For every j, $1\leqq j\leqq n$, we select $k_j(t)$, a piecewise constant approximation to $c_j(t)$ such that

$$\sup_{0\leqq t\leqq T}|c_j(t)-k_j(t)|<\frac{\delta}{2n}. \tag{D.4}$$

We then have

$$\begin{aligned}\int_0^T e^{Ls/\varepsilon}f(s)\,ds&=\int_0^T e^{Ls/\varepsilon}[f(s)-f^n(s)]\,ds\\&+\int_0^T e^{Ls/\varepsilon}\sum_{j=1}^n\left[c_j(s)-k_j(s)\right]\phi_j ds\\&+\int_0^T e^{Ls/\varepsilon}\sum_{j=1}^n k_j(s)\phi_j ds.\end{aligned} \tag{D.5}$$

Estimating in H,

$$\begin{aligned}\left\|\int_0^T e^{Ls/\varepsilon}f(s)\,ds\right\|_H&\leqq\frac{T\delta}{2}+Tn\frac{\delta}{2n}+\left\|\int_0^T e^{Ls/\varepsilon}\sum_{j=1}^n k_j(s)\phi_j ds\right\|_H\\&=T\delta+\left\|\int_0^T e^{Ls/\varepsilon}\sum_{j=1}^n\sum_{m=1}^{p_j}\chi_{j,m}k_{j,m}\phi_j ds\right\|_H\\&\leqq T\delta+\sum_{j=1}^n\sum_{m=1}^{p_j}\left\|\int_0^T e^{Ls/\varepsilon}\chi_{j,m}k_{j,m}\phi_j ds\right\|_H.\end{aligned} \tag{D.6}$$

Here, $k_j(t)=\sum_{m=1}^{p_j}\chi_{j,m}k_{j,m}$ is a piecewise constant approximation of $c_j(t)$ on $[0,T]$, where $\{\chi_{j,m}\}$ are characteristic functions of subintervals $\{I_{j,m}\}$ that can be chosen to be uniformly distributed in $[0,T]$. We note that the double-sum is finite and independent of ε. Since $f(s)\varepsilon N^{\perp}(L)$ for $0\leqq s\leqq T$, $\phi_j\varepsilon N^{\perp}(L)$. By the mean ergodic theorem [9], each term in this double-sum tends to zero as $\varepsilon\to 0$. Thus,

$$\frac{1}{T}\left\|\int_0^T e^{Ls/\varepsilon}f(s)\,ds\right\|_H\to\delta\quad\text{as }\varepsilon\to 0. \tag{D.7}$$

Since δ was arbitrary, the proof is complete.

Appendix E. A second proof of Proposition 2.7. This proof is based on the following general lemma.

LEMMA E.1. *Let A be a self-adjoint having exactly one negative eigenvalue, λ_0 with corresponding ground state eigenfunction $f_0\geqq 0$. Define*

$$-\infty<\alpha\equiv\min_f(Af,f),\quad\textit{where }\|f\|_2=1\textit{ and }(f,R)=0. \tag{E.1}$$

We assume $(R,f_0)\neq 0$ *and* $R\in N^{\perp}(A)$. *Then* $\alpha\geqq 0$ *if*

$$(A^{-1}R,R)\leqq 0. \tag{E.2}$$

Proof of Lemma E.1. If α is attained for the function f_*, then by the theory of Lagrange multipliers there is a triple (f_*,λ_*,β_*) satisfying

$$Af_*=\lambda_*f_*+\beta_*R,\quad \|f_*\|_2=1,\quad (f_*,R)=0. \tag{E.3}$$

Taking the inner product of (E.3) with f_* we get

$$\lambda_*=(Af_*,f_*). \tag{E.4}$$

Therefore, to prove that $\alpha\geq 0$ it suffices to preclude $\lambda_*\leqq 0$. First, if $\lambda_*=\lambda_0$, then taking the inner product of (E.3) with f_0 we conclude that either $\beta_*=0$ or $(R,f_0)=0$. Neither is possible since $(R,f_0)\neq 0$. If $\lambda\in(\lambda_0,0]$, we get from (E.3)

$$f_*=\beta_*(A-\lambda_*)^{-1}R. \tag{E.5}$$

Now λ_* is a critical point if

$$g(\lambda_*)=0, \tag{E.6}$$

where

$$g(\lambda)=\left((A-\lambda)^{-1}R,R\right). \tag{E.7}$$

Note that

$$g'(\lambda)=\left\|(A-\lambda)^{-1}R\right\|_2^2 \tag{E.8}$$

since A is self-adjoint. Therefore, g is increasing on $(\lambda_0,0]$. Moreover,

$$g(0)=(A^{-1}R,R). \tag{E.9}$$

It follows that if (E.2) holds, then $g(\lambda_*)\neq 0$ in $(\lambda_0,0)$. This proves the lemma.

Proof II *of Proposition* 2.7. We identify A with L_+. That L_+ has exactly one negative eigenvalue can be seen as follows (see also [30]):

Case (i) $N=1$: $R'\in N(L_+)$ has exactly one node at $x=0$ implying, by ODE oscillation theorems, that 0 is the second eigenvalue.

Case (ii) $N\geqq 2$, $\sigma<2/(N-2)$: Since the coefficient of the second term in (2.15) is nonpositive, we have that

$$\tilde{T}=L_++\left[(\sigma N-2)c_{\sigma N}(\Delta R,\cdot)-b_{\sigma N}(R,\cdot)\right]\Delta R \tag{E.10}$$

is a nonnegative operator. Therefore, L_+ is a rank one perturbation of a nonnegative operator. By the min-max principle, L_+ has at most one negative eigenvalue. Since $\nabla R\in N(L_+)$ is not positive it is not the ground state, implying that L_+ has exactly one negative eigenvalue.

As was remarked at the beginning of Proof I, we have that the infimum in (2.12) is nonpositive. By techniques used to prove Proposition 2.9, the infimum in (2.12) is actually a minimum. Therefore, L_+ and R satisfy the hypotheses of Lemma E.1 and it suffices to calculate $(L_+^{-1}R,R)$. By (B.1),

$$\left(L_+^{-1}R,R\right)=\left(\frac{N}{2}-\frac{1}{\sigma}\right)\|R\|_2^2. \tag{E.11}$$

This is nonpositive if $\sigma \leqq 2/N$, Proposition 2.7 now follows from Lemma E.1 □

We note that in the supercritical case $\sigma > 2/N$, the infimum in (2.12) will be negative since $(L_+^{-1}R, R) > 0$. This suggests modulational instability of the ground state for $\sigma > 2/N$.

Acknowledgments. I would like to thank Professors George C. Papanicolaou and David W. McLaughlin for their active interest and involvement in my work. I also wish to thank Professor Joseph B. Keller for many helpful criticisms and suggestions made during the preparation of this article.

Note added in proof. The author has proved, using the results presented here on the linearized NLS operator, nonlinear Lyapunov stability of ground states relative to small perturbations in initial data: *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, to appear.

REFERENCES

[1] S. A. Akhmanov, A. P. Sukhorukov and R. V. Khokhlov, *Self-focusing and self-trapping of intense light beams in a nonlinear medium*, Sov. Phys. JETP, 23 (1966), pp. 1025–1033.

[2] T. B. Benjamin, *The stability of solitary waves*, Proc. Roy. Soc. London A, 328 (1972), pp. 153–183.

[3] J. Bona, *On the stability of solitary waves*, Proc. Roy. Soc. London A, 344 (1975), pp. 363–374.

[4] H. Berestycki and P. L. Lions, *Existence of stationary states in nonlinear scalar field equations*, in Bifurcation Phenomena in Mathematical Physics and Related Topics, C. Bardos and D. Bessis, eds., D. Reidel Publ. Co. Dordrecht, Boston, London, 1976, pp. 269–292.

[5] T. Cazenave, *Stable solutions of the logarithmic Schrödinger equation*, preprint.

[6] T. Cazenave and P. L. Lions, *Orbital stability of standing waves for some nonlinear Schrödinger equations*, Comm. Math. Phys., 85 (1982), pp. 549–561.

[7] R. Y. Chiao, E. Garmire and C. H. Townes, *Self-trapping of optical beams*, Phys. Rev. Lett., 13 (1964), pp. 479–482.

[8] C. V. Coffman, *Uniqueness of the ground state solution for $\Delta u - u + u^3 = 0$ and a variational characterization of other solutions*, Arch. Rat. Mech. Anal., 46 (1972), pp. 81–95.

[9] N. Dunford and J. T. Schwartz, *Linear Operators. Part* I: *General Theory*, Wiley-Interscience, New York, 1958.

[10] D. B. Henry, J. F. Perez and W. F. Wreszinski, *Stability theory for solitary-wave solutions of scalar field equations*, Comm. Math. Phys., 85 (1982), pp. 351–361.

[11] E. Gagliardo, *Proprieta di alcune classi di funzioni in piu varibili*, Ricerche di Math. 7 (1958), pp. 102–137.

[12] ———, *Ulteriori proprieta di alcune classi di funzioni in piu variabili*, Ricerche di Math., 8 (1959), pp. 24–51.

[13] J. Ginibre and G. Velo, *On a class of nonlinear Schrödinger equations* I. *The Cauchy problem, general case*, J. Func. Anal., 32 (1979), pp. 33–71.

[14] R. T. Glassey, *On the blowing-up of solutions to the Cauchy problem for the nonlinear Schrödinger equations*, J. Math. Phys., 18 (1977), pp. 1794–1797.

[15] H. Hasimoto, *A soliton on vortex filament*, J. Fluid Mech., 51 (1972), pp. 477–485.

[16] D. J. Kaup, *A perturbation expansion for the Zakharov–Shabat inverse scattering transform*, SIAM J. Appl. Math., 31 (1976), pp. 121–133.

[17] J. P. Keener, and D. W. McLaughlin, *Solitons under perturbations*, Phys. Rev., A. 16 (1977), pp. 777–790.

[18] P. L. Kelley, *Self-focusing of optical beams*, Phys. Rev. Lett., 15 (1965), pp. 1005–1008.

[19] Y. Kodama and M. J. Ablowitz, *Perturbations of solitons and solitary waves*, Studies in Appl. Math., 64 (1981), pp. 225–245.

[20] D. W. McLaughlin, G. C. Papanicolaou and M. I. Weinstein, *Focusing and saturation of nonlinear beams*, in preparation.

[21] K. McLeod and J. Serrin, *Uniqueness of solutions of semilinear Poisson equations*, Proc. Nat. Acad. Sci., USA 78 #11 (1981), pp. 6592–6595.

[22] A. C. Newell, *Near-integrable systems, nonlinear tunneling and solitons in slowly changing media*, in Nonlinear Evolution Equations Solvable by the Inverse Spectral Transform, F. Calogero, ed., London, Pitman, 1978, pp. 127–179.

[23] L. Nirenberg, *Remarks on strongly elliptic partial differential equations*, Comm. Pure Appl. Math., 8 (1955), pp. 648–674.

[24] W. A. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149–162.

[25] C. Sulem, P. L. Sulem and H. Frisch, *Tracing complex singularities with spectral methods*, J. Comp. Phys., 50 (1983), pp. 138–161.

[26] P. L. Sulem, C. Sulem and A. Patera, *Numerical simulation of singular solutions to the two-dimensional nonlinear Schrödinger equation*, Comm. Pure Appl. Math., in press.

[27] V. I. Talanov, *Self-focusing of wave beams in nonlinear media*, JETP Lett., 2 (1965), p. 138.

[28] M. Tsutsumi, *Nonexistence of global solutions to nonlinear Schrödinger equations*, unpublished manuscript.

[29] M. I. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates*, Comm. Math. Phys., 87 (1983), pp. 567–576.

[30] ———, *Self-focusing and modulational analysis of nonlinear Schrödinger equations*, Ph. D. thesis, New York Univ., New York, 1982.

[31] V. E. Zakharov, *Collapse of Langmuir Waves*, Sov. Phys. JETP, 35 (1972), pp. 908–922.

[32] V. E. Zakharov and Yu. S. Sigov, *Strong turbulence and its computer simulation*, Journal de Physique, 40 (1979), pp. 7–63.

[33] V. E. Zakharov and V. S. Synakh, *The nature of the self-focusing singularity*, Sov. Phys. JETP, 41 (1976), pp. 465–468.

SIAM J. MATH. ANAL.
Vo. 16, No. 3, May 1985

CONVERGENCE OF ESSENTIAL SPECTRA FOR INTERMEDIATE HAMILTONIANS*

CHRISTOPHER BEATTIE[†]

Abstract. A variant of an intermediate Hamiltonian construction of Fox is shown to give convergent estimates from below to the lowest point of the essential spectrum for a large class of multielectron atomic Hamiltonians. An HVZ-type theorem is presented for such intermediate Hamiltonians and convergent lower bounds to all lower eigenvalues are shown to be attainable with this construction.

1. Introduction. An important theoretical and practical problem in quantum mechanics is the accurate estimation of Schrödinger operator eigenvalues. The task of assessing the accuracy of an estimate to any given eigenvalue of an atomic Hamiltonian, H, is tantamount to finding rigorous upper and lower bounds to that eigenvalue. Generally speaking, lower bound estimation requires more computational effort and the related analysis is commonly more subtle than that encountered in complimentary upper bound estimation. Due to varying needs in a priori information there are, in fact, relatively few lower bound procedures that are widely applicable, so selection of an appropriate method is quite sensitive to problem setting. We will focus our attention on intermediate operator methods.

Intermediate operator methods generally require a decomposition of the operator H as $H=H^0+\hat{H}$, where spectral information on the lower spectrum (i.e., spectrum *below* σ_{ess}) of the self-adjoint operator H^0 is explicitly known and $\hat{H}$ is some positive-definite symmetric operator. Then $H^0 \leq H$ in the usual sense of ordering for symmetric operators (cf. [15]) and the (known) lower eigenvalues of H^0 provide a priori lower bounds to the corresponding eigenvalues of H with a similar relationship holding for the lowest points of their respective essential spectra.

The lower bounds provided by H^0 generally tend to be quite crude and in order to improve them in such a way that they remain lower bounds, one must carefully approximate $\hat{H}$ from below. Aronszajn ([1]) developed a systematic procedure for doing this with an increasing chain of positive-semidefinite finite-rank operators and detailed the related theory necessary to resolve the spectral problem for what may then be viewed as a finite-rank perturbation of a spectrally resolvable operator. The long collaboration of Bazley and Fox (see e.g. [3], [4] and [5]) developed some of the tools introduced by Aronszajn into practical computational strategies that found application in many areas of fluid dynamics, mechanics and quantum theory.

Although these classical intermediate operator methods have had some success in application to Schrödinger operators (see [2] and [4]) one finds severe limitations imposed for complex atoms due to the stability of essential spectra under finite-rank perturbations. For atoms or ions larger then helium, part of the essential spectrum of the resolvable base operator H^0 generally lies below the lowest eigenvalue for the full operator H, rendering the usual finite-rank Aronszajn method useless in obtaining lower bounds to the level of accuracy required. Fox [10] discovered a modification of

* Received by the editors September 20, 1983, and in revised form December 16, 1983.

[†] Department of Mathematics, University of Arizona, Tucson, Arizona 85721. Current address. Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

Aronszajn's original method utilizing infinite-rank perturbations of the base operator yet retaining the critical property of producing computationally resolvable intermediate operators. The effectiveness of this technique was demonstrated by Fox and Sigillito ([11], [12]) in computing rigorous lower bounds to the lowest point of the essential spectrum and the first three eigenvalues of a simplified Hamiltonian related to lithium. Later computations by Reid ([13] and [14]) on a realistic Hamiltonian for lithium were unable to raise the intermediate essential spectrum sufficiently to make the eigenvalues of interest accessible to tight lower bounds and attested to the dangers of simplifying the computational procedure by ignoring difficult-to-use a priori spectral information. Building on Reid's experience, Freund ([9]) was able to carry out a computation demonstrating sufficient displacement of essential spectrum by incorporating more spectral information, however numerical instabilities have delayed the final acquisition of tight lower bounds to lithium bound state energies. The whole history of these efforts has underscored the computational sophistication required to successfully carry out Fox's method and has raised questions about the kind of convergence possible.

Various convergence results have been obtained recently for Fox's construction of intermediate Hamiltonians that are contingent upon the intermediate essential spectra moving up sufficiently to expose eigenvalues to convergent estimates ([6], [8]). What we are able to show here is that in the case of multiparticle Hamiltonians with Kato potentials, these convergence results in fact guarantee the convergence of the lowest point of the essential spectra of the intermediate Hamiltonians to the lowest point of the essential spectrum of the original Hamiltonian. A result like this appears in [6] for the special case of Coulombic potentials. Such a result is somewhat novel, since previously known convergence results for intermediate operators have been solely devoted to the convergence of isolated eigenvalues of finite multiplicity. The principal application of our result is to show that Fox's construction can be done in such a way as to guarantee convergent estimates to all the lower eigenvalues of a wide class of Schrödinger operators.

2. The Hamiltonian. We consider a model of an atomic system with m identical particles interacting with each other and with a fixed nucleus. With an appropriate choice of units the Hamiltonian is

$$H=\sum_{k=1}^{m}\left[-\frac{1}{2}\Delta_k+W(\mathbf{r}_k)\right]+\sum_{1<j}^{m}V(\mathbf{r}_i-\mathbf{r}_j)$$

operating on a domain dense in $L^2(\mathbb{R}^{3m})$. In this notation, $\{\Delta_k\}$ are Laplacians acting in copies of $L^2(\mathbb{R}^3)$, $\{\mathbf{r}_k\}$ are position vectors in $\mathbb{R}^3$, and both W and V are potential functions representing particle-nucleus and particle-particle forces, respectively.

The following three assumptions are fundamental to the construction of convergent eigenvalue estimates from below:

a) $V, W \in L^2(\mathbb{R}^3)+[L^\infty(\mathbb{R}^3)]_\varepsilon$.
b) $V \geq 0$ a.e. in $\mathbb{R}^3$.
c) The self-adjoint operator corresponding to the single-particle energy observable $-\frac{1}{2}\Delta+W$ in $L^2(\mathbb{R}^3)$ has the lower part of its spectrum consisting of eigenvalues of finite multiplicity, which may be computed to arbitrarily high precision together with the corresponding eigenfunctions.

Some observations may be made immediately as a consequence of these assumptions. We first note that the self-adjoint operator associated with the m-electron observable

$$H^0 = \sum_{k=1}^{m}\left[-\frac{1}{2}\Delta_k + W(\mathbf{r}_k)\right]$$

has discrete lower spectra that are computationally resolvable in terms of the spectra of $-\frac{1}{2}\Delta + W$ by separation of variables. Furthermore, the symmetric operator $\hat{H} = \Sigma_{i<j}^{m} V(\mathbf{r}_i - \mathbf{r}_j)$ is essentially self-adjoint on $C_0^\infty(\mathbb{R}^{3m})$ which is also a core for H and H^0. Since $\hat{H} \geq 0$, we see that $H^0 \leq H$ and hence intermediate operator methods for approximating the eigenvalues of H from below are applicable.

The m-particle space $L^2(\mathbb{R}^{3m})$ can evidently be written as an m-fold tensor product $L^2(\mathbb{R}^3) \otimes L^2(\mathbb{R}^3) \otimes \cdots \otimes L^2(\mathbb{R}^3)$. With this decomposition the original Hamiltonian may be represented as

$$\begin{aligned} H = {} & A_0 \otimes I_2 \otimes \cdots \otimes I_m + I_1 \otimes A_0 \otimes \cdots \otimes I_m \\ & + \cdots + I_1 \otimes \cdots \otimes A_0 + \sum_{i<j}^{m} U_{ij}^{-1}\left(A_{ij} \otimes \left[\bigotimes_{\nu \neq i,j} I_\nu\right]\right) U_{ij}. \end{aligned}$$

Here A_0 denotes the single-particle operator $-\frac{1}{2}\Delta + W$ on $L^2(\mathbb{R}^3)$ and A_{ij} denotes multiplication by $V(\mathbf{r}_i - \mathbf{r}_j)$ on the two-particle space $L^2(\mathbb{R}^3) \otimes L^2(\mathbb{R}^3)$. The operator U_{ij} is a unitary transformation on $L^2(\mathbb{R}^{3m})$ corresponding to the change of variables $\mathbf{r}_i \leftrightarrow \mathbf{r}_1$ and $\mathbf{r}_j \leftrightarrow \mathbf{r}_2$ and is introduced as a notational convenience to express coupling between particles that do not have adjacent labeling (i.e., $i \neq j-1$).

An s-electron ion has a Hamiltonian given by

$$H_s = \sum_{k=1}^{s}\left[-\frac{1}{2}\Delta_k + W(\mathbf{r}_k)\right] + \sum_{i<j}^{s} V(\mathbf{r}_i - \mathbf{r}_j)$$

on a domain dense in $L^2(\mathbb{R}^{3s})$. With the decomposition of $L^2(\mathbb{R}^{3s})$ into an s-fold tensor product of copies of $L^2(\mathbb{R}^3)$, there is an obvious parallel in structure with H as above.

3. The approximation. Fox's construction of intermediate Hamiltonians involves an approximation of $A_0 = -\frac{1}{2}\Delta + W$ and $A_{ij} = V(\mathbf{r}_i - \mathbf{r}_j)$ in such a way as to retain the "ionic" structure of the original Hamiltonian. The key notion in Fox's construction is the approximation of A_0 by a suitable bounded operator on the one-particle space $L^2(\mathbb{R}^3)$ and the interelectron coupling A_{ij} by a suitable finite-rank operator on the two-particle space $L^2(\mathbb{R}^3) \otimes L^2(\mathbb{R}^3)$. We consider a particular variant detailed in [6] and [7] for which convergence can be shown.

Since the lower spectral information for A_0 is known, the spectral truncation of A_0 may be explicitly constructed

$$A_0^n = \int_{-\infty}^{\lambda_n^0} \lambda \, dE(\lambda) + \lambda_{n+1}^0\left[I - E(\lambda_n^0)\right],$$

where the eigenvalues of A_0 have been denoted by $\{\lambda_k^0\}$ and $E(\lambda)$ is the associated spectral projection. If A_0 has only N eigenvalues below $\sigma_{\text{ess}}(A_0)$ (as might occur with W representing an effective screening potential) then $n = N$ is the largest allowable truncation index and by convention we assign $\lambda_{N+1}^0 = \inf \sigma_{\text{ess}}(A_0)$. The spectral information lost in approximating A_0 by the bounded operator A_0^n is contained in the truncation remainder

$$\tilde{A}_0^n = A_0 - A_0^n = \int_{\lambda_{n+1}^0-}^{\infty} (\lambda - \lambda_{n+1}^0)\, dE(\lambda).$$

$\tilde{A}_0^n$ is a nonnegative symmetric operator with null space $\mathcal{U}_n = E(\lambda_n^0)L^2(\mathbb{R}^3)$. A finite-rank operator approximating $\tilde{A}_0^n$ on $L^2(\mathbb{R}^3)$ may be constructed by choosing vectors $\{q_\nu\} \subset \mathrm{Dom}(A_0)$ and defining an idempotent operator for each integer α as

$$Q^\alpha = \sum_{i,j=1}^{\alpha} \langle \cdot, \tilde{A}_0^n q_i \rangle c_{ij} q_j,$$

where c_{ij} are elements of the generalized matrix inverse to $[\langle q_i, \tilde{A}_0^n q_j \rangle]$. The final approximation to A_0 then takes the form $A_0^n + \tilde{A}_0^n Q^\alpha$. It is straightforward to verify that this is a bounded self-adjoint operator and that any subspace of $L^2(\mathbb{R}^3)$ containing

$$\mathcal{U}_n \oplus \underset{\alpha}{\mathrm{span}}\, \{ \tilde{A}_0^n q_\nu \}$$

is a reducing space for the operator.

Since A_{ij} similarly is a nonnegative operator on $L^2(\mathbb{R}^3) \otimes L^2(\mathbb{R}^3)$, we may approximate A_{ij} in a way analogous to the way $\tilde{A}_0^n$ was approximated. We choose $\{p_\nu\} \subset \mathrm{Dom}(A_{ij})$ and define

$$P^\beta = \sum_{k,l=1}^{\beta \cdot \beta} \langle \cdot, A_{ij} p_k \rangle_{\otimes} b_{kl} p_l.$$

Here $\langle \cdot, \cdot \rangle_{\otimes}$ denotes the inner product on $L^2(\mathbb{R}^3) \otimes L^2(\mathbb{R}^3)$ and b_{kl} are elements of the generalized matrix inverse to $[\langle p_k, A_{ij} p_l \rangle_{\otimes}]$. In order to construct suitable reducing spaces for $A_{ij}P^\beta$ it is in general necessary to assume that

$$A_{ij} p_\nu \in \Pi^\beta \otimes \Pi^\beta \quad \text{for } \nu = 1, \cdots, \beta \cdot \beta,$$

where Π^β is a finite-dimensional subspace of $L^2(\mathbb{R}^3)$ for which an explicit basis $\{\pi_\nu\}_{\nu=1}^{\beta}$ is available. With this constraint there will be no more than $\beta \cdot \beta$ linearly independent p_n available for a given choice of Π^β. Any subspace of $L^2(\mathbb{R}^3) \otimes L^2(\mathbb{R}^3)$ containing $\Pi^\beta \otimes \Pi^\beta$ reduces $A_{ij}P^\beta$ and $[\Pi^\beta \otimes \Pi^\beta]^\perp \subset \ker(A_{ij}P^\beta)$.

With these one- and two-particle approximations we may build up the m-particle intermediate Hamiltonian as

$$\begin{aligned} H^{n\alpha\beta} = {} & (A_0^n + \tilde{A}_0^n Q^\alpha) \otimes I_2 \otimes \cdots \otimes I_m + I_1 \otimes (A_0^n + \tilde{A}_0^n Q^\alpha) \otimes \cdots \otimes I_m + \cdots \\ & + I_1 \otimes I_2 \otimes \cdots \otimes (A_0^n + \tilde{A}_0^n Q^\alpha) \\ & + \sum_{i<j}^{m} U_{ij}^{-1} \left(A_{ij} P^\beta \otimes \left[\bigotimes_{\nu \neq i,j} I_\nu \right] \right) U_{ij}. \end{aligned}$$

Written in this way, the similarity in algebraic structure to the original Hamiltonian is evident.

We can build up an approximation to the s-electron ionic Hamiltonian H_s in an analogous way. Explicitly,

$$\begin{aligned} H_s^{n\alpha\beta} = {} & (A_0^n + \tilde{A}_0^n Q^\alpha) \otimes I_2 \otimes \cdots \otimes I_s \\ & + I_1 \otimes (A_0^n + \tilde{A}_0^n Q^\alpha) \otimes \cdots \otimes I_s + I_1 \otimes I_2 \otimes \cdots \otimes (A_0^n + \tilde{A}_0^n Q^\alpha) \\ & + \sum_{i<j}^{s} U_{ij}^{-1} \left[A_{ij} P^\beta \otimes \Big(\bigotimes_{\nu \neq i,j} I_\nu \Big) \right] U_{ij}. \end{aligned}$$

In order to analyze the spectrum of $H^{n\alpha\beta}$ and $H_s^{n\alpha\beta}$ it is convenient to characterize the complete set of reducing subspaces. We introduce a subspace of $L^2(\mathbb{R}^3)$ defined by

$$\mathbf{M}_{n\alpha\beta} = \mathscr{U}_n \oplus \operatorname*{span}_{\alpha} \{ \tilde{A}_0^n q_v \} \oplus \Pi_\beta .$$

All the reducing spaces of $H^{n\alpha\beta}$ may then be characterized in terms of $\mathbf{M}_{n\alpha\beta}$ as

$$\mathbf{N}_1 \otimes \mathbf{N}_2 \otimes \cdots \otimes \mathbf{N}_m \quad \text{with each } \mathbf{N}_k = \mathbf{M}_{n\alpha\beta} \text{ or } \mathbf{N}_k = \mathbf{M}_{n\alpha\beta}^{\perp}.$$

The reducing spaces for $H_s^{n\alpha\beta}$ have the same structure over s-fold products of $\mathbf{N}_k$. Since $H^{n\alpha\beta} = H_m^{n\alpha\beta}$, in what follows we only need consider reducing spaces and the spectral problem for $H_s^{n\alpha\beta}$.

There are evidently 2^s reducing spaces for $H_s^{n\alpha\beta}$ having the structure given. Since interchange of variables defines an isomorphism of subspaces, there are only $s+1$ distinct classes of such reducing spaces and these may be indexed according to the number of factors $\mathbf{M}_{n\alpha\beta}$ appearing in the product $\mathbf{N}_1 \otimes \mathbf{N}_2 \otimes \cdots \otimes \mathbf{N}_s$. Following the terminology of Fox [10], a subspace of type r is built up from r copies of $\mathbf{M}_{n\alpha\beta}$ and $s-r$ copies of $\mathbf{M}_{n\alpha\beta}^{\perp}$. There are $\binom{s}{r}$ subspaces of type r generated for each value of $r = 0, 1, \cdots, s$.

The spectral problem for $H_s^{n\alpha\beta}$ will be resolved by considering the spectal problem for restrictions of $H_s^{n\alpha\beta}$ to subspaces of type r, denoted $H_s^{n\alpha\beta}|_r$. This is well-defined since the symmetry of $H_s^{n\alpha\beta}$ implies the equivalence of any two restrictions of $H_s^{n\alpha\beta}$ to subspaces of the same type. On the single subspace of type 0 $H_s^{n\alpha\beta}|_0$ reduces to $\lambda_{n+1}^0 I$. The single subspace of type s is finite-dimensional so that $H_s^{n\alpha\beta}|_s$ is isomorphic to a matrix operator of dimension no greater than $(n+\alpha+\beta)^s$ and hence is devoid of essential spectrum.

With this background we are now in a position to elaborate on the "ionic" structure of the intermediate Hamiltonians that was mentioned earlier. This comes out as what we view as an analogue to the HVZ theorem. If we let $\lambda_*(A) = \inf \sigma_{\text{ess}}(A)$ and $\lambda_1(A) = \inf \sigma(A)$ then the HVZ theorem asserts

$$\lambda_*(H_s) = \lambda_1(H_{s-1}), \qquad s = 1, 2, \cdots, m.$$

We find a similar correspondence for intermediate Hamiltonians.

THEOREM 1. $\lambda_*(H_s^{n\alpha\beta}) = \lambda_1(H_{s-1}^{n\alpha\beta}) + \lambda_{n+1}^0$ *for* $s = 1, 2, \cdots, m$.

Proof. We need only locate $\lambda_*(H_s^{n\alpha\beta}|_r)$ for each $r = 0, \cdots, s-1$, in order to find $\lambda_*(H_s^{n\alpha\beta})$. Since the composite span of these s reducing spaces has finite codimension

$$\lambda_*\big(H_s^{n\alpha\beta}\big) = \min_{r=0,\cdots,s-1} \lambda_*\big(H_s^{n\alpha\beta}|_r\big).$$

For $r=0$ separation of variables yields $\lambda_*(H_s^{n\alpha\beta}|_0)=s\lambda_{n+1}^0$. For $1\le r\le s-1$, we likewise obtain

$$\lambda_*\left(H_s^{n\alpha\beta}|_r\right)=\lambda_1\left(H_r^{n\alpha\beta}|_r\right)+(s-r)\lambda_{n+1}^0.$$

Thus with the convention $\lambda_1(H_0^{n\alpha\beta}|_0)=0$, we find

$$\begin{aligned}\lambda_*\left(H_s^{n\alpha\beta}\right)&=\min_{r=0,\cdots,s-1}\left[\lambda_1\left(H_r^{n\alpha\beta}|_r\right)+(s-r)\lambda_{n+1}^0\right]\\&=\min_{r=0,\cdots,s-1}\left[\lambda_1\left(H_r^{n\alpha\beta}|_r\right)+[(s-1)-r]\lambda_{n+1}^0\right]+\lambda_{n+1}^0=\lambda_1\left(H_{s-1}^{n\alpha\beta}\right)+\lambda_{n+1}^0.\end{aligned}$$

Recall that under the assumptions given on W, $\inf\sigma_{\text{ess}}(-\frac{1}{2}\Delta+W)=0$. If $\{\lambda_n^0\}$ is an infinite set then $\lim\lambda_n^0=0$, otherwise we adhere to the convention $\lambda_{N+1}^0=0$ if there are only N negative eigenvalues. In either case the deviation of the above theorem from a proper HVZ theorem is λ_{n+1}^0 and may be made arbitrarily small by making n sufficiently big.

4. Convergence. Due to the "ionic" structure persisting in Fox's intermediate Hamiltonians, the potential exists for displacement of intermediate essential spectra. Indeed one may see from Theorem 1 that improvement of lower bounds to the lowest eigenvalue for H_{s-1} relates directly to the movement of intermediate essential spectra approximating $\inf\sigma_{\text{ess}}(H_s)$. Along the same lines one might reasonably expect conclusions on the convergence of lower discrete eigenvalues of $H_{s-1}^{n\alpha\beta}$ to imply conclusions on the convergence of $\inf\sigma_{\text{ess}}(H_s^{n\alpha\beta})$.

The techniques rest on monotonicity principles derived from variational characterizations of the lower eigenvalues (see [15]). These principles may be used to identify an eigenvalue $\lambda_k(H^{n\alpha\beta})$ of the intermediate operator $H^{n\alpha\beta}$ as a lower bound to $\lambda_k(H)$ only if $\lambda_k(H^{n\alpha\beta})<\lambda_*(H^{n\alpha\beta})$. Hence convergent lower bounds to $\lambda_k(H)$ will be attainable only if $\lambda_k(H)<\sup_{n\alpha\beta}\lambda_*(H^{n\alpha\beta})$. Such eigenvalues of H are termed *accessible*. The goal of this work is to show that all lower eigenvalues of H are accessible to convergent lower bound estimates.

The construction of Fox's intermediate Hamiltonians requires the selection of two projecting families of vectors: $\{q_\nu\}$, with which to construct Q^α, and $\{\pi_\nu\}$, with which to construct $\{p_\nu\}$ and P^β. These families may be chosen in such a way as to guarantee convergence of intermediate estimates to the corresponding *accessible* eigenvalues of H. Density criteria on these families sufficient to guarantee such convergence was developed in [6] and [8], and may be succinctly stated as:

(1) $\operatorname{span}\{\pi_\nu\}$ is dense in $L^2(\mathbb{R}^3)$, and

(2) $\operatorname{span}\{q_\nu\}$ is dense in $W_2^2(\mathbb{R}^3)\setminus\operatorname{span}\{u_k^0\}$,

where W_2^2 denotes the second Sobolev space imbedded in $L^2(\mathbb{R}^3)$.

There may be an additional condition on the large-argument asymptotics of $\{\pi_\nu\}$ in order to guarantee that $p_{\nu_{kl}}=A_{ij}^{-1}(\pi_k\otimes\pi_l)$ is an element of $L^2(\mathbb{R}^3)\otimes L^2(\mathbb{R}^3)$ (e.g., exponential decrease at ∞).

As we mentioned, the one remaining issue and the final goal of this work is to characterize the eigenvalues of H that are accessible. The two-particle case is an appropriate starting place. For $s=2$, we find $\lambda_*(H_2)=\lambda_1^0$ due to the relative Δ-compactness of V and the HVZ theorem. A simple argument shows $\lambda_1(A_0^n)=\lambda_1(A_0^n+\tilde{A}_0^nQ^\alpha)=\lambda_1^0$. Then compactness of $A_{12}P^\beta$ and separation of variables show that $\lambda_*(H_2^{n\alpha\beta})=\lambda_1^0+\lambda_{n+1}^0$. Thus $\lambda_*(H_2)-\lambda_*(H_2^{n\alpha\beta})=-\lambda_{n+1}$ may be made arbitrarily small by choosing n sufficiently big, imply that every lower eigenvalue of H_2 is accessible to convergent estimates.

We make an induction hypothesis at this point: for some $s \leq m$ all the lower eigenvalues of H_k are accessible for $k = 1, \cdots, s-1$, i.e.,

$$\lim_{n\alpha\beta} \lambda_*(H_k^{n\alpha\beta}) = \lambda_*(H_k) \quad \text{for } k = 1, \cdots, s-1.$$

The density conditions provided then assert that $\lim_{n\alpha\beta}\lambda_1(H_k^{n\alpha\beta}) = \lambda_1(H_k)$ for each $k = 1, \cdots, s-1$. Theorem 1 implies that

$$\lim_{n\alpha\beta} \lambda_*(H_s^{n\alpha\beta}) = \lambda_1(H_{s-1}) = \lambda_*(H_s).$$

Thus each lower eigenvalue of H_s is accessible to convergent estimates—completing the induction step.

We have proved the final theorem of this work:

THEOREM 2. *Under the density hypotheses* (1) *and* (2), *Fox's intermediate Hamiltonian construction provides convergent lower bound estimates to the lowest point of the essential spectrum of H*:

$$\lim_{n\alpha\beta} \lambda_*(H^{n\alpha\beta}) = \lambda_*(H).$$

As a consequence every lower eigenvalue of H is accessible and Fox's construction provides convergent lower bounds to every lower eigenvalue of H.

In closing, we recall that physicists are primarily interested not in the full Hamiltonian H but in a restriction of H to a domain possessing certain symmetry properties that reflect the Pauli exclusion principle. In order to cover this interesting case it would suffice to develop an extension of Theorem 1 to the case where we restrict to a suitable symmetry subspace—an extension which does in fact hold for the original HVZ theorem.

Acknowledgments. It is a pleasure to thank Professor David Fox for the many stimulating discussions that laid the intuitive groundwork for this work.

REFERENCES

[1] N. ARONSZAJN, *Approximation methods for eigenvalues of completely continuous symmetric operators*, Proc. Symposium on Spectral Theory and Differential Problems, Stillwater, OK, 1951, pp. 179–202.

[2] N. W. BAZLEY, *Lower bounds for eigenvalues with applications to the helium atom*, Phys. Rev., 120 (1960), pp. 144–149.

[3] N. W. BAZLEY AND D. W. FOX, *Truncations in the method of intermediate problems for lower bounds to eigenvalues*, J. Res. Nat. Bur. Studs., 65B (2) (1961), pp. 105–111.

[4] ______, *Lower bounds for eigenvalues of Schrödinger's equation*, Phys. Rev., 124 (1961), pp. 483–492.

[5] ______, *Lower bounds to eigenvalues using operator decompositions of the form B*B*, Arch. Rational Mech. Anal., 10 (1962), pp. 352–360.

[6] C. BEATTIE, *Some Convergence Results for Intermediate Problems that Displace Essential Spectra.*, M.S.E. Research Center preprint series, John Hopkins Univ. Applied Physics Laboratory, Laurel, Maryland, No. 65, 1982.

[7] ______, *The computation of convergent lower bounds in quantum mechanical eigenvalue problems*, in Numerische Behandlung von Eigenwertaufgaben, ISNM, vol. 69, J. Albrecht, L. Collatz and W. Velte., eds., Birkhaüser, Zurich, 1984.

[8] C. BEATTIE AND W. M. GREENLEE, *Convergence theorems for intermediate problems*, in preparation.

[9] D. FREUND, *Upper and lower bounds to two and three electron systems*, Dissertation, Univ. Delaware, Newark, 1982.

[10] D. W. FOX, *Lower bounds for eigenvalues with displacement of essential spectra*, this Journal, 3 (1972), pp. 617–624.

[11] D. W. FOX AND V. G. SIGILLITO, *Bounds for energies of radial lithium*, Z. Angew. Math. Phys., 23 (1972), pp. 392–411.

[12] ______, *Lower and upper bounds to energies of radial lithium*, Chem. Phys. Lett., 13 (1972), pp. 85–87.

[13] C. E. REID, *Intermediate Hamiltonians for the lithium atom*, Int. J. Quantum Chem., 6 (1972), pp. 793–795.

[14] ______, *Lower bounds for the energy levels of the lithium atom*, Chem. Phys. Lett., 26 (1974), pp. 243–245.

[15] A. WEINSTEIN AND W. STENGER, *Methods for Intermediate Problems for Eigenvalues: Theory and Ramifications*, Academic Press, New York, 1972.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

DIFFUSION IN FISSURED MEDIA*

MICHAEL BÖHM† AND R. E. SHOWALTER‡

Abstract. The nonlinear initial-boundary value problem

$$\frac{\partial u}{\partial t}+\frac{1}{\varepsilon}(\alpha(u)-v)=f_1, \qquad -\operatorname{div}(k\operatorname{grad} v)+\frac{1}{\varepsilon}(v-\alpha(u))=f_2 \quad \text{in } G\times(0,T),$$
$$u(x,0)=u_0(x) \quad \text{in } G, \qquad v(s,t)=0 \quad \text{on } \partial G\times(0,T)$$

is a well-posed model of diffusion in a fissured porous medium. Special features of the solution include the perseverance of local spatial continuity or singularities in the concentration u, the instantaneous propagation of the partially-saturated region throughout G, the delayed and limited advance of the fully-saturated region into G, and the concentration discontinuity on the boundary of the fully-saturated region. Weak maximum and order-comparison principles are obtained as L^∞ and L^1 estimates on a solution and a difference of solutions, respectively.

1. Introduction. Our objectives here are to derive a system of partial differential equations as a model for nonstationary flow of a fluid through a fissured porous medium, to demonstrate that the appropriate initial-boundary-value problem is mathematically well-posed, and to describe special features of such a flow model which distinguish it from the classical porous medium equation. The system obtained is actually equivalent to a single evolution equation, the *fissured medium equation*, which can be regarded as a regularization of the porous medium equation. Also, the porous medium equation is known to be the homogeneous limit of the fissured medium equation with increasing degree of fissuring [7].

Section 2 contains the derivation of the system of differential equations for flow in fissured media. Initially we follow [1], where only a special linear case was considered, but we include in our model the nonlinearities arising from fluid type (liquid or gas), concentration (porosity, absorption or saturation), and the exchange rate [6], [11]. The essential requirement is that the fluid be compressible. The considerably more difficult case wherein permeability is concentration-dependent will be discussed in [2]. We briefly describe an analogous heat conduction model. In §3 we show the Cauchy problem for the fissured medium equation has a unique generalized solution and we give weak maximum and order-comparison principles in the form of L^∞ and L^1 estimates. In contrast to the case of (possibly degenerate) parabolic equations, we find that for the fissured medium equation the local spatial regularity or a singularity in the solution is stationary and may persevere for all time. We consider in §4 the evolution of a system originating with a uniform positive pressure in a portion G' of G and with null concentration in the complement of G' in G. It is shown that the partially-saturated region expands instantly to all of G, the positive-pressure set is nondecreasing, propagates only after a delay, and an upper bound is given for its measure.

*Received by the editors July 21, 1983, and in revised form January 30, 1984.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. On leave from Humboldt-Universität zu Berlin Sektion Mathematik, 1086 Berlin, PSF 1297, GDR.

‡ Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. On leave from the Department of Mathematics, University of Texas, RLM 8.100, Austin, Texas 78712.

Our notation is standard. G is a *bounded domain* in Euclidean space $\mathbf{R}^N$, $Q=G\times(0,T)$ is the indicated space-time cylinder, and ∂G denotes the boundary of G. $L^p(G)$ and $W^{m,p}(G)$ are the usual Lebesgue and Sobolev spaces, and $C^{m,\lambda}(G)$ is the Schauder space of functions whose derivatives of order m are Hölder-continuous with exponent λ, $0<\lambda<1$. For a Banach space B, we let $L^q(0,T;\ B)$ and $C(0,T;\ B)$ denote the spaces of q-summable or uniformly-continuous B-valued functions on $[0,T]$, respectively, and $W^{1,q}(0,T;\ B)$ denotes those strongly absolutely continuous functions whose derivatives belong to $L^q(0,T;\ B)$. The positive and negative parts of $u\in\mathbb{R}$ are given by $u^+=\max(u,0)$ and $u^-=\min(u,0)$, respectively, so $u=u^++u^-$. The Heaviside function is $H_0(u)=1$ for $u>0$ and $H_0(u)=0$ for $u\leqq 0$; its maximal monotone extension [3] is denoted by $H(u)=\{H_0(u)\}$ for $u\neq 0$ and $H(0)=[0,1]$. Likewise the sign function is $\operatorname{sgn}_0(u)=u/|u|$ for $u\neq 0$, $\operatorname{sgn}_0(0)=0$, and sgn denotes the maximal monotone extension. The gradient in $\mathbb{R}^N$ will be denoted by $\vec{\nabla}$ and similarly $\vec{\nabla}\cdot$ denotes the corresponding divergence operator.

2. Fissured medium equation. We consider the flow of a liquid or gas through a fissured porous medium, a structure consisting of porous permeable blocks separated by a system of fissures. The distribution of fissures prevents direct diffusion between adjacent blocks, and the system of fissures occupies a region of negligible relative volume. Thus the blocks provide for the local storage of fluid mass, and the fissures are the essential flow-paths for all the diffusion. The essential point in the construction of the fissured medium model is to introduce at each point in space two fluid pressures, the pressure p_1 in the blocks and the pressure p_2 in the fissures, where each is an average over a neighborhood which contains a substantial number of blocks.

The fluid under consideration may be any compressible liquid or gas whose density ρ and pressure p are related by an equation-of-state $\rho=s(p)$ for which the compressibility satisfies $s'(p)>0$ for $p\geqq 0$ and $s(0)\geqq 0$. The total concentration of fluid is given by $u=P(s(p_1)-s(0)+\xi(L+s(0)))$ where $P>0$ is porosity of the blocks, $L\geqq 0$ is that density of fluid which is immobilized due to absorption or chemical reaction with the medium, and the saturation level $0\leqq\xi\leqq 1$ is that fraction of $L+s(0)$ already immobilized or absorbed. Note $p_1(1-\xi)=0$ so $\xi\in H(p_1)$, the Heaviside graph. Thus u is a monotone graph of p_1 whose inverse $p_1\equiv\alpha(u)$ is a monotone function Lipschitz-continuous with constant K. The medium is *completely saturated* when $u\geqq P(L+s(0))$, hence, $\xi=1$, and *partially saturated* (*strictly partially saturated*) when $u>0$ (respectively, $0<u<P(L+s(0))$.)

The exchange of fluid between blocks and fissures occurs with a volume rate per volume of medium given by $(p_2-p_1)/\mu\varepsilon$ where μ is the viscosity of the fluid and $1/\varepsilon$ is a characteristic of the medium related to the degree of fissuring or the surface area common to the blocks and fissures. Thus the mass of fluid which flows from blocks to fissures per unit time is given by $\rho(p_1-p_2)/\mu\varepsilon$ where ρ is the average density on the pressure-interval $[p_1,p_2]$. Denoting by $S(p)\equiv\int_0^p s(r)\,dr$ the antiderivative of $s(p)$, or "flow potential" [6, p. 60], we have $\rho=(p_2-p_1)^{-1}\int_{p_1}^{p_2}s(p)\,dp=(p_2-p_1)^{-1}(S(p_2)-S(p_1))$. The fluid mass exchanged per unit time is $(S(p_1)-S(p_2))/\mu\varepsilon$. Thus the continuity equation for conservation of fluid mass in the blocks gives

$$\frac{\partial u}{\partial t}+\frac{1}{\varepsilon\mu}(S(p_1)-S(p_2))=f_1(x,t), \tag{2.1}$$

where f_1 is the volume-distributed source rate in the blocks.

We shall assume the velocity of the fluid in the fissures is given by Darcy's law. Thus, $V=-(k/\mu)\vec{\nabla}p_2$ where k is the permeability of the system of fissures. The flux in the fissures is computed by the chain rule as $\rho_2 V = -(k/\mu)\vec{\nabla} S(p_2)$. Since the relative volume of the fissure system is null, the enclosed concentration is negligible and the conservation of fluid mass in the fissure system gives

$$-\frac{1}{\mu}\vec{\nabla}\cdot k\vec{\nabla} S(p_2)+\frac{1}{\varepsilon\mu}(S(p_2)-S(p_1))=f_2(x,t). \tag{2.2}$$

Here f_2 denotes a volume-distributed source rate in the system of fissures.

The porosity and permeability may depend on the pressures. Given the small volume of the fissures the pressure p_2 will not appreciably affect the block porosity, so we may expect a mild dependence $P=P(p_1)>0$ of block porosity on block pressure. This does not alter the assumptions above on the relation $p_1=\alpha(u)$. Due to the relative volumes of blocks and fissures, any variation of the fissure permeability is essentially due to the block pressure p_1. This is equivalent to the assumption that the fluid in fissures does not participate in the support of the structure. In contrast to the slight variations of $k(p_1)$ for $p_1>0$, any swelling of the blocks due to saturation or absorption of fluid can result in a dramatic decrease of fissure permeability owing to their relative volumes. This sensitivity of permeability to saturation due to swelling is typical of consolidated sandstones containing clay or silt [6, p. 13]. We shall account for such phenomena by setting $k=k(u)$ in (2.2). The function $k(\cdot)$ is continuous, positive and nonincreasing on $0\leqq u$; furthermore, the model suggests $k(u)$ is essentially constant for $u\geqq P(L+s(0))$, the saturated zone.

In summary, the process of diffusion in a fissured medium is prescribed by the system of partial differential equations (2.1), (2.2) with $p_1=\alpha(u)$ and $k=k(u)$. The initial concentration $u(x,0)=u_0(x)$ is given over the region G of interest; this is equivalent to specifying initial block-pressure $p_1(x)$ and initial saturation $\xi_0(x)$ with $\xi_0(x)\in H(p_1(x))$. The description is completed by setting fissure-pressure $p_2=0$ on the boundary of the region G. Note that no boundary conditions are given for p_1, since all fluid flow in the blocks is accounted for in (2.1), even in a neighborhood of the boundary. Similarly, the initial pressure distribution in the fissures is determined by (2.2).

We shall write the above system as a single nonlinear evolution equation. Thus, for a function u on G of a type made precise below, let $A_u(v)=-(1/\mu)\vec{\nabla}\cdot k(u)\vec{\nabla} v$ be the indicated linear elliptic partial differential operator in divergence form subject to null Dirichlet boundary conditions. By adding (2.1) and (2.2), then substituting (2.2) we obtain

$$(I+\varepsilon\mu A_{u(t)})\left(\frac{\partial u}{\partial t}\right)+A_{u(t)}(S(\alpha(u)))=(I+\varepsilon\mu A_{u(t)})f_1(t)+f_2(t). \tag{2.3}$$

Alternatively, we may resolve (2.2) for $S(p_2)$ and substitute in (2.1) to obtain

$$\frac{\partial u}{\partial t}+\frac{1}{\varepsilon\mu}\left[I-(I+\varepsilon\mu A_{u(t)})^{-1}\right]S(\alpha(u))=f_1(t)+(I+\varepsilon\mu A_{u(t)})^{-1}f_2(t). \tag{2.4}$$

The equation (2.4) we shall call the *fissured medium equation*. It is actually equivalent to the above system. When $S(\alpha(u))$ is smooth and satisfies the Dirichlet boundary condition, i.e., belongs to the domain of $A_{u(t)}$, then (2.4) implies the stronger form (2.3). Note that formally taking $\varepsilon\to 0^+$ in either one leads to the classical *porous medium equation*

$$\frac{\partial u}{\partial t}-\frac{1}{\mu}\vec{\nabla}\cdot k(u)\vec{\nabla} S(\alpha(u))=f_1+f_2 \tag{2.5}$$

when $L=0$ and to the Stefan free-boundary problem in weak form when $L>0$. This corresponds to increasing the degree of fissuring, $1/\varepsilon$, and thereby approximating the homogeneous limiting case (2.5) [3], [7].

We shall briefly describe an analogous model for heat conduction in a heterogeneous medium consisting of two components. This thermal conduction model is formally equivalent to the fissured medium equation. Thus, assume the first component occurs in small blocks isolated by the second component which is distributed throughout the medium with negligible measure. We permit the first component material to undergo a phase change as in a Stefan free-boundary problem. A model for the situation is water (the first component) contained in a metal (second component) structure of thin walls forming a structure much like an ice-cube tray. Letting T_1 and T_2 denote temperatures (averaged) in the water and metal, respectively, we obtain the system

$$\begin{aligned} &\frac{\partial u}{\partial t}+\frac{1}{\varepsilon}(T_1-T_2)=f_1, \\ &-k\Delta T_2+\frac{1}{\varepsilon}(T_2-T_1)=f_2, \qquad (2.6) \\ &u\in C(T_1)+LH(T_1). \end{aligned}$$

Here the heat content u is given by the specific heat $C(T_1)$ in water and the latent heat L in the melted region ($T_1>0$), k is the conductivity of the second component material, and the heat exchange between water chambers and metal dividers is assumed proportional to the difference of their temperatures. The local description and derivation of the equations follows exactly as in [13]. This system is formally equivalent to a special case of (2.1) and (2.2). Unlike the diffusion model, we are interested in temperatures which are not necessarily nonnegative; these are permitted in our discussion below. A completely-saturated region in the diffusion model corresponds to a completely melted or water region ($u\geqq L$) in the conduction model, and a strictly-partially-saturated region corresponds to a region of *mush*, a mixture of ice and water in equilibrium at the freezing temperature. As we shall see below, the solution to such a conduction model is dramatically different from the classical Stefan problem solution. Specifically, (2.6) is *not* the Stefan problem for the pseudo-parabolic equation of heat conduction [5] as given in [8].

3. The Lipschitz case. We begin our discussion of (2.4) by considering the special case in which $k(u)$ is independent of u but is a function of $(x,t)\in Q$. In the diffusion problem this corresponds to the case of a rigid structure in which the permeability is not affected by the total concentration (density and saturation). We shall denote by

$$A(t)v\equiv-\overline{\nabla}\cdot\left(k(x,t)\overline{\nabla}v\right)$$

the indicated elliptic differential operator whose coefficient $k\in L^\infty(Q)$ is assumed to satisfy $0<k_0\leqq k(x,t)$, a.e. $(x,t)\in Q$. In the Banach space $L^1(G)$ the domain of $A(t)$ is $\operatorname{dom}(A(t))=\{v\in W_0^{1,1}(G)\colon A(t)v\in L^1(G)\}$, where $A(t)v$ is understood in the sense of distributions. This L^1-realization of $A(t)$ can be obtained as the L^1-closure of its restriction to $L^p(G)$, $1<p<+\infty$. Each such restriction, including $A(t)$ itself, is a linear m-accretive operator on the corresponding Banach sapce, $L^p(G)$. See [4], [10] for these and additional properties of these elliptic operators in L^p. Here we shall consider the realization of the fissured medium equation (2.4) in $L^1(G)$ in the form

$$u'(t)+\frac{1}{\varepsilon}\left(I-(I+\varepsilon A(t))^{-1}\right)\alpha(u(t))=f(t), \qquad 0\leqq t\leqq T. \qquad (3.1)$$

Without loss of generality we have set $\mu=1$, $S=I$, and $f(t)=f_1(t)+(I+\varepsilon A(t))^{-1}f_2(t)$ in $L^1(G)$. We assume hereafter that α is a nondecreasing Lipschitz continuous function on $\mathbb{R}$ with $\alpha(0)=0$. Thus the substitution operator $v\mapsto\alpha(v)$ is Lipschitz on each $L^p(G)$ and we easily obtain the following L^p-existence-uniqueness result.

LEMMA 1. *If $u_0\in L^p(G)$ and $f\in L^q(0,T;\ L^p(G))$, $1\leqq p,\ q\leqq+\infty$, then there is a unique $u\in W^{1,q}(0,T;\ L^p(G))$ which satisfies* (3.1) *and $u(0)=u_0$.*

Proof. Since each $(I+\varepsilon A(t))^{-1}$ is a contraction and $\alpha(\cdot)$ is Lipschitz on each $L^p(G)$, it follows that the u-dependence in (3.1) is Lipschitz, uniformly in t. From [12] it follows that the operator-valued map $t\mapsto(I+\varepsilon A(t))^{-1}$ is strongly-measurable into $\mathscr{L}(H^{-1}(G),H_0^1(G))$ and an elementary closure argument shows it is strongly-measurable into $\mathscr{L}(L^p(G))$. The classical successive-approximations finishes the proof.

In order to obtain "pointwise estimates" on solutions of (3.1) we write it in the form

$$u'(t)+\frac{1}{\varepsilon}\alpha(u(t))=\frac{1}{\varepsilon}(I+\varepsilon A(t))^{-1}\alpha(u(t))+f(t),\qquad 0\leqq t\leqq T. \tag{3.2}$$

This splitting of (3.1) displays explicitly its structure as an ordinary differential equation (in t) and an elliptic partial differential equation (in x). Moreover, it suggests we consider the ordinary initial-value problem

$$w'(t)+\frac{1}{\varepsilon}\alpha(w(t))=g(t),\qquad 0\leqq t\leqq T,\quad w(0)=w_0. \tag{3.3}$$

For each $g\in L^1(0,T)$ and $w_0\in\mathbb{R}$ there is a unique solution $w\in W^{1,1}(0,T)$. If $w_j(j=1,2)$ are solutions corresponding to data g_j, w_0^j, we subtract the equations, multiply by $H_0(w_1(t)-w_2(t))$ and integrate to obtain (since $(\alpha(w_1)-\alpha(w_2))H_0(w_1-w_2)\geqq0$)

$$[w_1(t)-w_2(t)]^+\leqq[w_0^1-w_0^2]^++\int_0^t[g_1(s)-g_2(s)]^+\,ds.$$

Moreover, if each $g_j\in L^1(Q)$ and $w_0^j\in L^1(G)$, the above holds for a.e. $x\in G$ and a further integration over G yields

$$\|[w_1(t)-w_2(t)]^+\|_{L^1(G)}\leqq\|[w_0^1-w_0^2]^+\|_{L^1(G)}+\int_0^t\|[g_1(s)-g_2(s)]^+\|_{L^1(G)}ds,\quad 0\leqq t\leqq T.$$

Thus, the operator W: $L^1(G)\times L^1(Q)\to C(0,T;\ L^1(G))$ defined by (3.3) with $w=W(w_0,g)$ is an order-preserving contraction. The elliptic operator $A(t)$ satisfies a similar estimate [4, Lemma 3*]

$$\left\|\left[(I+\varepsilon A(t))^{-1}g\right]^+\right\|_{L^1(G)}\leqq\|g^+\|_{L^1(G)},\qquad g\in L^1(G),$$

and trivially so also does α: $L^1(G)\to L^1(G)$.

The relevance of the preceding remarks is that a solution of (3.1) is characterized by

$$u=W\big(u_0,(1/\varepsilon)(I+\varepsilon A)^{-1}\alpha(u)+f\big). \tag{3.4}$$

The right side of (3.4) is Lipschitz with an integral bound implying it has a unique fixed point. This provides an alternate proof of Lemma 1 with $p=q=1$ but, more important, it yields the following comparison principle.

LEMMA 2. *Let u_1 and u_2 be the respective solutions of the initial-value problem for* (3.1) *with data $u_0^1,u_0^2\in L^1(G)$ and $f_1,f_2\in L^1(Q)$. If $u_0^1\geqq u_0^2$ a.e. in G and if $f_1\geqq f_2$ a.e. in Q, then $u_1\geqq u_2$ in Q.*

Proof. For $j=1,2$, we have $u_j=\lim_{n\to\infty}\mathscr{W}_j^{(n)}(u_0^j)$ in $C(0,T;\ L^1(G))$ where $\mathscr{W}_j(v)\equiv W(u_0^j,(1/\varepsilon)(I+\varepsilon A)^{-1}\alpha(v)+f_j)$. The preceding remarks show $\mathscr{W}_1(v_1)\geqq\mathscr{W}_2(v_2)$ whenever $v_1\geqq v_2$ and $u_0^1\geqq u_0^2$, so the desired follows.

In a similar manner, we can deduce an L^∞ estimate on the solution. However, this procedure is inefficient and does not yield the optimal estimates in either case; these will be obtained below. Although (3.2) has so far served only to motivate the comparison and maximum principles and to provide elementary proofs, it will be used below to directly obtain very distinctive and surprising results on local regularity of solutions. All of these we state as follows.

THEOREM 1. *Let $\{A(t):0\leqq t\leqq T\}$ be the uniformly elliptic family of elliptic operators on $L^1(G)$ as given above. Suppose $\varepsilon>0$ and α is monotone with $\alpha(0)=0$ and Lipschitz constant K.*

(a) *For each $u_0\in L^p(G)$ and $f\in L^q(0,T;\ L^p(G))$, $1\leqq p,\ q\leqq+\infty$, there is a unique solution $u\in W^{1,q}(0,T;\ L^p(G))$ of* (3.1) *with $u(0)=u_0$.*

(b) *This solution satisfies*

$$\|u(t)^+\|_{L^\infty(G)}\leqq\|u_0^+\|_{L^\infty(G)}+\int_0^t\|f(s)^+\|_{L^\infty(G)}ds,\qquad 0\leqq t\leqq T,$$

and similar estimates for $\|u(t)^-\|_{L^\infty(G)}$, $\|u(t)\|_{L^\infty(G)}$.

(c) *For $j=1,2$ let u_j be the solution with corresponding data $u_0^j\in L^1(G)$, $f_j\in L^1(Q)$. Then*

$$\|[u_1(t)-u_2(t)]^+\|_{L^1}\leqq\|[u_0^1-u_0^2]^+\|_{L^1}+\int_0^t\|[f_1(s)-f_2(s)]^+\|_{L^1}ds,\qquad 0\leqq t\leqq T,$$

and similarly for $\|[u_1(t)-u_2(t)]^-\|_{L^1}$ and $\|[u_1(t)-u_2(t)]\|_{L^1}$.

(d) *Assume $p>N/2$ and G' is a subdomain whose closure is contained in G. There are constants $C>0$, $\lambda>0$ such that*

$$[u(x_1,t)-u(x_2,t)]^+\leqq[u_0(x_1)-u_0(x_2)]^+$$
$$+\int_0^t[f(x_1,s)-f(x_2,s)]^+ds+C|x_1-x_2|^\lambda,$$
$$x_1,x_2\in G',\quad 0\leqq t\leqq T,$$

and similarly for $[u(x_1,t)-u(x_2,t)]^-$ and $|u(x_1,t)-u(x_2,t)|$.

(e) *Assume $p>N/2$, and let $x\in G$, $v\in\mathbb{R}^N$ be a unit vector, and denote the saltus or jump of a function w at x by $\sigma(w(x))\equiv\lim_{h\to0^+}(w(x+hv)-w(x))$. Assume there is a $g\in L^1(0,T)$ for which*

$$|f(x+hv,t)|\leqq g(t),\quad 0<h<h_0,\quad 0\leqq t\leqq T$$

and each of $\sigma(u_0(x))$, $\sigma(f(x,t))$ exists. Then $\sigma(u(x,t))$ exists for each $t\in[0,T]$ and

$$\sigma(u(x,t))^+\leqq\sigma(u_0(x))^++\int_0^t\sigma(f(x,s))^+ds,$$
$$\sigma(u(x,t))^+\geqq e^{-Kt/\varepsilon}\left\{\sigma(u_0(x))^++\int_0^te^{Ks/\varepsilon}\sigma(f(x,s))^-ds\right\}$$

with similar estimates for $\sigma(u(x,t))^-$ and $\sigma(u(s,t))$.

Proof. Part (a) is just Lemma 1. To prove (b) note first that the Yoshida approximation

$$A_\varepsilon(t)\equiv\frac{1}{\varepsilon}\left(I-(I+\varepsilon A(t))^{-1}\right),$$

satisfies the resolvent identity

$$(I+\lambda A_\varepsilon(t))^{-1}=(\varepsilon/(\varepsilon+\lambda))I+(\lambda/(\varepsilon+\lambda))(I+(\lambda+\varepsilon)A(t))^{-1}, \qquad \lambda>0,$$

which implies that $A_\varepsilon(t)$ satisfies the conditions in [4, Theorem 1]. From (3.1) in $L^\infty(G)$ subtract $\|f(t)^+\|_{L^\infty}$ and multiply by $H_0(u(x,t)-k-\int_0^t\|f^+(s)\|_{L^\infty}ds)$ where $k\geqq 0$ will be chosen below. Integrating the product gives

$$\int_G\left(u'(t)-\|f^+(t)\|_{L^\infty}\right)H_0\left(u-k-\int_0^t\|f^+\|_{L^\infty}ds\right)ds$$

$$+\int_G A_\varepsilon(t)(\alpha(u(t)))H_0\left(u-k-\int_0^t\|f^+\|_{L^\infty}ds\right)dx\leqq 0.$$

The second term is nonnegative by the fundamental [4, Lemma 2] and the first term is equal to $\frac{d}{dt}\int_G[u(x,t)-k-\int_0^t\|f(s)^+\|ds]^+\,dx$. Thus we obtain

$$\int_G\left[u(x,t)-k-\int_0^t\|f(s)^+\|_{L^\infty}ds\right]^+dx\leqq\int_G[u_0(x)-k]^+dx,$$

and choosing $k=\|u_0^+\|_{L^\infty}$ proves (b). Part (c) is proved similarly: subtracting (3.1) for $j=1,2$ and multiplying by $H_0(u_1-u_2)$ yields

$$\frac{d}{dt}\int_G[u_1(x,t)-u_2(x,t)]^+dx+\int_G A_\varepsilon(t)(\alpha(u_1)-\alpha(u_2))H_0(u_1-u_2)\,dx$$

$$=\int_G(f_1-f_2)H_0(u_1-u_2)\,dx\leqq\|(f_1-f_2)^+\|_{L^1}.$$

The second term is nonnegative as before and this leads to the end of proof of (c).

Consider the situation of part (d). Since the solution u is bounded in $L^p(G)$, so also is $\alpha(u)$ and it follows that $v(t)\equiv(I+\varepsilon A(t))^{-1}\alpha(u(t))$ is bounded in a Schauder space $C^{0,\lambda}(G')$ [10, p. 192]. Thus, there is a constant C_1 for which

$$|v(x_1,t)-v(x_2,t)|\leqq C_1|x_1-x_2|^\lambda, \qquad x_1,x_2\in G',\quad 0\leqq t\leqq T.$$

The splitting (3.2) and the estimates following (3.3) with $w_j(t)=u(x_j,t)$ lead directly to the proof of (d).

For (e) we difference (3.2) at $x_1=x+hv$ and $x_2=x$ and use the preceding estimates and the Lebesgue theorem to obtain

$$\frac{d}{dt}\sigma(u(x,t))+\frac{1}{\varepsilon}\sigma(\alpha(u(x,t)))=\sigma(f(x,t)).$$

Since $\sigma(\alpha(u))=\alpha(\sigma(u))$, the desired estimates follow as above or by Gronwall's inequality. This finishes the proof.

Estimates of the forms in (b) and (c) are known as weak maximum principles and as comparison principles, respectively. Those given are optimal as can be seen by taking $\alpha\equiv 0$. They imply that nonnegative data yield nonnegative solutions.

If the coefficients $k(\cdot,t)$ and the boundary of G are smooth, then in the situation of (d) we get $v(t)$ bounded in $W^{2,p}(G)$, hence, in $C^{1,\lambda/2}(G)$. This leads to pointwise estimates on smoothness of first-order spatial derivatives of the solution. Such estimates on higher (than first) order derivatives appear to require assumptions on the global regularity of the data.

From (d) it follows the solution is exactly as smooth in x as $u_0(\cdot)$ and $\int_0^T f(\cdot,s)\,ds$, up to Hölder continuity with constant λ in each neighborhood in G. Likewise, (e) shows any jump discontinuity in data persists at the same point for a positive time interval, and for all time if $\sigma(f^-)\sigma(u_0)^+=0$ at that point. This striking persistence of local regularity is a consequence of the form (3.2) of the fissured medium equation.

We consider the meaning of a jump discontinuity in the solution of (3.1) when the equation is used as a model for diffusion. First, recall that the variables introduced in the diffusion model were defined pointwise as averages over a neighborhood of an idealized variable, e.g., pressure. It follows for an integrable ideal variable that such averaged variables are necessarily absolutely continuous in their spatial dependence. Thus within the medium the data and hence the solution are continuous. Second, we note that a discontinuity in data can be induced by fitting together two regions with independently prescribed concentration distributions. This discontinuity along the common interface will then persist on that stationary interface. This is consistent with the fissured medium diffusion model, because the two regions are coupled only by way of the fissure system, a relatively weak coupling.

4. Propagation and saturation. We consider now the fissured medium equation (3.1) and assume for definiteness that $\alpha(u)=0$ for $0\leqq u\leqq L$ and that $\alpha(u)>0$ for $u>L$. The medium is called *partially saturated* (or *strictly partially saturated*) at $(x,t)\in Q$ if $u(x,t)>0$ (respectively, $0<u(x,t)<L$). From Theorem 1 it follows that each strictly-partially-saturated point remains so over some time interval. In order to follow the advance of the fluid through the medium we consider for each $t\in[0,T]$ the set $P(t)\equiv\{x\in G: u(x,t)>L\}=\{x\in G:\alpha(u(x,t))>0\}$ wherein the block-pressure is strictly positive and, hence, the medium is *completely saturated.*

THEOREM 2. *In the situation of Theorem* 1 *assume further that* $\alpha^{-1}(0)=[0,L]$, $p>N/2$ *and both* u_0 *and* f *are nonnegative. Thus* $u\geqq 0$ *and we also have the following*:

(a) *The set* $P(t)$ *is nondecreasing in* $t\in[0,T]$. *If* $P(t_0)$ *is nonempty then the medium is partially saturated at every* $(x,t)\in Q$ *with* $t\geqq t_0$.

(b) *Assume* $f\equiv 0$, *let* G_1 *be a measurable subset of* G, ρ_0 *and* L *be strictly positive, and set* $u_0(x)=\rho_0+L$ *for* $x\in G_1$, *and* $u_0(x)=0$ *for* $x\in G\sim G_1$. *Denoting Lebesgue measure by* $m(\cdot)$ *we have*

$$m(G_1)\leqq m(P(t))\leqq(1+\rho_0/L)m(G_1),\qquad 0\leqq t\leqq T.$$

(c) *Assume further that* $k=k(x)$ *is autonomous, there is a* $\delta>0$ *with* $\alpha(s)\geqq\delta(s-L)^+$, *and* $m(G_1)>0$. *Then for each* $x\in G$ *there is a* $C(\varepsilon,x)>0$ *such that* $\rho_0/L>C(\varepsilon,x)$ *implies that* $x\in P(t)$ *for all* t *sufficiently large.*

Proof. (a) Since K is the Lipschitz constant for α and $\alpha(L)=0$ we have

$$\varepsilon u_t+K(u-L)\geqq\varepsilon u_t+\alpha(u)-\alpha(L)=(I+\varepsilon A)^{-1}\alpha(u)+\varepsilon f(t)\geqq 0,$$

so there follows

$$u(x,t)-L\geqq e^{-(K/\varepsilon)(t-t_0)}(u(x,t_0)-L),\qquad t\geqq t_0,\quad x\in P(t_0).$$

This shows $P(t)\supset P(t_0)$. Similarly, we have

$$\varepsilon u_t+Ku\geqq(I+\varepsilon A)^{-1}\alpha(u)+\varepsilon f.$$

If for some $(x_0,t_0)\in\Omega$ we have $\alpha(u(x_0,t_0))>0$, then by the strong maximum principle [10, pp. 188–189] $((I+\varepsilon A)^{-1}\alpha(u))(x,t_0)>0$ for *all* $x\in G$ and there follows $u(x,t)>0$ for all $t\geqq t_0$.

(b) The first inequality follows from (a) since $P(0)=G_1$. The second is obtained from the L^1-estimate

$$Lm(P(t))\leqq\int_{P(t)}u(x,t)\,dx\leqq\|u(t)\|_{L^1}\leqq\|u_0\|_{L^1}=(\rho_0+L)m(G_1).$$

(c) For $x\in G\sim G_1$, $u_0(x)=0$ so by continuity the number $T(x)\equiv\sup\{\tau\geqq 0: u(x,t)\leqq L$ for all $0\leqq t\leqq\tau\}$ is strictly positive. We shall show $T(x)<\infty$. From the proof of (a) follows

$$\alpha(u(x,t))\geqq\delta(u(x_1,t)-L)\geqq\delta\rho_0e^{-Kt/\varepsilon},\qquad x_1\in G_1,\quad t\geqq 0.$$

Define χ_1 as the characteristic function of G_1 and $\varphi_1\equiv(I+\varepsilon A)^{-1}\chi_1$. By the strong maximum principle $\varphi_1(x)>0$ for every $x\in G$. Since $\alpha(u(x,t))\geqq\delta\rho_0e^{-Kt/\varepsilon}\chi_1(x)$ we obtain from the comparison principle

$$(I+\varepsilon A)^{-1}\alpha(u(x,t))\geqq\delta\rho_0e^{-Kt/\varepsilon}\varphi_1(x),\qquad x\in G,\quad t\geqq 0.$$

Thus, for $x\in G\sim G_1$ and $0\leqq t\leqq T(x)$ we have $\alpha(u(x,t))=0$ and from (3.1)

$$u_t(x,t)\geqq(\delta\rho_0/\varepsilon)e^{-Kt/\varepsilon}\varphi_1(x)$$

and therefore

$$u(x,t)\geqq(\delta\rho_0/K)(1-e^{-Kt/\varepsilon})\varphi_1(x),\qquad x\in G\sim G_1,\quad 0\leqq t\leqq T(x).$$

Thus, if $\rho_0/L\geqq K/\delta\varphi_1(x)$, then there is a $t^*=T(x)$ for which $u(x,t^*)=0$ and $u_t(x,t^*)>0$. This finishes the proof of the theorem.

The property expressed in (a), that every point in the medium is partially saturated as soon as any point has positive pressure, is a consequence of the instantaneous diffusion through the system of fissures. Although such infinite propagation speeds are standard for linear parabolic equations, the porous medium equation (2.5) is known to have finite propagation speeds for certain nonlinearities.

In the diffusion model leading to the situation described in (b), L is the amount of fluid required per volume to fill the voids or to overcome an absorption characteristic of the medium, and ρ_0 is the density of excess fluid available in the region G_1. The estimate in (b) is an explicit upper bound on the advance of the pressure set $P(t)$ in terms of the ratio ρ_0/L.

Similarly, (c) shows that for each point $x\in G$ there is a value of the ratio ρ_0/L which drives $P(t)$ to enclose x. The qualitative dependence of $C(\varepsilon,x)$ is clear from the proof and interesting. Specifically, for $x\in G\sim G_1$ we note $C(\varepsilon,x)$ increases as x approaches ∂G or as ε decreases to 0, $C(\varepsilon,x)$ decreases as x approaches G_1, and $C(\varepsilon,x)$ approaches $2k/\delta$ for x near G_1 and ε near 0.

Our knowledge of the regularity of the solution permits a description of its behavior along the "free-boundary" or interface Γ bounding the pressure set. Thus, let $Q_+\equiv\{(x,t)\in Q: u(x,t)>L\}$ and $Q_0\equiv\{(x,t)\in Q: 0\leqq u\leqq L\}$ in the situation of Theorem 2, and set $\Gamma=\partial Q_+$. At each point of Γ denote the unit normal by $(n_1,n_2,\cdots,n_N,n_t)$ and let n be the unit vector in $\mathbb{R}^N$ with direction of $(n_1,\cdots,n_N)$. Let σ_Γ denote the jump or saltus along Γ. The standard computation of (2.1) over Q_+, Q_0 and the divergence theorem lead to the interface condition

$$\sigma_\Gamma(u)n_t=0\quad\text{on }\Gamma.$$

Thus, at each point of Γ, either the concentration is continuous or the interface is stationary. A similar computation on (2.2) shows

$$\sigma_\Gamma\left(k\frac{\partial}{\partial n}(S(p_2))\right)=0 \quad \text{on } \Gamma.$$

Thus flux is continuous across Γ. Note that in the classical Stefan problem it is only the *sum* of the preceding values which vanishes, thereby giving a constraint on the velocity $n_t/\|(n_1,\cdots,n_N)\|$ of Γ. The regularity of a generalized solution of (3.1) will not permit a nonstationary singularity.

Finally, we note that the above remarks have physically meaningful consequences for the thermal conduction model (2.6). In contrast to the completely contrary property of the classical Stefan problem, the solution of (2.6) will permit the appearance of a mush zone even if one were not present initially and no outside sources are present. Moreover, Theorem 3.2(a) implies that such mush regions always form over large regions from initial conditions containing both pure ice ($u=0$) and water at positive temperature.

REFERENCES

[1] G. Barenblatt, I. Zheltov and I. Kochina, *Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks*, J. Appl. Math. Mech., 24 (1960), pp. 1286–1303.

[2] M. Böhm and R. E. Showalter, *A nonlinear pseudoparabolic diffusion equation*, this Journal, 16 (1985), to appear.

[3] H. Brezis, *Operateurs maximaux monotones et semigroupes de contractions dans les espaces d'Hilbert*, Math. Studies 5, North Holland-American Elsevier, New York, 1973.

[4] H. Brezis and W. Strauss, *Semi-linear second-order elliptic equations in L^1*, J. Math. Soc. Japan, 25 (1973), pp. 565–590.

[5] P. J. Chen and M. E. Gurtin, *On a theory of heat conduction involving two temperatures*, Z. Angew. Math. Phys., 19 (1968), pp. 614–627.

[6] R. E. Collins, *Flow of Fluids Through Porous Materials*, Reinhold, New York, 1961.

[7] E. DiBenedetto and R. E. Showalter, *Implicit degenerate evolution equations and applications*, this Journal, 12 (1981), pp. 731–751.

[8] ———, *A pseudo-parabolic variational inequality and Stefan problem*, Nonlinear Anal., Theo., Meth. & Appl., 6 (1982), pp. 279–291.

[9] ———, *A free-boundary problem for a degenerate parabolic system*, J. Differential Equations, 50 (1983), pp. 1–19.

[10] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1977.

[11] A. E. Scheidegger, *The Physics of Flow through Porous Media*, Macmillan, New York, 1957.

[12] R. E. Showalter, *Existence and representation theorems for a semilinear Sobolev equation in Banach space*, this Journal, 3 (1972), pp. 527–543.

[13] ———, *Mathematical formulation of the Stefan problem*, Int. J. Engng. Sci., 20 (1982), pp. 909–912.

[14] ———, *Local regularity, boundary values and maximum principles for pseudoparabolic equations*, Applicable Anal., 16 (1983), pp. 235–241.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

A NONLINEAR PROBLEM IN AGE-DEPENDENT POPULATION DIFFUSION*

MICHEL LANGLAIS[†]

Abstract. A nonlinear partial differential equation arising in a model of age-dependent population diffusion with random dispersal is analysed. Using the nonlinear parabolic equation satisfied by the spatial structure and a delay to handle nonlinearities we derive existence results; in simple situations we outline the asymptotic behaviour.

1. Introduction. We are interested in a mathematical model of an age dependent population moving in a limited environment Ω in $\mathbb{R}^N$. The age-space structure of the population is described through the age distribution $u(t,a,x)$ where $t>0$ is time, a is age: $0<a<A$ (A is the maximum life expectancy of the species) and x in Ω is the spatial position. An integration over all ages yields the spatial density:

$$P(t,x)=\int_0^A u(t,a,x)\,da. \tag{1.1}$$

Studying the rate of change of individuals of age a in the framework developed in Gurtin [8] leads to the balance law

$$u_t+u_a=-\operatorname{div} q-\mu(a,P)u,$$

where q is the flux of population due to dispersal and $s(t,a,x)=-\mu(a,P)u$ represents the supply of individuals due here only to deaths (μ is the mortality modulus).

We consider random dispersal and assume that $q(t,a,x)=-k(P)\operatorname{grad} u$, $k(P)>0$ is the dispersal modulus, so that the local flow of population lies in the direction of decreasing density (to contrast with the directed dispersal $q(t,a,x)=-uk(P)\operatorname{grad} P$ of [9], see also [10]). Hence u obeys the partial differential equation

$$u_t+u_a-\operatorname{div}(k(P)\operatorname{grad} u)+\mu(a,P)u=0. \tag{1.2}$$

The birth process takes the form:

$$u(t,0,x)=\int_0^A \beta(a,P)u(t,a,x)\,da \tag{1.3}$$

where β is the maturity function. We also assume that there is no migration through the boundary $\partial\Omega$ of Ω

$$k(P)\overline{\operatorname{grad}}\, u\cdot\vec{\eta}=0 \quad \text{on } \partial\Omega \quad (\eta=\text{normal vector}). \tag{1.4}$$

The problem is now to determine the evolution of u starting at time $t=0$ with the initial distribution

$$u(0,a,x)=u_0(a,x). \tag{1.5}$$

*Received by the editors December 30, 1981, and in final revised form October 31, 1983.

[†] U. E. R. de Mathématiques et Informatique, Université de Bordeaux I, 33405 Talence Cedex, France. This work was conducted while the author was at the Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

When the dispersal modulus k is a constant existence and uniqueness of solutions are analyzed in Di Blasio [4] for $A=+\infty$ and in Garroni and Langlais [6] for finite A, linear β and μ and u are subject to a constraint. See also Gopalsamy [7] and Marcati [17] (and Webb [20] and its bibliography for the spatially homogeneous case: $k=0$). Here we treat the case where $k(P)>0$ for $P\geqq 0$ (but see also Remark 3.2 in §3) and k continuous. Our main tool is the auxiliary equation satisfied by P: assuming either $A=+\infty$ or $A<+\infty$ and $u(t,A,x)=0$ and integrating (1.2) over all ages yields

$$P_t-\operatorname{div}(k(P)\operatorname{grad}P)=\int_0^A[\beta(P,a)-\mu(P,a)]u\,da, \tag{1.6}$$

or alternatively, introducing $K(P)$ the antiderivative of k vanishing at the origin:

$$P_t-\Delta K(P)=\int_0^A[\beta(P,a)-\mu(P,a)]u\,da. \tag{1.7}$$

Furthermore

$$\begin{aligned}&P(0,x)=\int_0^A u_0(a,x)\,da=p_0(x),\\ &k(P)\overline{\operatorname{grad}\vec{P}}\cdot\eta=0.\end{aligned} \tag{1.8}$$

If β and μ are independent of a the right-hand side of (1.7) reads $[\beta(P)-\mu(P)]P$; hence (1.7) becomes a nonlinear parabolic equation with a solution P. If that P is substituted into (1.2) one has a linear equation with a variational solution u. To conclude we must check (1.1).

The general case is more involved. From (1.7) the Hölder continuity of P is easy to derive; this allows one to get existence results under rather weak assumptions concerning the dependence of β and μ on P (namely continuity and some control on the growth).

2. Basic notation. We take $A=+\infty$.

Ω is a bounded open domain in $\mathbb{R}^N$ with smooth boundary $\partial\Omega$; η is the unit normal pointing outward, ∇ the gradient vector in $\mathbb{R}^N$ so that $\nabla\cdot\eta=\partial/\partial\eta$ is the normal derivative.

We set $\mathcal{O}=(0,T)\times(0,\infty)$, $Q=\mathcal{O}\times\Omega$ with generic element (t,a,x). Given τ and A_0 with $0<\tau\leqq T$, $0<A_0<\infty$, $\mathcal{O}_0$ will stand fror $(0,\tau)\times(0,A_0)$. Here T is a positive real number.

1_F is the characteristic function of the set F (used with $F=(0,\tau)$ or $\mathcal{O}_0$).

2.1. Main assumptions. The dispersal modulus k is continuous, $K(p)=\int_0^p k(\sigma)\,d\sigma$ and we assume

$$0<k_0\leqq k(p),\qquad p\in\mathbb{R}. \tag{2.1}$$

β is a measurable function $Q\times\mathbb{R}\to\mathbb{R}$, continuous with respect to p and

$$\begin{aligned}&0\leqq\beta(t,a,x,p)\leqq\bar{\beta}\quad\text{in } Q\times[0,\infty),\\ &0\leqq\beta(t,a,x,p)\leqq\beta_1(t,a,x)\cdot\beta_2(p),\ \beta_2 \text{ bounded for bounded } p.\end{aligned} \tag{2.2}$$

The initial distribution u_0 verifies

$$
(2.3)\qquad \begin{aligned} &u_0(a,x)\geqq 0, \qquad u_0\in L^2((0,\infty)\times\Omega),\\ &0\leqq p_0(x)=\int_0^\infty u_0(a,x)\,da\leqq M_0, \qquad \nabla p_0\in L^2(\Omega); \end{aligned}
$$

the boundedness of p_0 implies that u_0 lies in $L^1((0,\infty)\times\Omega)$. As to μ we assume it takes the form (see [10])

$$
\mu(t,a,x,p)=\mu_n(t,a,x,p)+\mu_e(t,x,p) \quad \text{on } Q\times\mathbb{R},
$$

where μ_n and μ_e are real measurable functions, continuous with respect to p and

$$
(2.4)\qquad \begin{aligned} &0\leqq \mu_n(t,a,x,p)\leqq \mu_{n1}(t,a,x)\mu_{n2}(p), \quad \mu_{n2} \text{ bounded for bounded } p,\\ &0\leqq \mu_e(t,x,p)\leqq \bar{\mu}_e(p), \quad \bar{\mu}_e \text{ bounded for bounded } p. \end{aligned}
$$

More precise assumptions concerning the behavior of β_1 and μ_{n1} with respect to the variable a are made in §3. The boundedness of β will ensure global existence.

2.2. Weak solutions. We must define a suitable notion of solution. By a weak solution of problem (1.1)–(1.5) we mean any nonnegative and integrable function u defined on Q satisfying

$$
\int_Q \left(u^2+|\nabla u|^2\right) dt\,da\,dx < +\infty, \quad P \text{ continuous in } (0,T)\times\Omega,
$$

$$
0\leqq P(t,x)=\int_0^\infty u(t,a,x)\,da\leqq M_0e^{\bar{\beta}t} \quad \text{in } (0,T)\times\Omega,
$$

and such that for any ϕ in $C^1(\bar{Q})$ vanishing at $t=T$ and for large a

$$
\begin{aligned} (2.5)\qquad &\int_Q \left[-(\phi_t+\phi_a)u+\mu(a,P)u\cdot\phi+k(P)\nabla u\cdot\nabla\phi\right] dt\,da\,dx\\ &=\int_{(0,A)\times\Omega} u_0(a,x)\phi(0,a,x)\,da\,dx\\ &\quad+\int_{(0,T)\times\Omega}\int_0^\infty \beta(a,P)\cdot u\,da\cdot\phi(t,0,x)\,dt\,dx. \end{aligned}
$$

This is derived from (1.1)–(1.5) upon integrating by parts. Throughout this work we write $\mu(a,P)$ (resp. $\beta(a,P)$) for $\mu(t,a,x,P(t,x))$ (resp. $\beta(t,a,x,P(t,x))$).

3. Main results. In the simple case where β and μ are age independent and for smooth data, we have

THEOREM 3.1. *Assume* (k,β,u_0,μ) *satisfy* (2.1)–(2.4). *Furthermore let* k *lie in* $C^2(\mathbb{R})$, *let* p_0 *be in* $C^{2+\delta}(\bar{\Omega})$ *with* $p_{0_\eta}=0$ *on* $\partial\Omega$, *assume* $\mu_n=0$, β *independent of the variable* a *and*

$$
\begin{aligned} &\mu_t,\ \mu_p,\ \beta_t,\ \beta_p \textit{ exist and are continuous on } [0,T]\times\bar{\Omega}\times\mathbb{R},\\ &\mu \textit{ and } \beta \textit{ are Hölder continuous in } x \textit{ with exponent } \delta. \end{aligned}
$$

Then problem (1.1)–(1.5) *has a unique weak solution and* P *belongs to* $C^{1+\delta/2,2+\delta}([0,T]\times\bar{\Omega})$.

The proof is found in §5.3; it can be shown that the solution is smoother than requested, namely u_t+u_a and Δu are square integrable over Q.

We now pass to the general case which is the main result of this paper.

THEOREM 3.2. *Assume* (k,β,u_0,μ) *satisfy* (2.1)–(2.4) *and*

$$0\leqq \mu_{n1}(t,a,x)\leqq \bar{\mu}_n<+\infty \quad in\ Q,$$
$$\int_0^{\infty}[\beta_1^2+\mu_{n1}^2](t,a,x)\,da\leqq C_1<+\infty \quad in\ (0,T)\times\Omega. \tag{3.1}$$

Then problem (1.1)–(1.5) *has a weak solution.*

The proof is found in §6. We also prove that P satisfies the initial boundary value problem (1.7), (1.8) and P_t, ∇P, $\Delta K(P)$ lie in $L^2((0,T)\times\Omega)$; further for any w in the Sobolev space of order 1 $H^1((0,T)\times\Omega)$

(3.2)

$$\int_{(0,T)\times\Omega}[P_tw+k(P)\nabla P\cdot\nabla w]\,dt\,dx=\int_{(0,T)\times\Omega}\int_0^{\infty}[\beta(a,P)-\mu(a,P)]u\,da\cdot w\,dt\,dx.$$

To this quasilinear parabolic equation the results of [12] apply; due to the properties assigned to k, β, μ and those obtained for P we conclude to the existence of a δ, $0<\delta<1$, such that P lies in the Hölder space $C^{\delta/2,\delta}((0,T)\times\bar{\Omega})$ and $C^{\delta/2,\delta}([0,T]\times\bar{\Omega})$ provided p_0 belong to $C^{\delta}(\bar{\Omega})$.

Remark 3.1. Let λ be a constant with

$$\beta(t,a,x,p)-\mu(t,a,x,p)+\lambda\geqq 0 \quad \text{in } Q\times\left[0,M_0e^{\bar{\beta}T}\right],$$

then P satisfies $P_t-\operatorname{div}(k(P)\nabla P)+\lambda P=F$ in $(0,T)\times\Omega$, $F(t,x)=\int_0^{\infty}(\beta(a,P)-\mu(a,P)+\lambda)u\,da\geqq 0$; using the minimum principle for parabolic equations we conclude that under the assumptions of Theorems 3.1, 3.2, whenever $0<m_0\leqq p_0(x)$ there exists a constant m_1 such that

$$0<m_1\leqq P(t,x) \quad \text{in } (0,T)\times\Omega. \tag{3.3}$$

Remark 3.2. When condition (2.1) is replaced by $k(P)=P^{m-1}$, $m>1$, the situation is quite different: (1.6) is no more parabolic but more commonly referred to as a porous medium type equation degenerating when $P=0$; see Oleinik [19], Aronson [2]. Obviously any additional assumptions ensuring that P stays away from 0 will lead to conclusions similar to the one drawn above: using Remark 3.1 and a device of [2] we can conclude that for $k(P)=P^{m-1}$, $m>1$, the conclusions of Theorems 3.1, 3.2 are still valid if $0<m_0\leqq p_0(x)$.

Lastly we briefly discuss the asymptotic behavior of P and u when

$$c(t,a,x,p)=\mu(t,a,x,p)-\beta(t,a,x,p)$$

takes simple forms ($-c$ is the growth rate).

THEOREM 3.3. *Under the assumptions of either Theorem* 3.1 *or* 3.2, *then* $c(t,a,x,p)\leqq c_0<0$ *implies* $u(t,\cdot,\cdot)\rightarrow_{t\rightarrow\infty}+\infty$ *in* $L^1((0,A)\times\Omega)$; $c(t,a,x,p)\geqq c_0>0$ *implies* $u(t,\cdot,\cdot)\rightarrow_{t\rightarrow\infty}0$ *in* $L^1((0,A)\times\Omega)$ *and* $P(t,\cdot)\rightarrow_{t\rightarrow\infty}0$ *in* $L^{\infty}(\Omega)$; $c(t,a,x,p)=0$ *implies*

$$P(t,\cdot)\underset{t\rightarrow\infty}{\rightarrow}\frac{1}{\operatorname{mes}\Omega}\int_{\Omega}p_0(x)\,dx \quad in\ L^{\infty}(\Omega).$$

The proof is found in §8. This is not surprising, the interesting situation occurring when c is not of one sign.

4. Abstract formulation of the problem. We need a more abstract notion of solution.

Let $H^1(\Omega)$ be the usual Sobolev space of order 1 over Ω; we denote by $\langle\ ,\ \rangle$ the duality pairing between $H^1(\Omega)$ and its dual space $[H^1(\Omega)]'$.

U being either $(0,T)$ or $\mathcal{O}$, and H either $H^1(\Omega)$ or its dual space, $L^2(U;\ H)$ is the Hilbert space of those measurable and square integrable functions $v: U \to H$. Set $\mathscr{V}= L^2(\mathcal{O};\ H^1(\Omega))$, $\mathscr{V}'=$ dual space of $\mathscr{V}=L^2(\mathcal{O};\ [H^1(\Omega)]')$, $\mathscr{W}=L^2(0,T;\ H^1(\Omega))$ and $\mathscr{W}'=$ dual space of $\mathscr{W}=L^2(0,T;\ [H^1(\Omega)]')$.

In the following ∂_t (resp. ∂_t, ∂_a) indicates partial differentiation in $\mathscr{D}'(0,T;\ [H^1(\Omega)]')$ (resp. $\mathscr{D}'(\mathcal{O};\ [H^1(\Omega)]')$): see Lions and Magenes [16], and D stands for $\partial_t+\partial_a$.

If we choose for test function in (2.5) an element φ in $\mathscr{D}(\mathcal{O};\ C^1(\bar{\Omega}))$—this is compactly supported in $\mathcal{O}$—the integrals on $t=0$ and $a=0$ disappear. It follows that the linear function

$$\varphi \to \int_Q \left[-\left(\frac{\partial\varphi}{\partial t} + \frac{\partial\varphi}{\partial a} \right) u + \mu u \varphi \right] dt\, da\, dx$$

is continuous on $\mathscr{D}(\mathcal{O};\ H^1(\Omega))$ equipped with the topology of $\mathscr{V}$and by density it can be extended to $\mathscr{V}$. This merely means that $Du+\mu u$ belongs to $\mathscr{V}'$. Hence a weak solution in the sense of §2 verifies

$$\begin{array}{lll} & u(t,a,x) \geqq 0 \quad \text{in } Q, & u \in L^1(Q) \cap \mathscr{V}, \\ (4.1) & 0 \leqq P(t,x) \leqq M_0 e^{\bar{\beta}T}, & P \in C^0((0,T)\times\Omega), \\ & Du+\mu(a,P)u \in \mathscr{V}', & \end{array}$$

and satisfies for any v in $\mathscr{V}$

$$(4.2) \qquad \int_{\mathcal{O}} \langle Du+\mu(a,P)u, v\rangle\, dt\, da + \int_Q k(P) \nabla u \cdot \nabla v\, dt\, da\, dx = 0.$$

Now we are in the situation described in [6, Lemma 0]: setting $\mathcal{O}_0=(0,A_0)\times(0,T)$, $0<A_0<\infty$ for any weak solution Du lies in $L^2(\mathcal{O}_0;\ [H^1(\Omega)]')$ so that the trace of a weak solution on $t=t_0$ (or $a=a_0$) makes sense in $L^2((0,A_0)\times\Omega)$ (or $L^2((0,T)\times\Omega)$). Furthermore the trace applications are continuous for the weak topologies and for any u, v in $L^2(\mathcal{O}_0;\ H^1(\Omega))$ such that Du, Dv belong to $L^2(\mathcal{O}_0;\ [H^1(\Omega)]')$ we have

(4.3)

$$\int_{\mathcal{O}_0} [\langle Du,v\rangle + \langle Dv,u\rangle]\, dt\, da = \int_{(0,A_0)\times\Omega} [uv(T,a,x)-uv(0,a,x)]\, da\, dx$$

$$+ \int_{(0,T)\times\Omega} [uv(t,A_0,x)-uv(t,0,x)]\, dt\, dx.$$

An alternative definition for a weak solution is now available: a function u is a weak solution in the sense of §2.2 if and only if it satisfies (4.1), (4.2)

$$\begin{array}{ll} & u(0,a,x)=u_0(a,x), \\ (4.4) & u(t,0,x)=\displaystyle\int_0^\infty \beta(t,a,x,P(t,x))u(t,a,x)\, da, \end{array}$$

as well as

$$(4.5) \qquad \int_0^\infty u(t,a,x)\,da = P(t,x).$$

One can pass back and forth between the two notions of weak solution upon using (4.3) and the density in $\mathscr{V}$of the φ used in §2.

5. Preliminary results and proof of Theorem 3.1. We derive various estimates for problems (1.1)–(1.5) and (1.6)–(1.8) (when not coupled), from which the proof of Theorem 3.1 follows at once.

5.1. The linear case. Let $m=m(t,a,x)$, $b=b(t,a,x)$ and $U=U(t,x)$, namely U is independent of the variable a, be given measurable functions with:

$$(5.1) \qquad \begin{aligned} &0 \leqq m(t,a,x) \leqq \bar{m} < +\infty, \qquad 0 \leqq b(t,a,x) \leqq \bar{\beta} < +\infty \quad \text{in } Q,\\ &0 < m_1 \leqq U(t,x) \leqq M_1 < +\infty \quad \text{in } (0,T)\times\Omega. \end{aligned}$$

We look for a function u satisfying

$$(5.2) \qquad \begin{aligned} &u \in \mathscr{V}, \quad Du \in \mathscr{V}', \quad \text{for any } v \text{ in } \mathscr{V},\\ &\int_{\mathcal{O}} \langle Du, v\rangle\,dt\,dx + \int_Q [U\nabla u\cdot\nabla v + muv]\,dt\,da\,dx = 0 \end{aligned}$$

and taking the initial values

$$(5.3) \qquad \begin{aligned} &u(0,a,x) = u_0(a,x) \quad \text{in } (0,\infty)\times\Omega,\\ &u(t,0,x) = \int_0^\infty b(t,a,x)\cdot u(t,a,x)\,da \quad \text{in } (0,T)\times\Omega. \end{aligned}$$

In some sense (5.2), (5.3) are a linearization of our problem (1.1), (1.5).

THEOREM 5.1. *Assume* (m,b,U) *verify* (5.1) *and*

$$(5.4) \qquad \int_0^\infty b^2(t,a,x)\,da \leqq C_2 < +\infty \quad \textit{in } (0,T)\times\Omega.$$

For any nonnegative u_0 *satisfying conditions* (2.3) *there exists a unique nonnegative and integrable solution of* (5.2), (5.3).

The proof is found in the Appendix.

We need qualitative properties of the integral with respect to the variable a of u. Set

$$z(t,x) = \int_0^\infty u(t,a,x)\,da$$

that is a nonnegative element of $L^1((0,T)\times\Omega)$.

COROLLARY 5.1. *Under the assumptions of Theorem* 5.1 *z belongs to* $\mathscr{W}$, $\partial_t z$ *to* $\mathscr{W}'$ *and* $z(0,x) = p_0(x) = \int_0^\infty u_0(a,x)\,da$. *Furthermore for any w in* $\mathscr{W}$

$$(5.5) \quad \int_0^T \langle \partial_t z, w\rangle\,dt + \int_{(0,T)\times\Omega} U\nabla z\cdot\nabla w\,dt\,dx = \int_{(0,T)\times\Omega}\int_0^\infty (b-m)u\,da\cdot w\,dt\,dx.$$

Formally the parabolic equation for z is obtained upon integrating (5.2) with respect to the variable a.

Proof. Let φ_n be a sequence of smooth function $[0,\infty)\to\mathbb{R}$ with

$$0\leqq\varphi_n(a)\leqq 1,\qquad \varphi_n(a)=1,\quad 0\leqq a\leqq n,$$

$$\varphi_n(a)=0,\quad n+1\leqq a,\qquad \left|\frac{d\varphi_n}{da}(a)\right|\leqq C_1\ \text{ independent of } a \text{ and } n.$$

The dominated convergence theorem ensures that strongly in $L^1((0,T)\times\Omega)$

$$z_n(t,x)=\int_0^\infty \varphi_n(a)\cdot u(t,a,x)\,da\underset{n\to\infty}{\to} z(t,x).$$

On the other hand z_n lies in $\mathscr{W}$ because φ_n is compactly supported. For any ψ in $\mathscr{D}(0,T;\,H^1(\Omega))$ one has

$$-\int_{(0,T)\times\Omega} z_n\cdot\frac{\partial\psi}{\partial t}\,dt\,dx=\int_{\mathscr{O}}\langle Du,\varphi_n\psi\rangle\,dt\,da+\int_{(0,T)\times\Omega}\int_0^\infty u\frac{d\varphi_n}{da}da\cdot\psi\,dt\,dx;$$

this is actually derived upon using (4.3) to integrate by parts the integral over $\mathscr{O}$. We conclude that the linear mapping

$$\psi\to-\int_{(0,T)\times\Omega} z_n\cdot\frac{\partial\psi}{\partial t}\,dt\,dx$$

is continuous over $\mathscr{D}(0,T;\,H^1(\Omega))$ equipped with the topology of $\mathscr{W}$and therefore $\partial_t z_n$ belongs to $\mathscr{W}'$.

Next, again using relation (4.3), we obtain for any w in $H^1((0,T)\times\Omega)$

$$\begin{aligned}\int_{\mathscr{O}}\langle Du,\varphi_n w\rangle\,dt\,da=&-\int_{(0,T)\times\Omega} z_n\frac{\partial w}{\partial t}\,dt\,dx\\&+\int_{(0,\infty)\times\Omega}[u(T,a,x)w(T,x)-u(0,a,x)w(0,x)]\,\varphi_n(a)\,da\,dx\\&-\int_{(0,T)\times\Omega}u(t,0,x)w(t,x)\,dt\,dx-\int_Q\frac{d\varphi_n}{da}wu\,dt\,da\,dx\\&=\int_0^T\langle\partial_t z_n,w\rangle\,dt-\int_{(0,T)\times\Omega}\int_0^\infty\left[b+\frac{d\varphi_n}{da}\right]u\,da\cdot w\,dt\,dx.\end{aligned}$$

Selecting $v=\varphi_n w$, w in $\mathscr{W}$, as test function in (5.2) the foregoing relation and the density of $H^1((0,T)\times\Omega)$ in $\mathscr{W}$imply

$$(*)\qquad \int_0^T\langle\partial_t z_n,w\rangle\,dt+\int_{(0,T)\times\Omega}U\nabla z_n\cdot\nabla w\,dt\,dx=\int_{(0,T)\times\Omega}f_n\cdot w\,dt\,dx,$$

where

$$f_n(t,x)=\int_0^\infty\left(b+\frac{d\varphi_n}{da}\right)u\,da-\int_0^\infty m\varphi_n u\,da.$$

Obviously $z_n(0,x)=\int_0^\infty\varphi_n(a)u_0(a,x)\,da\leqq p_0(x)$.

Taking $w=z_n\cdot 1_{(0,\tau)}$, $0<\tau<T$, as test function in $(*)$, integrating by parts, using the nonnegativity of m, φ_n, u, z_n and the properties required on b and φ_n yields

$$\begin{aligned}\frac{1}{2}\int_\Omega z_n^2(\tau,x)\,dx+\int_{(0,\tau)\times\Omega}U|\nabla z_n|^2\,dt\,dx&\\ \leqq \frac{1}{2}\int z_n^2(0,x)\,dx+\int_{(0,\tau)\times\Omega}\int_0^\infty\left(b+\frac{d\varphi_n}{da}\right)u\,da\cdot z_n\,dt\,dx&\\ \leqq \frac{1}{2}\int_\Omega P_0^2(x)\,dx+\frac{1}{2}(c_2+c_3^2)\int_Q u^2\,dt\,da\,dx+\int_{(0,\tau)\times\Omega}z_n^2\,dt\,dx.&\end{aligned}$$

With Gronwall's inequality we conclude that z_n is bounded in $\mathscr{W}$ and therefore z lies in $\mathscr{W}$.

Clearly

$$\int_0^\infty\frac{d\varphi_n}{da}u\,da\underset{n\to\infty}{\to}0\quad\text{strongly in } L^2((0,T)\times\Omega).$$

From (5.1) and the estimate for z_n

$$0\leqq\int_0^\infty m\varphi_n u\,da\leqq\overline{m}z_n\quad\text{is bounded in } L^2((0,T)\times\Omega);$$

but for any continuous function θ on $[0,T]\times\overline{\Omega}$, the dominated convergence theorem provides

$$\int_Q m\varphi_n u\theta\,dt\,da\,dx\underset{n\to\infty}{\to}\int_Q mu\theta\,dt\,da\,dx$$

(because u lies in $L^1(Q)$) and we conclude that $\int_0^\infty m\varphi_n u\,da$ converges to $\int_0^\infty mu\,da$ in $L^2((0,T)\times\Omega)$, weakly. Hence f_n is weakly converging to $\int_0^\infty(b-m)u\,da$ in $L^2((0,T)\times\Omega)$.

By letting n go to $+\infty$ we obtain the properties listed in Corollary 5.1.

COROLLARY 5.2. *Under the assumptions of Theorem* 5.1:

$$0\leqq z(t,x)\leqq M_0\cdot e^{\overline{\beta}t}\quad in\ (0,T)\times\Omega.$$

This implies that the uniform norm of z does not depend on U and m, and any b verifying (5.1).

Proof. From the nonnegativity of m, b and u we conclude that

$$\int_0^\infty(b-m)u\,da\leqq\overline{\beta}z\quad\text{in } (0,T)\times\Omega$$

and the conclusion follows from the weak maximum principle. Actually for any $\varepsilon>0$ let

$$r(t,x)=z(t,x)\cdot e^{-(\overline{\beta}+\varepsilon)t}\quad\text{in } (0,T)\times\Omega.$$

This new function r satisfies $r(0,x)=p_0(x)$ and for any w nonnegative function in $\mathscr{W}$

$$\int_0^T\langle\partial_t r,w\rangle\,dt+\int_{(0,T)\times\Omega}[U\nabla r\cdot\nabla w+\varepsilon r\cdot w]\,dt\,dx\leqq 0.$$

Defining $s(t,x)=r(t,x)-M_0$ (M_0 is defined in (2.3)) and choosing $w=s^+=\mathrm{Max}(s,0)$ in the inequality for r, an integration by parts leads to

$$\varepsilon\int_{(0,T)\times\Omega}|s^+|^2 dt\,dx \leqq 0$$

say $s^+=0$ or $r(t,x)\leqq M_0$ and the desired result is obtained by letting ε go to 0.

A similar form of the foregoing results is also needed.

THEOREM 5.2. *Let (m,B,U) be independent of the variable a and satisfy* (5.1).

For any u_0 as in (2.3) *there exists a unique nonnegative and integrable solution of* (5.2) *verifying*

$$\begin{aligned} &u(0,a,x)=u_0(a,x) && \textit{in } (0,\infty)\times\Omega,\\ &u(t,0,x)=B(t,x) && \textit{in } (0,T)\times\Omega. \end{aligned}$$

Furthermore the conclusions of Corollary 5.1 *still hold, the right-hand side of* (5.5) *being replaced by*

$$\int_{(0,T)\times\Omega}[B(t,x)-m(t,x)\cdot z]\,w\,dt\,dx.$$

Existence and uniqueness are derived in the Appendix. For qualitative properties we reproduce the proof of Corollary 5.1 with minor changes.

Remark 5.1. When $m(t,a,x)=m_1(t,a,x)+m_2(t,x)$ the right-hand side of (5.5) becomes

$$\int_{(0,T)\times\Omega}\left[m_2 z+\int_0^\infty(b-m_1)\cdot u\,da\right]w\,dt\,dx.$$

5.2. Remarks on a quasilinear parabolic equation. We will use special properties of the solution to the problem

$$(5.6)\qquad \begin{aligned} &P_t-\mathrm{div}(k(P)\nabla P)+\lambda P=F && \text{in } (0,T)\times\Omega,\\ &P(0,x)=p_0(x) && \text{in } \Omega,\\ &P_\eta=0 && \text{on } (0,T)\times\partial\Omega, \end{aligned}$$

where λ is a positive real number, F and p_0 being nonnegative and bounded functions. For classical solutions we have the next theorem.

THEOREM 5.3 ([12]). *Let F be in $C^1([0,T]\times\overline{\Omega})$ and nonnegative, p_0 be in $C^{2+\delta}(\overline{\Omega})$, $p_{0\eta}=0$ on $\partial\Omega$, and nonnegative, k be in $C^2([0,\infty))$ and $k(p)\geqq k_0>0$. Then there exists a unique nonnegative P in $C^{1+\delta/2,2+\delta}([0,T]\times\overline{\Omega})$ solution of* (5.6).

Equation (5.6) can be solved under much weaker assumptions. Recalling that K is the antiderivative of k vanishing at the origin we can rewrite our problem as

$$(5.7)\qquad \begin{aligned} &P_t-\Delta K(P)+\lambda P=F && \text{in } (0,T)\times\Omega,\\ &P(0,x)=p_0(x) && \text{on } \Omega,\\ &\frac{\partial}{\partial\eta}K(P)=0 && \text{on } (0,T)\times\partial\Omega. \end{aligned}$$

Nonlinear semigroup theory is available to handle (5.7); see Evans [5] for example, to which we refer for uniqueness.

THEOREM 5.4. *Assume k satisfies* (2.1), p_0 *verifies* (2.3) *and F is nonnegative and bounded. Then there exists a unique nonnegative and bounded P, continuous in* $(0,T)\times\Omega$, *with*

$$P_t,\ \nabla P,\ \Delta K(p) \quad in\ L^2((0,T)\times\Omega)$$

solution of (5.7) (*for* $\lambda>0$).

Proof. To get existence we approximate (k,p_0,F) by smooth functions (k_n,p_{0n},F_n) satisfying the conditions of Theorem 5.3, $k_0\leqq k_n(p)$ in $p\geqq 0$, $0\leqq p_{0n}(x)\leqq p_0(x)$ in Ω, $0\leqq F_n(t,x)\leqq F(t,x)$ in $(0,T)\times\Omega$ and

$$\begin{aligned} &k_n \underset{n\to\infty}{\to} k && \text{uniformly on compact subsets of } [0,\infty),\\ &F_n \underset{n\to\infty}{\to} F && \text{strongly in } L^2((0,T)\times\Omega),\\ &p_{0n} \underset{n\to\infty}{\to} p_0 && \text{strongly in } H^1(\Omega). \end{aligned}$$

Let P_n be the unique classical solution of (5.6) with (k,F,p_0) replaced by (k_n,F_n,p_{0n}). Keeping in mind that λ is a positive constant, we deduce from the maximum principle

$$0\leqq P_n(t,x)\leqq \operatorname{Max}\left(M_0,\lambda^{-1}\|F\|_{L^\infty((0,T)\times\Omega)}\right).$$

K_n being the antiderivative of k_n vanishing at the origin, P_n is easily seen to satisfy

$$P_{nt}-\Delta K_n(P_n)=F_n-\lambda P_n. \tag{5.8}$$

If we multiply this relation with $k_n(P_n)\Delta k_n(P_n)$ and integrate over $(0,\tau)\times\Omega$, $0<\tau<T$, we obtain

$$\begin{aligned} \frac{1}{2}\int_\Omega|\nabla K_n(P_n)|^2(\tau,x)\,dx+\int_{(0,\tau)\times\Omega}k_n(P_n)|\Delta K_n(P_n)|^2dt\,dx \\ =\frac{1}{2}\int_\Omega|\nabla K_n(p_{0n})|^2dx-\int_{(0,\tau)\times\Omega}(F_n-\lambda P_n)k_n(P_n)\Delta K_n(P_n)\,dt\,dx. \end{aligned}\tag{5.9}$$

P_n is bounded in $L^\infty((0,T)\times\Omega)$; it follows that $\nabla K_n(P_n)$ and $\Delta K_n(P_n)$ are bounded in $L^2((0,T)\times\Omega)$ and from relation (5.8) and the properties of k_n

$$P_{nt},\ \nabla P_n,\ \Delta K_n(P_n) \quad \text{are bounded in } L^2((0,T)\times\Omega).$$

Finally [12, Thm. 1.1, p. 419] applies to (5.8) providing an α, $0<\alpha<1$, such that P_n is bounded in $C^{\alpha/2,\alpha}((0,T)\times\Omega)$.

The existence is obtained by letting n go to $+\infty$, using the compactness of $C^{\alpha/2,\alpha}((0,T)\times\Omega)$ in $C^0((0,T)\times\Omega)$ to pass to the limit in the nonlinear term.

COROLLARY 5.3. *Under the assumptions of Theorem* 5.4 (*for* $\lambda>0$), *P verifies the estimate* $0\leqq P(t,x)\leqq\operatorname{Max}(M_0,\lambda^{-1}\|F\|_{L^\infty((0,T)\times\Omega)})$ *and satisfies the equation, for w in* $\mathscr{W}$,

$$\int_{(0,T)\times\Omega}[P_tw+k(P)\nabla P\cdot\nabla w+\lambda P\cdot w]\,dt\,dx=\int_{(0,T)\times\Omega}F\cdot w\,dt\,dx. \tag{5.10}$$

It suffices to pass to the limit in the corresponding relations for P_n.

COROLLARY 5.4. *Under the assumptions of Theorem* 5.4 *the following estimate holds, for* $0<\tau<T$,

$$\|\nabla P\|+\|P_t\|+\|\Delta K(P)\| \leqq \left(\|p_0\|_{H^1(\Omega)}+\|F-\lambda P\|\right)\cdot C\left(\|P\|_{L^\infty((0,T)\times\Omega)}\right)$$

where $\|\cdot\|$ *is the* $L^2((0,\tau)\times\Omega)$ *norm and* $C(\cdot)$ *a nondecreasing function of its argument.*

It is derived from (5.8) and estimate (5.9).

5.3. Proof of Theorem 3.2. When μ and β do not depend on the variable a P can be calculated directly from (1.6)–(1.8).

Under the conditions listed in Theorem 3.2 there exists a unique nonnegative P in $C^{1+\delta/2,2+\delta}([0,T]\times\bar{\Omega})$ solution of the quasilinear equation (see [12])

$$(5.11)\qquad \begin{aligned} &P_t-\operatorname{div}(k(P)\nabla P)+[\mu(P)-\beta(P)]P=0 && \text{in } (0,T)\times\Omega,\\ &P(0,x)=p_0(x) && \text{in } \Omega,\\ &P_\eta=0 && \text{on } (0,T)\times\partial\Omega. \end{aligned}$$

The boundedness of P implies that $U(t,x)=k(P(t,x))$, $m(t,x)=\mu(t,x,P(t,x))$ and $B(t,x)=\beta(t,x,P(t,x))$. $P(t,x)$ satisfy the assumptions of Theorem 5.2 supplying a unique u such that for any v in $\mathscr{V}$

$$(5.12)\qquad \begin{aligned} &\int_{\mathcal{O}}\langle Du,v\rangle\,dt\,du+\int_Q[k(P)\nabla u\cdot\nabla v+\mu(P)uv]\,dt\,da\,dx=0,\\ &u(0,a,x)=u_0(a,x),\\ &u(t,0,x)=\beta(t,x,P(t,x))P(t,x). \end{aligned}$$

Using (5.11) for P and Theorem 5.2 it is easily checked that P and $\int_0^\infty u(t,a,x)\,da$ are solutions of the same parabolic equation, namely, for any w in $\mathscr{W}$,

$$\int_0^T\langle\partial_t z,w\rangle\,dt+\int_{(0,T)\times\Omega}[k(P)\nabla z\cdot\nabla w+\mu(P)zw]\,dt\,dx=\int_{(0,T)\times\Omega}\beta(P)\cdot Pw\,dt\,dx.$$

Taking the same initial datum they coincide and we have proved existence.

To get uniqueness let u be a solution. From Theorem 5.2 $\int_0^\infty u(t,a,x)\,da$ satisfies (5.11) and is uniquely determined. Now (5.12) becomes a linear equation with a unique solution.

Remark 5.2. Along the lines of §5.2, (5.11) can be solved under weaker assumptions.

6. Proof of Theorem 3.2. We use a method involving a delay $h\geqq 0$; replacing $P(t,x)$ by $P^h(t-h,x)$ transforms (1.2), (1.5) into a linear equation as in §5. When $h\to 0$, $P^h\to P$ and by a continuity argument $\mu(P^h)$ and $\beta(P^h)$ converge to $\mu(P)$ and $\beta(P)$.

The proof given below requires P_0 to be Hölder continuous on $\bar{\Omega}$ to pass to the limit in terms containing a delay (namely R^h). Once the existence is granted for Hölder continuous p_0 the general case of p_0 satisfying (2.3) is obtained upon approximating u_0 by smooth functions; the proof of this last part is omitted due to its similarity with those detailed now and in the next section.

Let $M = M_0 e^{\bar{\beta} T}$ where M_0 and $\bar{\beta}$ are defined in §2 and let λ be a real number larger than 1 chosen so that

$$\begin{aligned}&\text{for any } (t,a,x) \in Q, \qquad 0 \leqq p \leqq \operatorname{Max}\big(M_0, (1+\bar{\beta})M\big),\\ &0 \leqq \beta(t,a,x,p) - \mu(t,a,x,p) + \lambda.\end{aligned}$$

Let h be a positive real number, the delay that will go to zero.

We claim that there exists a unique (u^h, P^h) having the following properties:

$$\begin{aligned}&0 \leqq P^h(t,x) \leqq \operatorname{Max}\big(M_0, (1+\bar{\beta})M\big),\\ (6.1)\qquad &P_t^h, \nabla P^h, \Delta K(P^h) \in L^2((0,T)\times\Omega),\\ (6.2)\qquad &u^h \in \mathscr{V} \cap L^1(Q), \qquad u^h \text{ nonnegative on } Q,\\ (6.3)\qquad &Du^h \in \mathscr{V}', \qquad u(0,a,x) = u_0(a,x) \quad \text{in } (0,\infty)\times\Omega,\\ (6.4)\qquad &0 \leqq z^h(t,x) = \int_0^\infty u^h(t,a,x)\,da \leqq M,\\ (6.5)\qquad &u^h(t,0,x) = \int_0^\infty \beta^h(t,a,x)u^h(t,a,x)\,da;\end{aligned}$$

where

$$\begin{aligned}&\beta^h(t,a,x) = \beta(t,a,x,R^h(t,x)) &&\text{in } Q,\\ &R^h(t,x) = P^h(t-h,x) &&\text{in } (0,T)\times\Omega;\end{aligned}$$

while u^h satisfies for any v in $\mathscr{V}$

$$(6.6)\qquad \int_{\mathscr{O}} \langle Du^n, v\rangle\,dt\,da + \int_Q \big[k(R^h)\nabla u^h \cdot \nabla v + \mu^h u^h v\big]\,dt\,da\,dx = 0;$$

where $\mu^h(t,a,x) = \mu(t,a,x,R^h(t,x))$ and P^h is the solution of

$$\begin{aligned}&P_t^h - \Delta K(P^h) + \lambda P^h = \int_{0 s \infty} \big[\beta^h - \mu^h + \lambda\big] u^h\,da,\\ (6.7)\qquad &P^h(t,x) = p_0(x) \quad \text{in } [-h,0]\times\Omega,\\ &\frac{\partial}{\partial\eta}K(P^h) = 0 \quad \text{on } (0,T)\times\partial\Omega.\end{aligned}$$

This is proved upon using repeatedly the results given in §5. First in $(0,h)\times\Omega$, $R^h(t,x) = p_0(x)$ and we can integrate the linear equations (6.5), (6.6) to get u^h in $(0,h)\times(0,\infty)\times\Omega$ satisfying (6.2), (6.3), (6.4); see Theorem 5.1 and its two corollaries. Next the right-hand side of (6.7) is nonnegative and bounded so that we can integrate it on $(0,h)\times\Omega$ and obtain P^h with (6.1); see Theorem 5.4 and its corollaries. Actually the only thing to be checked is the L^∞-norm but due to (6.4) and the nonnegativity of β^h, μ^h, and u^h:

$$0 \leqq \int_0^\infty (\beta^h - \mu^h + \lambda) u^h\,da \leqq (\bar{\beta} + \lambda)M$$

and the conclusion follows from Corollary 5.3 and $\lambda \geqq 1$. We can now iterate and integrate over Q (6.6) and (6.7).

In the course of proving the existence we have obtained that u^h is nonnegative, z^h satisfies (6.4) and P^h the uniform extimate in (6.1). We need more a priori estimates.

Set $\mathcal{O}_0=(0,A_0)\times(0,\tau)$, $0<A_0<+\infty$, $0<\tau<T$. With relation (4.3)

(6.8)

$$\int_{\mathcal{O}_0}\langle Du^h,u^h\rangle\,dt\,da=\frac{1}{2}\int_{(0,A_0)\times\Omega}\left[(u^h)^2(\tau,a,x)-(u^h)^2(0,a,x)\right]da\,dx$$
$$+\frac{1}{2}\int_{(0,\tau)\times\Omega}\left[(u^h)^2(t,A_0,x)-(u^h)^2(t,0,x)\right]dt\,dx.$$

We take $v=u^h.\ 1_{\mathcal{O}_0}$ as test function in (6.6). Using (6.8), the initial conditions in (6.3) and (6.5) and the nonnegativity of μ^h yields

$$\frac{1}{2}\int_{(0,A_0)\times\Omega}(u^h)^2(\tau,a,x)\,da\,dx+\int_{\mathcal{O}_0\times\Omega}k(R^h)|\nabla u^h|^2dt\,da\,dx$$
$$\leqq\frac{1}{2}\int_{(0,A_0)\times\Omega}u_0^2(a,x)\,da\,dx+\frac{1}{2}\int_{(0,\tau)\times\Omega}\left(\int_0^\infty\beta^hu^h\,da\right)^2dt\,dx.$$

The last term on the right-hand side is less than

$$C_4\int_{\mathcal{O}_0\times\Omega}(u^h)^2dt\,da\,dx.$$

This comes out from the boundedness of P^h and (2.2), (3.1). By letting A_0 go to ∞ we end up with, for any τ in $(0,T)$,

$$\int_{(0,\infty)\times\Omega}(u^h)^2(\tau,a,x)\,da\,dx+2k_0\int_{(0,\tau)\times(0,\infty)\times\Omega}|\nabla u^h|^2dt\,da\,dx \tag{6.9}$$
$$\leqq\int_{(0,\infty)\times\Omega}u_0^2\,da\,dx+2C_4\int_{(0,\tau)\times(0,\infty)\times\Omega}(u^h)^2dt\,da\,dx.$$

The Gronwall inequality implies that u^h is bounded in $L^2(Q)$ and it follows that u^h is also bounded in $\mathcal{V}$.

On the other hand the right-hand side of (6.7) has already been seen to be bounded in $L^\infty((0,T)\times\Omega)$ so that from Corollaries 5.3 and 5.4

$$P_t^h,\ \nabla P^h,\Delta K(P^h)\quad\text{are bounded in } L^2((0,T)\times\Omega).$$

Furthermore P^h is a solution of the parabolic equation, for any w in $\mathcal{W}$,

$$\int_{(0,T)\times\Omega}\left[P_t^hw+k(P^h)\nabla P^h\nabla w+\lambda P^hw\right]dt\,dx=\int_{(0,T)\times\Omega}F^hw\,dt\,dx$$

with $F^h=\int_0^\infty(\beta^h-\mu^h+\lambda)u^h\,da$ bounded in $L^\infty((0,T)\times\Omega)$. Hence from [12, Thm. 1.1, p. 419] P^h is uniformly bounded in the Hölder space $C^{\alpha/2,\alpha}([0,T)\times\Omega)$ for some α in $(0,1)$ if p_0 is Hölder continuous.

By construction R^h is also bounded in $C^{\alpha/2,\alpha}([0,T)\times\Omega)$.

Now we have enough estimates to pass to the limit. There exist a sequence h_n going to 0, we simply denote it h, and a bounded, continuous on $(0,T)\times\Omega$, function P such that

$$P^h\underset{h\to0}{\to}P\quad\text{uniformly on every compact subset of }(0,T)\times\Omega,$$

strongly in $L^2((0,T)\times\Omega)$ and weakly in $H^1((0,T)\times\Omega)$. We may also assume that

$$R^h \underset{h\to 0}{\to} R \quad \text{uniformly on every compact subset of } (0,T)\times\Omega.$$

The inequality

$$R^h(t,x)-P^h(t,x)=\int_t^{t-h}\frac{\partial P^h}{\partial\tau}(\tau,x)\,d\tau$$

and the boundedness of P_t^h show that R^h-P^h converges to 0 in $L^2((0,T)\times\Omega)$ and therefore $R=P$. Hence

$$k(P^h)\underset{h\to 0}{\to}k(P),\qquad \mu_e(t,x,R^h(t,x))\underset{h\to 0}{\to}\mu_e(t,x,P(t,x))$$

uniformly on compact subset of $[0,T)\times\Omega$, strongly in $L^2((0,T)\times\Omega)$ and

$$\mu_n^h(t,a,x)\underset{h\to 0}{\to}\mu_n(t,a,x,P(t,x)),$$
$$\beta^h(t,a,x)\underset{h\to 0}{\to}\beta(t,a,x,P(t,x)),$$

uniformly on compact subsets of Q and strongly in $L^2(Q)$ (for the latter the integrability condition (3.1) is used).

We may also suppose that $u^h\to_{h\to 0}u$ weakly in $\mathscr{V}$. But from (6.6) we derive that Du^h is bounded in $\mathscr{V}'$ and the continuity of the differential operator D in $\mathscr{D}'(\mathscr{O};[H^1(\Omega)]')$ implies that $Du^h\to_{h\to 0}Du$ weakly in $\mathscr{V}'$.

Lastly the continuity of the trace applications on $t=0$ and $a=0$ gives

$$u(0,a,x)=\lim_{h\to 0}u^h(0,a,x)=u_0(a,x),$$
$$u(t,0,x)=\lim_{h\to 0}u^h(t,0,x)=\lim_{h\to 0}\int_0^\infty\beta^h u^h\,da$$
$$=\int_0^\infty\beta(t,a,x,P(t,x))u(t,a,x)\,da.$$

We can also assume that $z^h\to_{h\to 0}z$ weakly in $L^2((0,T)\times\Omega)$.

Passing to the limit in (6.6), (6.7) provides (u,P,z) solution of:

For any v in $\mathscr{V}$,

$$\int_{\mathscr{O}}\langle Du,v\rangle\,dt\,da+\int_Q[k(P)\nabla u\cdot\nabla v+\mu(a,P)uv]\,dt\,da\,dx=0.$$
$$P_t-\Delta K(P)+\lambda P=\int_0^\infty[\beta(a,P)-\mu_n(a,P)]u\,da+[\lambda-\mu_e(P)]z.$$

The extraneous function z will help us to conclude that

$$\int_0^\infty u(t,a,x)\,da=z(t,x)=P(t,x).$$

From Corollary 5.1 we derive that z^h is a solution of the parabolic equation for any w in $\mathscr{W}$:

$$\int_0^T\langle\partial_t z^h,w\rangle\,dt+\int_{(0,T)\times\Omega}k(R^h)\nabla z^h\nabla w\,dt\,dx=\int_{(0,T)\times\Omega}\int_0^\infty(\beta^h-\mu^h)u^h\,da\cdot w\,dt\,dx.$$

Again $\int_0^\infty(\beta^h-\mu^h)u^h da$ is bounded in $L^2((0,T)\times\Omega)$. Using generic properties of weak solutions of parabolic equations we conclude that z^h is weakly converging to z in $\mathscr{W}$, $\partial_t z\in\mathscr{W}'$, a solution of:

For any w in $\mathscr{W}$,

$$\int_0^T\langle\partial_t z,w\rangle\,dt+\int_{(0,T)\times\Omega}[k(P)\nabla z\cdot\nabla w+\mu_e(P)zw]\,dt\,dx$$
$$=\int_{(0,T)\times\Omega}\int_0^\infty[\beta(a,P)-\mu_n(a,P)]\,u\,da\cdot w\,dt\,dx$$

and taking the initial value $z(0,x)=p_0(x)$.

Using Corollary 5.1 and Remark 5.1 we check that $\int_0^\infty u(t,a,x)da$ is also a solution; by uniqueness $\int_0^\infty u(t,a,x)da=z(t,x)$.

Now z is a solution of

$$\int_0^T\langle\partial_t z,w\rangle\,dt+\int_{(0,T)\times\Omega}[k(P)\nabla z\cdot\nabla w+\lambda zw]\,dt\,dx$$
$$=\int_{(0,T)\times\Omega}\int_0^\infty[\beta(a,P)-\mu(a,P)+\lambda]\,u\,da\,w\,dt\,dx$$

and so is P by Corollary 5.3; by uniqueness $z(t,x)=P(t,x)$.

7. Proof of Theorem 3.3. Denote $P(t,x,c)$, the solution of (1.7), (1.8).

When c_0 is nonnegative, setting $\chi(p)=\int_0^p\sigma\cdot K(\sigma)\,d\sigma$, then

$$V(\xi)=\int_\Omega\chi(\xi)\,d\xi,\qquad \xi\in L^1(\Omega)$$

is a Lyapunov function for $P_t-\Delta K(P)+c_0P=0$ in $(0,T)\times\Omega$, P, verifying (1.8). As in Alikakos and Rostamian [1, Lemma 2.1] it follows that in the sense of $L^\infty(\Omega)$

$$c_0=0\Rightarrow P(t,\cdot)\underset{t\to\infty}{\to}\frac{1}{\operatorname{mes}\Omega}\int_\Omega p_0(x)\,dx,$$
$$c_0>0\Rightarrow P(t,\cdot)\underset{t\to\infty}{\to}0.$$

The comparison principle for parabolic equations ensures that for $c(t,a,x,p)\geqq c_0$, $P(t,x,c)\leqq P(t,x,c_0)$. We conclude that when c is bounded from below by a positive constant, $P(t,\cdot)$ goes to 0 in $L^\infty(\Omega)$ and $u(t,\cdot,\cdot)$ goes to 0 in $L^1((0,A)\times\Omega)$.

When c_0 is a negative constant, $y(t)=\int_\Omega P(t,x,c_0)\,dx$ is a solution of the differential equation $y'+c_0y=0$, $y(0)=y_0>0$ and therefore $y(t)\to_{t\to\infty}+\infty$. From the comparison principle we derive that when c is bounded from above by a negative constant, $u(t,\cdot,\cdot)$ goes to $+\infty$ in $L^1((0,\infty)\times\Omega)$.

8. Appendix. In this last section we supply the proofs of Theorems 5.1 and 5.2.

It is convenient to perform a change of an unknown function: if u is a solution of (5.2), (5.3) then $v(t,a,x)=e^{-\lambda t}u(t,a,x)$ is also a solution with m replaced by $m+\lambda$. This is implicitly done with $\lambda\geqq C_2$, C_2 defined in (5.4) and v is still denoted u.

We first use some semigroup theory to solve (5.2) with prescribed initial data $u(t,0,x)=B(t,x)$ following the procedure used in Bardos [3] and Langlais [13]. Next along the lines of [6], [14], we obtain the results stated in Theorem 5.1 through a fixed point method and comparison theorem.

LEMMA 8.1. *Given f in $\mathscr{V}'$ there exists a unique u in $\mathscr{V}$, Du in $\mathscr{V}'$, such that for any v in $\mathscr{V}$:*

$$\int_{\mathscr{O}}\langle Du,v\rangle\,dt\,da+\int_{Q}[U\nabla u\cdot\nabla v+(m+\lambda)u\cdot v]\,dt\,da\,dx=\int_{\mathscr{O}}\langle f,v\rangle\,dt\,da$$

and verifying the initial data $u(0,a,x)=u(t,0,x)=0$.

Proof. Let e the bilinear form on $H^1(\Omega)$ be defined by

$$e(u,v)=\int_{\Omega}[U\nabla u\cdot\nabla v+(m+\lambda)u\cdot v]\,dx.$$

U and m being bounded, e is continuous on $H^1(\Omega)$. For any u in $H^1(\Omega)$

$$\operatorname{Min}(m_1,\lambda)\|u\|^2_{H^1(\Omega)}\leqq e(u,u)\leqq \operatorname{Max}(M_1,\lambda+\overline{m})\|u\|^2_{H^1(\Omega)},$$

so that e is also coercive on $H^1(\Omega)$. We denote E the linear bounded and coercive operator from $\mathscr{V}$to $\mathscr{V}'$ such that

$$\text{for any } u,\,v \text{ in } \mathscr{V},\qquad \langle\langle Eu,v\rangle\rangle=\int_{\mathscr{O}}e(u,v)\,dt\,da.$$

Now we consider Λ^0 the unbounded operator in $L^2(Q)$ with domain

$$D(\Lambda^0)=\{u\in L^2(Q),\,u_t+u_a\in L^2(Q),\,u(0,a,x)=u(t,0,x)=0\}$$

defined by $u\in D(\Lambda^0)$, $\Lambda^0u=u_t+u_a$. Then $-\Lambda^0$ is the infinitesimal generator of a contraction semigroup $(S(\tau),\tau\geqq 0)$ in $L^2(Q)$ (see [3]). Actually

$$(S(\tau)u)(t,a,x)=\begin{cases}u(t-\tau,a-\tau,x) & \text{if } (t-\tau,a-\tau,x)\in Q,\\ 0 & \text{otherwise,}\end{cases}$$

and $S(\tau)$ is a bounded and continuous semigroup in $\mathscr{V}$ and $\mathscr{V}'$. Let Λ^1 be the infinitesimal generator of $S(\tau)$ in $\mathscr{V}'$. Its domain $D(\Lambda^1,\mathscr{V}')$ is

$$D(\Lambda^1,\mathscr{V}')=\{u\in\mathscr{V}',\,Du\in\mathscr{V}',\,u(0,a,x)=u(t,0,x)=0\}$$

the traces of u on $t=0$ and $a=0$ make sense because u is continuous along the characteristics of the vector field $\partial/\partial t+\partial/\partial a$ with value in $[H^1(\Omega)]'$. Hence the unbounded operator Λ from $\mathscr{V}$to $\mathscr{V}'$ with domain

$$D(\Lambda)=\mathscr{V}\cap D(\Lambda^1,\mathscr{V}')=\{u\in\mathscr{V},\,Du\in\mathscr{V}',\,u(0,a,x)=u(t,0,x)=0\}$$

defined by $\Lambda u=Du$ for u in $D(\Lambda)$ is a maximal monotone operator in the sense that together with its adjoint Λ^*: $\mathscr{V}\to\mathscr{V}'$ they are nonnegative, closed with dense domain in $\mathscr{V}$; see [16, Chapter 3] and J. L. Lions [15, Chap. 3].

E being bounded and coercive and Λ being maximal monotone we conclude that for any f in $\mathscr{V}'$ there exists a unique u in $D(\Lambda)$ solution of $Eu+\Lambda u=f$ (see loc. cit.). This equation turns out to be an abstract formulation of our problem because for u in $D(\Lambda)$ and v in $\mathscr{V}$

$$\langle\langle\Lambda u,v\rangle\rangle=\int_{\mathscr{O}}\langle Du,v\rangle\,dt\,da.$$

Lemma 8.2. *Given u_0 in $L^2((0,\infty)\times\Omega)$ and B in $L^2((0,T)\times\Omega)$ there exists a unique solution to* (5.2) *satisfying the initial conditions*

$$u(0,a,x)=u_0(a,x), \qquad u(t,0,x)=B(t,x).$$

Proof. Uniqueness follows from uniqueness in the preceding lemma.

To prove the existence we introduce a sequence of smooth functions φ^n in $C^\infty(\overline{Q})$ such that

$$\varphi^n(0,a,x) \underset{n\to\infty}{\to} u_0(a,x) \quad \text{in } L^2((0,\infty)\times\Omega),$$
$$\varphi^n(t,0,x)\to B(t,x) \quad \text{in } L^2((0,T)\times\Omega).$$

From Lemma 8.1 we conclude that there exists a unique u^n in the $D(\Lambda)$ solution to $Eu^n+\Lambda u^n=-E\varphi^n-D\varphi^n$; therefore $u^n+\varphi^n=v^n$ is a solution to (5.2) and $v^n(0,a,x)=\varphi^n(0,a,x)$, $v^n(t,0,x)=\varphi^n(t,0,x)$. For any A_0, $0<A_0$, set $\mathscr{O}_0=(0,T)\times(0,A_0)$. We get

$$\int_{\mathscr{O}_0}\langle Dv^n,v^n\rangle\,dt\,da+\int_{\mathscr{O}_0}e(v^n,v^n)\,dt\,da=0, \qquad n\geqq 0.$$

The first term can be integrated by parts (see §4) to provide

$$\frac{1}{2}\int_{(0,A_0)\times\Omega}\left[(v^n)^2(T,a,x)-(v^n)^2(0,a,x)\right]da\,dx$$
$$+\frac{1}{2}\int_{(0,T)\times\Omega}\left[(v^n)^2(t,A_0,x)-(v^n)^2(t,0,x)\right]dt\,dx,$$

which is certainly larger than

$$-\frac{1}{2}\int_{(0,A_0)\times\Omega}(\varphi^n)^2(0,a,x)\,da\,dx-\frac{1}{2}\int_{(0,T)\times\Omega}(\varphi^n)^2(t,0,x)\,dt\,dx.$$

Substituting this inequality in the above relation for v^n, letting A_0 go to $+\infty$ and using the coercivity of the bilinear form e yields

$$\mathrm{Min}(m_1,\lambda)\|v^n\|^2_{\mathscr{V}}\leqq\frac{1}{2}\|\varphi^n(0,a,x)\|^2_{L^2((0,A)\times\Omega)}+\frac{1}{2}\|\varphi^n(t,0,x)\|^2_{L^2((0,T)\times\Omega)}.$$

By the choice of φ^n this implies that v^n is a bounded sequence in $\mathscr{V}$.

We can extract a sequence v^k of the sequence v^n such that

$$v^k \underset{k\to\infty}{\to} v \quad \text{weakly in } \mathscr{V},$$
$$Dv^k \underset{k\to\infty}{\to} f \quad \text{weakly in } \mathscr{V}'.$$

But the differential operator D is continuous in $\mathscr{D}'(\mathscr{O};\,[H^1(\Omega)]')$, thus $f=Dv$. On the other hand E is continuous and Ev^k goes to Ev. We first conclude that v is a solution to (5.2). The continuity of the trace applications on $t=0$ and $a=0$ implies that $v(t,0,x)=B(t,x)$ and $v(0,a,x)=u_0(a,x)$. Lastly the lower semicontinuity of the norm with respect to the weak convergence ensures that the solution satisfies

$$\mathrm{Min}(\lambda,m_1)\|v\|^2_{\mathscr{V}}\leqq\frac{1}{2}\|u_0\|^2_{L^2((0,\infty)\times\Omega)}+\frac{1}{2}\|b\|^2_{L^2((0,T)\times\Omega)},$$

so that the solution depends continuously on the data.

LEMMA 8.3. *There exists a unique solution to* (5.2), (5.3).

Proof. Let u be given in $\mathscr{V}$ and denote $Su = v$ the solution of (5.2) satisfying the initial conditions

$$v(0,a,x) = u_0(a,x),$$
$$v(t,0,x) = \int_0^\infty b(t,a,x)u(t,a,x)\,da = B(t,x).$$

S maps $\mathscr{V}$ into itself, is continuous and the solutions we are looking for are the fixed points of S. The inequality

$$\int_{(0,T)\times\Omega}\left[\int_0^\infty b(t,a,x)u(t,a,x)\,da\right]^2 dt\,dx \leqq C_2\int_Q u^2(t,a,x)\,dt\,da\,dx$$

(C_2 is defined in Theorem 5.1) and the estimate established for v by the end of the proof of Lemma 8.2 give

$$\operatorname{Min}(m_1,\lambda)\|v\|_{\mathscr{V}}^2 \leqq \frac{1}{2}\|u_0\|^2_{L^2((0,\infty)\times\Omega)} + \frac{C_2}{2}\|u\|^2_{L^2(Q)},$$

so that S is bounded from $L^2(Q)$ to $\mathscr{V}$. The conclusion will follow from the fact that S is a strict contraction in $L^2(Q)$.

Actually if u^1 and u^2 are chosen in $L^2(Q)$, letting $v^1 = Su^1$, $v^2 = Su^2$, the difference $v = v^1 - v^2$ vanishes on $t = 0$, is a solution to (5.2) and satisfies

$$v(t,0,x) = \int_0^\infty \beta(t,a,x)u(t,a,x)\,da, \qquad u = u^1 - u^2.$$

With the notation of the proof of Lemma 9.2 we still have

$$\int_{\mathscr{O}_0}\langle Dv, v\rangle\,dt\,da + \int_{\mathscr{O}_0} e(v,v)\,dt\,da = 0,$$

and in a quite similar way we arrive after some calculations at

$$\lambda\int_Q v^2\,dt\,da\,dx \leqq \frac{1}{2}C_2\|u\|^2_{L^2(Q)}.$$

Keeping in mind that $\lambda \geqq C_2$ this means that S is strictly contracting in $L^2(Q)$ because the latter inequality means

$$\|Su^1 - Su^2\|^2_{L^2(Q)} \leqq \frac{1}{2}\|u^1 - u^2\|^2_{L^2(Q)}.$$

LEMMA 8.4. *The solution of* (5.2), (5.3) *is nonnegative and integrable over Q.*

Proof. For any real function w we denote by w^+ its positive and by w^- its negative parts. Under the a.e. ordering ($v \leqq w$ if $v(x) \leqq w(x)$ a.e. in Ω) $H^1(\Omega)$ is a lattice and for any w in $H^1(\Omega) w^+, w^-$ belong to $H^1(\Omega)$ and $e(w^+, w^-) \leqq 0$; see Necas [18], D. Kinderlehrer and G. Stampacchia [11]. It follows that for any u in $\mathscr{V}$ u^+ and u^- are in $\mathscr{V}$ and

$$\int_{\mathscr{O}} e(u^+, u^-)\,dt\,da \leqq 0.$$

Let us prove now that provided u_0 and B be nonnegative then the solution of (5.2) obtained in Lemma 8.2 verifies

$$\int_{\mathcal{O}} \langle Du, u^- \rangle \, dt \, da \leqq 0.$$

Actually if u is smoother, say u lies in $H^1(Q)$, we have $u^+(0,a,x) = u_0(a,x)$, $u^+(t,0,x) = B(t,x)$ and $u^-(0,a,x) = u^-(t,0,x) = 0$ so that, for $\mathcal{O}_0 = (0,T)\times(0,A_0)$,

$$\begin{aligned}\int_{\mathcal{O}_0} \langle Du, u^- \rangle \, dt \, da &= -\int_{\mathcal{O}_0\times\Omega} \left(\frac{\partial u^-}{\partial t} + \frac{\partial u^-}{\partial a}\right) u^- \, dt \, da \\ &= -\frac{1}{2}\int_{(0,A_0)\times\Omega} (u^-)^2(T,a,x) \, da \, dx \\ &\quad -\frac{1}{2}\int_{(0,T)\times\Omega} (u^-)^2(t,A^0,x) \, dt \, dx \leqq 0.\end{aligned}$$

By letting A_0 go to $+\infty = A$ we end up with $\int_{\mathcal{O}} \langle Du, u^- \rangle \, dt \, da \leqq 0$ for smooth u. For the general case u in $\mathscr{V}$, Du in $\mathscr{V}'$, in general Du^- is not in $\mathscr{V}'$; nevertheless approximating u by smooth functions leads to the desired inequality which is preserved by passing to the limit.

For nonnegative u_0 and B substituting $v = u^-$ in (5.2) yields

$$\int_{\mathcal{O}} \langle Du, u^- \rangle \, dt \, da + \int_{\mathcal{O}} e(u, u^-) \, dt \, da = 0.$$

From the relation $e(u,u^-) = e(u^+,u^-) - e(u^- \; u^-)$ we get:

$$-\int_{\mathcal{O}} e(u^-, u^-) \, dt \, da \geqq 0$$

and the coercivity of e provides

$$\operatorname{Min}(\lambda, m_1)\|u^-\|^2_{\mathscr{V}} \leqq 0$$

namely $u^- = 0$ and u is nonnegative. Then letting $u^0 = 0$ and $u^{n+1} = Su^n$, $n \geqq 0$, where S is defined in the proof of Lemma 8.3, each u^n is nonnegative and the sequence u^n converges to the solution of (5.2), (5.3) which is therefore nonnegative.

To prove that the solution u of (5.2), (5.3) is integrable over Q we substitute in (5.2) the characteristic function of $(0,\tau)\times(0,A_0)\times\Omega$, $0<\tau<T$, $0<A_0<\infty$. Integrating by parts we obtain after some calculations:

$$0 \leqq \int_{(0,A_0)\times\Omega} u(\tau,a,x) \, da \, dx \leqq \int_{(0,A_0)\times\Omega} u_0(a,x) \, da \, dx + \int_{(0,\tau)\times(\Omega)}\int_0^\infty b\cdot u \, da \, dt \, dx.$$

The right-hand side is less or equal to:

$$\int_{(0,\infty)\times\Omega} u_0(a,x) \, da \, dx + K_1\left(\int_{(0,t)\times(0,\infty)\times\Omega} u^2(t,a,x) \, dt \, da \, dx\right)^{1/2}$$

where K_1 depends on b (through the constant C_2 defined in Theorem 5.1). It follows at once that u belongs to $L^1(Q)$ because u is nonnegative.

REFERENCES

[1] N. D. ALIKAKOS AND R. ROSTAMIAN, *Large time behaviour of solutions of Neumann boundary value problem*, Indiana Univ. Math. J., 30 (1981), pp. 749–785.

[2] D. G. ARONSON, *Regularity properties of flow through porous media, a counterexample*, SIAM J. Appl. Math., 19 (1970), pp. 299–307.

[3] C. BARDOS, *Problèmes aux limites pour les equations aux dèrivees partielles du premier ordre*, Ec. Norm. Sup., 4° série (1970), pp. 185–233.

[4] G. DI BLASIO, *Nonlinear age dependent population diffusion*, J. Math. Biol., 8 (1979), pp. 265–284.

[5] L. C. EVANS, *Application of nonlinear semigroup theory to certain partial differential equations*, in Nonlinear Evolution Equations, Academic Press, New York, 1978.

[6] M. G. GARRONI AND M. LANGLAIS, *Age dependent population diffusion with external constraints*, J. Math. Biol., 14 (1982), pp. 77–94.

[7] K. GOPALSAMY, *On the asymptotic age distribution in dispersive population*, Math. Biosc., 31 (1976), pp. 191–205.

[8] M. E. GURTIN, *A system of equations for age dependent population diffusion*, J. Theoret. Biol., 40 (1973), pp. 389–392.

[9] M. E. GURTIN AND R. C. MACCAMY, *On the diffusion of biological population*, Math. Biosci., 38 (1977), pp. 35–49.

[10] ———, *Product solutions and asymptotic behaviour in age dependent population diffusion*, Math. Biosci., 62 (1982), pp. 157–167.

[11] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational inequalities and Their Applications*, Academic Press, New York, 1980.

[12] O. A. LADYZHENSKAYA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and quasilinear equations of parabolic type*, Nauka, Moscow, 1967; English translation, Transl. Math. Monographs vol. 23, American Mathematical Society, Providence, RI, 1968.

[13] M. LANGLAIS, *Solutions fortes pour une classe de problèmes aux limits dégénérés*, Comm. Part. Diff. Equations, 4 (1979), pp. 869–897.

[14] ———, *On some linear age dependent population diffusion models*, Quart. Appl. Math., 40 (1983), pp. 447–460.

[15] J. L. LIONS, *Quelques méthodes de rèsolution des problèmes aux limites nonlinéaires*, Dunod, Paris, 1969.

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

[17] P. MARCATI, *Asymptotic behaviour in age dependent population dynamics*, this Journal, 12 (1981), pp. 904–916.

[18] J. NECAS, *Les méthodes directes en thèorie des èquations elliptiques*, Masson, Paris, 1967.

[19] O. A. OLEINIK, *On some degenerate quasilinear parabolic equations*, Istituto Nazionale di Alti Matematica Seminari 1962–63 Cremonese, Rome, 1965.

[20] G. F. WEBB, *Nonlinear semigroups and functional differential equations*, Lectures held in 1982, Scuola Normale Superiore, Pisa, Italy.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

FORMATION OF SINGULARITIES FOR A CONSERVATION LAW WITH MEMORY*

REZA MALEK-MADANI[†] AND JOHN A. NOHEL[‡]

Abstract. The formation of singularities in smooth solutions of the model Cauchy problem

$$\begin{aligned}&u_t+\phi(u)_x+a' * \psi(u)_x=0, \qquad x\in\mathbb{R}, \quad t\in[0,\infty),\\&u(x,0)=u_0(x),\end{aligned}$$

is studied. The constitutive functions $\phi, \psi:\mathbb{R}\to\mathbb{R}$ are smooth, $a:\mathbb{R}^+\to\mathbb{R}$ is a given memory kernel, subscripts denote partial derivatives, $'=d/dt$ and $*$ denotes the convolution on $[0,t]$. Under physically reasonable assumptions concerning the functions ϕ, ψ and a it is shown that the smooth solution u develops a singularity in finite time, whenever the smooth datum u_0 becomes "sufficiently large" in a precise sense.

AMS-MOS subject classifications (1980). Primary 35L65, 35L67, 45G10, 45K05, 45D05, 45M10, 45M99, 47H10, 47H15, 47H17, 73F15, 73H10, 76N15

Key words. conservation laws, Burgers' equation, nonlinear viscoelastic motion, materials with memory, stress-strain relaxation functions, nonlinear Volterra equations, hyperbolic equations, Riemann invariants, regularity, breakdown of smooth solutions

1. Introduction.

In this paper we study the model initial value problem

$$\begin{aligned}&u_t+\phi(u)_x+a' * \psi(u)_x=0, \qquad x\in\mathbb{R}, \quad t\in[0,\infty),\\&u(x,0)=u_0(x),\end{aligned} \tag{1.1}$$

where ϕ, $\psi:\mathbb{R}\to\mathbb{R}$ are given smooth constitutive functions, $a:\mathbb{R}^+\to\mathbb{R}$ is a given kernel, subscripts denote partial derivative, $'=d/dt$, and where $*$ denotes the usual convolution operator

$$(f*g)(t)=\int_0^t f(t-\tau)g(\tau)\,d\tau.$$

The goal is to investigate the formation of singularities in finite time of classical solutions of (1.1) when the datum u_0 is smooth. The motivation for studying equation (1.1) is provided by the more complex problem of the motion of a one-dimensional homogeneous viscoelastic body governed by the equation

$$u_{tt}-\sigma_x=0, \tag{1.2}$$

*Received by the editors May 26, 1983, and in revised form February 3, 1984.

[†] Department of Mathematics, United States Naval Academy, Annapolis, Maryland 21402. The work of this author was partially supported by the Mathematics Research Center and by Johns Hopkins University under a grant by the National Science Foundation.

[‡] Mathematics Research Center, University of Wisconsin-Madison, 610 Walnut Street, Madison, Wisconsin 53705. The work of this author was sponsored by the U. S. Army under contract DAAG29-80-C-0041.

together with appropriate initial and homogeneous boundary conditions; in (1.2) the stress σ is related to the strain u_x by the constitutive relation

$$\sigma(u_x)=\phi(u_x)+\int_0^t a'(t-\tau)\psi(u_x(x,\tau))\,d\tau. \tag{1.3}$$

Under appropriate physical assumptions concerning ϕ, ψ and a, the "memory" term in (1.3) induces a weak dissipation mechanism into the structure of solutions of (1.2). It has been shown (cf. Dafermos and Nohel [1]) that under physically proper assumptions on a, ϕ, ψ and on the initial data u_0 and u_1, the initial-boundary value problem (1.2) has a unique global C^2 solution, if the initial data are sufficiently smooth and "small" in an appropriate sense; moreover, this solution decays in a precise sense as $t\to\infty$. A similar behavior is exhibited by the solution u of (1.1) with u satisfying periodic boundary conditions (cf. Nohel [8]). Recently, Hrusa and Nohel [11] established a similar result for the Cauchy problem associated with (1.2), (1.3), and an analogous method leads to the same result for the Cauchy problem (1.1), if $\|u_0\|_{H^2(\mathbb{R})}$ is sufficiently small. These two results are of special interest since when $a'(t)\equiv 0$, (1.1) reduces to the Burgers equations, while (1.2), (1.3) reduce to the quasilinear wave equation $u_{tt}=\phi(u_x)_x$. For the wave equation it is well known (cf. Lax [5]) that under appropriate convexity assumptions on ϕ there are smooth solutions which develop a singularity in the highest derivatives in finite time, no matter how smooth and small one chooses the initial datum. Thus $a'(t)\not\equiv 0$ induces a weak dissipation mechanism which prohibits the breaking of waves when the initial amplitude of these waves is small.

This paper considers the natural question of how "large" one must choose the smooth initial datum in order that the shock forming structure of (1.1) overcomes this dissipation. Indeed, in Theorem 2.3 we show, under natural assumptions concerning the constitutive functions ϕ, ψ, the kernel a, and datum u_0, that the classical solution u of (1.1) develops a singularity in finite time for smooth and sufficiently "large" datum u_0, in the sense that first derivatives of u become unbounded in finite time while u itself remains bounded. Our ultimate objective is to prove such a result for the complicated problem (1.2), (1.3), and with $\psi\not\equiv\phi$.

Equation (1.1) has a simpler structure than (1.2) due to the fact that (1.1) has only one family of "genuinely nonlinear" characteristics and one "linearly degenerate" characteristic due to the convolution term. Our approach examines the variation of the solution of (1.1) along characteristics with the aid of Riemann invariants. A similar approach (under active study) appears promising for the more complicated higher order problem (1.2), (1.3); this latter equation has three families of characteristics (only two are genuinely nonlinear), and thus, in general (1.2), (1.3) does not have Riemann invariants. Introducing the generalized Riemann invariants (cf. John [4]) there is reason to expect that much of our analysis can be adapted for (1.2), (1.3).

Some experimental evidence for the breakdown of smooth solutions of model equations governing viscoelastic fluids can be found in the work of Tordella [10]. In addition some results on the loss of regularity in solutions of the equations governing viscoelastic fluids, for smooth and sufficiently large data, have been obtained by Slemrod [9], Gripenberg [2], and for dissipative hyperbolic Volterra problems by Hattori [3], all for the special cases of (1.2), (1.3) when $\psi\equiv\phi$. By methods similar to ours in spirit they also analyze the behavior of solutions along characteristics; however, they do not study the generalization to the more natural and more difficult situation in which $\psi\neq\phi$.

In §2 we state and discuss our assumptions and the main result; its proof is presented in §3. In §4 we prove two auxiliary results in the proof. We thank our colleagues, particularly C. M. Dafermos, R. Glassey, W. Hrusa, J. U. Kim, and M. Slemrod for helpful discussions.

2. Assumptions and statements of result. The basic constitutive assumption concerning ϕ is

$$\phi \in C^2(\mathbb{R}) \text{ and } \phi'(\cdot)>0, \quad \phi''(\cdot)>0, \quad \phi(0)=0. \tag{2.1}$$

The constitutive assumption concerning ψ is

$$\psi \in C^2(\mathbb{R}) \text{ and } \psi'(\cdot)>0, \quad \psi(0)=0. \tag{2.2}$$

In addition, we assume that ϕ and ψ are related as follows. There exists a constant $\beta>0$ such that

$$0<\psi'(u)<\beta\phi'(u), \qquad u\in\mathbb{R}. \tag{2.3}$$

Obviously, (2.3) is more restrictive than the assumption $\phi'(0)>a(0)\psi'(0)$ (i.e. (2.3) at $u=0$ with $\beta=a(0)^{-1}$) which was sufficient for the analysis of global solutions of (1.2), (1.3) in [1] for smooth and sufficiently small data. Assumption (2.3), automatically satisfies if $\psi\equiv\phi$, simplifies our relatively technical analysis of the development of singularities for solutions of (1.1); in Remark 2.5 below we point out how (2.3) can be removed. Concerning the memory kernel a we assume that it is positive, decreasing and convex in the sense

$$a\in C^2[0,\infty), \qquad (-1)^i a^{(i)}(t)\geqq 0 \qquad (i=0,1,2), \tag{2.4}$$

where the strict inequalities hold at $t=0$. Finally, we assume that the datum u_0 satisfies

$$u_0\in H^2(\mathbb{R}); \tag{2.5}$$

observe that $u_0\in H^2(\mathbb{R})$ implies $u_0\in C^1(\mathbb{R})$.

Under assumptions which include (2.1), (2.2), (2.4), (2.5) as special cases the Cauchy problem (1.1) has a unique classical local solution. For this argument (2.3) is not used. More precisely, the following local result, proved by an energy method coupled with a contraction mapping argument, holds (cf. Nohel [8]).

PROPOSITION 2.1. *Let $a',a''\in L^1_{\text{loc}}(0,\infty)$, $\phi,\psi\in C^3(\mathbb{R})$, $\phi(0)=\psi(0)=0$, $\psi'(\cdot)>0$, and let there exist a constant κ such that $\phi'(\xi)\geqq\kappa>0$ $(\xi\in\mathbb{R})$. If $u_0\in H^2(\mathbb{R})$, there exists $T>0$ and a unique solution $u\in C^1(\mathbb{R}\times[0,T])$ of* (1.1) *with u_x and u_t bounded on $\mathbb{R}\times[0,T]$, and $u_{tt},u_{tx},u_{xx}\in C([0,T];\ L^2(\mathbb{R}))$.*

Remark 2.2. It is also shown in [8] that the unique solution u exists on a maximal interval $[0,T_0)\times\mathbb{R}$; moreover, if $\sup_{\mathbb{R}\times[0,T_0)}[|u_x(x,t)|+|u_t(x,t)|]<\infty$, then $T_0=+\infty$. The last claim is established by combining the energy estimates obtained in the proof of Proposition 2.1 with a Gronwall inequality argument.

Our main result is the following theorem.

THEOREM 2.3. *Let the assumptions* (2.1)–(2.5) *be satisfied, and let $T_1>0$ be given. There exists a smooth initial datum u_0 such that no C^1-smooth solution u of* (1.1) *with u_x and u_t bounded can exist for $x\in\mathbb{R}$ and $t\geqq T_1$. More precisely, if $\sup_{x\in\mathbb{R}}|u_0(x)|$ is sufficiently small, and $u_0'(x)<0$ with $-\inf_{x\in\mathbb{R}}u_0'(x)$ is sufficiently large, then the life span of a C^1-solution u with u_x and u_t bounded is finite and cannot exceed T_1. In fact, there exists $0<t_1\leqq T_1$ such that*

$$\sup_{\mathbb{R}\times[0,t_1)}\big[|u_x(x,t)|+|u_t(x,t)|\big]=\infty.$$

A slight modification of the proof of Lemma 3.1 shows that the solution u constructed in the proof of Theorem 2.3 satisfies $\sup_{\mathbb{R}\times[0,t_1)}|u(x,t)|<\infty$ and that $t_1=T_0$, where $[0,T_0)$ is the maximal interval of existence in Remark 2.2.

Remark 2.4. While Theorem 2.3 establishes breakdown of smooth solutions of (1.1) for sufficiently large data, it does not prove the development of a shock front. Numerical evidence for this more complex phenomenon has been found by Markowich and Renardy [7] for the Cauchy problem associated with (1.2), (1.3) in the special case $\phi\equiv\psi$ when the smooth data are taken sufficiently large. The corresponding analytical problem is under active study.

Remark 2.5. Theorem 2.3 holds without assumption (2.3). Indeed, assumptions (2.1), (2.2) imply that $\psi'(\cdot)/\phi'(\cdot)$ is bounded away from zero, and (2.3) holds for some $\beta>0$ on every bounded interval. Then the analysis of §3 can be modified accordingly.

Remark 2.6. It is also clear from the proof (cf. proof of Lemma 3.2) that if the assumption $u_0'(x_0)<0$ and $-u_0'(x_0)$ sufficiently large holds at a single point x_0, then the conclusion of Theorem 2.3 holds.

3. Proof of Theorem 2.3. The proof is by contradiction. Let $T_1>0$ be given. Assume that for every datum u_0 satisfying (2.5) the unique smooth solution u of (1.1) exists for $(x,t)\in\mathbb{R}\times[0,T_1]$ and that $u_x(x,t)$ and $u_t(x,t)$ are bounded on $\mathbb{R}\times[0,T_1]$. We begin by transforming (1.1) to an equivalent system. Let u be a smooth solution of (1.1) on $\mathbb{R}\times[0,T_1]$ and introduce the dependent variable z by

$$z(x,t)=\int_0^t a'(t-\tau)\psi(u(x,\tau))\,d\tau,\qquad (x,t)\in\mathbb{R}\times[0,T_1]. \tag{3.1}$$

Equation (1.1) is then equivalent to the system

$$\begin{aligned} &u_t+\phi(u)_x+z_x=0,\\ &z_t=a'(0)\psi(u)+a''*\psi(u), \end{aligned}\qquad (x,t)\in\mathbb{R}\times[0,T_1], \tag{3.2}$$

together with the initial data $u(x,0)=u_0(x)$, $z(x,0)=0$. We next introduce $\mathbf{U}=[u,z]^T$ and the matrices

$$A[\mathbf{U}]=\begin{bmatrix}\phi'(u) & 1\\ 0 & 0\end{bmatrix},\qquad B(\mathbf{U},t)=\begin{bmatrix}0\\ -a'(0)\psi(u)-a''*\psi\end{bmatrix};$$

then (3.2) can be written as the equivalent quasilinear system

$$\mathbf{U}_t+A(\mathbf{U})\mathbf{U}_x+\mathbf{B}(\mathbf{U},t)=0,\qquad \mathbf{U}(x,0)=[u_0(x),0]^T.$$

The 2×2 matrix $A(\mathbf{U})$ has distinct eigenvalues $\phi'(u)>0$ and 0. A well-known theorem of Lax [6] guarantees the existence of two linearly independent Riemann invariants $r(u,z)$ and $s(u,z)$. By definition r and s satisfy

$$\mathbf{r}_1\cdot\nabla r=0,\qquad \mathbf{r}_2\cdot\nabla s=0, \tag{3.3}$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the right eigenvectors of $A(\mathbf{U})$. A simple calculation shows that $\mathbf{r}_1=[1,-\phi'(u)]^T$ and $\mathbf{r}_2=[1,0]^T$. It is then easy to show that

$$r(u,z)=z+\phi(u),\qquad s(u,z)=z, \tag{3.4}$$

satisfy (3.3), and moreover, by assumption (2.1), $\partial(r,s)/\partial(u,z)=\phi'(u)\neq0$.

We shall study the development of a singularity in the classical C^1-solution u of (1.1) along the *characteristic* $x=x(t,\xi)$ through any point $\xi\in\mathbb{R}$, defined to be the unique solution of the initial value problem

$$\frac{dx}{dt}=\phi'(u(x(t,\xi),t)),\qquad x(0,\xi)=\xi. \tag{3.5}$$

Assumption (2.1) and the classical theory of ODE guarantee that $x(t,\xi)$ exists for as long as the C^1-solution u of (1.1) exists and has $u_x(x,t)$ and $u_t(x,t)$ bounded. Under the present hypotheses $x(t,\xi)$ exists for $0\leqq t\leqq T_1$ for any $\xi\in\mathbb{R}$.

Let $x(t,\xi)$ denote the characteristic curve through ξ associated with (1.1) which satisfies (3.5). The derivative of r along this characteristic is

$$\begin{aligned}\frac{dr}{dt}&\equiv r_t+\phi'(u)r_x=z_t+\phi'(u)u_t+\phi'(u)\left[z_x+\phi'(u)u_x\right]\\&=z_t+\phi'(u)\left[-\phi'(u)u_x-z_x\right]+\phi'(u)\left[z_x+\phi'(u)u_x\right]\\&=z_t=s_t,\qquad 0\leqq t\leqq T_1.\end{aligned}$$

Thus, we may replace (3.2) by the system

$$\left.\begin{aligned}&\frac{dr}{dt}=s_t\\&s_t=a'(0)\psi(u)+a''*\psi(u)\end{aligned}\right.\qquad (0\leqq t\leqq T_1), \tag{3.6}$$

together with the initial data $r(u,z)(x,0)=\phi(u_0(x))$, $s(u,z)(x,0)=0$, and then by (3.4), $u=\phi^{-1}(r-s)$. It is clear that the above calculations are valid for as long as u is a classical solution of (1.1), i.e., for $(x,t)\in\mathbb{R}\times[0,T_1]$. To keep the notation simple it should be understood that when calculating derivatives along a characteristic $x=x(t,\xi)$, $r=r(x(t,\xi),t)=r(u(x(t,\xi),t),z(x(t,\xi),t))$ and similarly for s.

To proceed with the proof of Theorem 2.3, let $v(t,\xi)\equiv x_\xi(t,\xi)$, $0\leqq t\leqq T_1$. The function v measures the variation of two nearby characteristics at time t with respect to their initial positions and plays a key role in our analysis. When v is different from zero (1.1) and (3.6) are equivalent. Note that $v(0,\xi)=1$ for any $\xi\in\mathbb{R}$. We will show that if $|u_0(\xi)|$ is sufficiently small and $-u_0'(\xi)$ is sufficiently large, then $v(t,\xi)$ approaches zero at a finite time $t_1\leqq T_1$, while $u_\xi(x(t,\xi),t)$ remains finite and bounded away from zero. Observing that

$$u_x(x(t,\xi),t)=\frac{u_\xi(x(t,\xi),t)}{v(t,\xi)} \tag{3.7}$$

we then obtain a contradiction of the assumption that u_x remains bounded for all $t\in[0,T_1]$, and the proof will be complete.

Differentiation of (3.5) with respect to ξ yields

$$\frac{dv}{dt}=\phi''(u(x(t,\xi)),t)u_\xi(x(t,\xi),t),\qquad v(0,\xi)=1,\quad t\in[0,T_1]. \tag{3.8}$$

Since $\phi(u)=r-s$, we have

$$\phi'(u)u_\xi=r_\xi-s_\xi=r_\xi-s_xx_\xi; \tag{3.9}$$

thus

$$u_\xi=\frac{1}{\phi'(u)}r_\xi-\frac{1}{\phi'(u)}s_x v.$$

From (3.2) and (3.4) the derivative of u along the characteristic $x=x(t,\xi)$ is

$$\frac{du}{dt}=-z_x=-s_x,$$

so that

$$u_\xi=\frac{1}{\phi'(u)}r_\xi+\frac{1}{\phi'(u)}\frac{du}{dt}v$$

and (3.8) takes the form

$$\frac{dv}{dt}=\frac{\phi''(u)}{\phi'(u)}r_\xi+\frac{\phi''(u)}{\phi'(u)}\frac{du}{dt}v, \qquad v(0,\xi)=1, \quad t\in[0,T_1].$$

The above equation is an ODE for v along characteristic having $[\phi'(u)]^{-1}$ as an integrating factor. Thus

$$\frac{1}{\phi'(u)}v(t,\xi)-\frac{1}{\phi'(u_0(\xi))}=\int_0^t\frac{\phi''(u)}{[\phi'(u)]^2}r_\xi d\tau;$$

or equivalently

$$\begin{aligned}v(t,\xi)=&\frac{\phi'(u(x(t,\xi),t))}{\phi'(u_0(\xi))}\\ &\cdot\left[1+\phi'(u_0(\xi))\int_0^t\frac{\phi''(u(x(\tau,\xi),\tau))}{[\phi'(u(x(\tau,\xi),\tau))]^2}r_\xi(x(\tau,\xi),\tau)\,d\tau\right],\end{aligned}\tag{3.10}$$

for $t\in[0,T_1]$.

We will now use the following result which provides a bound for u, independent of $u_0'(\xi)$. Its proof is given in §4.

LEMMA 3.1. *Let the assumptions of Theorem* 2.3 *be satisfied and let u be a* C^1*-smooth solution of* (1.1) *with* u, u_x, u_t *bounded on* $\mathbb{R}\times[0,T_1]$. *Then for any* $\delta>0$ *there exists a number* $\eta=\eta(\delta,T_1)>0$ *such that*

$$\sup_{\mathbb{R}\times[0,T_1]}|u(x,t)|\leqq\delta, \quad \textit{whenever} \ \sup_{\mathbb{R}}|u_0(x)|<\eta.\tag{3.11}$$

For a given $\delta>0$ we choose u_0 and η in accordance with Lemma 3.1. Since $\phi'(\cdot)$ and $\psi'(\cdot)$ are continuous and $\sup_{\mathbb{R}\times[0,T_1]}|u(x,t)|\leqq\delta$, assumptions (2.1), (2.2), and (2.3) imply that there exists positive constants α_i, $i=1,\cdots,4$ such that

$$\begin{aligned}&\alpha_1\leqq\phi'(u(x(t,\xi),t))\leqq\alpha_2, \qquad \phi''(u(x(t,\xi),t))\geqq\alpha_3,\\ &\alpha_4\leqq\frac{\psi'(u(x(t,\xi),\tau))}{\phi'(u(x(t,\xi),\tau))}\leqq\beta\end{aligned}\tag{3.12}$$

for $0 \leqq \tau \leqq t \leqq T_1$, where β is the a priori constant in (2.3). We note that the constants α_i depend on δ but not on $u_0'(\xi)$.

To proceed with the proof we shall also need to estimate r_ξ in (3.10), as well as $r_\xi - s_\xi$ in (3.9). For this purpose note from (3.4), (3.9) that

$$r_\xi(x(0,\xi),0) = \phi'(u_0(\xi)u_0'(\xi)) \qquad (\xi \in \mathbb{R}).$$

Let $C(\xi)$ and C^* be defined by

$$\begin{aligned} C(\xi) &= -\phi'(u_0(\xi))u_0'(\xi), \\ C^* &= \sup_{\xi \in \mathbb{R}} |C(\xi)|. \end{aligned} \tag{3.13}$$

We note that $C(\xi)$ is positive whenever $u_0'(\xi)$ is negative. We will now use the following auxiliary result; its proof is given in §4.

LEMMA 3.2. *Let the assumptions of Lemma* 3.1 *be satisfied. Select the datum* u_0 *such that* $u_0'(\xi) < 0$, *and there is a point* ξ_0 *such that* $C(\xi_0) = C^*$. *Then there exists* $0 < T_2 \leqq T_1$, *independent of* C^* (*hence of* $u_0'(\xi_0)$), *such that*

$$-\frac{7C^*}{4} \leqq r_\xi(x(t,\xi_0),t) - s_\xi(x(t,\xi_0),t) \leqq -\frac{C^*}{4}, \qquad -\frac{3C^*}{2} \leqq r_\xi(x(t,\xi_0),t) \leqq -\frac{C^*}{4} \tag{3.14}$$

for $0 \leqq t \leqq T_2$.

To complete the proof use equation (3.10) and the inequalities (3.12), as well as the inequality for $r_\xi(x(t,\xi_0),t)$ in (3.14), to obtain the estimate

$$v(t,\xi) \leqq \frac{\phi'(u(x(t,\xi_0)))}{\phi'(u_0(\xi_0))}\left[1 - \frac{C^*}{4}\frac{\alpha_3\alpha_1}{\alpha_2^2}t\right] \tag{3.15}$$

for $0 \leqq t \leqq T_2$. By (3.12), $\phi'(u(x(t,\xi_0)))$ is finite and bounded away from zero for $0 \leqq t \leqq T_2$. Thus the right-hand side of (3.15) becomes zero at time $t_1^* = 4\alpha_2^2/\alpha_1\alpha_3 C^*$. Since T_2 is independent of $u_0'(\xi_0)$, we now choose C^* (i.e. $-u_0'(\xi_0) > 0$) so large (cf. (3.13)) that $t_1^* \leqq T_2$, while keeping $u_0(\xi_0)$ fixed and $|u_0(\xi_0)| < \eta$. Finally, by (3.9), (3.12) and the first inequality in (3.14) $u_\xi(x(t,\xi_0),t) < 0$, remains finite and bounded away from zero on $0 \leqq t \leqq T_2$. Used in (3.7) this contradicts the assumption that $u_x(x,t)$ and $u_t(x,t)$ remain bounded on $\mathbb{R} \times [0,T_1]$. The final conclusion of Theorem 2.3 follows by Remark 2.2.

4. Proofs of Lemmas 3.1 and 3.2.

a. *Proof of Lemma* 3.1. It follows from (2.3), (2.4), (3.4) and (3.6) that

$$\begin{aligned} \frac{dr}{dt}(x(t,\xi),t) &\leq \beta|a'(0)|\big[|r(x(t,\xi),t)| + |s(x(t,\xi),t)|\big] \\ &\quad + \beta\int_0^t a''(t-\tau)\big[|r(x(t,\xi),\tau)| + |s(x(t,\xi),\tau)|\big]\,d\tau, \\ s_t(x,t) &\leq \beta|a'(0)|\big[|r(x,t)| + |s(x,t)|\big] \\ &\quad + \beta\int_0^t a''(t-\tau)\big[|r(x,\tau)| + |s(x,\tau)|\big]\,d\tau \end{aligned} \tag{4.1}$$

for $0 \leqq t \leqq T_1$. Let $R(t)$ and $S(t)$ be defined by

$$R(t) = \sup_{x \in \mathbb{R}} |r(x,t)|, \quad S(t) = \sup_{x \in \mathbb{R}} |s(x,t)|. \tag{4.2}$$

Integrating the inequalities (4.1), taking the supremum on the right-hand side and using the definitions (4.2), we obtain

$$\begin{aligned} |r(x(t,\xi),t)| &\leqq \sup_{\xi \in \mathbb{R}} |r_0(\xi)| + \beta |a'(0)| \int_0^t [R(\tau)+S(\tau)]\, d\tau \\ &\quad + \beta \int_0^t \int_0^\eta a''(\eta-\tau)[R(\tau)+S(\tau)]\, d\tau\, d\eta, \\ |s(x,t)| &\leqq \beta |a'(0)| \int_0^t [R(\tau)+S(\tau)]\, d\tau \\ &\quad + \beta \int_0^t \int_0^\eta a''(\eta-\tau)[R(\tau)+S(\tau)]\, d\tau\, d\eta \end{aligned} \tag{4.3}$$

for $0 \leqq t \leqq T_1$, where $r_0(\xi) = r(x(0,\xi),0) = \phi(u_0(\xi)), s(x(0,\xi),0) = 0$. We note that the right-hand side of (4.3) is independent of x and ξ. Moreover, from the smoothness of u, u_t, and u_x, assumption (2.1), and the continuous dependence of solutions of equation (3.5) on the initial data, it follows readily that for each fixed t, $t < T_1$, there exists $\xi \in \mathbb{R}$ and $x(t,\xi)$ such $r(x(t,\xi),t) = R(t)$ and $s(x(t,\xi),t) = S(t)$ hold. Therefore, we can replace the left-hand sides of (4.3) by $R(t)$ and $S(t)$ respectively. Interchanging the order of integration in the double integrals in (4.3) yields

$$\begin{aligned} R(t) &\leqq \sup_{\mathbb{R}} |r_0(\xi)| + 2\beta |a'(0)| \int_0^t [R(\tau)+S(\tau)]\, d\tau \\ &\quad + \beta \int_0^t a'(t-\tau)[R(\tau)+S(\tau)]\, d\tau, \\ S(t) &\leqq 2\beta |a'(0)| \int_0^t [R(\tau)+S(\tau)]\, d\tau + \beta \int_0^t a'(t-\tau)[R(\tau)+S(\tau)]\, d\tau, \end{aligned} \tag{4.4}$$

for $0 \leqq t \leqq T_1$. We add the two inequalities in (4.4) to obtain

$$R(t)+S(t) \leqq \sup_{\mathbb{R}} |r_0(\xi)| + \int_0^t \left[4\beta|a'(0)| + 2\beta a'(t-\tau)\right][R(\tau)+S(\tau)]\, d\tau \tag{4.5}$$

for $0 \leqq t \leqq T_1$. Let $H(t) = \max_{0 \leqq \tau \leqq t}\{4\beta|a'(0)| + 2\beta a'(t-\tau)\}$, which is a nonnegative function by (2.4). Thus

$$R(t)+S(t) \leqq \sup_{\mathbb{R}} |r_0(\xi)| + H(t) \int_0^t [R(\tau)+S(\tau)]\, d\tau, \qquad 0 \leqq t \leqq T_1, \tag{4.6}$$

and the Gronwall inequality yields the estimate

$$R(t)+S(t) \leqq \sup_{\mathbb{R}} |r_0(\xi)| f(t), \qquad 0 \leqq t \leqq T_1, \tag{4.7}$$

where the positive function $f(\cdot)$ is defined by

$$f(t) = 1 + H(t) \int_0^t \left(\exp \int_s^t H(\tau)\, d\tau \right) ds, \qquad 0 \leqq t \leqq T_1. \tag{4.8}$$

Since $r_0(\xi)=\phi(u_0(\xi))$, inequality (4.7), equations (3.4), and the monotonicity of f imply that

$$|\phi(u(x,t))|\leq \sup_{\mathbb{R}} |\phi(u_0(\xi))|f(T_1) \tag{4.9}$$

for $(x,t)\in\mathbb{R}\times[0,T_1]$. We observe that (4.10) is equivalent to

$$|u(x,t)|\leq \max_{I} |\phi^{-1}(\tau)|, \tag{4.10}$$

where

$$I=\left[-\sup_{\mathbb{R}} |\phi(u_0(\xi))|f(T_1),\ \sup_{\mathbb{R}} |\phi(u_0(\xi))|f(T_1)\right] \quad \text{for } (x,t)\in\mathbb{R}\times[0,T_1].$$

The proof of the lemma now follows from the continuity of ϕ and ϕ^{-1} and the fact that $\phi(0)=0$.

b. *Proof of Lemma* 3.2. We write the system (3.6) in the equivalent from

$$\begin{aligned}&\frac{dr}{dt}=a'(0)\psi(u(x(t,\xi),t))+\int_0^t a''(t-\tau)\psi(u(x(t,\xi),\tau))\,d\tau,\\ &s(x,t)=\int_0^t a'(t-\tau)\psi(u(x,\tau))\,d\tau, \qquad t\in[0,T_1].\end{aligned} \tag{4.11}$$

Integrating (4.11) with respect to t, differentiating the outcome with respect to ξ and using (3.9), we obtain

$$\begin{aligned}r_\xi(x(t,\xi),t)=&-C(\xi)+a'(0)\int_0^t \frac{\psi'(u(x(\tau,\xi),\tau))}{\phi'(u(x(\tau,\xi),\tau))}\\ &\cdot\left[r_\xi(x(\tau,\xi),\tau)-s_\xi(x(\tau,\xi),\tau)\right]d\tau\\ &+\int_0^t\int_0^\tau a''(\tau-\eta)\frac{\psi'(u(x(\tau,\xi),\eta))}{\phi'(u(x(\tau,\xi),\eta))}\\ &\cdot\left[r_\xi(x(\tau,\xi),\eta)-s_\xi(x(\tau,\xi),\eta)\right]d\eta\,d\tau,\\ s_\xi(x,t)=&\int_0^t a'(t-\tau)\frac{\psi'(u(x,\tau))}{\phi'(u(x,\tau))}\left[r_\xi(x,\tau)-s_\xi(x,\tau)\right]d\tau.\end{aligned} \tag{4.12}$$

Define ρ and σ by

$$\rho(t)=\sup_{x\in\mathbb{R}}|r_\xi(x,t)|, \qquad \sigma(t)=\sup_{x\in\mathbb{R}}|s_\xi(x,t)|. \tag{4.13}$$

Next, we take absolute values of both sides in (4.12), use the definitions (4.13) and inequalities (3.12) to obtain the inequalities

$$\begin{aligned}|r_\xi(x(t,\xi),t)|\leq& C^*+\beta|a'(0)|\int_0^t[\rho(\tau)+\sigma(\tau)]\,d\tau\\ &+\beta\int_0^t\int_0^\tau a''(\tau-\eta)[\rho(\eta)+\sigma(\eta)]\,d\eta\,d\tau,\\ |s_\xi(x,t)|\leq&\beta\int_0^t|a'(t-\tau)|[\rho(\tau)+\sigma(\tau)]\,d\tau,\end{aligned} \tag{4.14}$$

where C^* is defined in (3.13). Let $\Sigma(t)=\rho(t)+\sigma(t)$. As in the proof of Lemma 3.1, we can replace the left-hand sides of (4.15) by $\rho(t)$ and $\sigma(t)$. After simplifying the first inequality in (4.14) by interchanging the order in the double integral and adding the two inequalities (4.14), we obtain

$$(4.15)\qquad \Sigma(t)\leqq C^*+2\beta\int_0^t\big[|a'(0)|+|a'(t-\tau)|\big]\Sigma(\tau)\,d\tau,\qquad 0\leqq t\leqq T_1.$$

Noting that $\max_{0\leqq\tau\leqq t}[|a'(0)|+|a'(t-\tau)|]=2|a'(0)|$ (cf. (2.4)), (4.15) becomes

$$(4.16)\qquad \Sigma(t)\leqq C^*+4\beta|a'(0)|\int_0^t\Sigma(\tau)\,d\tau,\qquad 0\leqq t\leqq T_1$$

which, by the Gronwall inequality, implies that

$$(4.17)\qquad \Sigma(t)\leqq C^*\exp\big(4\beta|a'(0)|t\big),\qquad 0\leqq t\leqq T_1.$$

We now choose $T_2^*\leqq T_1$ small enough so that

$$(4.18)\qquad \Sigma(t)\leqq\frac{3C^*}{2},\qquad t\in\big[0,T_2^*\big].$$

Note that T_2^* depends only on $u_0(\xi)$ and $a(\cdot)$, and T_2^* is independent C^*. Inequalities (4.14) and (4.18) combine to yield

$$(4.19)\qquad \big|s_\xi(x,t)\big|\leqq\frac{3C^*}{2}\beta\int_0^t|a'(\eta)|\,d\eta,\qquad 0\leqq t\leqq T_2^*.$$

We further restrict T_2^* so that

$$(4.20)\qquad \big|s_\xi(x,t)\big|\leqq\frac{C^*}{4},\qquad 0\leqq t\leqq T_2^*,\quad x\in\mathbb{R}.$$

We observe that up to this point the sign of $u_0'(\xi)$ plays no role and the estimates (4.18), (4.20) hold for any $\xi\in\mathbb{R}$.

We next turn to estimating $r_\xi(x(t,\xi),t)$; the estimate $|r_\xi(x(t,\xi),t)|\leqq\frac{3}{2}C^*$ for $0\leqq t\leqq T_2^*$, which follows trivially from (4.18), is too crude to establish Lemma 3.2. We now select the datum u_0 and a point ξ_0 as specified in the statement of Lemma 3.2. The goal is to obtain a *negative* upper bound for $r_\xi(x(t,\xi_0),t)$; this is obtained from the first equation in (4.12) as follows. Using (3.12) and estimating the two integrals on the right-hand side of (4.12) as in (4.14), (4.15), and then using (4.18), we obtain the estimate

$$\begin{aligned}&|a'(0)|\int_0^t\frac{\psi'(u(x(\tau,\xi),\tau))}{\phi'(u(x(\tau,\xi),\tau))}\big|r_\xi(x(\tau,\xi),\tau)-s_\xi(x(\tau,\xi),\tau)\big|\,d\tau\\(4.21)\qquad&+\int_0^t\int_0^\tau a''(\tau-\eta)\frac{\psi'(u(x(\tau,\xi),\eta))}{\phi'(u(x(\tau,\xi),\eta))}\big|r_\xi(x(\tau,\xi),\eta)-s_\xi(x(\tau,\xi,\eta))\big|\,d\eta\,d\tau\\&\qquad\qquad\leqq 3C^*\beta|a'(0)|t,\end{aligned}$$

for $0\leqq t\leqq T_2^*$. Putting $\xi=\xi_0$ in (4.12) and then using (4.21) gives

$$(4.22)\qquad r_\xi\big(x(t,\xi_0),t\big)\leqq -C^*+3C^*\beta|a'(0)|t,\qquad 0\leqq t\leqq T_2^*,$$

where T_2^* is independent of C^*. Then choosing $0 < T_2 \leqq T_2^*$ small enough and independently of C^* we obtain

$$r_\xi(x(t,\xi_0),t) \leqq -\frac{C^*}{4}, \qquad 0 \leqq t \leqq T_2. \tag{4.23}$$

This, together with the crude lower bound $(-\frac{3}{2}C^*)$ already mentioned proves the second set of desired inequalities in (3.14). These combined with (4.20) (which of course holds $0 \leqq t \leqq T_2 \leqq T_2^*$) yield the first set of inequalities in (3.14), and the proof of Lemma 3.2 is complete.

REFERENCES

[1] C. M. Dafermos and J. A. Nohel, *A nonlinear hyperbolic Volterra equation in viscoelasticity*, American J. Math., Supplement (1981), pp. 81–116.

[2] G. Gripenberg, *Nonexistence of smooth solutions for shearing flows in a nonlinear viscoelastic fluid*, this Journal, 13 (1982), pp. 954–961.

[3] H. Hattori, *Breakdown of smooth solutions in dissipative nonlinear hyperbolic equations*, Quart. Appl. Math., 40 (1982/83), pp. 113–127.

[4] F. John, *Formation of singularities in one-dimensional nonlinear wave propagation*, Comm. Pure Appl. Math., 27 (1974), pp. 377–405.

[5] P. D. Lax, *Development of singularities of solutions of nonlinear hyperbolic differential equations*, J. Math. Phys., 5 (1964), pp. 611–613.

[6] ———, *Hyperbolic systems of conservation laws*, II, Comm. Pure Appl. Math., 10 (1957), pp. 227–241.

[7] P. Markowich and M. Renardy, *Lax-Wendroff methods for hyperbolic history valued problems*, this Journal, 14 (1983), pp. 66–97.

[8] J. A. Nohel, *A nonlinear conservation law with memory*, *Volterra and Functional Differential Equations*, Lecture Notes in Pure and Applied Mathematics, 81, Marcel Dekker, New York, 1982, pp. 91–123.

[9] M. Slemrod, *Instability of steady shearing flows in a nonlinear viscoelastic fluid*, Arch. Rational Mech. Anal., 68 (1978), pp. 211–225.

[10] J. P. Tordella, *Unstable flows of molten polymers*, Rheology Theory and Appl., 5 (1969), pp. 57–92.

[11] W. J. Hrusa and J. A. Nohel, *Global existence and asymptotics in one-dimensional nonlinear viscoelasticity*, Trends in Applications of Pure Mathematics to Mechanics, P. Ciarlet, ed., Lecture Notes in Physics 195, Springer-Verlag, New York, 1984, pp. 165–187.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

AN EIGENVALUE PROBLEM FOR A VOLTERRA INTEGRAL OPERATOR WITH INFINITE DELAY*

GUSTAF GRIPENBERG†

Abstract. Conditions are given under which it is possible to construct a unique, up to a constant multiple, nontrivial function x that satisfies an integrability criterion and the equation

$$x(t)=\int_{-\infty}^{t} k(t,s)x(s)\,ds, \quad \text{a.e. } t\in(-\infty,0].$$

1. Introduction. The purpose of this note is to show how one can, in a constructive way, find a nontrivial solution of the Volterra integral equation

$$x(t)=\int_{-\infty}^{t} k(t,s)x(s)\,ds, \quad \text{a.e. } t\in(-\infty,0]\overset{\text{def}}{=}\mathbb{R}^{-}, \tag{1.1}$$

under certain assumptions on the kernel k. Moreover, the uniqueness of this solution, in a certain class of functions satisfying an integrability criterion, will also be considered. The difficulties are, of course, due to the infinite delay in the equation since the function x is not given on some initial interval of the form $(-\infty,T]$ and thus one has to solve an eigenvalue problem. Once one has found a solution of (1.1) on $(-\infty,0]$, (or some other interval of the form $(-\infty,T]$), one can extend the solution to the right by solving the equation

$$x(t)=\int_{0}^{t}k(t,s)x(s)\,ds+f(t), \qquad t\geq 0, \quad f(t)=\int_{-\infty}^{0} k(t,s)x(s)\,ds,$$

(see e.g. [2] and [3]). In the case of a convolution kernel $k(t,s)=K(t-s)$, the solutions of equation (1.1) are linear combinations of functions of the form $e^{\sigma t}$ where σ satisfies $\int_0^\infty e^{-\sigma t}K(t)\,dt=1$, so this case is not very interesting. One approach to the problem of solving (1.1) that is not taken here, would be to consider (1.1) as a limiting form of equations of the type

$$x(t)=\int_{T}^{t}k(t,s)x(s)\,ds+f_T(t), \qquad t\in[T,0],$$

where $f_T(t)\to 0$ as $T\to-\infty$. For other results on equations with infinitive delay, see [1] and the references mentioned there. In the last section an example is given.

2. Statement of results. We will prove the following theorem.

THEOREM. *Assume that*

(2.1) $k(t,s)$ *is measurable and nonnegative on the set* $\{(t,s)\mid -\infty<s\leq t\leq 0\}$,

(2.2) $k(t,t)$ *is locally integrable on* $\mathbb{R}^{-}$,

(2.3) *there exists a measurable function* $a:\mathbb{R}^{-}\to(0,\infty)$ *such that for a.e.* $t\in\mathbb{R}^{-}$, *every* $s\in(-\infty,t]$ *and* $v\in(-\infty,s]$,

$$k(t,v)/a(v)\leq k(t,s)/a(s)\leq k(t,t)/a(t),$$

* Received by the editors January 20, 1983, and in revised form September 22, 1983.

† Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo 15, Finland.

(2.4) *there exist a number $T_0 \in \mathbb{R}^-$ and a nonnegative, nontrivial function z such that $a(t)z(t)$ is integrable on $(-\infty, T_0]$ and*

$$z(t) \leq \int_{-\infty}^{t} k(t,s)z(s)\,ds, \quad a.e.\ t \in (-\infty, T_0].$$

Then there exists a unique solution x of equation (1.1) *such that $a(t)x(t)$ is locally integrable on $\mathbb{R}^-$ and* $\lim_{T\to -\infty} \int_T^0 a(t)x(t)\,dt = 1$. *Moreover, this solution x is nonnegative and can be found with the aid of an iteration procedure.*

In the case when the solution x is nonnegative there are no problems with defining the integral $\int_{-\infty}^{t} k(t,s)x(s)\,ds$, but in general we take this integral to be $\lim_{T\to -\infty} \int_T^t k(t,s)x(s)\,ds$, provided that the limit exists. To see when the assumption (2.4) is satisfied one can try to take $z(t) = e^{\sigma t}/a(t)$ and then one sees that a sufficient condition for (2.4) to hold is that

$$\lim_{\sigma \to 0+} \lim_{T\to -\infty} \operatorname{ess-inf}_{t \leq T} a(t) \int_{-\infty}^{t} k(t,s) e^{-\sigma(t-s)}/a(s)\,ds > 1.$$

The crucial assumption, however, is (2.3), but note that if $a(t) \to \infty$ as $t \to -\infty$, e.g. if $a(t) = e^{-\alpha t}$, $\alpha > 0$, then we may strongly restrict the class possible solutions by demanding that $\lim_{T\to -\infty} \int_T^0 a(s)x(s)\,ds = 1$.

An instructive prototype for the kernel in (1.1) consists of kernels of the form $k_0(t,s) = \alpha(t)\beta(t-s)\gamma(s)$, where α, β and γ are nonnegative functions defined on $\mathbb{R}^-, \mathbb{R}^+$ and $\mathbb{R}^-$ respectively and where moreover β is nonincreasing and γ positive. Then (2.1) and (2.3) hold, ($a(t) = \gamma(t)$), and it is straightforward to give conditions that imply that (2.2) and (2.4) hold true (with e.g. $z(t) = e^{\sigma t}/a(t)$ in (2.4)).

3. Proof of the theorem. The outline of the proof is as follows: There exists an operator Q (defined in (3.10) below), such that a fixed-point of Q gives rise to a solution of (1.1), (cf. (3.15)). In order for our iteration procedure involving Q to work, we need some monotonicity results (cf. (3.3)), and a lower bound for the fixed point and this lower bound involves the function z given in (2.4). The uniqueness of the solution is established through a contradiction argument.

Now we proceed to the technical details of the proof and we let E be the set of exceptional points for which the inequalities in (2.3) or (2.4) do not hold.

Define

$$(3.1) \qquad b(t) = \begin{cases} k(t,t) & \text{if } t \in \mathbb{R}^- \setminus E, \\ 1 & \text{if } t \in E, \end{cases}$$

$$(3.2) \qquad h(t,s) = \begin{cases} a(t)k(t,s)/(a(s)b(t)) & \text{if } -\infty < s \leq t \leq 0,\ t \notin E \text{ and } b(t) > 0, \\ 0 & \text{if } -\infty < s \leq t \leq 0,\ t \notin E \text{ and } b(t) = 0, \\ 1 & \text{if } t < s \leq 0 \text{ or } t \in E. \end{cases}$$

The function h is measurable and by (2.3) and (3.2) we have

(3.3) for every $t \in \mathbb{R}^-$, $s \in \mathbb{R}^-$ and $v \in (-\infty, s]$, $h(t,v) \leq h(t,s) \leq 1$.

We extend the function z as 0 on $(T_0, 0]$ and we observe that we may then just as well assume that $T_0 = 0$ in (2.4). Next we define the function w by

$$(3.4) \qquad w(t) = \begin{cases} a(t)z(t)/b(t) & \text{if } t \in \mathbb{R}^- \setminus E, b(t) > 0, \\ 0 & \text{if } t \in \mathbb{R}^-, b(t) = 0, \text{ or } t \in E. \end{cases}$$

From (2.3) and (2.4) we see that we cannot have $w(t)\equiv 0$ on $\mathbb{R}^-$ because then we would also have $z(t)=0$ a.e. on $\mathbb{R}^-$. By (2.4), (3.2) and (3.4) we deduce that

$$w(t)\leq\int_{-\infty}^{t} h(t,s)b(s)w(s)\,ds, \qquad t\in\mathbb{R}^-. \tag{3.5}$$

Since (3.3) holds and $b(t)w(t)$ is integrable, we see that w is bounded and also that

$$\lim_{t\to-\infty} w(t)=0. \tag{3.6}$$

If $\int_{-\infty}^{0} b(s)\,ds<\infty$, then we conclude from (3.3) and (3.5) that

$$\sup_{t\leq T} w(t)\leq\int_{-\infty}^{T} b(s)\,ds\times\sup_{t\leq T} w(t)$$

so that $w(t)\equiv 0$ on $(-\infty,T)$ if $T<0$ with $|T|$ sufficiently large and hence it follows from (3.5) that $w(t)\equiv 0$ on $\mathbb{R}^-$. Thus we must have

$$\int_{-\infty}^{0} b(s)\,ds=+\infty. \tag{3.7}$$

Define the function W by

$$W(t)=w(t)+\int_{t}^{0} b(s)w(s)\,ds, \qquad t\in\mathbb{R}^-. \tag{3.8}$$

Since we may without loss of generality assume that

$$\int_{-\infty}^{0} b(s)w(s)\,ds=1,$$

it follows from (3.3), (3.5), (3.6) and (3.8) that

$$\sup_{t\leq 0} W(t)=\lim_{t\to-\infty} W(t)=1. \tag{3.9}$$

For every measurable, bounded function Y such that $Y(-\infty)\overset{\text{def}}{=}\lim_{t\to-\infty}Y(t)$ exists, we define the function $Q(Y)$ by

(3.10)

$$Q(Y)(t)=Y(-\infty)\int_{-\infty}^{0} b(s)\exp\left(-\int_{s}^{0} b(v)\,dv\right)h(t,s)\,ds$$
$$+\int_{-\infty}^{0}\int_{-\infty}^{s} b(u)\exp\left(-\int_{u}^{s} b(v)\,dv\right)(h(t,s)-h(t,u))\,du\,b(s)(Y(s)-Y(-\infty))\,ds,$$
$$t\in\mathbb{R}^-.$$

Since it follows from (3.2), (3.5) and (3.8) that

$$W(t)\leq\int_{-\infty}^{0} h(t,s)b(s)w(s)\,ds \quad \text{on } \mathbb{R}^-$$

and since it follows from (3.8) that

$$w(t)=W(t)-\int_{t}^{0} b(s)\exp\left(-\int_{t}^{s} b(u)\,du\right)W(s)\,ds,$$

it is possible to deduce from (3.3), (3.7) and (3.10) that

$$W(t)\le Q(W(t)), \qquad t\in\mathbb{R}^-. \tag{3.11}$$

Let

$$Y_0(t)=1, \quad Y_{n+1}(t)=Q(Y_n)(t), \qquad t\in\mathbb{R}^-, \quad n\ge 0. \tag{3.12}$$

Since (3.2) and (3.3) hold we conclude that $Y_n(-\infty)=1$ for all n and from (3.3) we also deduce that

$$Y_1(t)=Q(1)(t)=\int_{-\infty}^{0} b(s)\exp\left(-\int_s^0 b(u)\,du\right)h(t,s)\,ds.$$

This inequality combined with (3.3), (3.9), (3.11) and (3.12) shows that

$$W(t)\le Y_{n+1}(t)\le Y_n(t)\le 1, \qquad t\in\mathbb{R}^-. \tag{3.13}$$

But as $W(-\infty)=1$ we conclude that there exists a function Y on $\mathbb{R}^-$ such that

$$\lim_{n\to\infty} Y_n(t)=Y(t), \qquad t\in\mathbb{R}^-$$

and therefore it follows from (3.10), (3.12) and (3.13) that

$$Y(t)=Q(Y)(t), \quad W(t)\le Y(t)\le 1, \qquad t\in\mathbb{R}^-, \quad Y(-\infty)=1. \tag{3.14}$$

We define the function y by

$$y(t)=Y(t)-\int_t^0 b(s)\exp\left(-\int_t^s b(u)\,du\right)Y(s)\,ds, \qquad t\in\mathbb{R}^-. \tag{3.15}$$

A straightforward calculation shows that as a consequence of (3.14) we have

$$y(t)=\lim_{T\to-\infty}\int_T^t h(t,s)b(s)y(s)\,ds, \qquad t\in\mathbb{R}^-. \tag{3.16}$$

Next we will show that y is nonnegative. Since $\lim_{t\to-\infty} Y(t)=1$ exists it follows from (3.15) that $\lim_{t\to-\infty}y(t)=0$ and

$$\lim_{T\to-\infty}\int_T^t b(s)y(s)\,ds=y(t)+1-Y(t), \qquad t\in\mathbb{R}^-. \tag{3.17}$$

If we use (3.17) and an integration by parts in (3.16) then we conclude that

$$\begin{aligned}(1-h(t,-\infty))y(t)&=h(t,-\infty)(1-Y(t))\\&\quad+\int_{(-\infty,t]}\int_s^t b(u)y(u)\,du\,d_sh(t,s), \qquad t\in\mathbb{R}^-,\end{aligned} \tag{3.18}$$

where $h(t,-\infty)\overset{\text{def}}{=}\lim_{T\to-\infty}h(t,T)$. Since $Y(t)\le 1$ it follows from (3.15) that

$$y(t)\ge Y(t)-1+\exp\left(-\int_t^0 b(s)\,ds\right), \qquad t\in\mathbb{R}^-. \tag{3.19}$$

On the other hand, since $Y(-\infty)=1$ and (3.7) holds, one can rewrite the equation $Y=Q(Y)$ as

$$Y(t)=h(t,-\infty)+\int_{-\infty}^{0}\int_{-\infty}^{s} b(v)\exp\left(-\int_0^s b(v)\,dv\right)\cdot(h(t,s)-h(t,u))\,du\,b(s)Y(s)\,ds, \qquad t\in\mathbb{R}^-,$$

and hence by (3.3) and the facts that b and Y are nonnegative we have

$$Y(t)\geq h(t,-\infty), \qquad t\in\mathbb{R}^-. \tag{3.20}$$

Let

$$D=\{\,t\in\mathbb{R}^-|y(t)<0\},$$

and assume that D is nonempty. If $t\in D$, then it follows from (3.19) and (3.20) that $h(t,-\infty)<1$ and since also $Y(t)\leq 1$ we are able to conclude from (3.18) that

$$\inf_{s\leq t}\int_s^t b(u)y(u)\,du\leq y(t)<0 \quad \text{if } t\in D. \tag{3.21}$$

If for some $t\in D$

$$\lim_{T\to-\infty}\int_T^t b(u)y(u)\,du=\inf_{s\leq t}\int_s^t b(u)y(u)\,du$$

then it follows from (3.17) and (3.21) that $Y(t)\geq 1$ and this is a contradiction by (3.19) since $y(t)<0$. This means that for each $t\in D$ there exists a number $\tau(t)$ such that

$$\int_{\tau(t)}^{t} b(u)y(u)\,du=\inf_{s\leq t}\int_s^t b(u)y(u)\,du \tag{3.22}$$

and

$$\int_s^{\tau(t)} b(u)y(u)\,du>0 \quad \text{for each } s<\tau(t). \tag{3.23}$$

Fix $t_0\in D$. From (3.22) and (3.23) we see that

$$\tau(t)=\tau(t_0) \quad \text{for each } t\in D\cap(\tau(t_0),t_0).$$

Therefore we obtain from (3.21) and (3.22) the inequality

$$\chi_D(t)|y(t)|\leq\int_{\tau(t_0)}^{t} b(u)\chi_D(u)|y(u)|\,du, \qquad t\in(\tau(t_0),t_0),$$

and by Gronwall's Lemma we have $\chi_D(t)y(t)\equiv 0$ on $(\tau(t_0),t_0)$ which is a contradiction in view of (3.21) and (3.22). Thus we have shown that $y(t)\geq 0$, $t\in\mathbb{R}^-$.

If we now define the function x by

$$x(t)=b(t)y(t), \qquad t\in\mathbb{R}^-$$

then x is nonnegative. By (3.2) and (3.16) we see that (1.1) holds and by (3.15) and (3.17) we have

$$\int_{-\infty}^{0} a(t)x(t)\,dt=1.$$

It remains for us to establish the uniqueness of this solution. For purposes of contradiction, assume that there exists a nontrivial function q such that $a(t)q(t)$ is locally integrable with

$$\lim_{T\to-\infty}\int_T^0 a(t)q(t)\,dt=0 \tag{3.24}$$

and such that

$$q(t)=\lim_{T\to-\infty}\int_T^t k(t,s)q(s)\,ds,\quad \text{a.e. } t\in\mathbb{R}^-. \tag{3.25}$$

It follows from (2.3) and (3.25) that for a.e. t for which $k(t,t)=0$ we have $q(t)=0$. Hence we can find a set $E\subset\mathbb{R}^-$ with measure 0 such that if b and h are defined by (3.1) and (3.2) and the function p is defined by (3.4) with z and w replaced by q and p respectively, then we see by invoking (3.24) and (3.25) that (3.3) holds and that

$$\begin{aligned} p(t)&=\lim_{T\to-\infty}\int_T^t h(t,s)b(s)p(s)\,ds,\qquad t\in\mathbb{R}^-,\\ &\lim_{T\to-\infty}\int_T^0 b(s)p(s)\,ds=0. \end{aligned} \tag{3.26}$$

Observe that since q is not identically 0 it follows that p is nontrivial. If we define

$$P(t)=p(t)+\int_t^0 b(s)p(s)\,ds,$$

then we have by (3.26) and the fact that $p\not\equiv 0$

$$\sup_{t\le 0}|P(t)|\overset{\text{def}}{=}c>0,\qquad \lim_{T\to-\infty}P(t)=0. \tag{3.27}$$

From (3.26) we also deduce that

(3.28)

$$P(t)=\int_{-\infty}^0\int_{-\infty}^s b(u)\exp\left(-\int_u^s b(v)\,dv\right)(h(t,s)-h(t,u))\,du\,b(s)P(s)\,ds,\qquad t\in\mathbb{R}^-.$$

Choose T to be so small that

$$\sup_{t\le T}|P(t)|\le\frac{c}{2}\quad\text{and}\quad 2\exp\left(-\int_T^0 b(u)\,du\right)\le 1.$$

Then it follows from (3.3), (3.27) and (3.28) that

$$\begin{aligned} |P(t)|&\le c\int_T^0\int_{-\infty}^s b(u)\exp\left(-\int_u^s b(v)\,dv\right)(h(t,s)-h(t,u))\,du\,b(s)\,ds\\ &\quad+\frac{c}{2}\int_{-\infty}^T\int_{-\infty}^s b(u)\exp\left(-\int_u^s b(v)\,dv\right)(h(t,s)-h(t,u))\,du\,b(s)\,ds\\ &\le c\int_{-\infty}^0 b(u)\exp\left(-\int_u^0 b(v)\,dv\right)h(t,u)\,du\\ &\quad-\frac{c}{2}\int_{-\infty}^T b(u)\exp\left(-\int_u^T b(v)\,bv\right)h(t,u)\,du\\ &\le c\left(1-\exp\left(-\int_T^0 b(v)\,dv\right)\right), \end{aligned}$$

and thus we have obtained a contradiction in view of (3.27). This completes the proof of the theorem.

4. An example. The equation

$$y(t)=k(t)\left(P-\int_{-\infty}^{t}\alpha(t-s)y(s)\,ds\right)\int_{-\infty}^{t}\beta(t-s)y(s)\,ds \tag{4.1}$$

can be used to describe a very large class of epidemics of so called SIS or SIR type. Here we consider a constant population of fixed size P, and let y denote the rate at which individuals susceptible to the disease in question become infected. Now $\beta(t-s)$ denotes the infectivity (i.e. ability to infect others) at time t of an individual that became infected at time $s\le t$, and $\alpha(t-s)$ denotes the fraction of those individuals that became infected at time s that have not lost their immunity by the time $t\ge s$. Finally it is assumed that the rate at which susceptible individuals become infected is proportional to the number of susceptibles and the "total infectivity", with a (perhaps time-varying) proportionality constant $k(t)$.

Equation (4.1) has a trivial solution $y_0(t)\equiv 0$ and if we linearize equation (4.1) around this solution, then we get the equation

$$x(t)=k(t)P\int_{-\infty}^{t}\beta(t-s)x(s)\,ds.$$

This equation is of the form (1.1). An interesting case is the one where k is nonnegative, continuous and periodic. If we moreover assume that β is nonnegative and bounded on $[0,\infty)$ and there exists a number $\delta\ge 0$ such that $e^{-\delta t}\beta(t)$ is nonincreasing then it is easy to show that the assumptions of the theorem hold if we also have

$$\min_{t\in R}\int_{0}^{\infty}k(t-s)\beta(s)e^{-\delta s}\,ds>\frac{1}{P}.$$

REFERENCES

[1] C. CORDUNEANU AND V. LAKSHMIKANTHAM, *Equations with unbounded delay: a survey*, Nonlinear Anal., 4 (1980), pp. 831–877.

[2] G. GRIPENBERG, *On the resolvents of nonconvolution Volterra kernels*, Funkcial. Ekvac., 23 (1980), pp. 83–95.

[3] R. K. MILLER, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Menlo Park, CA, 1971.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

011

ON THE SHARPNESS OF WEYL'S ESTIMATE FOR EIGENVALUES OF SMOOTH KERNELS*

J. B. READE†

Abstract. It is shown that Weyl's estimate of $o(1/n^{3/2})$ for the eigenvalues of any symmetric continuously differentiable kernel on a bounded region cannot be improved to $o(1/n^{3/2}\alpha_n)$ for any increasing $\alpha_n \to \infty$. The counter-example is constructed from Rudin–Shapiro polynomials.

1. Introduction. Let $K(x,t)=\overline{K(t,x)}\in L^2[a,b]^2$ be a symmetric square integrable kernel giving rise to a compact symmetric operator

$$Tf(x)=\int_a^b K(x,t)f(t)\,dt$$

on $L^2[a,b]$. The eigenvalues of T form a real sequence (λ_n) which converges to zero. It is therefore possible to arrange this sequence in descending order of modulus

$$|\lambda_1|\geqq|\lambda_2|\geqq\cdots\geqq|\lambda_n|\geqq\cdots,$$

and we shall always assume this has been done.

It is a classical result of H. Weyl (see [4]) that, if $K(x,t)\in C^1[a,b]^2$ has continuous partial derivatives, then $\lambda_n=o(1/n^{3/2})$. Recently, the present author has shown (see [2]) that, if in addition $K(x,t)$ is positive definite, then $\lambda_n=o(1/n^2)$.

We shall show that these results are best possible in the following sense. Suppose that (α_n) is any increasing positive real sequence which diverges to infinity. Then there exist symmetric kernels in $C^1[a,b]^2$ whose eigenvalues are not $o(1/n^{3/2}\alpha_n)$, and positive definite kernels in $C^1[a,b]^2$ whose eigenvalues are not $o(1/n^2\alpha_n)$.

2. Fourier series. Any $k(t)\in L^1[0,1]$ has a Fourier series $\Sigma_{-\infty}^{\infty}c_ne^{2\pi int}$ where

$$c_n=\int_0^1 k(t)e^{-2\pi int}dt$$

are the Fourier coefficients. If $k(-t)=\overline{k(t)}$, then the difference kernel $k(x-t)$ gives rise to a compact symmetric operator

$$Tf(x)=\int_0^1 k(x-t)f(t)\,dt$$

on $L^2[0,1]$ whose eigenvalues are (c_n), since

$$\int_0^1 k(x-t)e^{2\pi int}dt=\int_0^1 k(t)e^{2\pi in(x-t)}dt=c_ne^{2\pi inx}$$

and $(e^{2\pi int})$ form an orthonormal basis of $L^2[0,1]$.

* Received by the editors June 17, 1983.

† Department of Mathematics, The University, Manchester M13 9PL, England.

3. Rudin–Shapiro polynomials. The Rudin–Shapiro polynomials $P_n(z)$, $Q_n(z)$ are defined inductively for $n \geqq 0$ as follows.

$$\begin{aligned} &P_0(z)=Q_0(z)=1,\\ &P_{n+1}(z)=P_n(z)+z^{2^n}Q_n(z),\\ &Q_{n+1}(z)=P_n(z)-z^{2^n}Q_n(z) \end{aligned}$$

for all $n \geqq 0$. The Rudin–Shapiro signs are $\varepsilon_n = \pm 1$ where $P_n(z)=\Sigma_0^{2^n-1}\varepsilon_n z^n$. They have the remarkable property that

$$s_N(t)=\sum_0^N \varepsilon_n e^{2\pi i n t}=O(N^{1/2})$$

uniformly in t. (See [3].)

4. Construction of the counterexample. Let α_n increase and diverge to infinity. Let β_n be defined as follows. Choose n_k such that $\alpha_{n_k} > k^2$. Then let

$$\beta_n=\begin{cases} 1/n_1^{1/2}\alpha_{n_1} & \text{for all } 1 \leqq n \leqq n_1,\\ 1/n_k^{1/2}\alpha_{n_k} & \text{for all } n_{k-1}<n\leqq n_k, k \geqq 2. \end{cases}$$

Observe that (β_n) is a decreasing sequence which is $O(1/n^{1/2}\alpha_n)$ but not $o(1/n^{1/2}\alpha_n)$. Also $\Sigma_1^\infty(\beta_n-\beta_{n+1})n^{1/2}<\infty$ since

$$\begin{aligned} \sum_{n=1}^{\infty}(\beta_n-\beta_{n+1})n^{1/2} &= \sum_{k=1}^{\infty}(\beta_{n_k}-\beta_{n_{k+1}})n_k^{1/2}\\ &<\sum_{k=1}^{\infty}\beta_{n_k}n_k^{1/2}\\ &=\sum_1^\infty 1/\alpha_{n_k}<\sum_1^\infty 1/k^2<\infty. \end{aligned}$$

It follows that the series $\Sigma_1^\infty \varepsilon_n\beta_n e^{2\pi i n t}$, where (ε_n) are the Rudin–Shapiro signs, is uniformly convergent in t, since

$$\begin{aligned} \left|\sum_M^N \varepsilon_n\beta_n e^{2\pi i n t}\right| &= \left|\sum_M^N \beta_n(s_n(t)-s_{n-1}(t))\right|\\ &=\left|-\beta_M s_{M-1}(t)+\sum_M^{N-1}(\beta_n-\beta_{n+1})s_n(t)+\beta_N s_N(t)\right|\\ &\leqq A\left(\beta_M M^{1/2}+\sum_M^{N-1}(\beta_n-\beta_{n+1})n^{1/2}+\beta_N N^{1/2}\right), \end{aligned}$$

where A is an absolute constant,

$$\to 0$$

as M, $N\to\infty$ uniformly in t. Hence $k(t)=\Sigma_1^\infty c_n e^{2\pi i n t}$, where $c_n=\varepsilon_n\beta_n/n$, is in $C^1[0,1]$, and the kernel $k(x-t)\in C^1[0,1]^2$ has eigenvalues (c_n) which are not $o(1/n^{3/2}\alpha_n)$.

5. Positive definite kernels. If α_n increases and diverges to infinity, then the series $\sum_1^\infty \sin \pi nt/n\alpha_n$ is uniformly convergent in t, since $1/\alpha_n$ decreases and converges to zero and $\sum_1^\infty \sin \pi nt/n$ has uniformly bounded partial sums. (See e.g. [1, p. 6].) It follows that $k(t) = \sum_1^\infty \cos \pi nt/n^2\alpha_n \in C^1[-1,1]$, and therefore the kernel $K(x,t) = \sum_1^\infty \cos \pi nx \cos \pi nt/n^2\alpha_n \in C^1[-1,1]^2$ is positive definite having eigenvalues $(1/n^2\alpha_n)$.

REFERENCES

[1] J. P. Kahane, *Séries de Fourier absolument convergentes*, Ergebnisse der Mathematik und ihrer Grenzgebiete, 50, Paris 1970.

[2] J. B. Reade, *Eigenvalues of positive definite kernels*, this Journal, 14 (1983), pp. 152–157.

[3] W. Rudin, *Some theorems on Fourier coefficients*, Proc. Amer. Math. Soc., 10 (1959), pp. 855–859.

[4] H. Weyl, *Das Asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen.* Math. Ann., 71 (1912), pp. 441–479.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

012

LIMITS OF DILATED CONVOLUTION TRANSFORMS*

W. R. MADYCH[†]

Abstract. If k is a kernel so that the convolution transform $f \to k * f$ maps $L^p(R^n)$ into $L^p(R^n)$ we study the behavior in L^p of $k_t * f$, $t > 0$, as t goes to 0 or ∞; here k_t is the dilated kernel defined by $k_t(x) = t^{-n}k(t^{-1}x)$. In particular, we give conditions on k which imply that $\lim k_t * f(x) = \hat{k}(0)f(x)$ in L^p norm as t goes to 0 and conditions which imply $\lim k_t * f(x) = 0$ in L^p as t goes to ∞; these conditions are practically necessary. Generalizations to other notions of dilation are also indicated.

1. Introduction. Suppose k is a convolution kernel such that the transformation $f \to k * f$ maps $L^p(R^n)$ continuously into $L^p(R^n)$ for some p, $1 \leqq p \leqq \infty$, and integer $n \geqq 1$, where $k * f(x) = \int_{R^n} k(x-y)f(y)\,dy$ is the convolution of k with f. It is the purpose of this paper to record several results concerning the behavior for positive t of $k_t * f$ as t goes to 0 or ∞ where $k_t(x) = t^{-n}k(t^{-1}x)$. The point of view taken here is that, roughly speaking, certain aspects of this behavior can be determined from the behavior of the Fourier transform of k at the origin or infinity respectively. These results are motivated by the paper of Logan [2] where part of the case $n = p = 1$ and $t \to \infty$ was studied and questions concerning $p > 1$ were raised. (Note that $a = t^{-1}$ in the notation used there.)

In the general case considered here it is possible that such kernels k are tempered distributions which are not locally integrable. A classical example when $n = 1$ is the Hilbert transform. Thus it is convenient to think of the transforms studied here as general translation invariant operators on $L^p(R^n)$. The kernels k arise as follows: If the transformation $f \to Kf$ is a continuous translation invariant linear operator from $L^p(R^n)$ into $L^p(R^n)$ then there is a tempered distribution k on R^n so that $K\phi = k * \phi$ for all ϕ in $\mathscr{S}(R^n)$; $\mathscr{S}(R^n)$ denotes the space of infinitely differentiable and rapidly decreasing functions and $*$ denotes convolution in the distribution sense. Conversely, if k is a tempered distribution such that

$$\|k * \phi\|_p \leqq C\|\phi\|_p \tag{1}$$

holds for all ϕ in $\mathscr{S}(R^n)$, where $\|\cdot\|_p$ denotes the $L^p(R^n)$ norm and C is a constant independent of ϕ, then the transformation $\phi \to k * \phi$ can be extended continuously to all of $L^p(R^n)$.

Thus we think of the kernels in question as tempered distributions, k, which satisfy (1) for some p, $1 \leqq p \leqq \infty$, and all ϕ in $\mathscr{S}(R^n)$. This is roughly the point of view in [1] where the basic facts together with other very interesting material concerning such operators can be found. We also adopt the notation found in [1] which is quite standard.

Recall that $\hat{f}$ denotes the Fourier transform of f. The space L_p^p is the space of all those tempered distributions k for which (1) holds for all ϕ in $\mathscr{S}$ and the norm of an element k in L_p^p, denoted by $\|k\|_p^*$, is the smallest constant C in that inequality which is valid for all such ϕ. The space M_p^p is the Fourier transform of L_p^p, namely, f is in M_p^p if

* Received by the editors February 3, 1983, and in revised form, October 1, 1983. This work was partially supported by the National Science Foundation under grant MCS-8202147.

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

and only if $\hat{f}$ is in L_p^p and the M_p^p norm of f is the L_p^p norm of $\hat{f}$. Also recall that M_p^p is contained in L^∞, $1 \leqq p \leqq \infty$.

The following subclasses of M_p^p will play an important role in what follows:

(i) m_p^p is the closure of $\mathscr{S}$ in M_p^p.

(ii) $M_{p,c}^p$ is the class of those elements f in M_p^p for which there is a sequence $\{f_n\}$ of elements in M_p^p, each with compact support, such that $\{f_n\}$ converges to f in M_p^p.

(iii) $M_{p,0}^p$ is the class of those elements f in M_p^p for which the measure of $\{\xi: |f(\xi)| > \varepsilon\}$ is finite for each positive ε.

(iv) $M_{p,z}^p$ is the set of those elements f in M_p^p for which $f(t\xi)$ converges to 0 locally in measure as $t \to \infty$, namely for each ε, $0 < \varepsilon < \infty$, and r, $0 < r < \infty$, the measure of $\{\xi: |f(t\xi)| > \varepsilon\} \cap \{\xi: |\xi| < r\}$, $t > 0$, converges to 0 as t goes to ∞.

Observe that $m_p^p \subset M_{p,c}^p \subset M_{p,0}^p \subset M_{p,z}^p$. In the case $p=2$ it is clear that all the containments are proper. In the case $1 < p < 2$ elementary examples show that $m_p^p \subset M_{p,c}^p$ and $M_{p,0}^p \subset M_{p,z}^p$. For $p=1$ $m_1^1 = M_{1,c}^1 =$ the Fourier transform of L^1 and $M_{1,c}^1 \subset M_{1,0}^1$ where the containment is known to be proper, see [1, p. 111].

We say that a distribution f is in m_p^p in a neighborhood of some point ξ_0, if there is a positive ε such that $\psi(\varepsilon^{-1}(\xi - \xi_0))f(\xi)$ is in m_p^p. Here ψ is a nonnegative infinitely differentiable function such that $\psi(\xi) = 0$ for $|\xi| > 1$ and $\psi(\xi) = 1$ for $|\xi| \leqq 1/2$. A distribution f is said to be locally in m_p^p if it is in m_p^p in the neighborhood of every point in R^n.

Since the kernels k are, in general, tempered distributions which may not be locally integrable the pointwise definition of k_t, $t \to 0$, given in the first paragraph of this chapter need not necessarily make sense. However k_t, $t > 0$, can be defined by duality as the tempered distribution for which

$$(2) \qquad \langle k_t, \phi \rangle = \langle k, \phi^t \rangle$$

for all ϕ in $\mathscr{S}(R^n)$ where $\phi^t(x) = \phi(tx)$ and $\langle k, \phi \rangle$ denotes the distribution k evaluated at ϕ. Also, since the Fourier transform of $\phi(tx)$ is $t^{-n}\hat{\phi}(t^{-1}x)$ and $M_p^p \subset L^\infty(R^n)$, for k in L_p^p one can define k_t by the formula

$$(3) \qquad \hat{k}_t(\xi) = \hat{k}(t\xi)$$

which holds for almost all ξ in R^n. Note that the L_p^p norm of $k_{,} t > 0$, is the same as that of k, namely, the transformation $k \to k_t$ does not change L_p^p norm, $1 \leqq p \leqq \infty$.

2. The case $t \to 0$. The fact that functions whose Fourier transform have compact support are dense in L^p, $1 \leqq p < \infty$, is a key ingredient in proving the following.

THEOREM 1. *Suppose k is in L_p^p and f is in L^p for some p, $1 \leqq p < \infty$. If $\hat{k}$ is in m_p^p in a neighborhood of the origin then*

$$(4) \qquad \lim_{t \to 0} k_t * f(x) = \hat{k}(0) f(x)$$

in L^p norm.

Recall that $L_1^1 = \mathscr{M}$, the space of bounded Borel measures on R^n. Since it is easy to see that $\mathscr{M}$ is locally in m_1^1, application of Theorem 1 makes the following fact transparent.

COROLLARY. *If k is in $\mathscr{M}$ and f is in L^p for some p, $1 \leqq p < \infty$, then* (4) *holds in L^p norm.*

Various conditions in $\hat{k}$ imply that $\hat{k}$ is in m_p^p in a neighborhood of the origin. For example,

(i) $\hat{k}$ is infinitely differentiable in a neighborhood of zero or the more complicated,

(ii) all derivatives of order less than $n/2+1+\varepsilon$, $\varepsilon>0$, of $\hat{k}$ exist and are continuous in a neighborhood of zero

both imply that $\hat{k}$ is in m_1^1 in a neighborhood of the origin.

Observe that if k_t converges in the distribution sense as t goes to 0 then the limit must be a distribution homogeneous of degree $-n$. In particular, if $\hat{k}(t\xi)$ converges pointwise to $\hat{h}(\xi)$, then $\hat{h}$ is homogeneous of degree 0, namely $\hat{h}(t\xi)=\hat{h}(\xi)$ for $\xi\neq 0$ and $t>0$, and if $\hat{k}$ is in M^p, then so is $\hat{h}$. If $\hat{k}-\hat{h}$ is also in m_p^p in a neighborhood of the origin, then Theorem 1 implies that

$$\lim_{t\to 0} k_t * f(x) = h * f(x) \tag{5}$$

in L^p norm. However the following may be more interesting. (Recall that $\hat{k}(t\xi)$ converges locally in measure to $\hat{h}(\xi)$ as t goes to 0 means that for each ε, $0<\varepsilon<\infty$, and r, $0<r<\infty$, the measure of $\{\xi; |\hat{k}(t\xi)-\hat{h}(\xi)|>\varepsilon, |\xi|<r\}$ goes to 0 as t goes to zero).

THEOREM 2. *Suppose $\hat{k}$ is in M_p^p for some p, $1\leqq p<\infty$, and $\hat{k}(t\xi)$ converges locally in measure to $\hat{h}(\xi)$ as t goes to 0. Then $\hat{h}$ is homogeneous of degree 0, is in M_p^p and (5) holds weakly in L^p whenever f is in L^p. Furthermore*

(i) *if $p=2$ then (5) holds in L^2 norm whenever f is in L^2;*

(ii) *if $1<p<2$, then (5) holds in L^q norm whenever f is in L^q and $p<q\leqq p/(p-1)$;*

(iii) *if $2<p<\infty$, then (5) holds in L^q norm whenever f is in L^q and $p/(p-1)<q\leqq p$.*

If $\hat{k}$ is continuous at 0 then $\hat{h}$ is a constant or, Fourier transforming, h is a constant multiple of the Dirac measure at the origin. More precisely we have the following.

COROLLARY. *Suppose $\hat{k}$ is in M_p^p for some p, $1\leqq p\leqq\infty$, and $\hat{k}$ is continuous at 0. Then statements (i), (ii), and (iii) of Theorem 2 hold with (5) replaced by (4).*

The following indicates that Theorem 2 is best possible in some sense.

THEOREM 3. *Suppose k is in L_p^p and $k_t * f$ converges weakly in L^p as t goes to 0 for all f in L^p. Call this limit Kf. Then*

(i) *the transformation $f\to Kf$ is a continuous translation invariant linear operator on $L^p(R^n)$;*

(ii) *$Kf=h*f$ for all f in $\mathscr{S}$ where h is a tempered distribution homogeneous of degree $-n$, that is $h_t=h$ for all $t>0$;*

(iii) *$\hat{k}(t\xi)$ converges to $\hat{h}(\xi)$ locally in measure;*

(iv) *statements (i), (ii), and (iii) of Theorem 2 hold.*

Before closing this section we wish to mention that variants of the theorems and corollaries in this section and the next hold in the case $p=\infty$ if an appropriate substitute for L^∞ is used. We do not explicitly state these simple extensions in order not to complicate the statement of the theorems. For example, Theorem 1 and its corollary hold in the case $p=\infty$ if L^∞ is replaced by C_0, the space of continuous functions which have limit zero at infinity with the supremum norm. Using this same substitution when $q=\infty$ statements (ii) and (iii) of Theorem 2 hold when $p=1$ or ∞ respectively; the same is true of Theorems 4, 5 and 6 in the next section.

3. The case $t\to\infty$. The first theorem below is a consequence of the fact that functions whose Fourier transforms vanish in a neighborhood of the origin are dense in L^p, $1<p<\infty$.

THEOREM 4. *Suppose $\hat{k}$ is in $M_{p,c}^p$ and f is in L^p for some p, $1<p<\infty$. Then*

$$\lim_{t\to\infty} k_t * f(x) = 0 \tag{6}$$

in L^p norm.

Since $\widehat{L^1} \subset M^p_{p,c}$ for all p, Theorem 4 implies the n-dimensional analogue of [2, formula (9)].

COROLLARY. *Suppose k is in L^1 and f is in L^p for some p, $1<p<\infty$, then* (6) *holds in L^p norm.*

As in the case $t\to 0$, if $\hat{k}(t\xi)$ converges in some sense as t goes to ∞ then that limit must be homogeneous of degree 0. The same argument used to prove Theorem 2 can be used to prove the following.

THEOREM 5. *Suppose $\hat{k}$ is in M^p_p for some p, $1\leqq p\leqq\infty$, and $\hat{k}(t\xi)$ converges locally in measure to $\hat{h}(\xi)$ as t goes to ∞. Then $\hat{h}$ is homogeneous of degree* 0, *is in M^p_p, and*

$$\lim_{t\to\infty} k_t * f(x) = h * f(x) \tag{7}$$

weakly in L^p whenever f is in L^p. Furthermore, statements (i), (ii), *and* (iii) *of Theorem* 2 *hold with* (5) *replaced by* (7).

COROLLARY. *Suppose $\hat{k}$ is in $M^p_{p,z}$ and f is in L^p for some p, $1<p<\infty$, then statements* (i), (ii), *and* (iii) *of Theorem* 2 *hold with* (5) *replaced by* (6).

THEOREM 6. *Theorem* 3 *is true if the statement* "t goes to 0" *is replaced by the statement* "t goes to ∞" *throughout.*

The case $p=1$ is essentially settled by the theorem in [2, formulas (10) and (11)]. The proof given there also works in the n-dimensional case. However, we do record the following corollary.

THEOREM 7. *If k and f are both in L^1 then $k_t * f(x)$ converges to* 0 *in measure as t goes to ∞ and* $\lim_{t\to\infty}\|k_t * f\|_1 = |\hat{f}(0)|\|k\|_1$.

4. Details and further remarks. In what follows we will always use the symbol $\hat{\phi}$ to denote a function which is infinitely differentiable on R^n and satisfies $\hat{\phi}(\xi)=1$ for $|\xi|\leqq 1/2$ and $\hat{\phi}(\xi)=0$ for $|\xi|>1$; ϕ denotes its Fourier transform. Recall that $\phi_t(x)=t^{-n}\phi(t^{-1}x)$ and $\widehat{(\phi_t)}(\xi)=\hat{\phi}(t\xi)$, $t>0$.

As mentioned earlier, the following lemma is a key ingredient in the proof of both Theorems 1 and 4.

LEMMA 1. *$\mathscr{S}$ is dense in L^p, $1\leqq p<\infty$, and C_0. Furthermore,*

(i) *the subspace of $\mathscr{S}$ consisting of functions whose Fourier transforms have compact support is dense in L^p, $1\leqq p<\infty$, and in C_0, and*

(ii) *the subspace of $\mathscr{S}$ consisting of functions whose Fourier transform vanishes in a neighborhood of the origin is dense in L^p, $1<p<\infty$, and in C_0.*

Proof. Both the initial statement and (i) are well known and well documented. Statement (ii) seems to be less well known but is probably folklore; we outline its proof.

If f is any function in $\mathscr{S}$, write $g=f-\phi_t * f$. Then $\hat{g}$ vanishes in a neighborhood of the origin and

$$\|g-f\|_p = \|\phi_t * f\|_p \leqq \|\phi_t\|_p\|f\|_1 = t^{-n/p'}\|\phi\|_p\|f\|_1$$

where p' is the Hölder conjugate of p, i.e. $p'=p/(p-1)$. It follows that $\|g-f\|_p$ can be made arbitrarily small if t is sufficiently large and $1<p\leqq\infty$. Since $\mathscr{S}$ is dense in L^p, $1<p<\infty$, and C_0 the desired result follows.

Proof of Theorem 1. If k is in $\mathscr{S}$ the result is well known. In the general case if $\hat{f}$ has compact support write

$$k_t * f - \hat{k}(0)f = l_t * f + k_t * \phi_{t/\varepsilon} * f - g_t * f + g_t * f - \hat{g}(0)f + \hat{g}(0)f - \hat{k}(0)f$$

where $l_t = k_t - k_t * \phi_{t/\varepsilon}$ and g is in $\mathscr{S}$. Thus

$$\|k_t * f - k(0)f\|_p \leqq \|l_t * f\|_p + \|k * \phi_{1/\varepsilon} - g\|_p^* \|f\|_p$$
$$+ \|g_t * f - g(0)f\|_p + |\hat{g}(0) - \hat{k}(0)| \|f\|_p.$$

Now, since $\widehat{k * \phi_{1/\varepsilon}}$ is in m_p^p whenever ε is sufficiently small, we can choose g so that both $\|k * \phi_{1/\varepsilon} - g\|_p^*$ and $|\hat{g}(0) - \hat{k}(0)|$ are arbitrarily small. Since $\hat{f}$ has compact support and $\hat{k}(\xi) - \hat{k}(\xi)\hat{\phi}(\varepsilon^{-1}\xi)$ vanishes for $|\xi| < \varepsilon/2$ it follows that $l_t * f = 0$ if t is sufficiently small. Furthermore since g is in $\mathscr{S}$ $\|g_t * f - \hat{g}(0)f\|_p$ can be made arbitrarily small for t small enough.

Altogether we see that whenever $\hat{f}$ has compact support, $\|k_t * f - \hat{k}(0)f\|_p$ can be made arbitrarily small if k is in L_p^p, $\hat{k}$ is in m_p^p in a neighborhood of the origin, and t is sufficiently small. Since such f 's are dense in L^p, this is true for all f in $L^p(R^n)$.

Most of Theorem 2 can be restated in the following more general lemma. Notice that in the statement below the parameter α is simply an index for the family of distributions k_α; it does not necessarily relate to any notion of dilation as the parameter t does.

LEMMA 2. *Suppose* k_α, $0 < \alpha < 1$, *is a family of distributions in* L_p^p, $1 \leqq p < \infty$, *such that* $\|k_\alpha\|_p^*$ *is uniformly bounded for all* α.

(A) *If* $\hat{k}_\alpha$ *converges locally in measure to a distribution* $\hat{k}$ *as* α *goes to* 0 *then* $\hat{k}$ *is in* M_p^p *and*

$$\lim_{\alpha \to 0} k_\alpha * f(x) = k * f(x) \tag{8}$$

weakly in L^p *for all* f *in* L^p. *Furthermore*

(i) *if* $p = 2$ *then* (8) *holds strongly in* L^p *norm*;

(ii) *if* $1 < p < 2$ *and* f *is in* L^q *then* (8) *holds strongly in* L^q *norm whenever* $p < q \leqq p/(p-1)$;

(iii) *if* $2 < p < \infty$ *and* f *is in* L^q, *then* (8) *holds strongly in* L^q *norm whenever* $p/(p-1) < q \leqq p$.

(B) *Conversely, if* $k_\alpha * f$ *converges weakly in* L^p *for all* f *in* L^p *as* α *goes to* 0, *then there is a* $\hat{k}$ *in* M_p^p *such that* $\hat{k}_\alpha$ *converges locally in measure to* $\hat{k}$ *as* α *goes to* 0. *Furthermore statements* (i)–(iii) *above also hold.*

Statements (ii) and (iii) also hold in the cases $p = 1$ and ∞ if L^∞ is replaced by C_0.

Proof. Suppose $\hat{k}_\alpha$ converges locally in measure to $\hat{k}$ as α goes to 0. Then the fact that

$$\lim_{\alpha \to 0} \int_{R^n} (k_\alpha * f(x) - k * f(x))\overline{g(x)}\, dx = 0 \tag{9}$$

holds for all f in L^p and all g in $L^{p/(p-1)}$ follows from (a) the fact that Plancherel's theorem and the Lebesgue dominated convergence theorem imply that (9) holds whenever f and g are both in $\mathscr{S}$ and have compactly supported Fourier transforms together with (b) the fact that such functions are dense in L^p and $L^{p/(p-1)}$. By writing

$$\|k_\alpha * f - k * f\|_2^2 = \int_{R^n} |[\hat{k}_\alpha(\xi) - \hat{k}(\xi)]\hat{f}(\xi)|^2 d\xi$$

we see that statement (i) also follows by essentially the same argument.

If $\hat{f}$ is in $\mathscr{S}$ and has compact support and $2 \leqq p < \infty$ then using the Hausdorff–Young inequality we may write

$$\|k_\alpha * f - k * f\|_p \leqq \left\{ \int_{R^n} |(\hat{k}_\alpha(\xi) - \hat{k}(\xi))\hat{f}(\xi)|^{p/(p-1)} dx \right\}^{(p-1)/p}$$

and applying the Lebesgue dominated convergence theorem gives us $\lim_{\alpha \to 0} \|k_\alpha * f - k * f\|_p = 0$. If $1 < p < 2$ then Hölder's inequality and Plancherel's formula imply that if $p < q < 2$ then

$$\|k_\alpha * f - k * f\|_q^q \leqq A^\theta B^{1-\theta},$$

where $\theta = (2-q)/(2-p)$, $A = \|k_\alpha * f - k * f\|_p$, and

$$B^2 = \int_{R^n} |(\hat{k}_\alpha(\xi) - \hat{k}(\xi))\hat{f}(\xi)|^2 d\xi.$$

Thus, since A is uniformly bounded, the dominated convergence theorem applied to B implies that $\lim_{\alpha \to 0} \|k_\alpha * f - k * f\|_q = 0$ for such q. The density of such f's in L^p, $1 \leqq p < \infty$, and duality imply both statements (ii) and (iii).

Suppose $k_\alpha * f$ converges weakly to Kf in L^p as α goes to 0. Then the transformation $f \to Kf$ is a continuous, translation invariant operator on L^p and thus, [1, Thm. 1.2], there is a k in L_p^p such that $Kf = k * f$ for all f in $\mathscr{S}$. Now

$$m\{\xi: |k_\alpha(\xi) - k(\xi)| > \varepsilon, |\xi| < r\} \leqq I_\alpha(\varepsilon, r) = \varepsilon^{-1} \int_{R^n} |(\hat{k}_\alpha(\xi) - \hat{k}(\xi))| |\hat{\phi}(\xi/2r)|^2 d\xi,$$

where m denotes the Lebesgue measure of the set. Since

$$I_\alpha(\varepsilon, r) = \varepsilon^{-1} \int_{R^n} (k_\alpha * \phi_{1/2r}(x) - k * \phi_{1/2r}(x)) \overline{\phi_{1/2r}(x)}\, dx$$

it follows that for fixed ε and r, $\lim_{\alpha \to 0} I_\alpha(\varepsilon, r) = 0$ which implies the initial statement in (B). The rest of (B) now follows immediately from (A). This completes the proof.

Proof of Theorems 2 *and* 3. The statement concerning the homogeneity of $\hat{h}$ follows from the fact that $\lim_{t \to 0} \hat{k}(st\xi) = \lim_{t \to 0} \hat{k}(t\xi)$ for all $s > 0$. The rest of Theorem 2 is an immediate consequence of Lemma 2(A).

Concerning Theorem 3 the fact that h is homogeneous of degree $-n$ follows from the fact that $\lim_{t \to 0} k_{st} * f = \lim_{t \to 0} k_t * f$ for all $s > 0$. The rest is an immediate consequence of Lemma 2(B).

Proof of Theorem 4. Write $k_t * f = k_t * f - g_t * f + g_t * f$ so that

$$\|k_t * f\|_p \leqq \|k - g\|_p^* \|f\|_p + \|g_t * f\|_p,$$

where g is in L_p^p and $\hat{g}$ has compact support. By hypothesis such a g can be chosen so that $\|k - g\|_p^*$ is as small as desired. If $\hat{f}$ vanishes in a neighborhood of the origin, then for t sufficiently large $g_t * f = 0$ for such g. Thus if $\hat{f}$ vanishes in a neighborhood of the origin it follows that $\|k_t * f\|_p$ can be made arbitrarily small by choosing t large enough. Since such f's are dense in L^p, $1 < p < \infty$, the desired result follows.

Proofs of Theorems 5 *and* 6. These theorems are the same as Theorems 2 and 3 respectively with t replaced by $1/t$ appropriately.

Proof of Theorem 7. Write

$$\hat{k}(t\xi)\hat{f}(\xi)=[\hat{f}(\xi)-\hat{f}(0)]\hat{k}(t\xi)+\hat{f}(0)\hat{k}(t\xi)$$

or

$$k_t * f(x)=\mu * k_t(x)+\hat{f}(0)k_t(x).$$

Since $\|\mu * k_t\|_1=\|\mu_{1/t} * k\|_1$ and μ satisfies the hypothesis on k in the corollary to Theorem 1 with $\hat{\mu}(0)=0$, it follows that $\lim_{t\to\infty}\|\mu * k_t\|_1=\lim_{t\to 0}\|\mu_t * k\|_1=0$. Now $\|k_t\|_1=\|k\|_1$ so we can conclude that $\lim_{t\to\infty}\|k_t * f\|=|\hat{f}(0)|\|k\|_1$.

To see the statement concerning convergence in measure pick any ε, $0<\varepsilon<\infty$, and write

$$\begin{aligned} m\{x\colon |k_t * f(x)|>\varepsilon\} &\leqq m\{x\colon |\mu * k_t(x)|>\varepsilon/2\}+m\{x\colon |\hat{f}(0)k_t(x)|>\varepsilon/2\}\\ &\leqq \frac{2}{\varepsilon}\|\mu * k_t\|_1+m\{x\colon |\hat{f}(0)k_t(x)|>\varepsilon/2\}. \end{aligned}$$

Since $\lim_{t\to\infty}\|\mu * k_t\|_1=0$ and $\lim_{t\to\infty}k_t(x)=0$ in measure, the desired conclusion follows.

5. Generalizations. It should not be difficult to see that appropriate analogues of the results in this paper are valid when more general notions of dilation are used.

For example, if A is a linear transformation of R^n into R^n we may consider a family of linear transformation $x\to t^A x$, $t>0$, where $t^A=\exp(A\log t)$. Note that in the case that $A=I$, the identity, then $t^A x=tx$, the case considered above. Since the determinant of t^A is t^α, where α is the trace of A, we see that the appropriate analogue of k_t, when the pointwise definition makes sense, is

$$t^{-\alpha}k(t^{-A}x). \tag{10}$$

Such nonisotropic dilates arise in many different contexts including the study of certain nonelliptic partial differential equations; for example, see [3] for an application to the study of the heat equation.

Observing that the Fourier transform of (10) is $\hat{k}(t^A\xi)$ and that $s^At^Ax=(st)^Ax$, it is clear how to generalize the notion of homogeneity to this setting. In order that t^Ax behave reasonably as t goes to 0 or ∞, i.e. consider the case $A=0$, a restriction on A is needed. Such a restriction is the condition that

$$\langle Ax,x\rangle \geqq \varepsilon\langle x,x\rangle \tag{11}$$

for some positive ε, where $\langle\cdot,\cdot\rangle$ denotes the scalar product in R^n. If (11) holds, then $t^\varepsilon|x|\leqq|t^Ax|\leqq t^\delta|x|$ for all x, where δ is the norm of A. Thus it is clear that all the results mentioned above hold in this setting.

As another example consider the multi-parameter situation where k_t is replaced by

$$(t_1\cdots t_n)^{-1}k(x_1/t_1,\cdots,x_n/t_n) \tag{12}$$

where $t_i>0$, $i=1,\cdots,n$, and $(t_1,\cdots,t_n)$ goes to 0 or ∞. Such multi-parameter dilates also arise in many different contexts; for example, see [4, Chaps. 2, 3] for an application to the study of boundary behavior of multiply harmonic functions and Hardy space theory in tubes. It is not difficult to see that all the results mentioned above hold in this setting also.

REFERENCES

[1] L. Hormander, *Estimates for translation invariant operators in L^p spaces*, Acta Math., 104 (1960), pp. 93–139.

[2] B. F. Logan, *Limits in L_p of convolution transforms with kernels $aK(at)$, $a \to 0$*, this Journal, 10 (1979), pp. 733–740.

[3] B. F. Jones, *A class of singular integrals*, Amer. J. Math., 86 (1964), pp. 441–462.

[4] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton Univ. Press, Princeton, NJ, 1971.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

013

ORDER STARS, APPROXIMATIONS AND FINITE DIFFERENCES I. THE GENERAL THEORY OF ORDER STARS*

A. ISERLES†

Abstract. Order stars are certain level sets in the complex plane, whose geometry helps to examine various approximation-theoretic features. The present paper develops the general theory which pertains to this complex-analytic concept, unifying the different forms of order stars that were so far used in numerical analysis and approximation theory. Subsequent papers (SIAM J. Math. Anal., 16 (1985), to appear) in the present sequence will use this general framework to prove various results that are of interest in numerical mathematics.

1. Introduction. The present paper is devoted to the development of a general theory for the investigation of analytic approximations to analytic functions. Given a function f which is analytic and single valued in the closed complex plane $\mathrm{cl}\,\mathbb{C} = \mathbb{C} \cup \{\infty\}$ with the possible exception of at most countable set of poles and a finite set of essential singularities and an approximation R of similar form, we show that many important features of the approximation are reflected in various properties of the level set of the function $\sigma(z) := R(z)/f(z)$. More precisely, approximation-theoretic features—loci and multiplicities of interpolation points, contractivity etc.—are connected to analytic properties, like the positions of zeros, poles and essential singularities.

It is elementary that one cannot separate local and global behaviour of an analytic function. In this sense the present work is a natural extension of both the content and the spirit of the theory of analytic functions to the domain of approximation theory.

Wanner, Hairer and Nørsett [21] were the first to introduce order stars, for the special case of $f(z) = \exp(z)$ and R being a rational function. This led to the proof of many outstanding conjectures and open problems in the analysis of numerical methods for first-order ordinary differential equations. Their results were generalized to a large extent by Nørsett and Wanner [19], Iserles and Powell [12], Iserles [7] and Hairer [5]. Extension of order stars to new applications were fast to come: Hairer [4] generalized some results to second-order ordinary differential equations, Iserles [8] used order starts to study numerical methods for first-order hyperbolic differential equations (this was extended by Iserles and Strang [13], [20], Iserles and Williamson [14] and Jeltsch and Strack [18]) and, finally, Iserles and Nørsett [11] recently applied order starts to give a new proof to the first Dahlquist barrier. This list justifies an attempt to develop a general and cohesive theory of order stars, which embraces all the present applications and offers a framework for further extension. Parts of that theory have been already presented elsewhere, sometimes under a different disguise, and we include them here for the sake of completeness.

This paper is Part I of a three part series. Parts II [9] and III [10] will use the theory of order stars to analyse general approximation-theoretic properties like contractive approximations and upper bounds on the block-size in Padé tableaux [9], and to investigate order and stability of optimal full discretizations for linear parabolic differential equations [10].

In §2 we formally define the order star and prove some of its properties that pertain to its behaviour near interpolation points, relation among the loci of zeros, poles and interpolation points and the connection between the geometry of the order

*Received the the editors April 5, 1983, and in revised form October 24, 1983.

† King's College, University of Cambridge, Cambridge CB2 1ST, England.

star and contractivity. Section 3 is devoted to the behaviour of the order star in the neighbourhood of essential singularities. Not surprisingly, this is the most theoretically demanding part of our analysis. In §4 we obtain a different kind of order star and explore its properties. Finally, in §5 we explore some changes of variable that cater for functions with branch cuts.

An important generalization of order stars is left out of the present theory: Wanner, Hairer and Nørsett [21] used order stars on a Riemann surface to investigate multistep methods for ordinary differential equations. Their work was generalized by Jeltsch and Nevanlinna [15], [16], [17]. Since all these results have to do with an approximation of a single function, exp(z), by algebraic functions, it is the feeling of the present author that the time is not yet ripe to develop a general theory of order stars on Riemann surfaces.

2. Main properties of order stars. We say that a complex function is *essentially analytic* if it is analytic and single-valued in $\mathrm{cl}\,\mathbb{C}$ with the possible exception of either a finite or countable set of poles and of a finite set of essential singularities. Of course, each of these sets may be empty.

Given a function f and an approximation R, both essentially analytic, we set

$$\sigma(z):=\frac{R(z)}{f(z)}. \tag{1}$$

Note that also σ is essentially analytic. We divide the closed complex plane into three sets:

$$\begin{aligned} A &:= \{z\in \mathrm{cl}\,\mathbb{C}:\ |\sigma(z)|>1\},\\ D &:= \{z\in \mathrm{cl}\,\mathbb{C}:\ |\sigma(z)|<1\},\\ \partial &:= \{z\in \mathrm{cl}\,\mathbb{C}:\ |\sigma(z)|=1\}. \end{aligned}$$

This decomposition of $\mathrm{cl}\,\mathbb{C}$ is called the *order star* of σ. Note that, unless $|\sigma|\equiv 1$, ∂ is a union of (at most) countable number of closed simple Jordan curves.

Several features of the approximation are reflected in the geometry of the order star. We commence with the analysis of the pattern of interpolation.

A point $z_0\in \mathrm{cl}\,\mathbb{C}$ is said to be an interpolation point of *degree* $p\geqq 1$ if f is analytic at z_0 and

$$\begin{aligned} R(z)&=f(z)+c(z-z_0)^p+O\left(|z-z_0|^{p+1}\right) && (z_0 \text{ finite}),\\ R(z)&=f(z)+cz^{-p}+O\left(\frac{1}{|z|^{p+1}}\right) && (z_0=\infty), \end{aligned}$$

where $c\neq 0$. Let $z_0\in \mathrm{cl}\,\mathbb{C}$ belong to ∂. We define

$$\begin{aligned} \mathbb{P}(z_0):=\{k:\ &\text{for every } \varepsilon>0 \text{ there exists } 0<\delta\leqq\varepsilon \text{ such that}\\ &\text{there are exactly } k \text{ arcs of } D \text{ on } \{z\in\mathbb{C}:\ |z-z_0|=\delta\}\} \end{aligned}$$

for $z_0\in\mathbb{C}$ and

$$\begin{aligned} \mathbb{P}(\infty):=\{k:\ &\text{for every } r>0 \text{ there exists } r_1\geqq r \text{ such that}\\ &\text{there are exactly } k \text{ arcs of } D \text{ on } \{z\in\mathbb{C}:\ |z|=r_1\}\} \end{aligned}$$

if $z_0 = \infty$. Then the number

$$\operatorname{ind}(z_0) := \min\{k\colon k \in \mathbb{P}(z_0)\}$$

is called the *index of* z_0. Intuitively speaking, $\operatorname{ind}(z_0)$ equals the number of sectors of D (or, for that matter, sectors of A) that approach $z_0 \in \partial$. The definition caters also for the case when the number of such vectors is not well defined—for example when z_0 is an essential singularity that is an accumulation point of poles of σ (cf. Fig. 3 below). By $\operatorname{ind}(z_0) = 0$ we mean that it is impossible to approach z_0 by curves in both A and D. If the set $\mathbb{P}(z)$ includes just one element and if all the sectors of A and of D approach z with an equal asymptotic angle, we say that z is *regular*.

PROPOSITION 1 (interpolation property). *Let*

$$f(z) = d(z - z_0)^k + O\left(|z - z_0|^{k+1}\right) \qquad (z_0 \in \mathbb{C})$$

or

$$f(z) = dz^{-k} + O\left(|z|^{-k-1}\right) \qquad (z_0 = \infty),$$

where k is an integer and $d \neq 0$. If z_0 is an interpolation point of degree $p \geqq \max(1, k+1)$, then $z_0 \in \partial$, $\operatorname{ind}(z_0) = p - k$ and z_0 is regular.

Proof. Without loss of generality we may assume that $z_0 \in \mathbb{C}$ — if $z_0 = \infty$, we conformally map $z \to 1/z$. It follows from (1) that

$$\sigma(z) = 1 + \frac{c}{d}(z - z_0)^{p-k} + O\left(|z_0|^{p-k+1}\right).$$

Therefore indeed $z_0 \in \partial$. Furthermore, σ is analytic in a neighbourhood of z_0.

Letting $z = z_0 + re^{i\theta}$, $r > 0$, we obtain

$$|\sigma(z)| = 1 + r^{p_1}\operatorname{Re}\{c_1 e^{ip_1\theta}\} + O(r^{p_1+1}) := \phi(r, \theta), \tag{2}$$

where $p_1 = p - k \geqq 1$, $c_1 = c/d \neq 0$. Thus, for any given $\varepsilon > 0$ and sufficiently small $r > 0$, $z \in A$ if $\operatorname{Re}\{c_1 e^{ip_1\theta}\} > \varepsilon$ and $z \in D$ if $\operatorname{Re}\{c_1 e^{ip_1\theta}\} < -\varepsilon$. $\operatorname{Re}\{c_1^{ip_1\theta}\}$ changes sign $2p_1$ times for equally spaced values of $0 \leqq \theta < 2\pi$. Hence the proposition is true, except for the possibility that A and D contain some additional sectors which are so thin that they fit between the sectors that have been shown to exist.

Such sectors may occur only if, as $r \to 0$, some values of θ that satisfy the equation $\phi(r, \theta) = 1$ tend to coalesce. In this case a zero of $d\phi(r, \theta)/d\theta$ becomes arbitrarily close to a root of $\phi(r, \theta)$. This is impossible, since the analyticity of σ in the neighbourhood of z_0 and (2) imply

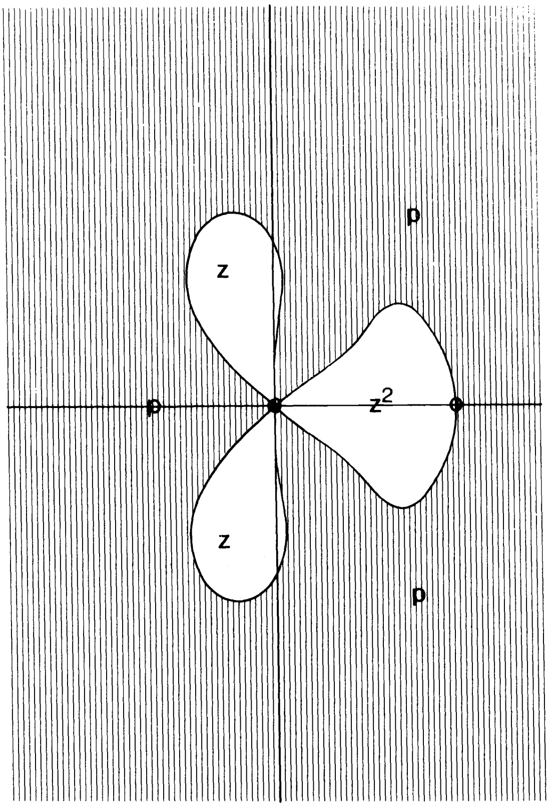
$$\frac{d}{d\theta}\phi(r, \theta) = -p_1 r^{p_1}\operatorname{Im}\{c_1 e^{ip_1\theta}\} + O(r^{p_1+1})$$

and $\operatorname{Re}\{c_1 e^{ip_1\theta}\}$, $\operatorname{Im}\{c_1 e^{ip_1\theta}\} = O(r)$ cannot both hold as $r \to 0$. Therefore no extra sectors exist, $\operatorname{ind}(z_0) = p_1 = p - k$ and z_0 is regular. □

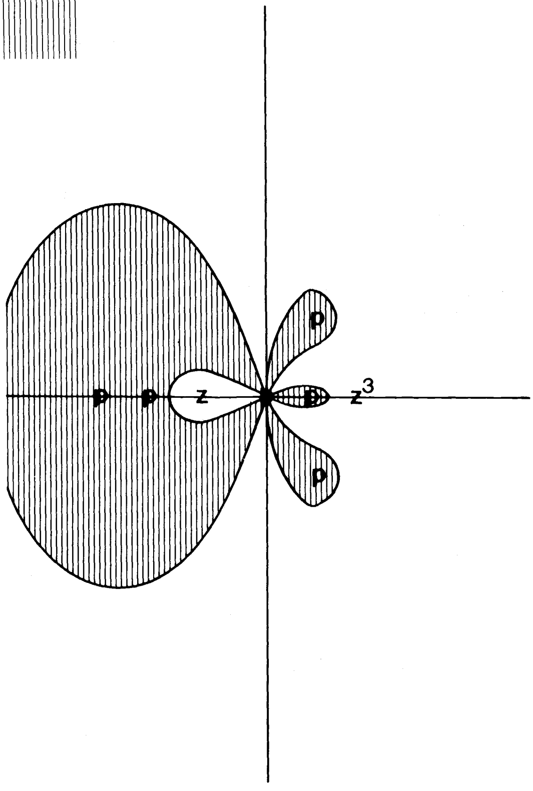
The last proof is a straightforward generalization of [21, Prop. 3] and [12]. Figure 1 gives examples of Padé approximations to the rational function

$$f(z) = \frac{1 - z + z^3}{(1 - z)^2}. \tag{3}$$

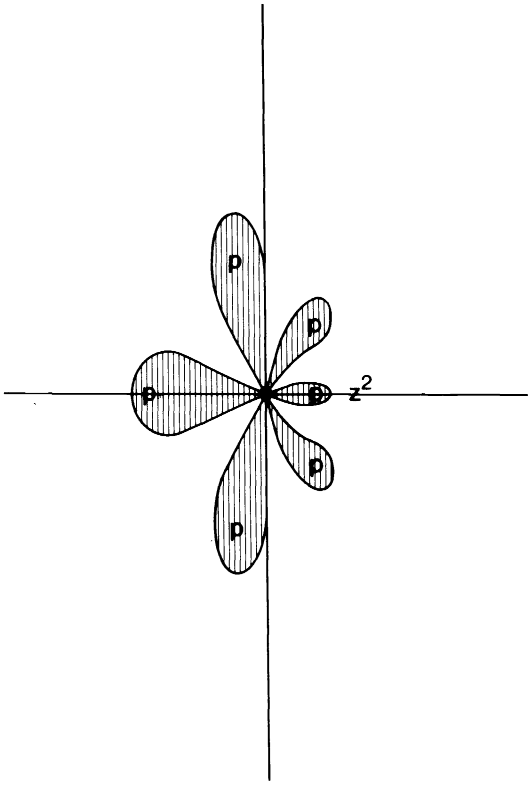
Note the nontrivial block-structure of the Padé tableau (cf. [3]).



(a) $R_{2/0}(z)=1+z+z^2$, *order* 2.



(b) $R_{2/2}(z)=(1-z^2)/(1-z-z^2)$, *order* 4.



(c) $R_{0/3}(z)=1/(1-z-z^3)$, *order* 5.

FIG. 1. *Order stars of Padé approximations to* (3), *for* $|\mathrm{Re}\, z|, |\mathrm{Im}\, z| \leqq 5$. *The dark-shaded regions denote* A_0. *"P", "Z" and "O" denote poles, zeros and interpolation points respectively. Note that* ind(0) = *order* + 1. *Both here and later the figures are elongated along the imaginary axis—this is done to emphasize various thin regions, which would have disappeared altogether otherwise.*

It is a consequence of the last lemma that the geometry of the order star reflects the pattern of interpolation (this was, indeed, the reason for the name "order star"). There is, however, yet another expression of interpolation in the order star, via its connection with the loci of zeros and poles of σ.

It is obvious that poles of f and zeros of R belong to D, whereas zeros of f and poles of R belong to A—all this unless a pole and a zero of σ coalesce, leading to a removable singularity. In what follows we establish a relationship between the number of poles of σ, say, in portions of A and the interpolation pattern.

We call the connected components of A and D *A-regions* and *D-regions* respectively. Such a region is *analytic* if σ is analytic along its boundary (in other words, no essential singularities occur along the boundary). We say that an A-region or a D-region is of *multiplicity L* if its directed boundary passes through exactly L points z where $\sigma(z)=1$, which need not be distinct. Note that each such z is an interpolation point.

PROPOSITION 2 (multiplicity property). *Let* $|\sigma| \not\equiv 1$. *The multiplicity L of an analytic A-region* (*D-region*) *equals the number of poles* (*zeros*) *of σ, counted with their multiplicity, inside the region. Furthermore, it is always true that* $1 \leqq L < \infty$.

Proof. We prove the proposition for A-regions, since the proof for D-regions is similar.

Let f be an analytic A-region of multiplicity L. We parametrize ∂F with positive orientation as $\gamma(t)=\gamma_1(t)+i\gamma_2(t)$, $0\leqq t\leqq 1$, where both γ_1 and γ_2 are real. The analyticity of σ implies that γ is differential almost everywhere in $[0,1]$.

We denote by $v:=(\gamma_1(t),\gamma_2(t))$ and $n:=(\gamma_2'(t),-\gamma_1'(t))$ the tangent and the outward-pointing normal at $\gamma(t)$, respectively. Since, by the definition of A, $|\sigma(z)|$ decreases locally along n in the vicinity of the boundary, also $\ln|\sigma(z)|$ decreases there. Hence, given the polar representation

$$(4)\qquad \sigma(z)=r(x,y)e^{i\phi(x,y)},\qquad z=x+iy,$$

it is true for every $x+iy\in\partial F$ that

$$(5)\qquad \frac{\partial}{\partial n}\ln r(x,y)<0.$$

The function σ is analytic along γ. Therefore it satisfies there the Cauchy–Riemann equations. In polar coordinates we have

$$\frac{\partial}{\partial x}\ln r=\frac{\partial}{\partial y}\phi,\qquad \frac{\partial}{\partial y}\ln r=-\frac{\partial}{\partial x}\phi.$$

Therefore

$$\frac{\partial}{\partial n}\ln r=\gamma_2'\frac{\partial}{\partial x}\ln r-\gamma_1'\frac{\partial}{\partial y}\ln r=\gamma_2'\frac{\partial}{\partial y}\phi-\gamma_1'\left(-\frac{\partial}{\partial x}\phi\right)$$
$$=\gamma_1'\frac{\partial}{\partial x}\phi+\gamma_2'\frac{\partial}{\partial y}\phi=\frac{\partial}{\partial v}\phi.$$

This, together with (4) and (5) implies that $\arg\sigma$ decreases strictly monotonically along γ. Since σ is analytic and γ is a union of closed curves in $\mathrm{cl}\,\mathbb{C}$, the variation of $\arg\sigma$ along γ is a negative integer multiple of 2π, $-2\pi K$ say. Since $\sigma(\tilde z)=1$, $\tilde z\in\gamma$, means that $\arg\sigma(\tilde z)$ is an integer multiple of 2π, necessarily $K=L$, the multiplicity of the region. $L>1$ follows at once by $\partial\phi/\partial v<0$, whereas unless $L<\infty$ we would have an accumulation point of zeros of $\sigma-1$. This is impossible, since σ is not identically 1.

Finally, we use the argument principle,

$$Z_F-P_F=\frac{1}{2\pi i}\int_\gamma\frac{d}{dz}\ln\sigma(z)\,dz$$
$$=\text{the number of rotations of }\arg\sigma\text{ along }\gamma=-L_1$$

where Z_F and P_F denote the number of zeros and of poles of σ in F respectively. By the definition of an A-region $Z_F=0$, and so $P_F=L$ and the proposition is true. □

The last proposition is a straightforward extension of [12, Prop. 5] and [21, Prop. 4] and its proof follows that in [21].

Another property of the approximation which is of interest in many applications is the *contractivity*. Let V be an open sub-set of $\mathrm{cl}\,\mathbb{C}$ such that f is analytic in V and $|f|\equiv 1$ along ∂V, with the possible exception of a finite number of points, $\tilde z_1,\cdots,\tilde z_m$, say. These points must necessarily be essential singularities of f. We further assume that there exists $\varepsilon_1>0$ such that

$$\sup\left\{|f(z)|:z\in V\cap\left\{z\in\mathbb{C}:|z-\tilde z_k|=\varepsilon\right\}\right\}=1$$

for every $1 \leqq k \leqq m$ and $0 < \varepsilon < \varepsilon_1$. Then it follows at once from the maximal modulus principle that $|f| \leqq 1$ in $\operatorname{cl} V$. Given an approximation R, it is sometimes important whether it satifies a similar boundedness condition. We say that R is a *V-contraction of f* if $|R| \leqq 1$ in $\operatorname{cl} V$. The V-contractivity can be expressed easily by the geometry of the order star. It is of great interest in many applications to the numerical analysis of differential equations, since it is equivalent to stability of some numerical methods.

PROPOSITION 3 (stability property). *R is a V-contraction of f if and only if it is analytic in V and $A \cap \partial V = \varnothing$.*

Proof. Follows at once by the definition of the order star, $|f| \equiv 1$ along ∂V (with the possible exception of a finite set) and the maximal modulus principle. □

It is a consequence of that last proposition that ∂V separates between A-regions in the event of a V-contractive approximation. This is crucial in many proofs that are based on the order star theory—cf. [5], [7], [8], [9], [10], [12], [19], [21]. For example, the familiar case of A-acceptable rational approximations to the exponential leads to $V = \mathbb{C}^- := \{z \in \mathbb{C}: \operatorname{Re} z < 0\}$, $m = 1$, $\tilde{z}_1 = \infty$.

3. Behaviour near essential singularities. Essential singularities of both f and R, unless they cancel each other, are also essential singularities of σ. It is obvious by virtue of the Weierstrass theorem that every essential singularity of σ belongs to ∂. However, different types of singularities give rise to different patterns of behaviour of the order star in their neighbourhoods. We give a partial classification of this behaviour. It turns out, as in §2, that the geometry of the order star is in strong relationship to the global behaviour of the function σ.

Let $\tilde{z}_1, \cdots, \tilde{z}_m$ be all the essential singularities of σ. It follows from the standard theory of analytic functions that σ can be represented as

$$\sigma(z) = \sigma_1(z)\sigma_2(z) \cdots \sigma_m(z),$$

where each $\sigma_j(z)$ has a single essential singularity at $\tilde{z}_j$. Since the conformal mapping $z \to 1/(\tilde{z}_j - z)$ takes $\tilde{z}_j$ to infinity, we may assume, without loss of generality, that the underlying function σ is *entire* and has an essential singularity at infinity. This can be done because only σ_j determines the behaviour of the order star in the neighbourhood of $\tilde{z}_j$.

We recall the concept of a *perfect order of growth* $\rho(\sigma)$ of an entire function σ: given

$$M(r) := \max_{|z|=r} \{|\sigma(z)|\}, \qquad r > 0,$$

set

$$\rho(\sigma) := \limsup_{r \to \infty} \frac{\ln \ln M(r)}{\ln r}.$$

PROPOSITION 4. *If $\rho(\sigma) < \infty$, then* $\operatorname{ind}(\infty) \leqq 2\rho(\sigma)$.

Proof. We recall from the theory of analytic functions the theorem of Ahlfors [2, p. 352]: an entire, nonconstant function of order ρ has at most 2ρ finite asymptotic values. Therefore at most $2\rho(\sigma)$ sectors of D may approach infinity and the proposition follows. □

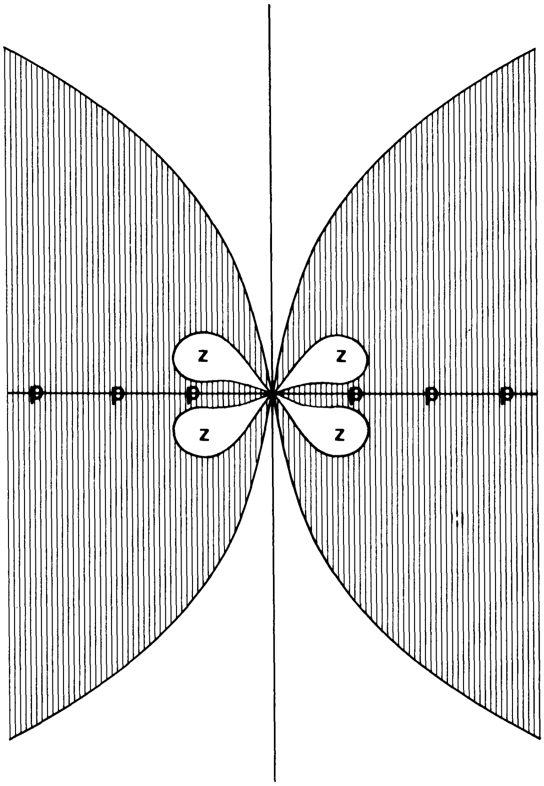
The upper bound of Proposition 4 is often too generous and, indeed, in the remainder of this section we give further results which lower it for certain types of functions. However, it is attainable for every choice of $\rho(\sigma)$: if $1 \leqq K = \rho(\sigma) < \infty$, then it is attained by $\sigma(z)=(\sin z^K)/z^K$. It is easy to verify that $2K$ sectors of A approach infinity with an asymptotic angle of π/K, bisected by $2K$ cusps of D along the asymptotics $\{r\exp(i\pi k/K)\colon r\gg 0\}$, $0\leqq k\leqq 2K-1$. Figure 2 displays order stars that correspond to Padé approximations to $(\sin z)/z$ and demonstrate this type of behaviour.

A more complicated example is required for $K=0$. We set

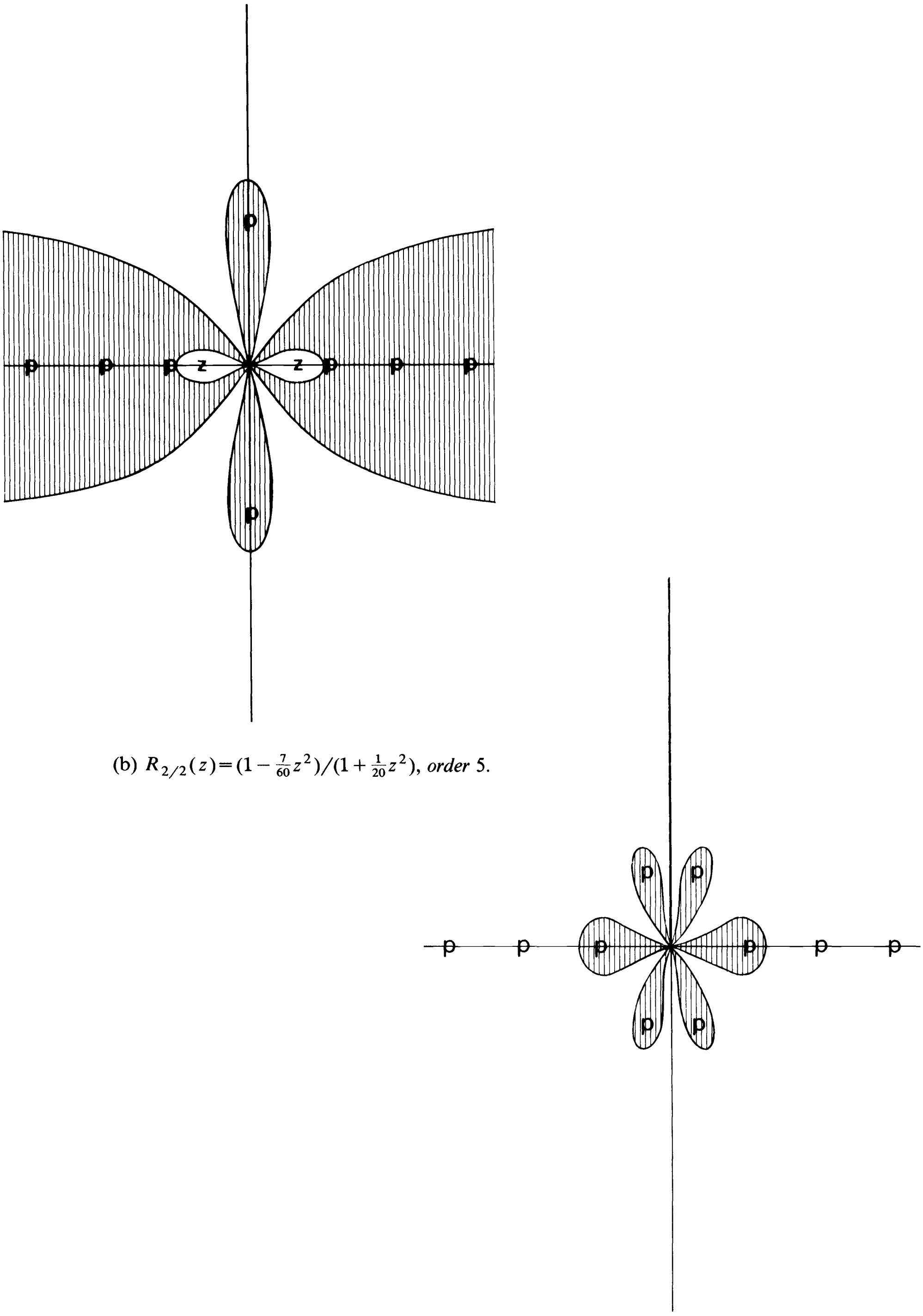
$$f(z)=\prod_{n=1}^{\infty}\left(1-\frac{z}{2^n}\right),\qquad R(z)=1-z+\frac{1}{3}z^2-\frac{1}{21}z^3. \tag{6}$$

Then, by [6, p. 183] $\sigma(z)=R(z)/f(z)$ is of perfect order of growth 0. It has poles at 2^n, $n\geqq 1$, which belong, for increasing n, to progressively smaller A-regions. We have $\mathbb{P}(\infty)=\{0,1\}$ and $\operatorname{ind}(\infty)=0$.

A natural question arises regarding the geometry of the order star when $\rho(\sigma)=\infty$. In this case it may well happen that an infinite number of sectors of A and D tend to infinity and $\operatorname{ind}(\infty)=\infty$. A simple example is $\sigma(z)=\exp(e^z)$. In that case the order star consists of alternating parallel strips of A and D.



(a) $R_{4/0}(z)=1-\frac{1}{6}z^2\frac{1}{120}z^4$, *order* 5.



(b) $R_{2/2}(z)=(1-\frac{7}{60}z^2)/(1+\frac{1}{20}z^2)$, *order* 5.

(c) $R_{0/4}(z)=1/(1+\frac{1}{6}z^2+\frac{7}{310}z^4)$, *order* 5.

FIG. 2. *Order stars of Padé approximations to* $(\sin z)/z$, *with* $|\mathrm{Re}\, z|, |\mathrm{Im}\, z| \leqq 10$. "*Z*" *and* "*P*" *denote zeros and poles of* σ, *respectively. Note that two cusps of A tend to infinity through* $\mathbb{R}$ *and that the A-regions that surround some of the poles in* (c) *are too small to be discerned.*

Let us now assume that $d,\alpha,\beta\in\mathbf{C}$, d, $\beta\neq 0$ and a natural number M exist so that the entire function σ has the asymptotic behaviour

$$\sigma(z)=dz^{\alpha}e^{\beta z^M}(1+o(1)) \qquad (z\to\infty) \tag{7}$$

We say that σ is of *exponential type M.*

PROPOSITION 5. *If σ is of exponential type $M\geqq 1$,then* $\operatorname{ind}(\infty)=M$ *and ∞ is a regular point of* ∂.

Proof. Let $z=re^{i\theta}$, $r>0$. Equation (7) gives

$$\begin{aligned}\ln|\sigma(z)|&=\ln|d|-\theta\operatorname{Im}\alpha+(\operatorname{Re}\alpha)\ln r\\&\quad+r^M\operatorname{Re}\{\beta e^{iM\theta}\}+o(1):=\psi(r,\theta)\qquad(r\to\infty).\end{aligned}$$

Therefore, for sufficiently large r and for every $\varepsilon>0$ $z\in A$ if $\operatorname{Re}\{\beta e^{iM\theta}\}>\varepsilon$ and $z\in D$ if $\operatorname{Re}\{\beta e^{iM\theta}\}<-\varepsilon$. Hence M sectors of A and M sectors of D approach infinity, each with an asymptotic arc-length π/M.

As in the proof of Proposition 1, there is still a possibility of additional sectors approaching infinity as cusps. It is ruled out, as before, by showing that the zeros of $\psi(r,\theta)$ and $\partial\psi(r,\theta)/\partial\theta$ are separated for $r\gg 0$. □

An important instance of a function of exponential type occurs when $1\leqq\rho(\sigma)<\infty$ and σ possesses only a finite number of zeros: by the Hadamard factorization theorem [6, p. 199] every entire function σ of bounded perfect order of growth has a representation of the form

$$\sigma(z)=e^{g(z)}z^K\prod_{n=1}^{\infty}E\left(\frac{z}{z_n},p\right),$$

where z_1, $z_2,\cdots$ are all the zeros of σ away from the origin, g is a polynomial, $\deg g\leqq\rho(\sigma)$, p and K are nonnegative integers, $p\leqq\rho(\sigma)$ and

$$E(z,0):=1-z,\qquad E(z,p):=(1-z)\exp\left(\sum_{k=1}^{p}\frac{1}{k}z^k\right),\quad p\geqq 1.$$

Given that σ has $K+L$ zeros, we have

$$\sigma(z)=e^{g(z)}z^K\prod_{n=1}^{L}E\left(\frac{z}{z_n},p\right)=e^{\tilde g(z)}z^K\prod_{n=1}^{L}\left(1-\frac{z}{z_n}\right),$$

where

$$\tilde g(z)=\begin{cases}g(z)+\sum\limits_{k=1}^{p}\frac{1}{k}\left(\sum\limits_{n=1}^{L}\frac{1}{z_n^k}\right)z^k, & p\geqq 1,\\ g(z), & p=0.\end{cases}$$

Therefore (7) is satisfied with $M=\deg\tilde g=\rho(f)$, σ if of exponential type M and we can use Proposition 5 to determine the geometry of the order star from $|z|\gg 0$.

A refinement of Proposition 4 can be also obtained for certain entire functions which have an infinite number of real zeros. Let σ be entire and have only real zeros $p_1\geqq p_2\geqq p_3\geqq\cdots$. Further, we stipulate that $C>0$ and $0<\alpha<1$ exist such that the number of zeros in every interval of the form $[-r,\infty)$, is $Cr^{\alpha}(1+o(1))$ for every $r \gg 0$. It follows from the Hadamard factorization that

$$\sigma(z)=e^{g(z)}\prod_{n=1}^{\infty}\left(1+\frac{z}{p_n}\right), \tag{8}$$

where g is a polynomial, $\deg g = M$, say.

PROPOSITION 6. *Given an entire function* (8), *it is true that* $\operatorname{ind}(\infty) \leqq M+1$. *Moreover, if* $0<\alpha<\frac{1}{2}$, *then* $\operatorname{ind}(\infty)=M$ *and* ∞ *is a regular point of* ∂.

Proof. By the Polya–Szegö theorem [6, p. 206]

$$\ln \prod_{n=1}^{\infty}\left(1+\frac{z}{p_n}\right)=\frac{C\pi}{\sin \alpha\pi} z^{\alpha}(1+o(1))$$

uniformly for every $|\arg z| \leqq \pi-\varepsilon$, $\varepsilon>0$. Hence, given $z=re^{i\theta}$, $r>0$, $0 \leqq |\theta|<\pi$, (8) gives

$$\ln|\sigma(z)|=\operatorname{Re} g(z)+\frac{C\pi}{\sin\alpha\pi} r^{\alpha}(1+o(1)).$$

Therefore $\operatorname{ind}(\infty) \leqq M+1$, by an argument similar to the proof of Proposition 4. Furthermore, either $\operatorname{ind}(\infty)=M$ and ∞ is regular or $\operatorname{ind}(\infty)=M+1$ and there is a cusp along the negative half-axis.

Let $0<\alpha<\frac{1}{2}$. Then [6, p. 207] there exists a sequence of values $\{x_m\}_{m=1}^{\infty}$, tending to ∞ through the negative half-axis, such that

$$\lim_{m\to\infty}\prod_{n=1}^{\infty}\left|1+\frac{x_m}{p_n}\right|=\infty.$$

Therefore $\mathbb{P}(\infty)=\{M, M+1\}$, implying that $\operatorname{ind}(\infty)=M$ and ∞ is a regular point. □

An example of a function that satisfies the conditions of the last proposition is

$$\sigma(z)=\frac{1}{\Gamma(z^K)}$$

where $K \geqq 2$ is a natural number. In fact, the order star of $\sigma(z)=1/\Gamma(z)$ also has similar geometry. Although this can be readily ascertained from its plot (cf. Fig. 3), it is instructive to work out the structure of the order star analytically.

We have $\sigma(z)=e^{-z}(1+o(1))$ uniformly for every $z\in\mathbb{C}$ such that $|\arg z| \leqq \pi-\varepsilon$, $\varepsilon>0$. Therefore $z\in A$ for $z\in\mathbb{C}^-$, $|z| \gg 0$, away from the real axis, and $z \in D$ for $z\in\mathbb{C}^+ := \{z\in\mathbb{C}\colon \operatorname{Re} z>0\}$, $|z|\gg 0$. It follows that $1\in\mathbb{P}(\infty)$ and $\operatorname{ind}(\infty)=1$.

The recurrence relation $\Gamma(z+1)=z\Gamma(z)$, together with $\Gamma(\frac{1}{2})=\pi^{1/2}$, gives

$$|\sigma(-m-1/2)|=\frac{(2m+1)!}{z^{2m+1}m!\pi^{1/2}}$$

for every integer $m \geqq 0$. Hence

$$\frac{|\sigma(-(m+1)-1/2)|}{|\sigma(-m-1/2)|}=\frac{2m+3}{2m}>1, \qquad \left|\sigma\left(-\frac{5}{2}\right)\right|=\frac{15}{8\pi^{1/2}}>1$$

imply that $-m-\frac{1}{2}\in A$ for every $m \geqq 2$. Since σ has no poles, it follows by Proposition 2 that there exists just one A-region, that it must be unbounded and that for $m \geqq 3$ the zeros of $-m$ of σ belong to analytic D-regions of multiplicity 1. On the other hand

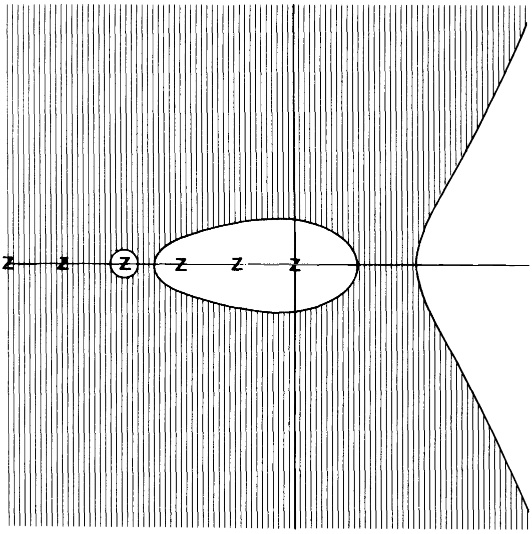


FIG. 3. *The order star of* $\sigma(z)=1/\Gamma(z)$, *with* $-5\leqq\operatorname{Re}z\leqq 4$, $|\operatorname{Im}z|\leqq 4$.

$|\sigma(x)|<1$ for every $-2\leqq x<1$, $|\sigma(x)|>1$ for every $1<x<2$ [1, p. 255], and so the zeros $0, -1, -2$ belong to a single analytic D-region of multiplicity 3. Finally, the Weierstrass formula for $\Gamma(z)$ gives

$$|\sigma(it)|^2=t^2\prod_{n=1}^{\infty}\left(1+\frac{t^2}{n^2}\right)=\frac{t\sin(\pi t)}{\pi},\qquad t\in\mathbb{R}.$$

Hence $|\sigma(it)|<1$ for small $t>0$, $|\sigma(it)|>1$ for $t\gg 0$ and $|\sigma(it)|=1$ has exactly one solution for $t>0$ — the intersection of the boundary of the D-region of mulitplicity three with $\mathbb{R}$. All this analysis determines in a unique way the shape and the geometry of the order star in Fig. 3. It also implies, in conjunction with Proposition 3, that the equation $\Gamma(z)=1$ has at most two complex conjugate solutions in $\mathbb{C}$ away from the negative half-axis.

In order to complete this section we reiterate that the results of Propositions 4–6 can be translated in a natural way from infinity to any $z\in\mathbb{C}$ and can also cater for functions σ that possess several essential singularities.

4. Order stars of the second kind. Let R be an approximation to f, both functions bring the same type as in §2. We set

$$\tilde{f}(z):=e^{f(z)},\qquad \tilde{R}(z):=e^{R(z)}$$

and consider the order star of $\tilde{\sigma}=\tilde{R}/\tilde{f}$. Since

$$|\tilde{\sigma}(z)|\gtrless 1\Leftrightarrow \operatorname{Re}\sigma(z)\gtrless 0,\quad z\in\operatorname{cl}\mathbb{C}, \tag{9}$$

where $\sigma=R-f$, this leads to the *order star of the second kind* with

$$\begin{aligned}\tilde{A}&:=\{z\in\operatorname{cl}\mathbb{C}:\ \operatorname{Re}\sigma(z)>0\},\\ \tilde{D}&:=\{z\in\operatorname{cl}\mathbb{C}:\ \operatorname{Re}\sigma(z)<0\},\\ \partial&:=\{z\in\operatorname{cl}\mathbb{C}:\ \operatorname{Re}\sigma(z)=0\}.\end{aligned}\tag{10}$$

This order star, which was introduced in [8] and used in [11], [13], and [14], has several similar properties to the order star from §2. However, there are major differences that make it a more appropriate tool for the stability analysis of certain finite differences.

The following results are a trivial extension of Propositions 1–3.

PROPOSITION 7. *If $z_0 \in \mathrm{cl}\,\mathbf{C}$ is an interpolation point of degree p,*

$$R(z)=f(z)+c(z-z_0)^p+O\left(|z-z_0|^{p+1}\right), \qquad c\neq 0,$$

then $z_0 \in \partial$, $\mathrm{ind}(z_0)=p$ and z_0 is regular.

Note that the behaviour of order stars of the second kind near an interpolation point is not sensitive to this point being a zero or a pole.

Both interpolation points and singularities (poles and essential singularities alike) belong to ∂. The following relationship holds:

PROPOSITION 8. *Between any two singularities that are connected by an arc of ∂ there is an interpolation point. Between any two interpolation points that are connected by an arc of ∂ there is a singularity.*

Proof. It follows from the method of proof of Proposition 2 that $\mathrm{Im}\,\sigma$ is a strictly monotone function along the oriented boundary of ∂ (decreasing along the positively oriented boundary of $\tilde{A}$ and increasing along the positively oriented boundary of $\tilde{D}$). The present proposition follows, since $\mathrm{Im}\,\sigma$ vanishes at an interpolation point and $\mathrm{Im}\,\sigma$ is unbounded at a singularity. □

The last proposition highlights yet another difference between the two kinds of order stars: while the relationship between singularities and interpolation points is somewhat stronger than in Proposition 2, the zeros of f and R have no similar role. In some cases it is possible to show that various conditions impose restrictions on the pattern of zeros [14]. Elsewhere [11] the role of the zeros and the poles is reversed by taking $\sigma=1/R-1/f$.

It is seen at once, by virtue of (9), that a Kth order pole of σ leads to a local pattern of an order star of the second kind that is equivalent to an essential singularity of exponential type K in the standard order star.

It now follows from Proposition 5 that

PROPOSITION 9. *If z_0 is a pole of σ of order $K \geqq 1$,*

$$\sigma(z)=\alpha(z-z_0)^{-K}(1+o(z-z_0)), \qquad \alpha\neq 0,$$

then $z_0 \in \partial$, $\mathrm{ind}(z_0)=K$ and z_0 is regular.

By the same token an essential singularity of a bounded perfect order of growth in σ is equivalent to an essential singularity of an unbounded perfect order of growth in $\tilde{\sigma}$. This leads to infinite indices. Fortunately, if σ is entire, then it is often possible to restrict its investigation to a strip of the form $0 \leqq |\mathrm{Im}\,z| \leqq a$, where only a finite number of "strips" of A and D tend to infinity—cf. [10], [11], [13], [14] and §5 of the present paper. In that case $\mathrm{ind}(\infty)=K$, say, and ∞ being regular imply that the asymptotic width of each "strip" of A and D is π/K.

The geometry of an order star of the second kind is not very helpful as far as contractivity is concerned. However, it displays a conceptually similar feature, which is very useful in the analysis of semi-discretizations of partial differential equations of evolution. Let Ω be a set of complex points such that $\mathrm{Re}\,f(z)=0$ for every $z\in\Omega$ (obviously, unless f is a real constant, Ω does not contain an open neighbourhood). Given such a function f we say that R has *property R* if $\mathrm{Re}\,R(z)\leqq 0$ in Ω.

PROPOSITION 10. *R has property R if and only if $\Omega \cap \tilde{A} = \varnothing$.*

Proof. Follows at once from the definition (10). □

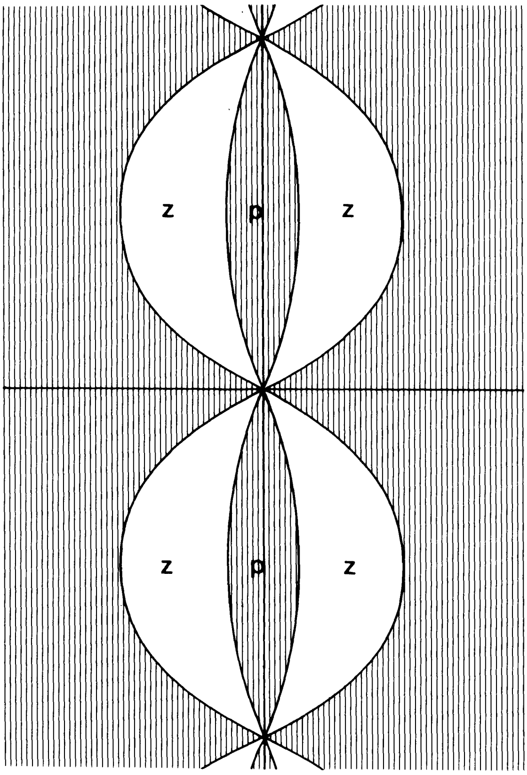
5. Multivalued functions. The most natural extension of the present theory to multivalued functions is by considering order stars on Riemann surfaces. This approach obscures, however, the geometric intuition that is necessary to transform a complex-analytic problem to a combinatorial one. Nonetheless, order stars on Riemann surfaces led to important new results [14], [15], [16], [17] and [21]. However, the subject of the present paper being an exposition of a theory of order stars in the complex plane, we restrict our attention to an approach which avoids Riemann surfaces altogether.

Let f and R be of the form

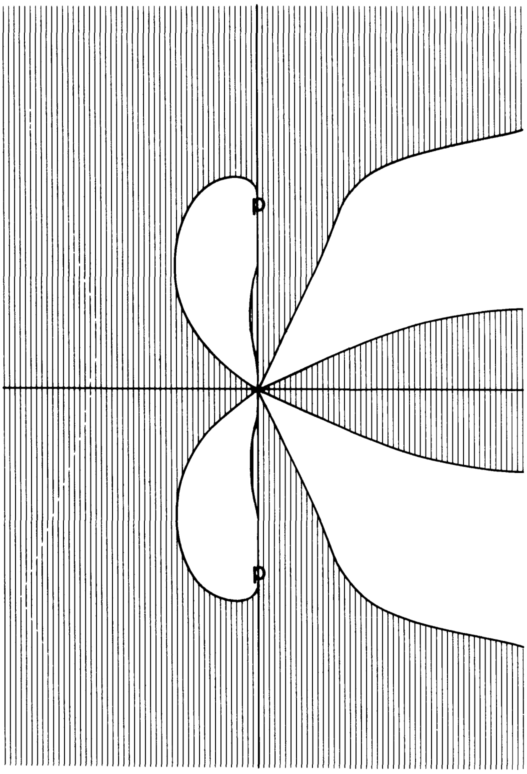
$$f(z)=f^*(z^{\mu_1},z^{\mu_2},\cdots,z^{\mu_M}),\qquad R(z)=R^*(z^{\eta_1},z^{\eta_2},\cdots,z^{\eta_N}), \tag{11}$$

where f^* and R^* are rational functions of their arguments and $\mu_1,\cdots,\mu_M,\eta_1,\cdots,\eta_N\in \mathbf{C}/\{0\}$. We set

$$\tilde{f}(z):=f(e^z),\qquad \tilde{R}(z):=R(e^z). \tag{12}$$



(a) $R_{2/1}(z)=(1+6z+z^2)/(4(1+z))$, *order* 3.



(b) $R_{2/1}(z)=(1+6z+z^2)/(4(z+1))$, *order* 3.

FIG. 4. *Order stars for Padé approximations to* $z^{1/2}$ *about* $z=1$. *The standard order stars are given first, to the scale* $|\mathrm{Re}\, z|$, $|\mathrm{Im}\, z|\leqq 5$, *order stars of the second kind being displayed subsequently to the scale* $|\mathrm{Re}\, z|\leqq 6$, $|\mathrm{Im}\, z|\leqq 7$. *"Z" and "P" denote zeros and poles, respectively.*

Both $\tilde{f}$ and $\tilde{R}$ are single-valued functions and so we can consider order stars of both kinds, with respect to $\sigma(z)=\tilde{R}(z)/\tilde{f}(z)$ and $\sigma(z)=\tilde{R}(z)-\tilde{f}(z)$ respectively. The replacement of z by a periodic function leads to an inflation in the number of zeros, poles and essential singularities, as well as to a change in the nature of essential singularities. In many cases this added complexity can be easily handled. For example, if $\mu_1,\cdots,\mu_M$, $\eta_1,\cdots,\eta_M$ are real rational numbers,

$$\mu_k=\frac{p_k}{q_k},\qquad 1\leqq k\leqq M,$$

$$\eta_k=\frac{s_k}{t_k},\qquad 1\leqq k\leqq N,$$

say, then the order star is periodic with period $2\pi iK$, where K is the lowest common multiple of $|p_1|,\cdots,|p_M|$, $|q_1|,\cdots,|q_M|$, $|s_1|,\cdots,|s_N|$, $|t_1|,\cdots,|t_N|$.

Figure 4 displays order stars of both kinds for Padé approximations to $f(z)=z^{1/2}$ which were obtained by the transformation (12). Only the strip $0 \leqq |\mathrm{Im}\, z| \leqq 2\pi$ is given, since the order stars have period $4\pi i$.

A similar change of variable is suitable for

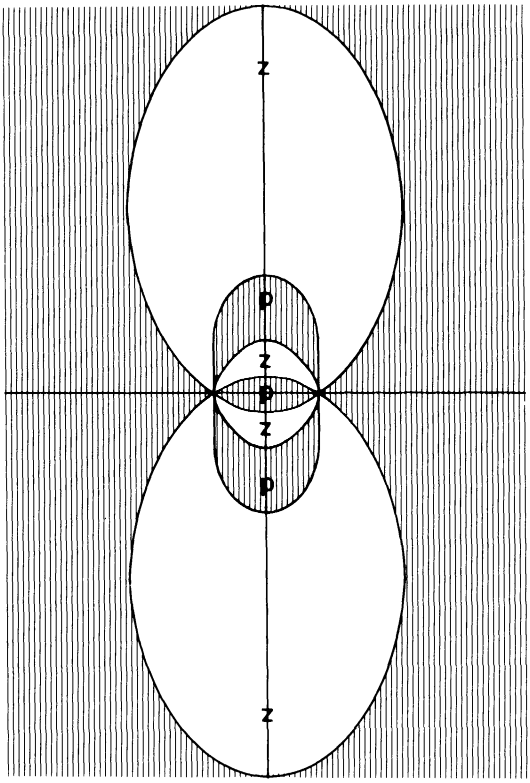
$$(13) \qquad f(z)=f^*(z,\ln z), \qquad R(z)=R^*(z,\ln z).$$

Functions and approximations of the forms (11) and (13) appear in the analysis of fully-discretized and semi-discretized finite difference (cf. [8], [10], [11], [13], [14], [18] and [20]).

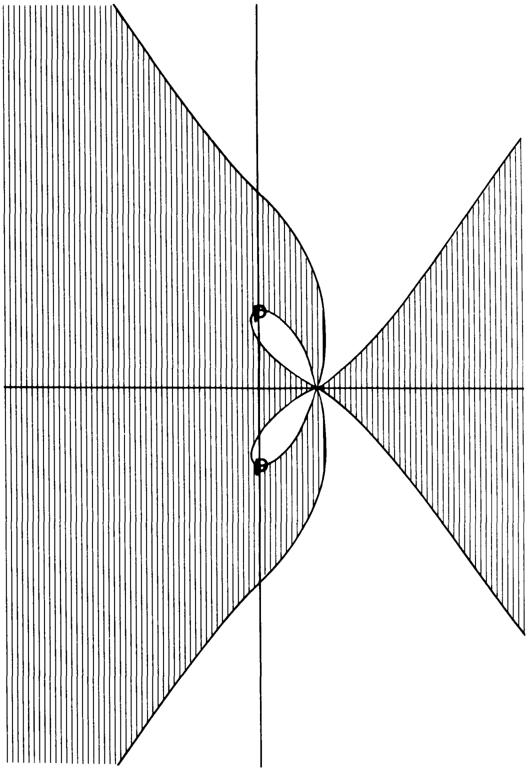
Another change of variable can successfully cope with functions and approximations of the form (11) if $\mu_1,\cdots,\mu_M, \eta_1,\cdots,\eta_N$ are rational numbers. We set

$$(14) \qquad \tilde{f}(z):=f(z^K), \qquad \tilde{R}(z):=R(z^K),$$

where K was given before, and consider order stars with respect to $\sigma(z)=\tilde{R}(z)/\tilde{f}(z)$ or $\sigma(z)=\tilde{R}(z)-\tilde{f}(z)$. Figure 5 gives the counterparts of order stars from Fig. 4 which are obtained by the present transformation.



(a) $R_{2/1}(z)=(1+6z+z^2)/(4(1+z))$, *order* 3.



(b) $R_{2/1}(z)=(1+6z+z^2)/(4(1+z))$, *order* 3.

FIG. 5. *Order stars of the Padé approximations to* $z^{1/2}$ *from Fig.* 4 *by using the transformations* (14). *See Fig.* 4 *for notation and scale.*

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1970.

[2] G. M. GOLUZIN, *Geometric Theory of Functions of Complex Variable*, AMS Trans. Math. Monographs, 26, 1969.

[3] W. B. GRAGG, *The Padé table and its relation to certain algorithms of numerical analysis*, SIAM Rev., 14 (1972), pp. 1–62.

[4] E. HAIRER, *Unconditionally stable methods for second-order differential equations*, Numer. Math., 32 (1979), pp. 373–379.

[5] ______, *Constructive characterization of A-stable approximations to* exp(z) *and its connection with algebraically stable Runge–Kutta methods*, Numer. Math., 39 (1982), pp. 247–258.

[6] E. HILLE, *Analytic Function Theory*, Vol. II, Blaisdell, Waltham, MA, 1962.

[7] A. ISERLES, *Generalized order star theory*, in Padé Approximation and Its Applications, Amsterdam 1980, M. G. de Bruin and H. van Rossum, Lecture Notes in Mathematics 888, Springer-Verlag, Berlin, 1981, pp. 228–238.

[8] ______, *Order stars and a saturation theorem for first-order hyperbolics*, IMA J. Numer. Anal., 2 (1982), pp. 49–61.

[9] ______, *Order stars, approximations and finite differences* II, *Theorems in approximation theory*, this Journal, 16 (1985), to appear.

[10] ______, *Order stars, approximations and finite differences* III, *Finite differences for* $u_t=\omega u_{xx}$, this Journal, 16 (1985), to appear.

[11] A. ISERLES AND S. NØRSETT, *A proof of the first Dahlquist barrier by order stars*, BIT, to appear.

[12] A. Iserles and M. J. D. Powell, *On the A-acceptability of rational approximations that interpolate the exponential function*, IMA J. Numer. Anal., 1 (1981), pp. 241–251.

[13] A. Iserles and G. Strang, *The optimal accuracy of difference schemes*, Trans. Amer. Math. Soc., 227 (1983), pp. 774–803.

[14] A. Iserles and R. Williamson, *Order and accuracy of semidiscretized finite differences*, IMA J. Numer. Anal., 4 (1984), pp. 289–307.

[15] R. Jeltsch and O. Nevanlinna, *Stability of explicit time discretizations for solving initial value problems*, Numer. Math., 37 (1981), pp. 61–91.

[16] ———, *Stability and accuracy of time discretizations for initial value problems*, Numer. Math., 40 (1982), pp. 245–296.

[17] ———, *Stability of semidiscretizations of hyperbolic problems*, SIAM J. Numer. Anal., 20 (1983), pp. 1210–1218.

[18] R. Jeltsch and K. G. Strack, *Accuracy bounds for semidiscretizations of hyperbolic problems*, Math. Comp. (1984), to appear.

[19] S. P. Nørsett and G. Wanner, *The real-pole sandwich for rational approximations and oscillation equations*, BIT, 19 (1979), pp. 89–94.

[20] G. Strang and A. Iserles, *Barriers for stability*, SIAM J. Numer. Anal., 20 (1983), pp. 1251–1257.

[21] G. Wanner, E. Hairer and S. P. Nørsett, *Order stars and stability theorems*, BIT, 18 (1978), pp. 475–489.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

014

RADAR AMBIGUITY FUNCTIONS AND GROUP THEORY*

L. AUSLANDER[†] AND R. TOLIMIERI[†]

Abstract. P. M. Woodward in the early 1950's introduced a mapping from a radar signal f to a function of two variables $W(f)$, called the ambiguity function, that plays a central role in the radar design problem. We may think of W as a nonlinear operator from $L^2(\mathbb{R})$ into $L^2(\mathbb{R}^2)$. The description of the range of W has been an open problem. This paper provides, in terms of special functions in $L^2(\mathbb{R})$ and $L^2(\mathbb{R}^2)$ a fairly complete description of $W(L^2(\mathbb{R}))$. We show also that $W(L^2(\mathbb{R}))$ is a closed subset of $L^2(\mathbb{R}^2)$ and if $W(f)+W(g)=W(h), f,g,h\in L^2(\mathbb{R})$ then $f=\lambda g$, λ a constant.

1. Introduction. Because radar computations are not familiar to the general mathematical community, we have begun this introduction with a brief simplified version of how ambiguity functions are used in radar computations. We will follow this with the familiar listing of what we consider our important new results.

Let $X_1,\cdots,X_N$ be N objects or targets and assume the radar is at the origin. Let $r_j(t)$, $j=1,\cdots,N$, denote the range (distance from the origin) of X_j and $v_j(t)$ denotes the velocity of X_j at time t. The problem is to transmit an electromagnetic wave or pulse for $-T<t<T$ and from the echo determine the quantities $r_j(0)$ and $v_j(0), j=1,\cdots,N$. Let $s(t)$ denote the pulse, where $s(t)$ is real valued, and let $e(t)$ denote the echo.

We will now briefly outline how information is extracted from $e(t)$. The computational process depends on a "representation" of $s(t)$ and some simplifying assumptions. The first step is to pass from the pulse to a complex valued function (representation) called the waveform of the pulse. If $g(t)\in L^2(\mathbb{R})$ we will use $\hat{g}(f)$ to denote the Fourier transform of g and call the variable f, frequency. Because $s(t)$ is real valued we have

$$\hat{s}(-f)=\hat{s}^*(f)$$

where we will (following electrical engineering notation) use * to denote the complex conjugate. Hence $s(t)$ is completely determined by its positive spectrum. Define

$$\Psi_s(t)=\int_0^\infty \hat{s}(f)e^{2\pi i f t}\,df.$$

Then

$$\Psi_s(t)=s(t)+i\sigma(t)$$

where σ is the Hilbert transform of s. Explicitly, using principal part integrals,

$$\sigma(t)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s(\tau)}{t-\tau}\,d\tau,$$

$$s(t)=-\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\sigma(\tau)}{t-\tau}\,d\tau.$$

*Received by the editors June 16, 1983, and in revised form, January 26, 1984.

[†] IBM T. J. Watson Research Center, Yorktown Heights, New York 10598. Present address, Department of Mathematics, Graduate Center, City University of New York, New York, New York 10036.

Using $\|f\|$, $f \in L^2(\mathbb{R})$, to denote the norm of f, we have

$$\|\Psi_s(t)\|^2 = 2\|s(t)\|^2.$$

It is customary to call $\|f\|^2$ the "energy" of the signal f. For the rest of the motivational discussion we will assume $\|\Psi_s(t)\|^2 = 1$, $t_0 = \int_{-\infty}^{\infty} t|\Psi_s(t)|^2 dt < \infty$ and

$$f_0 = \int_{-\infty}^{\infty} f|\hat{\Psi}_s(t)|^2 df < \infty.$$

It is usual to call t_0 the epoch and f_0 the carrier frequency.

DEFINITION. The waveform $u_s(t)$ of the pulse $s(t)$ is defined by

$$u_s(t) = \Psi_s(t+t_0)e^{-2\pi i f_0(t+t_0)}.$$

It follows that $s(t) = \mathrm{Re}\{\Psi_s(t)\} = \mathrm{Re}\{u_s(t-t_0)e^{2\pi i f_0 t}\}$, where $\mathrm{Re}\{\cdot\}$ denotes the real part of the function in the bracket, and $\|u_s(t)\|^2 = 1$. The function $u_s(t)$ is "slowly varying" in the sense that its spectrum is centered about the 0-frequency.

We would like the echo $e(t)$ to be "as much like" $s(t)$ as possible. If we have one target and the physical assumptions listed later are satisfied then

$$e(t) = \mathrm{Re}\{e^{-2\pi i f_0 x_0}u_s(t-t_0-x_0)e^{2\pi i(f_0-y_0)t}\} = \mathrm{Re}\{\Psi_e(t)\}$$

where $x_0 = (2/c)r_1(0)$, $y_0 = (2f_0/c)v_1(0)$ and c is the velocity of light. Hence for one target

x_0 = time delay of the echo,
y_0 = doppler or frequency shift of echo

completely determine $r_1(0)$ and $v_1(0)$. One estimates x_0, y_0 by the following method originally suggested by P. M. Woodward [W] and motivated by probabilistic considerations. Consider

$$\Psi_{xy}(t) = e^{-2\pi i f_0 x}u_s(t-t_0-x)e^{-2\pi i y t}e^{2\pi i f_0 t}$$

and form

$$I(x,y) = \left|\int_{-\infty}^{\infty} \Psi_e(t)\Psi_{xy}^*(t)\,dt\right|^2,$$

because $\|u_s(t)\|^2 = I$, $I(x_0,y_0) = 1$ and $I(x,y) \leq 1$ for all x,y. Thus if we plot $I(x,y)$ by light intensity on a screen the brightest point should be (x_0,y_0) and so we can determine $r_1(0)$ and $v_1(0)$ or the range and velocity of the target. It is crucial for us to observe that

$$I(x,y) = |A_u(x_0-x, y_0-y)|^2$$

where

$$A_u(x,y) = \int_{-\infty}^{\infty} u\left(t-\frac{x}{2}\right)u^*\left(t+\frac{x}{2}\right)e^{-2\pi i y t}dt.$$

We will now list our physical assumptions and then state the results for several targets.

Physical assumptions.

1. Radar cross sections of targets are independent of frequency.
2. All targets are in the far field of the radar.
3. Multiple reflecting waves among the targets are negligible.
4. The functions $r_j(t), j=1,\cdots,N$ are approximately linear for $-T<t<T$.
5. The velocity of the targets is small compared to the speed of electromagnetic propagation.

Then from several targets we have approximately $I(x,y)=M_1^2|A_u(x_1-x,y_1-y)|^2 + \cdots + M_N^2|A_u(x_N-x,y_N-y)|^2$ where the M_j depend on the range and the radar cross sections of the targets.

Actually $I(x,y)$ does not determine the number of targets, their range or velocity uniquely and, of course, $I(x,y)$ depends on the form of $A_u(x,y)$. Because of this $A_u(x,y)$ is called the ambiguity function of radar.

Woodward concludes his fundamental book [W] published in 1953 with the following paragraph. (We have changed notation, but nothing else, to fit with our conventions.)

> The reader may feel some disappointment, not unshared by the writer, that the basic question of what to transmit (choice of s) remains unanswered. One might have hoped that practical requirements of range and velocity resolution in any particular problem could be sketched in an x-y diagram and the waveform $u(t)$ then calculated to satisfy the requirements. It seems that this is not possible because the form of $|A_u(x,y)|^2$ cannot be arbitrarily chosen. The precise nature of the restrictions which must be placed in $|A_u(x,y)|^2$ has not been fully investigated.

Calvin H. Wilcox [W1] in 1960 took up the detailed study of ambiguity functions and called the problem posed above by Woodward the "synthesis problem of radar design." Wilcox used only Abelian harmonic analysis in his work. However, it turns out that there is a great deal to be gained by using the representation theory of the Heisenberg group and considering ambiguity functions as special functions on the Heisenberg group. This is not surprising because of the radar uncertainty principle and the deep relation between the Heisenberg group and the Heisenberg uncertainty principle (see [Wg] and [Wyl]). The desire to use the non-Abelian results forces us to operate in a slightly more general setting then Wilcox and so we will have to give slightly different treatments of many of his results.

We will now introduce notation that we will follow for the rest of this paper. It is intentionally slightly different from that used up to now.

If $\langle f,g\rangle$ denotes the usual inner product of functions $f,g\in L^2(\mathbb{R})$, defined by

$$\langle f,g\rangle=\int_{\mathbb{R}} f(t)g^*(t)\,dt,$$

then we can write

$$\mathscr{A}(f)(u,v)=\int_{\mathbb{R}} f\left(t-\frac{u}{2}\right)f^*\left(t+\frac{u}{2}\right)e^{-2\pi ivt}dt \tag{1}$$

as

$$\mathscr{A}(f)(u,v)=\left\langle f\left(t-\frac{u}{2}\right)e^{-\pi ivt},f\left(t+\frac{u}{2}\right)e^{\pi ivt}\right\rangle. \tag{2}$$

For $F(u,v)$ and $G(u,v)$ we define

$$\langle F,G\rangle_2=\int_{\mathbb{R}}\int_{\mathbb{R}}F(u,v)G^*(u,v)\,du\,dv$$

and $\|F\|_2^2=\langle F,F\rangle_2$. We will use $L^2(\mathbb{R}^2)$ to denote the above Hilbert space of square summable functions on $\mathbb{R}^2$.

We can now state two results from the paper.

THEOREM A. *The set of ambiguity functions $\mathscr{A}(f)$, $f\in L^2(\mathbb{R})$, is a closed subset of $L^2(\mathbb{R}^2)$.*

THEOREM B. *For f, $g\in L^2(\mathbb{R})$ let $\mathscr{A}(f)$ and $\mathscr{A}(g)$ be the corresponding ambiguity functions. Then $\mathscr{A}(f)+\mathscr{A}(g)$ is an ambiguity function if and only if $f=\lambda g$, λ a constant.*

The last part of this paper is devoted to ways of describing all ambiguity functions. In order to state some of these results we will need the following definition.

DEFINITION. Let $f\in L^2(\mathbb{R})$ and define

$$f_{ab}=e^{2\pi ibt}f(t+a),\qquad a,b\in\mathbb{Z},$$

and let $\mathscr{F}$ denote the set $\{f_{ab}|a,b\in\mathbb{Z}\}$. We will say that f generates an L^2-basis of $L^2(\mathbb{R})$ if $L^2(\mathbb{R})$ is the closure of linear combinations of elements of $\mathscr{F}$, but no proper subset of $\mathscr{F}$ has this property.

Theorem 6 of §4, due to R. Sacksteder, gives necessary and sufficient conditions for $f\in L^2(\mathbb{R})$ to generate an L^2-basis.

THEOREM C. *Let $f\in L^2(\mathbb{R})$ generate an L^2-basis. The set of ambiguity functions is the closure in $L^2(\mathbb{R}^2)$ of the set of functions*

$$\sum_{a,b,c,d\in\mathbb{Z}}\alpha(a,b)\alpha^*(c,d)K(a,b,c,d)A(f)(u+c-a,v+d-b),$$

where

$$K(a,b,c,d)=(-1)^{(a+c)(b+d)}e^{-\pi i[(b+d)u-(a+c)v]}$$

and $\alpha(a,b)$ is a function on $\mathbb{Z}\times\mathbb{Z}$ taking a finite number of nonzero values.

The importance of Theorem C can perhaps best be illuminated by the following special case.

Let

$$r(t)=\begin{cases}1, & |t|<\dfrac{1}{2},\\[2mm] 0, & |t|\geqq\dfrac{1}{2},\end{cases}$$

and let $r_{ab}=e^{2\pi ibt}r(t+a)$, $a,b\in\mathbb{Z}$. then the set $\{r_{ab}|a,b\in\mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$ and

$$A(r)=\begin{cases}\dfrac{\sin(\pi v(1-|u|))}{\pi v}, & |u|\leqq 1,\\[2mm] 0, & \text{otherwise.}\end{cases}$$

THEOREM D. *Let Ψ be the set of complex valued functions α on $\mathbb{Z}\times\mathbb{Z}$ such that*

$$\sum|\alpha(a,b)|^2<\infty.$$

Then the set of all ambiguity functions is given by

$$F(\alpha)(u,v)=\sum_{a,b,c,d\in\mathbf{Z}}\alpha(a,b)\alpha^*(c,d)K(a,b,c,d)A(r)(u+c-a,v+d-b)$$

where $K(a,b,c,d)$ *is defined in Theorem* C *and* $\alpha\in\Psi$. *Further, if* $f=\sum_{a,b\in\mathbf{Z}}\alpha(a,b)r_{ab}$, *then* $\mathscr{A}(f)(u,v)=F(\alpha)(u,v)$.

Thus we can describe the set of all ambiguity functions in terms of well-known functions.

Several theorems will be given different proofs. The techniques used in §2 will probably be accessible to most readers, as they require only the most basic results from Abelian harmonic analysis. Orthonormal bases play an important role, especially in the proofs of Theorems A and B, and essentially, translate the problem under consideration, into a problem of infinite matrices satisfying certain conditions (see the discussion following Theorem 2.4). The inversion formula, given in Lemma 2.2, is the main tool in earlier parts of the section and could be applied to prove these results, as well, Theorem C is about a special kind of orthonormal bases.

In §3, ideas arising from unitary representation theory of the Heisenberg group are applied to the study of ambiguity functions. The definitions and results stated at this point can be discussed within the framework of locally compact groups, but we will not do so.

Equally powerful and related ideas can be introduced from the theory of Hilbert–Schmidt operators. Certain of these ideas have been previously considered in [W1] and [S] and applied to the problem of synthesizing ambiguity functions which best approximate, in the L^2-norm, a given function in $L^2(\mathbf{R}^2)$. This theory will play no direct part in this work.

An interesting aspect of ambiguity theory is that it finds itself within the scope of several mathematical disciplines. However, it should be emphasized, that radar theory and more generally image processing create special classes of problems not usually encountered in these general mathematical theories.

2. Ambiguity functions. The elementary properties of ambiguity functions will be established in this section using methods of Abelian harmonic analysis. Our main reference will be [K].

Consider f, $g\in L^1(\mathbf{R})$ and define

$$\mathscr{A}(f,g)(u,v)=\int f\left(t-\frac{u}{2}\right)g^*\left(t+\frac{u}{2}\right)e^{-2\pi ivt}dt.$$

We call $\mathscr{A}(f,g)$ the cross-ambiguity function of f with g. The ambiguity function $\mathscr{A}(f)$ of f is given by

$$\mathscr{A}(f)\equiv\mathscr{A}(f,f).$$

A closely related expression $\mathscr{B}(f,g)$ is sometimes also called the cross-ambiguity function of f with g and for some purposes is easier to work with. Set

$$\mathscr{B}(f,g)\equiv e^{\pi iuv}\circ\mathscr{A}(f,g).$$

A simple change of variables argument shows that we can write

$$\mathscr{B}(f,g)(u,v)=\int f(t)g^*(t+u)e^{-2\pi ivt}dt.$$

In this paper we will study $\mathscr{B}(f,g)$, but to avoid confusion we will only call $\mathscr{A}(f,g)$ the cross-ambiguity function of f with g.

There are two obvious ways to consider $\mathscr{B}(f,g)$. The first begins by setting

$$h(u,t)=f(t)g^*(t+u)$$

and viewing $h(u,t)$ as a family of functions in t, parameterized by u. In general, if $F(x,y)$ is any function of two variables x and y, for any fixed $x\in\mathbb{R}$, we set

$$F_x(y)=F(x,y)$$

and consider F_x as a function of y. Using this notation, we can write

$$\mathscr{B}(f,g)_u(v)=\hat{h}_u(v).$$

The behavior of $h(u,t)$ determines to a large extent the behavior of $\mathscr{B}(f,g)$. The following elementary result provides the necessary information upon which a great deal of ambiguity function theory rests.

LEMMA 2.1. *For $f,g\in L^2(\mathbb{R})$, the function $h(u,t)=f(t)g^*(t+u)$ is in $L^2(\mathbb{R}^2)$ and*

$$\|h\|_2^2=\|f\|^2\|g\|^2.$$

Proof. By Fubini's theorem and the positivity of $|h(u,t)|$

$$\iint |h(u,t)|^2\,du\,dt=\int\left[\int |h(u,t)|^2\,du\right]dt$$
$$=\int |f(t)|^2\left[\int |g^*(t+u)|^2\,du\right]dt.$$

But

$$\int |g^*(t+u)|^2\,du=\|g\|^2 \quad \forall t.$$

Hence

$$\|h\|_2^2=\|g\|^2\int |f(t)|^2\,dt=\|g\|^2\|f\|^2.$$

LEMMA 2.1′. *Let $f_1, f_2, g_1, g_2\in L^2(\mathbb{R})$ and let*

$$h_1(u,t)=f_1(t)g_1^*(t+u),$$
$$h_2(u,t)=f_2(t)g_2^*(t+u).$$

Then $\langle h_1,h_2\rangle_2=\langle f_1,f_2\rangle\langle g_2,g_1\rangle$.

Proof. Formally

$$\iint h_1(u,t)h_2^*(u,t)\,du\,dt=\int\left[\int h_1(u,t)h_2^*(u,t)\,du\right]dt$$
$$=\int f_1(t)f_2^*(t)\left[\int g_1^*(t+u)g_2(t+u)\,du\right]dt.$$

But $\int g_1^*(t+u)g_2(t+u)du=\langle g_2,g_2\rangle$ all t. And so Lemma 2.1′ follows.

To make this rigorous, note that if

$$\iint |h_1(u,t)h_2^*(u,t)|\,du\,dt < \infty$$

then we may replace the double integral with the iterated integral in any order. But, if h_1; $h_2 \in L^2(\mathbb{R}^2)$, so are $|h_1|$, $|h_2|$, and we know that the dot product $\langle |h_1|, |h_2| \rangle_2 < \infty$. Thus our formal manipulations are legitimate.

It follows, also by Fubini's theorem, that for almost every $u \in \mathbb{R}$, the function $h_u \in L^2(\mathbb{R})$. Since h_u is the product of two $L^2(\mathbb{R})$ functions, it is in $L^2(\mathbb{R})$ by the Schwarz inequality. The formula

$$\mathscr{B}(f,g)_u(v) = \hat{h}_u(v)$$

implies that $\mathscr{B}(f,g)_u$ is the Fourier transform of $h_u \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. By standard Abelian harmonic analysis (see [K]) we have the following corollary.

COROLLARY. *Let $C(\mathbb{R})$ denote the continuous functions on $\mathbb{R}$. For almost every $u \in \mathbb{R}$,*

$$\mathscr{B}(f,g)_u(v) \in C(\mathbb{R}) \cap L^2(\mathbb{R})$$

and

$$\lim_{|v| \to \infty} \mathscr{B}(f,g)_u(v) = 0.$$

THEOREM 2.1. *$\mathscr{B}(f,g) \in L^2(\mathbb{R}^2)$, whenever f, $g \in L^2(\mathbb{R})$. Moreover,*

$$\|\mathscr{B}(f,g)\|_2^2 = \|f\|^2 \|g\|^2.$$

Proof. By Fubini's theorem,

$$\|\mathscr{B}(f,g)\|_2^2 = \int \left[\int |\hat{h}_u(v)|^2 dv \right] du$$

which by the Plancherel theorem becomes

$$\int \left[\int |h_u(t)|^2 du \right] du = \|f\|^2 \|g\|^2.$$

THEOREM 2.1′. *Let $f_1, f_2, g_1, g_2 \in L^2(\mathbb{R})$. Then*

$$\langle \mathscr{B}(f_1,g_1), \mathscr{B}(f_2,g_2) \rangle_2 = \langle f_1, f_2 \rangle \langle g_2, g_1 \rangle.$$

Proof. Since $\mathscr{B}(f_\alpha, g_\alpha)$, $\alpha = 1,2$ are in $L^2(\mathbb{R}^2)$ so are $|\mathscr{B}(f_\alpha, g_\alpha)|$ and so we may apply Fubini's theorem and write

$$\iint \mathscr{B}(f_1,g_1)\mathscr{B}^*(f_2,g_2)\,du\,dv = \int \left[\int h_1 u(v) h_2^* u(v)\,dv \right] du$$
$$= \langle f_1, f_2 \rangle \langle g_2, g_1 \rangle.$$

The following "inversion" formulas provide important tools for the further study of $\mathscr{B}(f,g)$.

LEMMA 2.2. *For any f, $g \in L^2(\mathbb{R})$.*

$$f(t)g^*(t+u)=\int \mathscr{B}(f,g)(u,v)e^{2\pi itv}dv,$$
$$f(t)\hat{g}(x)^*e^{-2\pi ixt}=\check{\mathscr{B}}(f,g)(x,t),$$

for almost every $(u,t)\in\mathbb{R}^2$ and almost every $(x,t)\in\mathbb{R}^2$, where $\check{\mathscr{B}}$ denotes the inverse Fourier transform in $L^2(\mathbb{R}^2)$.

Proof. Since for almost every $u, \mathscr{B}(f,g)_u \in L^2(\mathbb{R})$ we can apply the inverse Fourier transform to $\mathscr{B}(f,g)_u$, in the sense given by the Plancherel theorem. Thus, for almost every u, the first formula holds, for almost every t. By Fubini's theorem, applied to the characteristic function of the set of $(u,t)\in\mathbb{R}^2$ for which the first formula *does not* hold, we get that it holds except on a set in $\mathbb{R}^2$ of measure zero.

To prove the second formula, take the inverse Fourier transform of the first formula with respect to the u variable.

Let $\mathscr{B}(f)=\mathscr{B}(f,f)$. Then

$$\mathscr{B}(f)(-u,-v)=\int f(t)f^*(t-u)e^{2\pi ivt}dt.$$

Let $s=t-u$. Then

$$(**)\qquad \mathscr{B}(f)(-u,-v)=\int f(s+u)f^*(s)e^{2\pi iv(s+u)}ds=e^{2\pi ivu}\mathscr{B}(f)^*(u,v).$$

Now consider the change of variables

$$\tau=t+u,\qquad t=t.$$

and let $H(t,\tau)=f(t)f^*(\tau)$. Then

$$H(t,\tau)=\int \mathscr{B}(f)(\tau-t,v)e^{2\pi itv}dv \quad\text{and}\quad H^*(t,\tau)=\int \mathscr{B}(f)^*(\tau-t,v)e^{-2\pi itv}dv.$$

Using formula $(**)$ we have

$$H^*(t,\tau)=\int \mathscr{B}(f)(t-\tau,-v)e^{-2\pi itv}e^{-2\pi i(\tau-t)v}dv=H(\tau,t).$$

Consider the mapping U: $\mathscr{B}(f)(u,v)\to H(t,\tau)$. This has the property that it is 1 to 1 and norm preserving. Further the $H(t,\tau)$ are easily seen to satisfy the functional equations

1. $H^*(t,\tau)=H(\tau,t)$,
2. $H(t,t)\geqq 0$,
3. $H(\xi,\xi)H(t,\tau)=H(t,\xi)H(\xi,\tau)$.

THEOREM 2.2. *Let $F(t,\tau)\in L^2(\mathbb{R}^2)$ and satisfy equations* 1,2 *and* 3 *above. Then there exists a $\mathscr{B}(f)$ such that $U(\mathscr{B}(f))=F(t,\tau)$.*

Proof. Equations 1 and 3 combine to yield

$$F(t,t)F(\xi,\xi)=|F(t,\xi)|^2.$$

By hypothesis, $F(t,\xi)\in L^2(\mathbb{R}^2)$ and so

$$\iint F(t,t)F(\xi,\xi)\,dt\,d\xi=\iint |F(t,\xi)|^2\,dt\,d\xi$$

$$=\|F(t,\xi)\|_2^2<\infty.$$

Since $F(t,t)\geqq 0$, we may apply Fubini's theorem to conclude that

$$\left(\int F(t,t)\,dt\right)^2=\|F(t,\xi)\|_2^2\geqq 0.$$

The only interesting case is when $\|F(t,\xi)\|_2>0$ and so there exists ξ_0 such that $F(\xi_0,\xi_0)>0$ and $F(t,\xi_0)\in L^2(\mathbb{R})$.

Define $f(t)=F(t,\xi_0)/(F(\xi_0,\xi_0))^{1/2}$. Then

$$f(t)f^*(\tau)=F(t,\tau).$$

It is clear that $U(\mathscr{B}(f))=F(t,\tau)$ and so we have proven our theorem.

Consider the mapping

$$\mathscr{B}\colon L^2(\mathbb{R})\times L^2(\mathbb{R})\to L^2(\mathbb{R}^2).$$

It is clearly bilinear.

THEOREM 2.3. *$\mathscr{B}$ is continuous and the image of $\mathscr{B}$ spans a dense subspace of* $L^2(\mathbb{R}^2)$.

Proof. If $f_n\to f$ and $g_n\to g$ in $L^2(\mathbb{R})$ then, by the continuity of the Fourier transform $\hat{g}_n\to\hat{g}$ in $L^2(\mathbb{R})$ and

$$f_n(y)\hat{g}_n(x)^*e^{-2\pi ixy}\to f(y)\hat{g}(x)^*e^{-2\pi ixy}$$

in $L^2(\mathbb{R}^2)$. Lemma 2 implies

$$\check{\mathscr{B}}(f_n,g_n)\to\check{\mathscr{B}}(f,g)$$

in $L^2(\mathbb{R}^2)$ and hence, $\mathscr{B}$ is continuous where $\check{}$ denotes the inverse Fourier transform in $L^2(\mathbb{R}^2)$.

Suppose $F\in L^2(\mathbb{R}^2)$ is orthogonal to the span of the image of $\mathscr{B}$. Then, $\check{F}(x,y)e^{2\pi ixy}$ is orthogonal to the span of the space of all products $f(y)\hat{g}^*(x)$ which is known to be dense in $L^2(\mathbb{R}^2)$. It follows $F=0$ and the theorem is proved.

COROLLARY. *The collection of functions $\mathscr{B}(f)$, $f\in L^2(\mathbb{R})$, spans a dense subspace of* $L^2(\mathbb{R}^2)$.

Proof. Suppose $F\in L^2(\mathbb{R})$ is orthogonal to every $\mathscr{B}(f)$, $f\in L^2(\mathbb{R})$. Then, since

$$\mathscr{B}(f+g)=\mathscr{B}(f)+\mathscr{B}(f,g)+\mathscr{B}(g,f)+\mathscr{B}(g),$$

we have F orthogonal to $\mathscr{B}(f,g)+\mathscr{B}(g,f)$. Also, since

$$\mathscr{B}(f+ig)=\mathscr{B}(f)+i\mathscr{B}(g,f)-i\mathscr{B}(f,g)+\mathscr{B}(g)$$

we have F orthogonal to $\mathscr{B}(g,f)-\mathscr{B}(f,g)$. Thus, F is orthogonal to $\mathscr{B}(f,g)$, f, $g\in L^2(\mathbb{R})$, and by the theorem is zero almost everywhere.

The function $\mathscr{B}(f,g)$ can also be viewed as a cross-correlation. For fixed $v\in\mathbb{R}$, set

$$G_v(t)=g(t)e^{2\pi ivt}.$$

Form the cross-correlation $f \circ G_v$ defined by

$$f \circ G_v(u) = \int f(t) G_v^*(u+t)\,dt.$$

and upon writing out the integral, observe that

$$\mathscr{B}(f,g)(u,v) = e^{2\pi iuv} f \circ G_v(u).$$

LEMMA 2.3. $\mathscr{B}(f,g)(u,v) = e^{2\pi iuv}\mathscr{B}(\hat{f},\hat{g})(-v,u)$.

Proof. Since

$$\mathscr{B}(f,g)(u,v) = e^{2\pi iuv}\langle f(t), G_v(u+t)\rangle$$

it follows that

$$\mathscr{B}(f,g)(u,v) = e^{2\pi iuv}\langle \hat{f}, G_v(u+t)\rangle$$

But

$$\int g(u+t)e^{2\pi ivt}e^{-2\pi itx}\,dt = e^{-2\pi iuv}e^{2\pi iux}\hat{g}(x-v)$$

which proves the lemma.

COROLLARY. $\hat{f}(x)\hat{g}(x-v)^* = \int \mathscr{B}(f,g)(u,v)e^{2\pi iu(x-v)}\,du$.

We will now show $\mathscr{B}(f,g)$ is continuous. The first step is the next lemma.

LEMMA 2.4. *If $f_n \to f$ and $g_n \to g$ in $L^2(\mathbb{R})$ then*

$$\mathscr{B}(f_n,g_n) \to \mathscr{B}(f,g)$$

uniformly over $\mathbb{R}^2$.

Proof. (R. Sacksteder independently suggested this proof to one of the authors.) Set $\mathscr{B}_n(u,v) = \mathscr{B}(f,g)(u,v) - \mathscr{B}(f_n,g_n)(u,v)$. Then

$$\mathscr{B}_n(u,v) = \int \left((f(t)-f_n(t))g^*(t+u) + f_n(t)(g^*(t+u)-g_n^*(t+u))\right)e^{-2\pi ivt}\,dt$$

and by the Schwarz inequality,

$$|\mathscr{B}_n(u,v)| \leq \|f-f_n\|\,\|g\| + \|f_n\|\,\|g-g_n\|.$$

The lemma follows.

THEOREM 2.4. *$\mathscr{B}(f,g)$ is a continuous bounded function which achieves its maximum $\langle f,g\rangle$ at the origin.*

Proof. The Schwarz inequality proves everything except for the continuity. By the preceding theorem it is sufficient to prove continuity for f and g taken from a dense subspace of $L^2(\mathbb{R})$. The set of functions

$$\{e^{-\pi(t+r)^2} : r \in \mathbb{R}\}$$

spans a dense subspace of $L^2(\mathbb{R})$. Taking f and g from this span it is easy to see that we are done once we show

$$\mathscr{B}\left(e^{-\pi t^2}, e^{-\pi(t+r)^2}\right)$$

is continuous, for all r. But,

$$\mathscr{B}\left(e^{-\pi t^2}, e^{-\pi(t+r)^2}\right)=\frac{\sqrt{2}}{2}e^{-\pi/2(u+r)^2}e^{-\pi/2v^2}e^{\pi i(r+u)v}$$

which is clearly continuous.

Let $\mathscr{I}_i$ be an orthonormal basis of $L^2(\mathbb{R})$. Then the set of functions $\psi_{ik}=\mathscr{B}(\mathscr{I}_i,\mathscr{I}_k)$, $i,k\in\mathbb{Z}$ is orthonormal by Lemma 1′ and complete by Theorem 3. Now $f(t)\in L^2(\mathbb{R})$ can be written as

$$f(t)=\sum_{-\infty}^{\infty} a_i\mathscr{I}_i, \qquad \sum_{-\infty}^{\infty}|a_i|^2=\|f\|^2<\infty.$$

Similarly, $F(u,v)\in L^2(\mathbb{R}^2)$ can be written as

$$F(u,v)=\sum_{m,n\in\mathbb{Z}} c_{mn}\psi_{mn}, \qquad \sum_{m,n\in\mathbb{Z}}|c_{mn}|^2<\infty.$$

Now consider $\mathscr{B}(f)\in L^2(\mathbb{R}^2)$. Then if

$$c_{mn}=\int\int\mathscr{B}(f)\bar{\psi}_{mn}du\,dv,$$
$$\mathscr{B}(f)=\sum c_{mn}\psi_{mn}.$$

By Lemma 2.1′

$$c_{mn}=\langle f,\mathscr{A}_m\rangle\langle f,\mathscr{I}_n\rangle^*=a_m a_n^*.$$

Conversely, if $c_{mn}=a_m a_n^*$ then $f=\sum a_m\mathscr{I}_m\in L^2(\mathbb{R})$ and $H(t,\tau)=f(t)f(\tau)^*$ satisfies the hypothesis of Theorem 2.2. Hence $F(u,v)\in L^2(\mathbb{R}^2)$ is an ambiguity function if and only if $c_{mn}=a_m\bar{a}_n$.

COROLLARY. *Let $F(u,v)\in L^2(\mathbb{R}^2)$ and $F=\sum_{m,n\in Z}c_{mn}\psi_{mn}$. Then $F(u,v)$ is an ambiguity function if and only if $c_{kk}c_{mn}=c_{mk}c_{kn}$, $c_{mn}=c_{mm}^*$ and $c_{kk}\geq 0$ all m,n, $k\in\mathbb{Z}$.*

THEOREM A. *The set of ambiguity functions is a closed subset of $L^2(\mathbb{R}^2)$.*

Proof. Let $F(u,v)$ be the limit of sequence of ambiguity functions $\mathscr{B}(f_i)$, $i=1,\cdots,n\cdots$. Let

$$c_{mn}(i)=\langle\mathscr{B}(f_i),\psi_{mn}\rangle_2,$$
$$c_{mn}=\langle F,\psi_{mn}\rangle_2.$$

Hence for each i, $c_{mn}(i)$ satisfy the conditions in the above corollary. Since $\lim_{i\to\infty}\mathscr{B}(f_i)=F$, we have for all $m,n\lim_{i\to\infty}c_{mn}(i)=c_{mn}$. Hence the c_{mn} satisfy the equations of the above corollary and F is an ambiguity function.

THEOREM B. *For f, $g\in L^2(\mathbb{R})$ let $\mathscr{B}(f)$ and $\mathscr{B}(g)$ be the corresponding ambiguity functions. Then $\mathscr{B}(f)+\mathscr{B}(g)$ is an ambiguity function if and only if $f=\lambda g$, λ a constant.*

Now consider $\mathscr{B}(f)$ and $\mathscr{B}(cf)$ where c is a constant. Then by direct computation

$$\mathscr{B}(f)+\mathscr{B}(cf)=\mathscr{B}\left(\sqrt{1+|c|^2f}\right).$$

Now let f, $g\in L^2(\mathbb{R})$ and consider $\mathscr{B}(f)+\mathscr{B}(g)$. If

$$\mathscr{B}(f)=\sum_{a,b,c,d\in\mathbb{Z}}\alpha_1(a,b,c,d)F_{abcd}, \quad \mathscr{B}(g)=\sum_{a,b,c,d\in\mathbb{Z}}\alpha_2(a,b,c,d)F_{abcd}.$$

Then

$$\mathscr{B}(f)+\mathscr{B}(g)=\sum_{a,b,c,d\in\mathbb{Z}}\left[\alpha_1(a,b,c,d)+\alpha_2(a,b,c,d)\right]F_{abcd}.$$

Now assume that $\mathscr{B}$ $(f)+\mathscr{B}$ (g) is an ambiguity function. Then the corollary to Theorem 2.4 implies that

$$\begin{aligned}\alpha_1(r,s,r,s)\alpha_2(a,b,c,d)+\alpha_1(a,b,c,d)\alpha_2(r,s,r,s)\\=\alpha_1(r,s,c,d)\alpha_2(a,b,r,s)+\alpha_1(a,b,r,s)\alpha_2(r,s,c,d).\end{aligned}\tag{2.1}$$

Because $\mathscr{B}(f)$ and $\mathscr{B}(g)$ are ambiguity functions, we know that

$$\begin{aligned}\alpha_1(a,b,c,d)&=\alpha_1(a,b)\alpha_1^*(c,d),\\\alpha_2(a,b,c,d)&=\alpha_2(a,b)\alpha_2^*(c,d).\end{aligned}$$

Hence we can rewrite (2.1) as

$$\begin{aligned}\alpha_1(r,s)\alpha_1^*(r,s)\alpha_2(a,b)\alpha_2^*(c,d)+\alpha_1(a,b)\alpha_1^*(c,d)\alpha_2(r,s)\alpha_2^*(r,s)\\=\alpha_1(r,s)\alpha_1^*(c,d)\alpha_2(a,b)\alpha_2^*(r,s)+\alpha_1(a,b)\alpha_1^*(r,s)\alpha_2(r,s)\alpha_2^*(c,d).\end{aligned}\tag{2.2}$$

Assume f is not the zero function, then $\alpha_1(r_0,s_0)\neq 0$ for some r_0 and s_0. It is easy to see that there is no loss in generality in assuming that $\alpha_1(0,0)\neq 0$. Then setting $v=s=0$ in (2.2) we obtain

$$\left[\alpha_1(0,0)\alpha_2(a,b)-\alpha_2(0,0)\alpha_1(a,b)\right]\left[\alpha_1^*(0,0)\alpha_2^*(c,d)-\alpha_2^*(0,0)\alpha_1^*(c,d)\right]=0;$$

this implies that

$$\alpha_1(0,0)\alpha_2(a,b)=\alpha_2(0,0)\alpha_1(a,b)$$

or

$$\alpha_2(a,b)=\frac{\alpha_2(0,0)}{\alpha_1(0,0)}\alpha_1(a,b),\quad\text{for all } a,b\in\mathbb{Z}.$$

Thus if $c=\alpha_2(0,0)/\alpha_1(0,0)$ we have

$$g=cf.$$

This proves our assertion.

THEOREM 2.5. *Let f, $g\in L^2(\mathbb{R})$ and assume*

$$\mathscr{B}(f)=\mathscr{B}(g).$$

Then $f=c\circ g$ almost everywhere, where c is a constant and $|c|=1$.

Proof. By Lemma 2.2.

$$f(y)\hat{f}(x)^*=g(y)\hat{g}(x)^*$$

for almost all $(x,y)\in\mathbb{R}^2$. If f does not vanish on a set of positive measure, then for some y_0 we have $f(y_0)\neq 0$ and

$$f(y_0)\hat{f}(x)^*=g(y_0)\hat{g}(x)^*$$

holds for almost every x. Thus, there is a constant $c=g(y_0)^*/f(y_0)^*$ such that

$$f(y)=c\circ g(y),$$

for almost every y. The constant must have modulus one since $\mathscr{B}(f)=\mathscr{B}(g)$.

If f vanishes almost everywhere that $\mathscr{B}(f)=0$ and $g(y)=\hat{g}(x)^*=0$ almost everywhere in $(x,y)\in\mathbb{R}^3$. It is easy to see that $g(y)=0$ almost everywhere.

The same argument proves the following.

COROLLARY. *For f, $g\in L^2(\mathbb{R})$, if $\mathscr{B}(f,g)=0$ then $f=0$ almost everywhere or $g=0$ almost everywhere.*

We will return now to the ambiguity functions $\mathscr{A}(f)$ and denote by $\mathscr{R}$ the collection of all ambiguity functions.

Let $SL(2,\mathbb{R})$ denote the group of all 2×2 real matrices of determinant one, acting on $\mathbb{R}^2$ by the rule

$$T(u,v)=(au+bv,cu+dv)$$

where

$$T=\begin{bmatrix} a & b\\ c & d\end{bmatrix}.$$

As a group $SL(2,\mathbb{R})$ is generated by the following matrices:

$$J=\begin{bmatrix} 0 & 1\\ -1 & 0\end{bmatrix},\quad t(a)=\begin{bmatrix} 1 & 0\\ a & 1\end{bmatrix},\quad a\in\mathbb{R},\quad m(b)=\begin{bmatrix} b & 0\\ 0 & 1/b\end{bmatrix},\quad b>0.$$

To see this for $c>0$, write

$$\begin{bmatrix} a & b\\ c & d\end{bmatrix}=t\left(\frac{a}{c}\right)J^{-1}m(c)t\left(\frac{d}{c}\right).$$

We will now state how $SL(2,\mathbb{R})$ acts on $\mathscr{R}$.

THEOREM 2.6. *$\mathscr{R}$ is invariant under the action of $SL(2,\mathbb{R})$, and*

1. $\mathscr{A}(f)\circ J=\mathscr{A}(\hat{f})$,
2. $\mathscr{A}(f)\circ t(a)=\mathscr{A}(g)$ *where* $g(t)=e^{\pi i a t^2}f(t)$,
3. $\mathscr{A}(f)\circ m(b)=\mathscr{A}(h)$ *where* $h(t)=f(bt)$.

Proof. Statement 1 follows from Lemma 2.3. The last two statements can easily be proved by direct substitution.

3. Ambiguity functions and the Heisenberg group. In this section, we will work with the ambiguity functions $\mathscr{A}(f)$. A unitary operator on $L^2(\mathbb{R})$ is a linear mapping U of $L^2(\mathbb{R})$ satisfying

$$\langle Uf,Ug\rangle=\langle f,g\rangle,$$

for all f, $g\in L^2(\mathbb{R})$. The collection of all unitary operators U on $L^2(\mathbb{R})$ forms a group under composition which will be denoted by $\mathscr{U}$. An implication of the Plancherel theorem is that the Fourier transform, denoted by $\mathscr{F}$, is a unitary operator on $L^2(\mathbb{R})$.

The following two unitary operators on $L^2(\mathbb{R})$ play an important role in Abelian harmonic analysis and hence, the development of the theory of the ambiguity function given in the preceding section.

For $f\in L^2(\mathbb{R})$ and $a\in\mathbb{R}$, set

$$(S(a)f)(t)=f(t+a),\qquad t\in\mathbb{R},$$
$$(M(a)f)(t)=e^{2\pi i a t}f(t),\qquad t\in\mathbb{R},$$

and observe that the mappings $S(a)$ and $M(a)$ are unitary operators of $L^2(\mathbb{R})$.

Consider, now, S: $\mathbb{R} \to \mathcal{U}$ and M: $\mathbb{R} \to \mathcal{U}$ as mappings from $\mathbb{R}$ into $\mathcal{U}$. We set

$$\mathcal{S} = S(\mathbb{R}), \qquad \mathcal{M} = M(\mathbb{R}),$$

and call $\mathcal{S}$ the shift operators and $\mathcal{M}$ the multiplication operators. Both $\mathcal{S}$ and $\mathcal{M}$ are subgroups of $\mathcal{U}$ and in fact, we have the following lemma.

LEMMA 3.1. *S and M are group isomorphisms of $\mathbb{R}$ into $\mathcal{U}$.*

Proof. Immediate from the definition.

The reason that non-Abelian group theory enters into the study of ambiguity functions is contained in the next result.

LEMMA 3.2. $M(v)S(u) = e^{-2\pi iuv}S(u)M(v)$.

Proof. For $f \in L^2(\mathbb{R})$,

$$(M(v)S(u)f)(t) = e^{2\pi ivt}(S(u)f)(t) = e^{2\pi ivt}f(t+u),$$
$$(S(u)M(v)f)(t) = (M(v)f)(t+u) = e^{2\pi iv(t+u)}f(t+u),$$

which verifies the truth of the lemma.

Thus, the operators $M(v)$ and $S(u)$ do not commute. This observation is the mathematical basis for the introduction of the Heisenberg group in quantum mechanics and is an expression of the uncertainty principal. We will now define the Heisenberg group and study its implications in ambiguity function theory.

Let I denote the identity operator on $L^2(\mathbb{R})$ and set

$$C(\lambda) = \lambda I, \qquad \lambda \in \mathbb{C}, \quad |\lambda| = 1.$$

Then, $C(\lambda)$ is a unitary operator and the mapping

$$C\colon \mathbb{C}(1) \to \mathcal{U}(L^2(\mathbb{R})) = \mathcal{U},$$

where $\mathbb{C}(1)$ denotes the multiplicative group of complex numbers of modulus 1, is a group monomorphism. We set $\mathcal{C}$ equal to the range of C.

Clearly, $\mathcal{C}$ is a subgroup of $\mathcal{U}$ and is, in fact, the center of $\mathcal{U}$.

Let

$$\mathcal{H} = \mathcal{C} \circ \mathcal{M} \circ \mathcal{S}$$

denote the set of operators of the form

$$C(\lambda)M(b)S(a), \qquad |\lambda| = 1, \quad a, b \in \mathbb{R}.$$

THEOREM 3.1. *$\mathcal{H}$ is a subgroup of $\mathcal{U}$.*

Proof. By Lemma 3.2, we can write

$$\begin{aligned} &C(\lambda_1)M(b_1)S(a_1)C(\lambda_2)M(b_2)S(a_2) \\ &\quad = C(\lambda_1)C(\lambda_2)C(e^{2\pi ia_1b_2})M(b_1)M(b_2)S(a_1)S(a_2), \end{aligned}$$

which by Lemma 3.1, becomes

$$C(\lambda_1\lambda_2 \circ e^{2\pi ia_1b_2})M(b_1+b_2)S(a_1+a_2).$$

Thus, the product of two operators in $\mathcal{H}$ is again in $\mathcal{H}$.

It follows that

$$I = C(\lambda_1)M(b_1)S(a_1)C(\lambda_2)M(-b_1)S(-a_1)$$

if and only if $\lambda_2 = \lambda_1^{-1}e^{2\pi ia_1b_1}$ and hence, the inverse of an operator in $\mathcal{H}$ is again in $\mathcal{H}$.

An alternate definition of $\mathscr{H}$ can be taken to be the group generated by $\mathscr{M}$ and $\mathscr{I}$.

$\mathscr{H}$ is sometimes called the Heisenberg group, however, we will reserve this term for the abstractly defined group N given as follows.

As a set N consists of all points $\mathbf{x}=(x_1,x_2,x)\in\mathbb{R}^3$. The multiplication rule on N is given by the formula

$$\mathbf{x}\circ\mathbf{y}=\left(x_1+y_1,x_2+y_2,x+y+\tfrac{1}{2}(x_2y_1-x_1y_2)\right).$$

It is easy to verify that N is a group having centext X consisting of all points $(0,0,x)$, $x\in\mathbb{R}$.

For future use, we will single out two especially important automorphisms of N. Let $\mathscr{J}$ denote the mapping of N given by

$$\mathscr{J}(\mathbf{x})=(x_2,-x_1,x).$$

Clearly, $\mathscr{J}$ is an automorphism on N which acts by the identity mapping when restricted to the center.

Define D: $N\to\mathscr{U}$ by setting

$$D(\mathbf{x})=C(e^{2\pi i\lambda(\mathbf{x})})M(x_1)S(x_2)$$

where $\lambda(\mathbf{x})=x+\frac{1}{2}x_1x_2$. Equivalently,

$$(D(\mathbf{x})f)(t)=C(e^{2\pi i\lambda(\mathbf{x})})e^{2pix_1t}f(t+x_2).$$

Using Lemma 3.2, the next result is easily proved.

THEOREM 3.2. *D: $N\to\mathscr{U}$ is a group homomorphism satisfying*

1. $\ker D=\{(0,0,x)\colon x\in\mathbb{Z}\}$,

2. $\operatorname{im} D=\mathscr{H}$.

The group homomorphism D has, by necessity, been built in a non-Abelian fashion from the group homomorphisms S and M. In a sense, examined more closely in the next section, the Fourier transform $\mathscr{F}$ is closely related to these group homomorphisms and hence, to the Heisenberg group N. For a more complete discussion see [A–T]. At this time, the formulas of the next theorem will suffice.

THEOREM 3.3.

$$\mathscr{F}S(x)\mathscr{F}^{-1}=M(x),\quad \mathscr{F}M(x)\mathscr{F}^{-1}=S(-x),\quad \mathscr{F}D(\mathbf{x})\mathscr{F}^{-1}=D(\mathscr{J}\mathbf{x}).$$

Proof. The first two formulas are easily proved by Abelian harmonic analysis methods. The last formula comes from the definition of D and Lemma 3.2.

The ambiguity function $\mathscr{A}(f)$ can be expressed in terms of the group homomorphism D. This is accomplished in the next theorem.

THEOREM 3.4. *For* $\mathbf{x}\in N$, *and* $f\in L^2(N)$,

$$\mathscr{A}(f)(x_2,x_1)=e^{2\pi ix}\langle f,D_{\mathbf{x}}f\rangle.$$

Proof. Since

$$D_{\mathbf{x}}f(t)=e^{2\pi i\lambda(\mathbf{x})}e^{2\pi ix_1t}f(t+x_2),$$

we can write,

$$\langle f,D_{\mathbf{x}}f\rangle = e^{-2\pi i\lambda(\mathbf{x})}\int f(t)f^*(t+x_2)e^{-2\pi i x,t}\,dt,$$

$$= e^{-2\pi i\lambda(\mathbf{x})}\mathscr{B}(f)(x_2,x_1),$$

$$= e^{-2\pi i x}\mathscr{A}(f)(x_2,x_1),$$

which proves the theorem.

The significance of this result is that we can view ambiguity functions as well-known objects in the theory of unitary representations of the Heisenberg group. For the most part, the theory we develop can be generalized to the theory of unitary representations of locally compact groups on Hilbert spaces but we will restrict our analysis to what we need to study ambiguity functions.

A unitary representation of N is a homomorphism U of N into $\mathscr{U}$. Let U be a unitary representation of N and $f\in L^2(\mathbb{R})$. Consider the function on N defined by

$$p(\mathbf{x}) = \langle U_{\mathbf{x}}f,f\rangle, \qquad \mathbf{x}\in N.$$

THEOREM 3.5. *The function p is positive define on N, in the sense that, for any finite number of elements $g_1,\cdots,g_n$ in N and complex numbers $\lambda_1,\cdots,\lambda_n$ we have,*

$$\sum_{k=1}^{n}\sum_{j=1}^{n} p\left(g_j^{-1}g_k\right)\lambda_j^*\lambda_k \geqq 0.$$

Proof. Let $g=\sum_{k=1}^{n}\lambda_k U_g f$. A direct calculation shows,

$$0\leqq\langle g,g\rangle = \sum_{k=1}^{n}\sum_{j=1}^{n} p\left(g_j^{-1}g_k\right)\lambda_j^*\lambda_e,$$

which proves the theorem.

We note that, by Theorem 3.4, we have

$$e^{2\pi i x}\mathscr{A}(f)(x_2,x_1)^*$$

is positive definite. This enables us to translate general results about positive definite functions into assertions about ambiguity functions.

The following results are well-known about positive definite functions. Observe the relationship of these results to the corresponding results about $\mathscr{A}(f)$ coming from Theorem 2.4. We list them without proof.

1. $p(\mathbf{0})\geqq 0$,
2. $p(g^{-1})=p(g)^*$, $g\in N$,
3. $|p(g)|\leqq p(\mathbf{0})$, $g\in N$.

The unitary representation U of N is called continuous if, for each $f\in L^2(\mathbb{R})$, the mapping,

$$\mathbf{x}\rightarrow U_{\mathbf{x}}f,$$

is continuous from N into the Hilbert space $L^2(\mathbb{R})$. We give N the topology of the underlying Euclidean space. If U is a continuous unitary representation of N and $f\in L^2(\mathbb{R})$, then $p(\mathbf{x})=\langle U_{\mathbf{x}},f,f\rangle$ is continuous. Since D can be shown to be continuous, we can prove, by this approach, that $\mathscr{A}(f)$ is continuous.

A deeper result is that D is irreducible, in the sense of the following definition. The unitary representation U is irreducible if, for any closed subspace V of $L^2(\mathbb{R})$ such that

$$U_{\mathbf{x}} f \in V,$$

wherever $f \in V$, we have $V = L^2(\mathbb{R})$.

A proof that D is irreducible can be found in [Wyl]. The first implication of D being irreducible is that, for any $f \in L^2(\mathbb{R})$ which does not vanish on a set of positive measure, the span of the set of functions,

$$\{D_{\mathbf{x}} f : \mathbf{x} \in N\},$$

is dense in $L^2(\mathbb{R})$. As we see, in the proof of the following result, the uniqueness Theorem 2.5 of §2, the density of this span in $L^2(\mathbb{R})$, can be viewed as the key in the uniqueness theorem.

THEOREM 3.6 *If* $f, g \in L^2(\mathbb{R})$ *and* $\mathscr{A}(f) = \mathscr{A}(g)$ *then*

$$f = \lambda g,$$

for some constant λ, $|\lambda| = 1$.

Proof. From $\langle D_{\mathbf{x}} f, f\rangle = \langle D_{\mathbf{x}} g, g\rangle$, $\mathbf{x} \in N$ it easily follows that $\langle D_{\mathbf{x}} f, D_{\mathbf{y}} f\rangle = \langle D_{\mathbf{x}} g, D_{\mathbf{y}} g\rangle$ for all $\mathbf{x}, \mathbf{y} \in N$.

Note $\|f\|_2 = \|g\|$ implies $f = 0$ almost everywhere if and only if $g = 0$ almost everywhere. We will assume, therefore, that both f and g are nonzero on set of positive measure. From the irreducibility of D it follows that each of the set

$$A = \{D_{\mathbf{x}} f : \mathbf{x} \in N\}$$

and

$$B = \{D_{\mathbf{x}} g : \mathbf{x} \in N\}$$

spans a dense subspace of $L^2(\mathbb{R})$.

Define the mapping $U: A \to B$ by setting $U(D_{\mathbf{x}} f) = D_{\mathbf{x}} g$, $\mathbf{x} \in N$. We have to show U is well defined. Suppose $D_{\mathbf{u}} f = D_{\mathbf{v}} f$. Then

$$\langle D_{\mathbf{u}} f, D_{\mathbf{x}} f\rangle = \langle D_{\mathbf{u}} g, D_{\mathbf{x}} g\rangle,$$

for all $\mathbf{x}$, by the remarks above. This implies by the assumption $D_{\mathbf{u}} f = D_{\mathbf{v}} f$ that

$$\langle D_{\mathbf{u}} g, D_{\mathbf{x}} g\rangle = \langle D_{\mathbf{v}} g, D_{\mathbf{x}} g\rangle, R$$

for all $\mathbf{x} \in N$. Since B spans a dense subspace of $L^2(\mathbb{R})$, $D_{\mathbf{u}} g = D_{\mathbf{v}} g$. Thus, U is well defined. The condition $\langle D_{\mathbf{x}} f, D_{\mathbf{v}} f\rangle = \langle D_{\mathbf{x}} g, D_{\mathbf{v}} g\rangle$ immediately implies, along with the previous described property of B, that U extends to a unitary operator of $L^2(\mathbb{R})$.

It is trivial to see that

$$U D_{\mathbf{x}} U^{-1} = D_{\mathbf{x}}, \qquad \mathbf{x} \in N$$

and so $U = \lambda I$, $|\lambda| = 1$ which proves the theorem.

Another consequence of the condition of irreducibility will now be discussed. Consider two positive definite functions, p_1 and p_2, on N. We say that p_2 dominates p_1 if $p_2 - p_1$ is positive definite. A positive definite function p on N is called indecomposable if every positive definite function on N which is dominated by p is a scalar multiple of p.

The following theorem can be found in [A], in a slightly different setting, and will be asserted without proof.

THEOREM 3.7. *If U is an irreducible unitary representation of N and $f \in L^2(\mathbb{R})$ which does not vanish on a set of positive measure, then the corresponding positive definite function, p,*

$$p(\mathbf{x}) = \langle U_{\mathbf{x}} f, f \rangle,$$

is indecomposable.

An immediate implication is that $\langle D_{\mathbf{x}} f, f \rangle$ is indecomposable for every $f \in L^2(\mathbb{R})$ which does not vanish on a set of positive measure.

We will now reprove Theorem B using these ideas from unitary representation theory. For $f \in L^2(\mathbb{R})$, we write,

$$p_f(\mathbf{x}) = \langle D_{\mathbf{x}} f, f \rangle, \qquad \mathbf{x} \in N.$$

Suppose $f, g, h \in L^2(\mathbb{R})$ and

$$\mathscr{A}(h) = \mathscr{A}(f) + \mathscr{A}(g).$$

Then,

$$p_h = p_f + p_g.$$

Since p_h is indecomposable and p_h dominates both p_f and p_g, neglecting the trivial case, we can write,

$$p_f = c p_g,$$

where $c \neq 0$ is constant. From $p_f(\mathbf{0}) \geqq 0$ and $p_g(\mathbf{0}) \geqq 0$, we can infer $c > 0$. Let $g' = \sqrt{c}\, g$. Then,

$$p_f = p_{g'},$$

and

$$\mathscr{A}(f) = \mathscr{A}(g'),$$

from which it follows, by Theorem 3.6, that,

$$f = \lambda g' = \lambda \sqrt{c}\, g,$$

which is the conclusion of Theorem B.

4. Another unitary representation of N. A "piece" of another unitary representation of N will be defined which is unitarily equivalent to the representation D defined in the proceeding section. We will avoid as many technical details as possible. For further details see [A–T].

Let Γ be the subgroup of N generated by $(1,0,0)$ and $(0,1,0)$ and denote by H the space of all functions F on N which satisfies the following conditions:

1. $F(\gamma \mathbf{x}) = F(\mathbf{x})$, $\gamma \in \Gamma$, $\mathbf{x} \in N$,
2. $\|F\|_H^2 = \int_0^1 \int_0^1 \int_0^1 |F(\mathbf{x})|^2 d\mathbf{x} < \infty$,
3. $F(\mathbf{x}\mathbf{z}) = e^{2\pi i z} F(\mathbf{x})$, $\mathbf{x} \in N$, $\mathbf{z} \in Z$.

One can prove that H is a Hilbert space and that for $\mathbf{x} \in N$ and $F \in H$, the function

$$(\mathscr{D}(\mathbf{x})F)(\mathbf{y}) = F(\mathbf{yx}), \qquad \mathbf{y} \in N$$

is again in H. In fact, we can prove the following

THEOREM 4.1. *$\mathscr{D}$ is a unitary representation of N on H.*

We will tie together D and $\mathscr{D}$ by the Weil–Brezin mapping

$$W\colon L^2(\mathbb{R}) \to H$$

defined by setting

$$W(f)(\mathbf{x}) = e^{2\pi i(x + x_1x_2/2)} \sum_{m \in \mathbb{Z}} f(x_2 + m)e^{2\pi i m x_1}.$$

THEOREM 4.2. *W is an isometry from $L^2(\mathbb{R})$ onto H satisfying*

$$W^{-1}\mathscr{D}(\mathbf{x})W = D(\mathbf{x}), \qquad \mathbf{x} \in N.$$

Proof. Complete details of the proof can be found in [A–T]. We will prove the formula. Since

$$(\mathscr{D}(\mathbf{x})W(f))(\mathbf{y}) = W(f)(\mathbf{yx}) = W(f)(y_1 + x_1, y_2 + x_2, y + x + \tfrac{1}{2}(y_2x_1 - y_1x_2)),$$

it follows that

$$(\mathscr{D}(\mathbf{x})W(f))(\mathbf{y}) = e^{2\pi i(y + x + y_2x_1/2 - y_1x_2)} e^{2\pi i(y_1 + x_1)(y_2 + x_2)} \sum_{m \in \mathbb{Z}} f(y_2 + x_2 + m)e^{2\pi i m(y_1 + x_1)}.$$

Upon expanding the right-hand side we get

$$(\mathscr{D}(\mathbf{x})W(f))(\mathbf{y}) = W(D(\mathbf{x})F)(\mathbf{y}).$$

We say that W is an intertwining operator between $D(\mathscr{J}) = D \circ \mathscr{J}$and $\mathscr{D}$.

Consider $\mathbf{a} \in \Gamma$. Then, a_1, $a_2 \in \mathbb{Z}$ and $a \equiv \frac{1}{2}a_1a_2 \bmod \mathbb{Z}$. Let $F \in H$. Recall $F(\mathbf{ay}) = F(\mathbf{y})$. It is easy to see that

$$\mathbf{y} \cdot \mathbf{a} = \mathbf{a} \cdot \mathbf{y}[\mathbf{y}, \mathbf{a}]$$

where $[\mathbf{y}, \mathbf{a}] = \mathbf{y}^{-1}\mathbf{a}^{-1}\mathbf{ya} = (0, 0, y_2a_1 - y_1a_2)$.

THEOREM 4.3. *For $\mathbf{a} \in \Gamma$ and $F \in H$,*

$$\mathscr{D}(\mathbf{a})F(\mathbf{y}) = (WD(\mathbf{a})W^{-1}F)(\mathbf{y}) = e^{2\pi i(a_2y_1 - a_1y_2)}F(\mathbf{y}).$$

Proof. By definition,

$$(\mathscr{D}(\mathbf{a})F)(\mathbf{y}) = F(\mathbf{y} \cdot \mathbf{a}) = F(\mathbf{a} \cdot \mathbf{y} \cdot [\mathbf{y}, \mathbf{a}]) = e^{2\pi i(a_2y_1 - a_1y_2)}F(\mathbf{y}).$$

COROLLARY. *$e^{2\pi i(a_2y_1 - a_1y_2)}W(f)(\mathbf{y}) = W(g)(\mathbf{y})$ where*

$$g(y) = D(\mathbf{a})f(y).$$

Consider $H_0 = L^2(\mathbf{R}^2/\mathbf{Z}^2)$. For $F, G \in H$,

$$F(\mathbf{x})G^*(\mathbf{x}) = F_0(x_1, x_2)G_0^*(x_1, x_2)$$

where $F_0(x_1, x_2) = F(x_1, x_2, 0)$. Thus, F_0, $G_0 \in H_0$ and

$$\langle F, G \rangle_H = \langle F_0, G_0 \rangle_{H_0}.$$

For a_1, $a_2 \in \mathbb{Z}$ and x_1, $x_2 \in \mathbb{R}$, define

$$\chi_{a_1 a_2}(x_1, x_2) = e^{2\pi i(a_1 x_1 + a_2 x_2)}.$$

Clearly $\chi_{a_1,a_2} \in H_0$. We also have, for $F, G \in H$,

$$\left\langle \chi_{a_1,a_2} F, G \right\rangle_H = \left\langle \chi_{a_1,a_2}, F_0^* G_0 \right\rangle_{H_0}.$$

THEOREM 4.4. *The set of functions*

$$\left\{ \chi_{a_1,a_2} F \colon a_1, a_2 \in \mathbb{Z} \right\}$$

is an orthonormal basis of H if and only if

$$|F(\mathbf{x})| \equiv 1, \quad \textit{almost everywhere}.$$

Proof. Clearly, if $|F(\mathbf{x})| \equiv 1$, almost everywhere, then the set of functions is orthonormal since

$$\left\langle \chi_{a_1,a_2} F, \chi_{b_1,b_2} F \right\rangle_H = \left\langle \chi_{a_1,a_2}, \chi_{b_1,b_2} \right\rangle_{H_0} = 0.$$

Moreover, if $G \in H$ satisfies

$$\left\langle \chi_{a_1,a_2} F, G \right\rangle_H = \left\langle \chi_{a_1,a_2}, F^* G \right\rangle_{H_0} = 0,$$

for all a_1, $a_2 \in \mathbb{Z}$ then by the completeness of χ_{a_1,a_2}, a_1, $a_2 \in \mathbb{Z}$ in H_0, $F^*G \equiv 0$, almost everywhere. Since $|F| \equiv 1$, almost everywhere, $G \equiv 0$, almost everywhere which implies the set $\{\chi_{a_1,a_2} F \colon a_1, a_2 \in \mathbb{Z}\}$ is an orthonormal basis in H.

Conversely, if $\{\chi_{a_1,a_2} F \colon a_1, a_2 \in \mathbb{Z}\}$ is an orthonormal basis in H, then

$$\left\langle \chi_{a_1,a_2} F, F \right\rangle = \left\langle \chi_{a_1,a_2}, |F|^2 \right\rangle = 0,$$

whenever both a_1 and a_2 are not both 0. Thus, $|F|^2$ is constant almost everywhere. But

$$\langle F, F \rangle_H = 1$$

implies $|F| \equiv 1$, almost everywhere.

Theorem 4.2, the corollary to Theorem 4.3, and Theorem 4.4 immediately imply the next result.

THEOREM 4.5. *For* $f \in L^2(\mathbb{R})$, *satisfying*

$$|W(f)(\mathbf{y})| \equiv \left| \sum_{l \in \mathbb{Z}} f(y_2 + l) e^{2\pi i l y_1} \right| \equiv 1,$$

almost everywhere, the collection of functions

$$f_{a_1,a_2}(y) = e^{2\pi i a_2 y} f(y + a_1),$$

as a_1, a_2 *run over* $\mathbb{Z}$, *forms an orthonormal basis of* $L^2(\mathbb{R})$.

If $F(\mathbf{x})$ does not satisfy the condition $|F(\mathbf{x})| = 1$, almost everywhere, then the collection of functions

$$W = \{\chi_{a,b} \cdot F(\mathbf{x}) \colon a, b \in \mathbb{Z}\}$$

will not be an orthonormal basis but could be an L^2-basis of H, in the sense that, the linear span of W is dense in H and no proper subset of W has this property. It will be convenient to discuss the problem of when W determines an L^2-basis of H by considering the analogous problem on $\pi^2=\mathbb{R}^2/\mathbb{Z}^2$.

Consider $F(u,v)\in L^2(\pi^2)$ and set

$$W_0=\{\chi_{a,b}(u,v)\cdot F(u,v)\colon a,b\in\mathbb{Z}\}.$$

Let

$$g(t)=m\{(u,v)\in\pi^2\colon |F(u,v)|\leqq t\},$$

where m denotes Lebesgue measure on π^2.

Observe that $g(t)$ is the distribution function of $|F(u,v)|$, and hence determines a probability measure on $\mathbb{R}$.

THEOREM 4.6. *W_0 is a minimal basis of $L^2(\pi^2)$ if and only if*

1. $g(0)=0$,

2. $\int_{0^+}^{\infty}(1/t^2)\,dg(t)<\infty$. (*This includes* 1.)

Proof. Take $G\in L^2(\pi^2)$ satisfying

$$\langle G,W_0\rangle=0.$$

Then

$$\langle\chi_{a,b}F,G\rangle=\langle\chi_{a,b},F^*G\rangle=0,\qquad a,b\in\mathbb{Z},$$

which by the completeness of the set $\{\chi_{a,b}\colon a,b\in\mathbb{Z}\}$ in $L^2(\pi^2)$ implies $F^*G=0$ almost everywhere. Thus, $g(0)=0$ implies $G=0$ almost everywhere. We have proved that $g(0)=0$ implies W_0 spans a dense subspace of $L^2(\pi^2)$. The converse is trivial, for if $g(0)\neq 0$, let G be the function which is identically one where F vanishes and zero otherwise. Then G is orthogonal to W_0 but is not the zero function in $L^2(\pi^2)$.

We will now show the equivalence of minimality to Theorem 4.6, statement 2. The argument includes the above discussion.

Suppose $a_0b_0\in\mathbb{Z}$ and that the closure V in $L^2(\pi^2)$ of the set $F\cdot(\mathbb{C}\cdot\chi_{a_0,b_0})^{\perp}$ is proper in $L^2(\pi^2)$. As is standard $(\mathbb{C}\cdot\chi_{a_0,b_0})^{\perp}$ denotes the orthogonal complement of $\mathbb{C}\cdot\chi_{a_0,b_0}$ in $L^2(\pi^2)$. Choose G_1 orthogonal to V. Then, for every function G_2 orthogonal to χ_{a_0,b_0}, we have

$$\langle G_1,F\cdot G_2\rangle=\langle F^*G_1,G_2\rangle=0.$$

Thus, $F^*G_1=\lambda\cdot\chi_{a_0,b_0}$ for some constant $\lambda\neq 0$. This implies $F^{*-1}=\lambda^{-1}\chi_{a_0,b_0}^{-1}\cdot G_1\in L^2(\pi^2)$ and hence $F^{-1}\in L^2(\pi^2)$. The converse is obvious. Thus, we have proved that W_0 is a minimum L^2-basis of $L^2(\pi^2)$ if and only if $F^{-1}\in L^2(\pi^2)$.

We will now show that $F^{-1}\in L^2(\pi^2)$ and only if Theorem 4.6, statement 2 holds. We simply observe that

$$\int_{0^+}^{\infty}\frac{1}{t^2}dg(t)=\int_{\pi^2}|F(u,v)|^{-1}du\,dv,$$

and hence $F^{-1}\in L^2(\pi^2)$ if and only if

$$\int_{0^+}^{\infty}\frac{1}{t^2}dg(t)<\infty.$$

5. Examples of ambiguity functions. In this section, we will build ambiguity functions which include the standard ambiguity functions dealt with in radar theory along with an example coming from Heisenberg group theory. We begin with a few general remarks.

An orthonormal basis of $L^2(\mathbb{R})$ is a set of functions f_n, $n\in\mathbb{Z}$, in $L^2(\mathbb{R})$ such that

$$\langle f_n,f_m\rangle=\begin{cases}1, & n=m,\\ 0, & n\neq m,\end{cases}$$

and the closure of the linear span of these functions in $L^2(\mathbb{R})$. More generally, a countable subset M of $L^2(\mathbb{R})$ will be called an L^2-basis of $L^2(\mathbb{R})$ if the closure of its linear span equals $L^2(\mathbb{R})$ and no proper subset of M has this property.

The L^2-basis of $L^2(\mathbb{R})$ we construct will be of the following form. A fixed function $f\in L^2(\mathbb{R})$ will be taken and we define

$$f_{a,b}(t)=(M(b)S(a)f)(t)=e^{2\pi ibt}f(t+a),\qquad a,b\in\mathbb{Z}.$$

We will consider examples where the set of functions

$$\{f_{a,b}\colon a,b\in\mathbb{Z}\}$$

is an L^2-basis and use Theorem 4 of §5 to show we have an orthonormal basis.

In the original manuscript, the authors believed that the Gaussian $g(t)=e^{-\pi t^2}$ leads to a minimal basis. As pointed out by the referee, this is not the case. A proof can be seen by showing that $G=|W(g)|$ does not satisfy condition 2 of Theorem 4.6.

The two L^2-bases we consider will be orthonormal. Consider the rectangular function

$$r(t)=\begin{cases}1, & |t|<\frac{1}{2},\\ 0, & |t|>\frac{1}{2}.\end{cases}$$

It is easy to see that $r(t)$ satisfies the hypothesis of Theorem 4.5. Thus, the collection of functions

$$\mathscr{R}=\{r_{a,b}\colon a,b\in\mathbb{Z}\}$$

is an orthonormal basis of $L^2(\mathbb{R})$, called the rectangular basis of $L^2(\mathbb{R})$.

The rectangle function r is a standard signal processing function. The next basis we consider is more exotic and comes from Heisenberg group theory, especially Theorem 4.5 and [A–T, pp. 81–82]. Applying the Weil–Brezin mapping W to the Gaussian g gives the Heisenberg group theory analogue of the classical theta function. Explicitly,

$$W(g)(\mathbf{x})=e^{2\pi i(x+x_1x_2/2)}\sum_{l\in\mathbb{Z}}e^{-\pi(x_2+l)^2}e^{2\pi ilx_1}$$

where $g(t)=e^{-\pi t^2}$. Consider

$$F(\mathbf{x})=\frac{W(g)(\mathbf{x})}{|W(g)(\mathbf{x})|}$$

and observe $F(\mathbf{x})$ satisfies the conditions needed for Theorem 4.5 to assert that the set of functions

$$\{\chi_{a_1,a_2}\cdot F\colon a_1,a_2\in\mathbb{Z}\}$$

is an orthonormal basis of H. By Theorem 4.5, if

$$t(y)=W^{-1}(F)(y)$$

then the set of functions

$$T=\{t_{a,b}\colon a,b\in\mathbb{Z}\}$$

is an orthonormal basis. The only facts we will need are given in the following lemma.

LEMMA 5.1. *Let* $\theta(z)=\sum_{l\in\mathbb{Z}}e^{-\pi l^2}e^{2\pi i l z}$, $z=x+iy$, *be the classical theta function and set*

$$t(y)=\int_{x=0}^{1}\frac{\theta(z)}{|\theta(z)|}\,dx,\qquad y\in\mathbb{R}.$$

Then the set of functions

$$T=\{t_{a,b}\colon a,b\in\mathbb{Z}\}$$

is an orthonormal basis of $L^2(\mathbb{R})$.

We call T the theta basis of $L^2(\mathbb{R})$.

We will now state, without proof, how the Fourier transform acts on the three bases considered. First,

$$\hat{t}=t,\qquad \hat{r}(v)=\frac{\sin\pi V}{\pi V}.$$

Since, by Theorem 2.6

$$\hat{f}_{a,b}=(\hat{f})_{+b,-a}$$

we have that T is invariant under the action of the Fourier transform and $\mathscr{R}$ maps onto sinusoidals.

We will now relate the cross-ambiguity function $\mathscr{A}(f_{a,b},f_{c,d})$ to $\mathscr{A}(f)$.

THEOREM 5.1. *Let* $f\in L^2(\mathbb{R})$ *and* $f_{a,b}=M(b)S(a)\cdot f$. *Then*

$$\mathscr{A}(f_{a,b},f_{c,d})(u,v)=K\cdot\mathscr{A}(f)(u+c-a,v+d-b)$$

where $K=(-1)^{(a+c)(b+c)}e^{-\pi[(b+d)u-(a+c)v]}$.

Proof. Consider

$$\mathscr{B}(f_{a,b},f_{c,d})(u,v)=\langle f_{a,b},M(v)(u)f_{c,d}\rangle$$

which we can write

$$\langle M(b)S(a)f,M(v)S(u)M(d)S(c)f\rangle=\langle f,S(-a)M(-b)M(v)S(u)M(d)S(c)f\rangle.$$

Using Lemma 3.2, this becomes

$$e^{-2\pi i u c}e^{2\pi i a(v+c-b)}\mathscr{B}(f)(u+c-a,v+d-b).$$

The theorem follows once we observe $\mathscr{B}(f,g)=e^{\pi i u v}\mathscr{A}(f,g)$.

The ambiguity function of r is easy to compute and for convenience we give the answer in the next lemma. The ambiguity function of t does not have a simple form.

LEMMA 5.2.

$$\mathscr{A}(r)(u,v)=\begin{cases}-\sin((u-1)\pi v)/\pi v, & 0<u<1,\\ \sin((u-1)\pi v)/\pi v, & -1\leqq u<0,\end{cases}$$

and vanishes elsewhere.

The corresponding cross-ambiguity functions can be determined by Theorem 5.1. Suppose $f \in L^2(\mathbb{R})$ and

$$F=\{f_{a,b}\colon a,b\in\mathbb{Z}\}$$

is an L^2-basis of $L^2(\mathbb{R})$. Then, any $h\in L^2(\mathbb{R})$ can be written as

$$h=\sum_{a,\,b\in\mathbb{Z}}\alpha(a,b)f_{a,b}$$

where

$$\langle h,h\rangle=\sum\alpha(a,b)\alpha^*(c,d)\langle f_{a,b},f_{c,d}\rangle<\infty.$$

Of course, if F is an orthonormal basis the above condition reduces to

$$\sum_{a,\,b\in\mathbb{Z}}|\alpha(a,b)|^2<\infty.$$

By Theorem 5.1 and Lemma 5.2, we immediately have the following result.

THEOREM C. *Let $f\in L^2(\mathbb{R})$ generate an L^2 basis. Let Φ_f denote the set of functions α: $\mathbb{Z}\times\mathbb{Z}\to\mathbb{C}$ such that*

$$\sum\alpha(a,b)\alpha^*(c,)\langle f_{a,b},f_{c,d}\rangle<\infty.$$

The set of functions

$$F(\alpha)=\sum_{a,\,b,\,c,\,d}\alpha(a,b)\alpha^*(c,d)K(a,b,c,d)A(f)(u+c-a,v+d-b)$$

where

$$K(a,b,c,d)=(-1)^{(a+c)(b+d)}e^{-\pi i[(b+d)u-(a+c)v]}$$

$\alpha\in\Phi_f$, is the set of ambiguity functions.

Theorem D follows easily from Theorem C and the discussion in this section.

We close with an interesting example.

Example. Let $p(t)$ be a periodic function of period 1 and consider

$$f(t)=p(t)e^{-\pi t^2}.$$

Write $p(t)=\sum_{a\in\mathbb{Z}}\alpha(a)e^{2\pi iat}$.

Then, if

$$\mathscr{B}(p)(u,l)=e^{\pi ilu}\int_0^1 p\left(t-\frac{u}{2}\right)p^*\left(t+\frac{u}{2}\right)e^{2\pi ilt}\,dt$$

we have

$$\mathscr{B}(f)(u,v)=\mathscr{B}(e^{-\pi t^2})(u,v)\sum_{l\in\mathbb{Z}}\mathscr{B}(p)(u,l)e^{-\pi l^2/2}e^{2\pi il(u+iv)/2}.$$

Thus we can interpret the ambiguity function of f as having a continuous part $\mathscr{B}(e^{-\pi t^2})$ and a "discrete" part which is a theta-like function with coefficients given by the periodic version of the ambiguity function of $p(t)$.

REFERENCES

[A] L. Auslander, *Unitary representations of locally compact groups*, Yale Univ. Lecture Notes, New Haven, CT, 1961.

[A–B] L. Auslander and T. Brezin, *Fiber bundle structures and harmonic analysis of compact Heisenberg manifolds*, Conference on Harmonic Analysis, Lecture Notes in Mathematics 266, Springer-Verlag, New York, 1971.

[A–T] L. Auslander and R. Tolimieri, *Abelian harmonic analysis, theta functions and function algebras on a nilmanifold*, Lecture Notes in Mathematics 436, Springer-Verlag, New York, 1975.

[B] R. E. Blahut, Lecture notes on radar theory.

[H–S] H–M Helsaker and W. Schempp, *Radar detections, quantum mechanics and nilpotent harmonic analysis*, preprint.

[K] Y. Katznelson, *Harmonic Analysis*, John Wiley, New York, 1968.

[S] S. Sussman, *Least-square synthesis of radar ambiguity functions*, IRE Trans. Info. Theory, (April 1962), pp. 246–254.

[W1] C. H. Wilcox, *The synthesis problem for radar ambiguity functions*, MRC Technical Report 157, Mathematics Research Center, U. S. Army, Univ. Wisconsin, Madison, 1960.

[Wg] E. P. Wigner, *On the quantum correction for thermodynamic equilibrium*, Phys. Rev., (1932), pp. 749–759.

[W] P. M. Woodward, *Probability and Information theory, with Applications to Radar*, Pergamon Press, New York, 1953.

[Wyl] H. Weyl, *Quantenmechanik und Gruppentheorie*, Z. Phys., 46 (1928), p. 1–46.

[V] J. Ville, *Théorie et applications de la notion de signal analytique*, Cables et Transmissions, 2 (1948), pp. 1–74.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

015

MATRIX ELEMENTS OF IRREDUCIBLE REPRESENTATIONS OF $SU(2)\times SU(2)$ AND VECTOR-VALUED ORTHOGONAL POLYNOMIALS*

TOM H. KOORNWINDER[†]

Abstract. The matrix elements of irreducible representations of $SU(2)\times SU(2)$ in a diag$(SU(2)\times SU(2))$-basis are expressed in terms of vector-valued orthogonal polynomials, which generalize the Jacobi polynomials.

0. Introduction. It is well known (cf. Vilenkin [11, Chap. 3]) that the matrix elements of the irreducible representations of $SU(2)$ in $S(U(1)\times U(1))$-basis can be expressed in terms of Jacobi polynomials, such that the orthogonality relations for these polynomials are equivalent to Schur's orthogonality relations for the matrix elements. More generally, let G be a compact Lie group with closed subgroup K such that each irreducible representation of G, restricted to K, is multiplicity free. Consider the matrix elements of the irreducible representations of G in a K-basis. Is it possible to express them in terms of some kind of orthogonal polynomials? For the case $G=SU(2)\times SU(2)$, $K=$ diagonal in G, this paper will give a positive answer. (Note that this case is a covering of the pair $(G,K)=(SO(4),SO(3))$.) The resulting polynomials are vector-valued and orthogonal on $[-1,1]$ with respect to a positive definite matrix-valued weight function. It would be of interest to generalize these results to the cases $(G,K)=(SO(n),SO(n-1))$ or $(U(n),U(n-1))$.

The topic of this paper originated from work on the global approach to the representation theory of a noncompact semisimple Lie group G (cf. [7]) for $SL(2,\mathbb{R})$, Kosters [8] for $SL(2,\mathbb{C})$). In this approach one needs some knowledge of the matrix elements of the principal series representations of G in a K-basis (K maximal compact subgroup of G). These matrix elements have integral representations in terms of the matrix elements of irreducible representations of K (cf. (4.1) in the case $G=SL(2,\mathbb{C})$). Manipulation of these integral representations will be simplified if one can express the matrix elements for K in terms of orthogonal polynomials. Thus the results of the present paper will be useful for the analysis on $SO_0(4,1)$.

It is the author's feeling that the highly nontrivial example of vector-valued orthogonal polynomials presented here is interesting for its own sake. Hopefully this paper will also be useful for physicists, who have already studied the matrix elements for $SO(4)$ for a long time (cf. for instance Freedman and Wang [3], Smorodinskiĭ and Shepelev [10], Basu and Srinvasan [1]). Many authors start with the matrix elements of the principal series representations of $SO_0(3,1)$ (cf. [1], [10]) and then obtain the matrix elements for the compact case by analytic continuation. In the present paper, with its emphasis on orthogonal polynomials, it seemed more natural to start with the compact case, but in the final §4 the noncompact analogue is briefly discussed.

Sections 1 and 2 are of a preliminary nature. In §1 matrix elements for $SU(2)$ are reviewed, both as a tool needed later and as a motivating example. In §2 Schur's orthogonality relations for matrix elements for $SU(2)\times SU(2)$ are expressed as an

*Received by the editors June 7, 1983.

[†] Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, The Netherlands.

orthogonality for vector-valued functions on $[0,\pi]$ and good candidates are selected for the expected vector-valued orthogonal polynomials. In §3 these polynomials are really obtained together with an integral representation and a power series expansion. There are two further matters of particular interest in §3: First, a trick to deform the integral of an analytic function over $SU(2)$ into the complexification $SL(2,\mathbb{C})$ by multiplication on the right of the integration variable with a particular element of $SL(2,\mathbb{C})$ (cf. the transition (3.3)→(3.6)) and, second, an unexpected symmetry (3.11) for the vector-valued polynomials.

1. The matrix elements for $SU(2)$. Let $l\in\frac{1}{2}\mathbb{Z}_+:=\{0,\frac{1}{2},1,\frac{3}{2},\cdots\}$. Let H_l be the space of homogeneous polynomials of degree $2l$ in two complex variables, made into a Hilbert space by the choice of orthonormal basis $\{\psi_n^l|n=-l,-l+1,\cdots,l\}$

$$\psi_n^l(x,y):=\binom{2l}{l-n}^{1/2}x^{l-n}y^{l+n}. \tag{1.1}$$

Define a representation T^l of $GL(2,\mathbb{C})$ on H_l by

$$\left(T^l\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}f\right)(x,y):=f(\alpha x+\gamma y,\beta x+\delta y). \tag{1.2}$$

The T^l's form a complete system of representatives for $(SU(2))^\wedge$ (cf. Vilenkin [11, Chap. 3]).

Write $T^l(g)(g\in GL(2,\mathbb{C}))$ as a matrix $(t_{mn}^l(g))$ with respect to the basis functions ψ_n^l

$$T^l(g)\psi_n^l=\sum_{m=-l}^{l}t_{mn}^l(g)\psi_m^l,\qquad g\in GL(2,\mathbb{C}). \tag{1.3}$$

If g is a diagonal matrix then so is $(t_{mn}^l(g))$. It follows from (1.1), (1.2), (1.3) that

$$\binom{2l}{l-n}^{1/2}(\alpha x+\gamma y)^{l-n}(\beta x+\delta y)^{l+n}=\sum_{m=l}^{l}t_{mn}^l\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}\binom{2l}{l-m}^{1/2}x^{l-m}y^{l+m}. \tag{1.4}$$

Expansion of the left-hand side of (1.4) yields

$$t_{mn}^l\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}=((l-m)!(l+m)!(l-n)!(l+n)!)^{1/2} \tag{1.5}$$
$$\cdot\sum_{r=0\vee(-n-m)}^{(l-n)\wedge(l-m)}\frac{\alpha^l\beta^{l-m-r}\gamma^{l-n-r}\delta^{m+n+r}}{r!(l-m-r)!(l-n-r)!(m+n+r)!}.$$

This implies the symmetries

$$\beta^m\gamma^n t_{mn}^l\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}=\beta^n\gamma^m t_{nm}^l\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}, \tag{1.6}$$

$$t_{mn}^l\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}=t_{nm}^l\begin{pmatrix}\alpha & \gamma\\ \beta & \delta\end{pmatrix}. \tag{1.7}$$

From (1.4) and (1.7) we obtain the integral representation

$$(1.8)\qquad t^l_{mn}\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}=\left(\frac{(l-n)!(l+n)!}{(l-m)!(l+m)!}\right)^{1/2}$$

$$\cdot\frac{1}{2\pi}\int_0^{2\pi}\left(\alpha e^{i\phi}+\beta e^{-i\phi}\right)^{l-m}\left(\gamma e^{i\phi}+\delta e^{-i\phi}\right)^{l+m}e^{2in\phi}\,d\phi.$$

The following symmetry is apparent from (1.8).

$$(1.9)\qquad t^l_{mn}\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}=t^l_{-m,-n}\begin{pmatrix}\delta & \gamma\\ \beta & \alpha\end{pmatrix}.$$

Now specialize to $SU(2)$. We will use the notation

$$(1.10)\qquad k(\alpha,\beta):=\begin{pmatrix}\alpha & \beta\\ -\bar{\beta} & \bar{\alpha}\end{pmatrix}\quad\text{where } |\alpha|^2+|\beta|^2=1,$$

$$(1.11)\qquad b_\theta:=k\left(\cos\frac{\theta}{2},\,\sin\frac{\theta}{2}\right),$$

$$(1.12)\qquad m_\phi:=k(e^{i\phi/2},0).$$

Note that

$$(1.13)\qquad t^l_{mn}(m_\phi)=e^{-in\phi}\delta_{mn}.$$

By the Cartan decomposition each element of $SU(2)$ can be written as $m_\phi b_\theta m_\psi$ and the corresponding integration formula reads

$$(1.14)\qquad \int_{SU(2)}f(g)\,dg=\frac{1}{2}\int_0^{\pi}\int_0^{4\pi}\int_0^{4\pi}f(m_\phi b_\theta m_\psi)\sin\theta\,d\theta\,\frac{d\phi}{4\pi}\frac{d\psi}{4\pi},\qquad f\in C(SU(2)).$$

By Schur's orthogonality relations, (1.13) and (1.14) we obtain

$$\int_0^{\pi}t^l_{mn}(b_\theta)t^{l'}_{m,n}(b_\theta)\sin\theta\,d\theta=0,\qquad l\neq l'.$$

Suppose that $m+n\geqq 0$, $m-n\geqq 0$. Then the "lowest" element of the orthogonal system $\{t^l_{mn}\,|\,l=m,m+1,\cdots\}$ is t^m_{mn}. From (1.5) we obtain

$$(1.15)\qquad t^m_{mn}(b_\theta)=(-1)^{m-n}\binom{2m}{m-n}^{1/2}\left(\sin\frac{\theta}{2}\right)^{m-n}\left(\cos\frac{\theta}{2}\right)^{m+n}.$$

Hence, if $l\neq l'$, then

$$\int_0^{\pi}\frac{t^l_{mn}(b_\theta)}{t^m_{mn}(b_\theta)}\,\frac{t^{l'}_{mn}(b_\theta)}{t^m_{mn}(b_\theta)}\left(\sin\frac{\theta}{2}\right)^{2m-2n+1}\left(\cos\frac{\theta}{2}\right)^{2m+2n+1}d\theta=0.$$

By (1.5) $t^l_{mn}(b_\theta)/t^m_{mn}(b_\theta)$ is a polynomial in $\cos\theta$ of degree $\leqq l-m$. It follows that

$$t^l_{mn}(b_\theta)/t^m_{mn}(b_\theta)=\text{const.}\;P^{(m-n,\,m+n)}_{l-m}(\cos\theta),$$

where the *Jacobi polynomial* $P^{(m-n,m+n)}_{l-m}$ is an orthogonal polynomial of degree $l-m$ with respect to the weight function $(1-x)^{m-n}(1+x)^{m+n}$ on the interval $(-1,1)$. Of course, this result has been derived in many other ways (cf. Vilenkin [11, Chap. 3]).

2. The matrix elements for $SU(2)\times SU(2)$. Let $K := SU(2)$, $G := K\times K$, $K^* := \operatorname{diag}(K\times K)$, $A := \{a_\theta := (m_\theta, m_{-\theta})\}$ (m_θ is defined by (1.12)). Then $G = K^*AK^*$ is a Cartan decomposition. The corresponding integral formula is

$$(2.1)\qquad \int_G f(g)\,dg = \frac{1}{2\pi}\int_0^\pi \int_{K^*}\int_{K^*} f(k_1 a_\theta k_2)\sin^2\theta\, d\theta\, dk_1\, dk_2, \qquad f\in C(G),$$

which is a special case of Helgason [5, Prop. X.1.19].

A complete system of representatives for $\hat{G}$ is given by the representations $T^{l_1,l_2}(l_1, l_2 \in \frac{1}{2}\mathbf{Z}_+)$:

$$(2.2)\qquad T^{l_1,l_2}(k_1,k_2) := T^{l_1}(k_1)\otimes T^{l_2}(k_2), \qquad k_1,k_2\in K.$$

The representation space $H_{l_1}\otimes H_{l_2}$ of T^{l_1,l_2} can be identified with the space of polynomials in four complex variables x, y, u, v, homogeneous of degree $2l_1$ in x, y and homogeneous of degree $2l_2$ in u, v. An orthonormal basis of $H_{l_1}\otimes H_{l_2}$ is given by the polynomials

$$(x,y,u,v)\mapsto \psi^{l_1}_{j_1}(x,y)\,\psi^{l_2}_{j_2}(u,v).$$

PROPOSITION 2.1 (cf. [6, Thms. 3.1, 3.2]). *The functions* $\phi^{l_1,l_2}_{l,j}(|l_1 - l_2| \leqq l \leqq l_1 + l_2$, $|j| \leqq l)$ *defined by*

$$(2.3)\qquad \phi^{l_1,l_2}_{l,j}(x,y,u,v) := (-1)^{l_1+l_2-l}\left(\frac{(2l+1)(2l_1)!(2l_2)!}{(l_1+l_2-l)!(l_1+l_2+l+1)!}\right)^{1/2} \cdot (xv-yu)^{l_1+l_2-l}\, t^l_{l_2-l_1,j}\begin{pmatrix} x & y\\ u & v\end{pmatrix}$$

form an orthonormal basis of $H_{l_1}\otimes H_{l_2}$ *such that*

$$(2.4)\qquad T^{l_1,l_2}(k,k)\phi^{l_1,l_2}_{l,j'} = \sum_{j=-l}^{l} t^l_{j,j'}(k)\phi^{l_1,l_2}_{l,j}, \qquad k\in K.$$

Define the matrix elements of T^{l_1,l_2} with respect to this K^*-basis $\{\phi^{l_1,l_2}_{l,j}\}$ by

$$(2.5)\qquad T^{l_1,l_2}(g)\phi^{l_1,l_2}_{l',j'} = \sum_{l=|l_1-l_2|}^{l_1+l_2}\ \sum_{j=-l}^{l} t^{l_1,l_2}_{l,j;l',j'}(g)\phi^{l_1,l_2}_{l,j}, \qquad g\in G.$$

Since the elements of A commute with the elements (m_θ, m_θ) in K^* and since

$$T^{l_1,l_2}(m_\theta,m_\theta)\phi^{l_1,l_2}_{l,j} = e^{-ij\theta}\phi^{l_1,l_2}_{l,j}$$

by (2.4) and (1.12), we conclude that

$$(2.6)\qquad t^{l_1,l_2}_{l,j;l',j'}(a_\theta) = 0 \quad \text{if } j\neq j'.$$

By (2.4), (2.6) and the decomposition $G = K^*AK^*$ the matrix elements $t^{l_1,l_2}_{l,j;l',j'}$ will be known if we know the functions $t^{l_1,l_2}_{l,j;l',j}|_A$.

PROPOSITION 2.2 (cf. [3]). *There are the orthogonality relations*

(2.7)

$$\frac{1}{2\pi}\sum_{j=-(l\wedge m)}^{l\wedge m}\int_0^\pi t^{l_1,l_2}_{l,j;m,j}(a_\theta)\,t^{l_1',l_2'}_{l,j;m,j}(a_\theta)\,\sin^2\theta\, d\theta = \frac{(2l+1)(2m+1)}{(2l_1+1)(2l_2+1)}\delta_{l_1,l_1'}\delta_{l_2,l_2'}.$$

Proof. It follows from Schur's orthogonality relations, (2.1), (2.4) and (2.6) that

$$\frac{\delta_{l_1,l_1'}\delta_{l_2,l_2'}}{(2l_1+1)(2l_2+1)}$$

$$=\frac{1}{2\pi}\int_0^\pi\int_{K^*}\int_{K^*} t^{l_1,l_2}_{l,p;m,p}(k_1a_\theta k_2)\overline{t^{l_1',l_2'}_{l,p;m,p}(k_1a_\theta k_2)}\,\sin^2\theta\,d\theta\,dk_1\,dk_2$$

$$=\sum_{j=-(l\wedge m)}^{l\wedge m}\ \sum_{j'=-(l\wedge m)}^{l\wedge m}\frac{1}{2\pi}\int_0^\pi\int_K\int_K t^l_{p,j}(k_1)t^{l_1,l_2}_{l,j;m,j}(a_\theta)t^m_{j,p}(k_2)$$

$$\cdot\overline{t^l_{p,j'}(k_1)t^{l_1',l_2'}_{l,j';m,j'}(a_\theta)t^m_{j',p}(k_2)}\,\sin^2\theta\,d\theta\,dk_1\,dk_2$$

$$=\frac{1}{(2l+1)(2m+1)}\sum_{j=-(l\wedge m)}^{l\wedge m}\frac{1}{2\pi}\int_0^\pi t^{l_1,l_2}_{l,j;m,j}(a_\theta)\overline{t^{l_1',l_2'}_{l,j;m,j}(a_\theta)}\,\sin^2\theta\,d\theta.$$

It follows from (2.5) and (2.3) that $t^{l_1,l_2}_{l,j;m,j}(a_\theta)$ is real. □

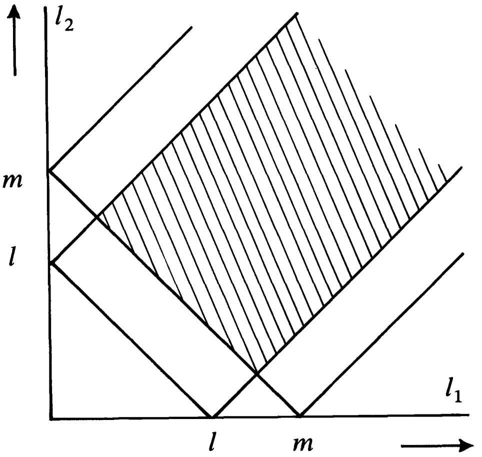


FIG. 1.

From now on fix l and m ($l,m\in\frac{1}{2}\mathbb{Z}_+, l-m\in\mathbb{Z}$) such that $l\leq m$. (Because of unitariness of T^{l_1,l_2} this last condition is not an essential restriction). Then the indices l_1, l_2 in $t^{l_1,l_2}_{l,j;m,j}(a_\theta)$ can assume all values in $\frac{1}{2}\mathbb{Z}_+$ such that (cf. Fig. 1)

$$(2.8)\qquad l_1+l_2\geq m,\quad |l_1-l_2|\leq l,\quad l_1+l_2-l\in\mathbb{Z}$$

and $j\in\{-l,-l+1,\cdots,l\}$. Thus, (2.7) can be viewed as the orthogonality relations for the vector-valued functions

$$(2.9)\qquad \theta\mapsto\Big(\big(t^{l_1,l_2}_{l,-l;m,-l}\big)(a_\theta),\,t^{l_1,l_2}_{l,-l+1;m,-l+1}(a_\theta),\cdots,t^{l_1,l_2}_{l,l;m,l}(a_\theta)\Big),$$

where (l_1,l_2) run through all values satisfying (2.8). Like at the end of §1 we pick the "lowest" elements of this orthogonal family. Candidates for these elements are all functions of the form (2.9) with $l_1+l_2=m$. Suppose that we can prove that for all θ in $(0,\pi)$ the matrix

$$\left(t_{l,j;m,j}^{(m+p)/2,(m-p)/2}(a_\theta)\right)_{j,p=-l,-l+1,\cdots,l} \tag{2.10}$$

is nonsingular. Then, for $n=0,1,2,\cdots$ and $k=-l,-l+1,\cdots,l$ we can define the real vector-valued functions

$$x\mapsto P_{n,k}^{l,m}(x)=\left(P_{n,k,-l}^{l,m}(x),P_{n,k,-l+1}^{l,m}(x),\cdots,P_{n,k,l}^{l,m}(x)\right) \tag{2.11}$$

on $(-1,1)$ by

$$t_{l,j;m,j}^{l_1,l_2}(a_\theta)=\sum_{p=-l}^{l}t_{l,j;m,j}^{(m+p)/2,(m-p)/2}(a_\theta)P_{l_1+l_2-m,l_2-l_1,p}^{l,m}(\cos\theta). \tag{2.12}$$

Also define

$$W_{p,q}^{l,m}(\cos\theta):=\sin\theta\sum_{j=-l}^{l}t_{l,j;m,j}^{(m+p)/2,(m-p)/2}(a_\theta)t_{l,j;m,j}^{(m+q)/2,(m-q)/2}(a_\theta). \tag{2.13}$$

Then

$$W^{l,m}(\cos\theta):=\left(W_{p,q}^{l,m}(\cos\theta)\right)_{p,q=-l,\cdots,l} \tag{2.14}$$

is a positive definite real symmetric matrix for all θ in $(0,\pi)$ and it follows from (2.7), (2.12), (2.13) that the vector-valued functions $P_{n,k}^{l,m}$ satisfy the orthogonality relations

$$\frac{1}{2\pi}\sum_{p,q=-l}^{l}\int_{-1}^{1}P_{n,k,p}^{l,m}(x)P_{n',k',q}^{l,m}(x)W_{p,q}^{l,m}(x)\,dx=\frac{(2l+1)(2m+1)}{(n+m+1)^2-k^2}\delta_{n,n'}\delta_{k,k'}. \tag{2.15}$$

In this paper we will show that the matrix (2.10) is indeed nonsingular for θ in $(0,\pi)$ and that $P_{n,k,p}^{l,m}$ is a polynomial of degree $n-|p+k|$. Hence the orthogonality relations (2.15) will characterize the vector-valued functions $P_{n,k}^{l,m}$ up to constant factors.

3. The vector-valued orthogonal polynomials. First we derive an integral representation for the canonical matrix elements. Consider (2.5) with $g=a_\theta$ and evaluate both sides for $(x,y,u,v)=(\alpha,\beta,-\bar{\beta},\bar{\alpha})$, where $|\alpha|^2+|\beta|^2=1$. In view of (2.3) and (2.6) we obtain

$$\begin{aligned}
&(-1)^{l_1+l_2-m}\left(\frac{(2m+1)(2l_1)!(2l_2)!}{(l_1+l_2-m)!(l_1+l_2+m+1)!}\right)^{1/2}\\
&\quad\cdot t_{l_2-l_1,j}^{m}\begin{pmatrix} e^{i\theta/2}\alpha & e^{-i\theta/2}\beta \\ -e^{-i\theta/2}\bar{\beta} & e^{i\theta/2}\bar{\alpha}\end{pmatrix}\left(e^{i\theta}|\alpha|^2+e^{-i\theta}|\beta|^2\right)^{l_1+l_2-m}\\
&\qquad=\sum_{l=|l_1-l_2|}^{l_1+l_2}(-1)^{l_1+l_2-l}\left(\frac{(2l+1)(2l_1)!(2l_2)!}{(l_1+l_2-l)!(l_1+l_2+l+1)!}\right)^{1/2}\\
&\qquad\quad\cdot t_{l,j;m,j}^{l_1,l_2}(a_\theta)t_{l_2-l_1,j}^{l}(k(\alpha,\beta)).
\end{aligned}$$

Hence, by Schur's orthogonality relations:

$$(3.1)\quad t^{l_1,l_2}_{l,j;m,j}(a_\theta)=(-1)^{l-m}\left(\frac{(2l+1)(2m+1)(l_1+l_2-l)!(l_1+l_2+l+1)!}{(l_1+l_2-m)!(l_1+l_2+m+1)!}\right)^{1/2}$$

$$\cdot\int_K\left(e^{i\theta}|\alpha|^2+e^{-i\theta}|\beta|^2\right)^{l_1+l_2-m}$$

$$\cdot t^m_{l_2-l_1,j}\begin{pmatrix} e^{i\theta/2}\alpha & e^{-i\theta/2}\beta \\ -e^{-i\theta/2}\bar{\beta} & e^{i\theta/2}\bar{\alpha}\end{pmatrix} t^l_{l_2-l_1,j}\left(k(\bar{\alpha},\bar{\beta})\right)dk(\alpha,\beta).$$

Next, by some manipulations we will modify this integral representation into a form which is more suitable for our purpose. Substitution of (1.7) into (3.1) yields

$$t^{l_1,l_2}_{l,j;m,j}(a_\theta)=c^{l_1,l_2}_{l,j;m,j}\frac{1}{2\pi}\int_K\int_0^{2\pi}\left(e^{i\theta}|\alpha|^2+e^{-i\theta}|\beta|^2\right)^{l_1+l_2-m}$$

$$\cdot\left(\alpha e^{i(\phi+\theta/2)}+\beta e^{-i(\phi+\theta/2)}\right)^{m+l_1-l_2}$$

$$\cdot\left(-\bar{\beta}e^{i(\phi-\theta/2)}+\bar{\alpha}e^{i(-\phi+\theta/2)}\right)^{m-l_1+l_2}$$

$$\cdot e^{2ij\phi}t^l_{l_2-l_1,j}\left(k(\bar{\alpha},\bar{\beta})\right)dk(\alpha,\beta)\,d\phi,$$

where

(3.2)

$$c^{l_1,l_2}_{l,j;m,j}=(-1)^{l-m}\left(\frac{(2l+1)(2m+1)(l_1+l_2-l)!(l_1+l_2+l+1)!(m-j)!(m+j)!}{(l_1+l_2-m)!(l_1+l_2+m+1)!(m+l_1-l_2)!(m-l_1+l_2)!}\right)^{1/2}.$$

In this last integral representation consider the K-integral as the inner integral and make the transformation of integration variable $k(\alpha,\beta)\mapsto k(\bar{\alpha},\bar{\beta})m_{-2\phi}$. Then the integrand no longer depends on ϕ and we obtain

(3.3)

$$t^{l_1,l_2}_{l,j;m,j}(a_\theta)=c^{l_1,l_2}_{l,j;m,j}\int_K\left(e^{i\theta}|\alpha|^2+e^{-i\theta}|\beta|^2\right)^{l_1+l_2-m}$$

$$\cdot\left(\alpha e^{i\theta/2}-\beta e^{-i\theta/2}\right)^{m-l_1+l_2}\left(\bar{\alpha}e^{i\theta/2}+\bar{\beta}e^{-i\theta/2}\right)^{m+l_1-l_2}$$

$$\cdot t^l_{l_2-l_1,j}\left(k(\alpha,\beta)\right)dk(\alpha,\beta).$$

LEMMA 3.1. *Let K be a connected compact Lie group which has a complexification K_c. Let f be a complex analytic function on an open connected left-K-invariant subset V of K_c containing K. Then*

$$(3.4)\qquad \int_K f(k)\,dk=\int_K f(kk')\,dk,\qquad k'\in V.$$

Proof. The right-hand side is a complex analytic function of k' on V which is constant on K. □

Now observe that the integrand in (3.3) is the restriction to $SU(2)$ of the complex analytic function

$$\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}\mapsto\left(e^{i\theta}\alpha\delta-e^{-i\theta}\beta\gamma\right)^{l_1+l_2-m}\left(\alpha e^{i\theta/2}-\beta e^{-i\theta/2}\right)^{m-l_1+l_2}$$

$$\cdot\left(-\gamma e^{-i\theta/2}+\delta e^{i\theta/2}\right)^{m+l_1-l_2}t^{l}_{l_1-l_2,j}\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}\quad\text{on } SL(2,\mathbb{C}).$$

For $0<\theta<\pi$ apply Lemma 3.1 to this function with K' chosen as

$$g_\theta:=e^{i\pi/4}(2\sin\theta)^{-1/2}\begin{pmatrix}e^{-i\theta/2} & e^{i\theta/2}\\ e^{i\theta/2} & e^{-i\theta/2}\end{pmatrix}. \tag{3.5}$$

We obtain

(3.6)

$$t^{l_1,l_2}_{l,j;m,j}(a_\theta)=c^{l_1,l_2}_{l,j;m,j}e^{3\pi im/2}(2\sin\theta)^m$$

$$\cdot\sum_{p=-l}^{l}t^{l}_{pj}(g_\theta)\int_K\left(2|\beta|^2\cos\theta+\alpha\bar\beta-\bar\alpha\beta\right)^{l_1+l_2-m}$$

$$\cdot\beta^{m-l_1+l_2}\left(-\bar\beta\right)^{m+l_1-l_2}t^{l}_{l_2-l_1,p}(k(\alpha,\beta))\,dk(\alpha,\beta).$$

PROPOSITION 3.2. *We have*

$$t^{(m+p)/2(m-p)/2}_{l,j;m,j}(a_\theta)=\left(\frac{(2l+1)(m-j)!(m+j)!(m-p)!(m+p)!}{(2m)!(m-l)!(m+l+1)!}\right)^{1/2} \tag{3.7}$$

$$\cdot(-1)^{l+m}e^{3\pi im/2}(2\sin\theta)^m t^{l}_{pj}(g_\theta).$$

For $0<\theta<\pi$ the matrix $(t^{(m+p)/2,(m-p)/2}_{l,j;m,j}(a_\theta))_{j,p=-l,\cdots,l}$ is nonsingular.

Proof. Formula (3.6), together with (1.13) and the invariance of the integral in (3.6) under right multiplication by m_ϕ yields

$$t^{(m+p)/2(m-p)/2}_{l,j;m,j}(a_\theta)=c^{(m+p)/2(m-p)/2}_{l,j;m,j}e^{3\pi im/2}(2\sin\theta)^m$$

$$\cdot t^{l}_{pj}(g_\theta)\int_K\beta^{m-p}\left(-\bar\beta\right)^{m+p}t^{l}_{-p,p}(k(\alpha,\beta))\,dk(\alpha,\beta).$$

The integral can be evaluated by using (1.5), (1.14), the beta integral and the *Chu–Vandermonde sum*

$${}_2F_1(-n,b;c;1)=\frac{(c-b)_n}{(c)_n},\qquad n=0,1,\cdots;\quad c-b,c\neq 0,-1,\cdots,-n+1. \tag{3.8}$$

Finally use (3.2). □

THEOREM 3.3. *Formula* (2.12) *holds with*

(3.9)

$$P^{l,m}_{n,k,p}(x)=A^{l,m}_{n,k,p}\int_K\left(2|\beta|^2x+\alpha\bar\beta-\bar\alpha\beta\right)^n\beta^{m+k}\left(-\bar\beta\right)^{m-k}t^{l}_{k,p}(k(\alpha,\beta))\,dk(\alpha,\beta),$$

where

(3.10)

$$A_{n,k,p}^{l,m}:=(-1)^{2l}\left(\frac{(2m+1)!(n+m-l)!(n+m+l+1)!(m-l)!(m+l+1)!}{n!(n+2m+1)!(m-k)!(m+k)!(m-p)!(m+p)!}\right)^{1/2}.$$

There are the symmetries

$$P_{n,k,p}^{l,m}=P_{n,p,k}^{l,m}=P_{n,-k,-p}^{l,m}=P_{n,-p,-k}^{l,m}, \tag{3.11}$$

$$P_{n,k,p}^{l,m}(-x)=(-1)^{n+k+p}P_{n,k,p}^{l,m}(x). \tag{3.12}$$

Proof. Formula (3.9) follows from (3.7), (3.6) and (3.2). The symmetries are derived from (3.9) by the use of (1.6) and (1.9) in the case of (3.11) and by (1.13) in the case of (3.12). □

Of course, by the use of (2.12) and (3.7), the symmetries (3.11) imply certain symmetries for the matrix elements $t_{l,j;m,j}^{l_1,l_2}|_A$. It would be interesting to get a deeper understanding of the first of these symmetries.

Now expand the integrand in (3.9) with respect to x and use the invariance of the integral under right multiplication with m_ϕ and (1.13). We obtain

$$P_{n,k,p}^{l,m}(x)=A_{n,k,p}^{l,m}\sum_{\substack{q=|p+k|\\ q+k+k\text{ even}}}^{n} d_{n,k,p,q}^{l,m}x^{n-q}, \tag{3.13}$$

where

$$d_{n,k,p,q}^{l,m}=\frac{(-1)^{m-k+(q-k-p)/2}2^{n-q}n!}{((q-k-p)/2)!((q+k+p)/2)!(n-q)!} \tag{3.14}$$
$$\cdot\int_K \alpha^{(q+k+p)/2}\bar{\alpha}^{(q-k-p)/2}\beta^{m+n+(k-p-q)/2}$$
$$\cdot\bar{\beta}^{m+n+(-k+p-q)/2}t_{kp}^{l}(k(\alpha,\beta))\,dk(\alpha,\beta).$$

By using (1.5), (1.14) and the beta integral we obtain, for $k+p\geqq 0$

(3.15)

$$d_{n,k,p,q}^{l,m}=d_{n,-k,-p,q}^{l,m}=\frac{(-1)^{l+m+(q+k+p)/2}2^{n-q}n!(l+m+n-(q+k+p)/2)!}{((q-k-p)/2)!(n-q)!(k+p)!(l+m+n+1)!}$$
$$\cdot\sqrt{\frac{(l+k)!(l+p)}{(l-k)!(l-p)}}\,{}_3F_2\left(\begin{matrix}-l+k,\,-l+p,\,(q+k+p)/2+1\\ k+p+1,\,-l-m-n+(q+k+p)/2\end{matrix}\middle|1\right).$$

For $q=p+k$ use (3.8). Then, for $k+p\geqq 0$

(3.16)

$$d_{n,k,p,k+p}^{l,m}=d_{n,-k,-p,k+p}^{l,m}=\frac{(-1)^{l+m+p+k}2^{n-p-k}n!(m+n-k)!(m+n-p)!}{(m-l+n)!(m+l+n+1)!(p+k)!(n-p-k)!}$$
$$\cdot\left(\frac{(l+k)!(l+p)!}{(l-k)!(l-p)!}\right)^{1/2}\neq 0.$$

Hence $P_{n,k,p}^{l,m}$ is a polynomial of degree $n-|p+k|$.

THEOREM 3.4. *The vector-valued polynomial $P_{n,k}^{l,m}$ satisfies the conditions*

$$P_{n,k,p}^{l,m}(x)=\frac{(-1)^{l-m}2^n(m-k+1)_n(m+k+1)_n\delta_{k,-p}x^n}{(n!(2m+2)_n(m-l+1)_n(m+l+2)_n)^{1/2}} \tag{3.17}$$

$$+\textit{polynomial of degree less than } n,$$

$$\sum_{p=-l}^{l}\int_{-1}^{1}P_{n,k,p}^{l,m}(x)x^{n'}W_{p,q}^{l,m}(x)\,dx=0 \tag{3.18}$$

for all q in $\{-l,\cdots,l\}$ and all n' in $\{0,\cdots,n-1\}$.

Proof. Use (3.13), (3.16) and (3.10) for (3.17), and (2.15) together with (3.17) for (3.18). □

Note that (3.17) and (3.18) completely determine $P_{n,k}^{l,m}$. They also imply (2.15) for $n\neq n'$. However, from the point of view of Theorem 3.4, the orthogonality relations (2.15) for $n=n'$, $k\neq k'$ are rather unexpected.

Remark 3.5. Lemma 3.1 can also be applied in order to extract the factor $t_{mn}^{m}(b_\theta)$ from the integral representation (1.8) for $t_{mn}^{l}(b_\theta)$. Substitute $\alpha:=\cos(\theta/2)$, $\beta:=\sin(\theta/2)$ in (1.8) and make the successive transformations of integration variable $\phi\mapsto z\mapsto\psi\mapsto\chi$, where $e^{2i\phi}=z=e^{i\psi}\cot(\theta/2)$, $\chi=2\psi$:

$$\left(\frac{(l-m)!(l+m)!}{(l-n)!(l+n)!}\right)^{1/2}t_{mn}^{l}(b_\theta)$$

$$=\frac{1}{2\pi i}\oint_{(0)}(z\cos(\theta/2)+\sin(\theta/2))^{l-m}(-z\sin(\theta/2))+\cos(\theta/2)^{l+m}z^{n-l-1}\,dz$$

$$=(\sin(\theta/2))^{m-n}(\cos(\theta/2))^{m+n}$$

$$\cdot\frac{1}{2\pi}\int_0^{2\pi}\left(e^{i\psi}\cos^2\frac{\theta}{2}+\sin^2\frac{\theta}{2}\right)^{l-m}e^{i\psi(n-l)}(1-e^{i\psi})^{l+m}d\psi$$

$$=(-2i)^{l+m}\left(\sin\frac{\theta}{2}\right)^{m-n}\left(\cos\frac{\theta}{2}\right)^{m+n}$$

$$\cdot\frac{1}{\pi}\int_0^{\pi}(\cos\chi+i\sin\chi\cos\theta)^{l-m}e^{2ni\chi}(\sin\chi)^{l+m}d\chi.$$

Now assume $m\geqq n$ and use [2, 1.5(29)]. Then

$$t_{mn}^{l}(b_\theta)/t_{mn}^{m}(b_\theta)=\text{const.}\int_0^{\pi}(\cos\chi+i\sin\chi\cos\theta)^{l-m}e^{2ni\chi}(\sin\chi)^{l+m}d\chi \tag{3.19}$$

with nonzero constant. Again by [2, 1.5 (29)], the right-hand side of (3.19) is a polynomial of degree $l-m$ in $\cos\theta$ which takes a nonzero value if $\cos\theta=1$. In Greiner and Koornwinder [4, §1.3] the integral representation for Jacobi polynomials resulting from (3.19) is obtained in a quite different context.

4. The noncompact analogue. Let now $G:=SL(2,C)$ with Iwasawa decomposition $G=KAN$ such that

$$K=SU(2),\quad A=\left\{a_t:=\begin{pmatrix}e^{t/2}&0\\0&e^{-t/2}\end{pmatrix}\middle|\,t\in\mathbb{R}\right\},\quad N:=\left\{\begin{pmatrix}1&z\\0&1\end{pmatrix}\middle|\,z\in\mathbb{C}\right\}.$$

Let $k(\alpha,\beta)$ in K be defined by (1.10) and m_ϕ by (1.12). $M:=\{m_\phi|0\leqq\phi<4\pi\}$ is the centralizer of A in K.

Let $\pi^{\lambda,k}(\lambda\in\mathbb{C},k\in\frac{1}{2}\mathbb{Z})$ be the representation of G which is induced by the representation $m_\phi a_t n\mapsto e^{-ik\phi}e^{\lambda t}$ of MAN: a principal series representation. Then $\pi^{\lambda,k}|_K$ is unitary and decomposes as $\oplus_{l=k,k+1,\dots}T^l$. Choose a K-basis for which $\pi^{\lambda,k}$ has matrix elements $\pi^{\lambda,k}_{l,p;m,q}(l,m=k,k+1,\dots;\ p=-l,\cdots,l;\ q=-m,\cdots,m)$ such that

$$\pi^{\lambda,k}_{l,p;m,q}(k)=\delta_{l,m}t^l_{p,q}(k),\qquad k\in K.$$

Then

$$\text{(4.1)}\qquad \pi^{\lambda,k}_{l,j;m,j}(a_t)=(2l+1)^{1/2}\int_K\left(e^{-t}|\alpha|^2+e^t|\beta|^2\right)^{-\lambda-m-1}$$
$$\cdot t^m_{kj}\begin{pmatrix} e^{-t/2}\alpha & e^{t/2}\beta\\ -e^{t/2}\bar{\beta} & e^{-t/2}\bar{\alpha}\end{pmatrix}t^l_{kj}\left(k(\bar{\alpha},\bar{\beta})\right)dk(\alpha,\beta),$$

cf. Rühl [9, §3–5], Kosters [8, §3.1].

Similarly to (3.3) we derive from (4.1) that

$$\text{(4.2)}\qquad \pi^{\lambda,k}_{l,j;m,j}(a_t)=c_{k,l,m,j}\int_K\left(e^{-t}|\alpha|^2+e^t|\beta|^2\right)^{-\lambda-m-1}\left(e^{-t/2}\alpha-e^{t/2}\beta\right)^{m+k}$$
$$\cdot\left(e^{-t/2}\bar{\alpha}+e^{t/2}\bar{\beta}\right)^{m-k}t^l_{kj}\left(k(\alpha,\beta)\right)dk(\alpha,\beta),$$

where

$$\text{(4.3)}\qquad c_{k,l,m,j}:=\left(\frac{(2l+1)(2m+1)(m-j)!(m+j)!}{(m-k)!(m+k)!}\right)^{1/2}.$$

For $s>0$ let

$$\text{(4.4)}\qquad h_s:=(2\operatorname{sh}s)^{-1/2}\begin{pmatrix} e^{s/2} & e^{-s/2}\\ e^{-s/2} & e^{s/2}\end{pmatrix}.$$

Then we can apply Lemma 3.1 to (4.2) with $k':=h_s$ for $0<t<s$. We obtain

(4.5)

$$\pi^{\lambda,k}_{l,j;m,j}(a_t)=c_{k,l,m,j}2^m(\operatorname{sh}s)^{-m}$$
$$\cdot\sum_{p=-l}^{l}t^l_{pj}(h_s)\int_K\left(\operatorname{ch}t-\operatorname{coth}s\operatorname{sh}t\left(|\alpha|^2-|\beta|^2\right)+(\alpha\bar{\beta}-\beta\bar{\alpha})\frac{\operatorname{sh}t}{\operatorname{sh}s}\right)^{-\lambda-m-1}$$
$$\cdot\left(\alpha\operatorname{sh}\tfrac{1}{2}(s-t)-\beta\operatorname{sh}\tfrac{1}{2}(s+t)\right)^{m+k}$$
$$\cdot\left(\bar{\alpha}\operatorname{sh}\tfrac{1}{2}(s-t)+\bar{\beta}\operatorname{sh}\tfrac{1}{2}(s+t)\right)^{m-k}$$
$$\cdot t^l_{kp}\left(k(\alpha,\beta)\right)dk(\alpha,\beta),\qquad 0<t<s.$$

If $\operatorname{Re}\lambda \leqq m-1$ then the limit passage $s \downarrow t$ is certainly allowed in (4.5).

(4.6)

$$\pi^{\lambda,k}_{l,j;m,j}(a_t) = c_{k,l,m,j}(-1)^{2m}(2\,\mathrm{sh}\,t)^m$$

$$\cdot \sum_{p=-l}^{l} t^l_{pj}(h_t)\int_K \left(2|\beta|^2\mathrm{ch}\,t + \alpha\bar{\beta} - \beta\bar{\alpha}\right)^{-\lambda-m-1}$$

$$\cdot \beta^{m+k}(-\bar{\beta})^{m-k} t^l_{kp}(k(\alpha,\beta))\,dk(\alpha,\beta).$$

Closer examination of the integral, using (1.14), shows that (4.6) holds with convergent integral if $\operatorname{Re}\lambda < 0$. Thus it is meaningful to study the vector-valued function $x \mapsto (P^{l,m}_{n,k,p}(x))_{p=-l,\cdots,l}$, defined by (3.9), for complex n, $\operatorname{Re} n > 0$, and for $x > 1$. In particular, this function has a nice asymptotics as $x \to \infty$.

REFERENCES

[1] D. Basu and S. Srinivasan, *A unified treatment of the groups $SO(4)$ and $SO(3,1)$*, Czech. J. Phys., B 27 (1977), pp. 629–635.

[2] A. Erdelyi, et al., *Higher Transcendental Functions, Vol.* I, McGraw-Hill, New York, 1953.

[3] D. Z. Freedman and J.-M. Wang, *$O(4)$ symmetry and Regge-pole theory*, Phys. Rev., 160 (1967), pp. 1560–1571.

[4] P. C. Greiner and T. H. Koornwinder, *Variations on the Heisenberg spherical harmonics*, Math. Centrum Report, ZW 186, Mathematisch Centrum, Amsterdam, 1983.

[5] S. Helgason, *Differential Geometry and Symmetric Spaces*, Academic Press, New York, 1962.

[6] T. H. Koornwinder, *Clebsch–Gordan coefficients for $SU(2)$ and Hahn polynomials*, Nieuw Arch. Wisk., (3) 29 (1981), pp. 140–155.

[7] ______, *The representation theory of $SL(2,\mathbb{R})$, a noninfinitesimal approach*, Enseignement Math., 28 (1982), pp. 53–90.

[8] M. T. Kosters, *A study of the representations of $SL(2,\mathbb{C})$ using noninfinitesimal methods*, Math. Centrum Report TW 190, Mathematisch Centrum, Amsterdam, 1979.

[9] W. Rühl, *The Lorentz Group and Harmonic Analysis*, Benjamin, Menlo Park, CA, 1970.

[10] Ya. A. Smorodinskiĭ and G. I. Shepelev, *Boost matrix elements in $O(3,1)$ and continuation to $O(4)$*, Soviet J. Nuclear Phys., 13 (1971), pp. 248–253.

[11] N. J. Vilenkin, *Special Functions and the Theory of Group Representations*, AMS Transl. Math. Monographs, 22, American Mathematical Society, Providence, RI, 1968.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

016

ON THE CONVEXITY OF THE ZEROS OF BESSEL FUNCTIONS*

ÁRPÁD ELBERT† AND ANDREA LAFORGIA‡

Abstract. Let $c_{\nu k}$ be the kth positive zero of the cylinder function $C_\nu(x)=\cos\alpha J_\nu(x)-\sin\alpha Y_\nu(x)$. We prove among other things that the function $c_{\nu 1}$ is convex on $\frac{1}{2}\leqq\nu<\infty$ if $\pi-\frac{1}{2}+\varepsilon_0<\alpha<\pi$, where $\varepsilon_0=0.163302\cdots$ is specified in Lemma 2.1.

1. Introduction and background. For $\nu\geqq 0$ and $k=1,2,\cdots$, we use $j_{\nu k}$ and $c_{\nu k}$ to denote the kth positive zeros of the Bessel function $J_\nu(x)$ of the first kind and of the general cylinder function

$$C_\nu(x)=\cos\alpha J_\nu(x)-\sin\alpha Y_\nu(x)$$

respectively, where α is fixed, $0\leqq\alpha<\pi$ and $Y_\nu(x)$ is the Bessel function of the second kind.

The definitions may be extended to negative values of ν in such a way that $c_{\nu k}$ varies continuously with ν, and $c_{\nu k}\to 0$, when $\nu\to\alpha/\pi-k$ and on the interval

$$\frac{\alpha}{\pi}-k<\nu<\frac{\alpha}{\pi}-k+1.$$

$c_{\nu_i k}$ is the first positive zero of $C_\nu(x)$ [5, pp. 508–509].

Since the notation $c_{\nu k}$ does not reflect the dependence on the values α, it is useful to define the function $j_{\nu\kappa}$, where κ is a real positive parameter, as in [2].

The sequence $j_{\nu 1}, j_{\nu 2},\cdots$ has been already defined. For any κ with $k-1<\kappa<k$, where k is some natural number, let $j_{\nu\kappa}=c_{\nu k}$ with $\alpha=(k-\kappa)\pi$. It is clear that this correspondence between $j_{\nu\kappa}$ and $c_{\nu k}$ is one to one. Moreover by this notation the above mentioned limit relation for $c_{\nu k}$ reads as

$$\lim_{\nu\to -\kappa+0} j_{\nu\kappa}=0. \tag{1.1}$$

The function $c_{\nu k}$ satisfies the differential equation [5, p. 508]

$$\frac{d}{d\nu}c_{\nu k}=2c_{\nu k}\int_0^\infty K_0(2c_{\nu k}\sinh t)e^{-2\nu t}\,dt,$$

where $K_0(x)$ is the modified Bessel function of order zero. Thus the function $j_{\nu\kappa}$ is the solution of the differential equation

$$\frac{d}{d\nu}j=2j\int_0^\infty K_0(2j\sinh t)e^{-2\nu t}\,dt \tag{1.2}$$

for all $\kappa>0$ with the boundary condition (1.1).

Since the right-hand side of (1.2) is Lipschitzian with respect to j for $j>0$, the solution of any initial value problem is unique. Concerning the boundary condition

* Received by the editors November 24, 1982, and in revised form June 20, 1983. This work was sponsored by the Consiglio Nazionale delle Ricerche,—Italy.

† Mathematical Institute of the Hungarian Academy of Sciences, 1053 Budapest, Réaltanoda u. 13–15, Hungary.

‡ Department of Mathematics University of Torino, Via Carlo Alberto, 10 10123 Torino, Italy.

(1.1) the uniqueness is ensured by the fact that for *every* $\kappa>0$ we have $\lim_{\nu\to-\kappa+0} j_{\nu\kappa}=0$. So the relation $\lim_{\nu\to-\kappa+0} j_{\nu\kappa'}=0$ implies $\kappa'=\kappa$. By the uniqueness we have that, if $0<\kappa'<\kappa''$, then

$$j_{\nu\kappa'}<j_{\nu\kappa''} \quad \text{for } \nu>-\kappa'. \tag{1.3}$$

An interesting special case is when $\nu=\frac{1}{2}$. Since $J_{1/2}(x)=\sqrt{2/\pi x}\,\sin x$, $Y_{1/2}(x)=-\sqrt{2/\pi x}\,\cos x$, the general cylinder function $C_{1/2}(x)$ is

$$C_{1/2}(x)=\sqrt{\frac{2}{\pi x}}\,\sin(x+\alpha), \qquad 0\leqq\alpha<\pi.$$

Consequently $c_{1/2,k}=k\pi-\alpha$, hence with our notation

$$j_{1/2,\kappa}=\kappa\pi \quad \text{for } \kappa>0. \tag{1.4}$$

In the case $\nu=-\frac{1}{2}$ there is another similar relation, namely

$$j_{-1/2,\kappa}=\left(\kappa-\frac{1}{2}\right)\pi \quad \text{for } \kappa>\frac{1}{2}. \tag{1.5}$$

An alternative proof of the monotonicity property of $j_{\nu\kappa}$ given in (1.3) is as follows. Let us consider the initial value problem for the differential equation (1.2) with the initial condition $j_0=\kappa\pi$ at $\nu_0=\frac{1}{2}$. Then by (1.4) and by uniqueness the solution obtained is $j_{\nu\kappa}$. Making use again of uniqueness we have for any two solutions $j_{\nu\kappa'}, j_{\nu\kappa''}$ of (1.2) with $0<\kappa'<\kappa''$: $j_{1/2,\kappa'}=\kappa'\pi<\kappa''\pi=j_{1/2,\kappa''}$ and consequently (1.3) holds.

The relations (1.1), (1.4), (1.5) show that the notation $j_{\nu\kappa}$ is reasonable.

We observe that the property (1.3) implies that the function $c_{\nu k}$ with fixed ν, k decreases when α increases and $0\leqq\alpha<\pi$. This result has been proved by L. Lorch and L. J. Newman in [4, p. 362] using more sophisticated arguments.

In [1] the first author proved that the functions $j_{\nu 1}, j_{\nu 2}, \ldots$ are concave on the whole domain of existence. This result was extended by M. E. Muldoon and by the second author [3] to show that $j_{\nu\kappa}$ is concave for $\kappa\geqq\frac{1}{2}$ on $\nu\geqq 0$. The question arises naturally whether the concavity of $j_{\nu\kappa}$ holds for every $\kappa>0$. The answer is in the negative and our aim here is to show that for sufficiently small κ the function $j_{\nu\kappa}$ is convex at least on $\nu\geqq\frac{1}{2}$.

Now we recall some known results which will be useful in the next section. We start with the integral formula [5, p. 388]

$$\int_0^\infty K_0(x)e^{-x}dx=1. \tag{1.6}$$

The second derivative $j''=d^2j_{\nu\kappa}/d\nu^2$ has already been computed in [1, p. 87]

$$j''=2j\int_0^\infty K_0(2j\sinh t)e^{-2\nu t}I(t)\,dt \tag{1.7}$$

where

$$I(t)=2\nu\frac{j'}{j}\tanh t+\frac{j'}{j}\tanh^2 t-2t. \tag{1.8}$$

Despite the fact that the formula (1.7), (1.8) was derived for $j_{\nu 1}, j_{\nu 2}, \cdots$, it is not difficult to check that all steps used in [1] to derive (1.7) are valid also for all $\kappa>0$.

Finally we recall the integral formula for $K_0(x)$ [5, p. 446]

$$K_0(x)=\int_0^\infty e^{-x\cosh t}dt.$$

From this formula we obtain that $K_0(x)$ decreases as x increases, and also the stronger property that $e^x K_0(x)$ decreases on $0<x<\infty$. This property will be exploited in the proof of Lemma 2.3. For the sake of the later reference we express this property in the form

$$e^x K_0(x) < e^y K_0(y) \quad \text{if } 0<y<x. \tag{1.9}$$

2. The main result. Our main result will be proved as a simple consequence of the next four lemmas.

LEMMA 2.1. *There is a value* $\varepsilon_0>0$ *such that the inequality*

$$e^{2\varepsilon(\sinh t - t)+2\varepsilon\sinh t} > \cosh t$$

holds for $\varepsilon>\varepsilon_0$ *and* $t>0$. *The value of* ε_0 *lies between* 0.16330286 *and* 0.16330298.

Proof. Let us consider the function $f(t,\varepsilon)$ defined by

$$f(t,\varepsilon) = 4\varepsilon\sinh t - 2\varepsilon t - \log\cosh t.$$

Then we must find the value ε_0 so that $f(t,\varepsilon)>0$, for all $t>0$ and $\varepsilon>\varepsilon_0$. Since $2\sinh t - t>0$ for $t>0$, we obtain for $t>0$ that

$$\varepsilon > \frac{1}{2}\frac{\log\cosh t}{2\sinh t - t} = g(t).$$

The function $g(t)$ satisfies the relations $\lim_{t\to+0} g(t)=0$; $\lim_{t\to+\infty} g(t)=0$; and $g(t)>0$ for $t>0$. Therefore $g(t)$ is bounded from above and

$$\varepsilon_0 = \max_{t>0} g(t) = g(t_0) > 0 \quad \text{with some } t_0 \in (0,\infty).$$

Hence $f(t_0,\varepsilon_0)=0$ and, owing to the restriction $f(t,\varepsilon_0)\geqq 0$, we find $(\partial/\partial t)f(t_0,\varepsilon_0) = 0$. To determine the values ε_0, t_0 we make the following observation. The function

$$\frac{\partial}{\partial t} f(t,\varepsilon) = 4\varepsilon\cosh t - 2\varepsilon - \tanh t$$

is convex on $0\leqq t<\infty$, $(\partial/\partial t)f(0,\varepsilon)=2\varepsilon>0$, $\lim_{t\to\infty}(\partial/\partial t)f(t,\varepsilon)=\infty$ for every $\varepsilon>0$. We claim that the function $(\partial/\partial t)f(t,\varepsilon_0)$ has exactly two single zeros. Suppose the contrary. Then by the convexity $(\partial/\partial t)f(t,\varepsilon_0)$ would have double zero at $t=t_0$. Consequently $(\partial/\partial t)f(t,\varepsilon_0)$ would be nonnegative and $f(t,\varepsilon_0)$ would be nondecreasing. But $f(0,\varepsilon_0)=f(t_0,\varepsilon_0)=0$; hence $f(t,\varepsilon_0)\equiv 0$ for $0\leqq t\leqq t_0$ which contradicts $(\partial/\partial t)f(0,\varepsilon_0)=2\varepsilon_0>0$.

Suppose that ε is in the vicinity of ε_0. Let $t_1(\varepsilon)$ and $t_2(\varepsilon)$ be the zeros of the function $(\partial/\partial t)f(t,\varepsilon)$ with $t_1(\varepsilon)<t_2(\varepsilon)$. Then $f(t,\varepsilon)$ has a local maximum at $t=t_1(\varepsilon)$ and a local minimum at $t=t_2(\varepsilon)$. At $t=t_2(\varepsilon)$ the second derivative $(\partial^2/\partial t^2)f(t,\varepsilon)$ should be positive.

Thus we have the following algorithm for computing the values t_0, ε_0. We choose a value $t_2>0$ as $t_2(\varepsilon)$. Then $\varepsilon=\varepsilon(t_2)=g(t_2)$. We should check whether

$$\left.\frac{\partial^2}{\partial t^2} f(t,\varepsilon)\right|_{t=t_2;\,\varepsilon=\varepsilon(t_2)} > 0.$$

By a trivial calculation we find that the latter inequality certainly holds if $t_2\geqq 1$. Let $F(t_2)=f(t_2,\varepsilon(t_2))$. Then we must solve the equation $F(t)=0$.

For $t^{(1)}=1.165513$ and $t^{(2)}=1.165514$ we find $F(t^{(1)})=1.5\times 10^{-7}$, $F(t^{(2)})=-2.8\times 10^{-7}$ and hence the value of t_0 in question lies between $t^{(1)}$ and $t^{(2)}$; consequently $\varepsilon(t^{(2)})=0.16330286<\varepsilon_0<\varepsilon(t^{(1)})=0.16330298$ as we stated.

The next three lemmas are of local type characterising the local behaviour of the function $j_{\nu\kappa}$.

LEMMA 2.2. *If, for any real number $\kappa>0$ and fixed $\nu_0>0$,*

$$\frac{\nu_0}{j_{\nu_0\kappa}}\frac{dj_{\nu_0\kappa}}{d\nu}\geqq 1,\quad \textit{then}\quad \left.\frac{d^2}{d\nu^2}j_{\nu\kappa}\right|_{\nu=\nu_0}>0.$$

Proof. Here and in the sequel, we shall use the notation

$$j_0=j_{\nu_0\kappa},\quad j_0'=\left.\frac{d}{d\nu}j_{\nu\kappa}\right|_{\nu=\nu_0}\quad\text{and}\quad j_0''=\left.\frac{d^2}{d\nu^2}j_{\nu\kappa}\right|_{\nu=\nu_0}.$$

From (1.8) we get

$$I(0)=0,\qquad I'(t)=\frac{2\nu_0 j_0'}{j_0}\frac{1}{\cosh^2 t}+\frac{2j_0'}{j_0}\frac{\tanh t}{\cosh^2 t}-2.$$

Hence $\lim_{t\to\infty}I'(t)=-2$ and by our assumption, $I'(0)=2(\nu_0 j_0'/j_0-1)>0$. On the other hand $(j_0/2j_0')\cosh^4 t\cdot I''(t)=-2\nu_0\sinh t\cosh t+1-2\sinh^2 t$. Since $\nu_0>0$ the function on the right-hand side decreases from 1 to $-\infty$ as t increases from 0 to ∞. Thus the function $I(t)$ is convex for small t's and then ultimately concave.

These properties of $I(t)$ show that there is a value t_0 such that $I(t)>0$ for $0<t<t_0$ and $I(t)<0$ for $t>t_0$.

Now we use the fact that $K_0(x)$ decreases for $x>0$. From (1.7) we obtain

$$\begin{aligned}\frac{j_0''}{2j_0'}&=\int_0^\infty K_0(2j_0\sinh t)I(t)e^{-2\nu_0 t}dt\\ &>\int_0^\infty K_0(2j_0\sinh t_0)I(t)e^{-2\nu_0 t}dt=K_0(2j_0\sinh t_0)(I_1+I_2),\end{aligned}$$

where

$$I_1=\int_0^\infty\frac{j_0'}{j_0}\tanh^2 t\cdot e^{-2\nu_0 t}dt$$

and

$$I_2=\int_0^\infty\left(\frac{2\nu_0 j_0'}{j_0}\tanh t-2t\right)e^{-2\nu_0 t}dt.$$

An integration by parts of I_2 gives

$$\begin{aligned}I_2=&-\left[\left(\frac{2\nu_0 j_0'}{j_0}\tanh t-2t\right)\frac{e^{-2\nu_0 t}}{2\nu_0}\right]_0^\infty\\ &+\frac{1}{2\nu_0}\int_0^\infty\left(\frac{2\nu_0 j_0'}{j_0}\frac{1}{\cosh^2 t}-2\right)e^{-2\nu_0 t}dt.\end{aligned}$$

For $\nu_0 > 0$ the first term on the right-hand side is zero; hence

$$\frac{j_0''}{2j_0} > K_0(2j_0\sinh t_0)\int_0^\infty \left(\frac{j_0'}{j_0} - \frac{1}{\nu_0}\right) e^{-2\nu_0 t}\,dt \geqq 0,$$

i.e. $j_0'' > 0$ and Lemma 2.2. is proved.

LEMMA 2.3. *For* $\nu_0 > \varepsilon_0$, $\kappa > 0$ *suppose* $0 < j_{\nu_0\kappa} < \nu_0 - \varepsilon_0$, *where* ε_0 *is specified in Lemma* 2.1. *Then*

$$\frac{\nu_0}{j_{\nu_0\kappa}}\,\frac{d}{d\nu}\,j_{\nu\kappa}\bigg|_{j=j_0;\,\nu=\nu_0} > 1.$$

Proof. By using the inequality (1.9) with $x = 2\nu_0\sinh t$ and $y = 2j_0\sinh t$ in (1.2), we obtain

$$\begin{aligned}\frac{j_0'}{j_0} &= \int_0^\infty K_0(2j_0\sinh t)\,e^{-2\nu_0 t}\,dt\\ &> 2\int_0^\infty K_0(2\nu_0\sinh t)\,e^{2\nu_0(\sinh t - t) + 2(\nu_0 - j_0)\sinh t - 2\nu_0\sinh t}\,dt.\end{aligned}$$

Since Lemma 2.1 obviously implies

$$e^{2\varepsilon_0(\sinh t - t) + 2\varepsilon_0 \sinh t} \geqq \cosh t,$$

then by the substitution $x = 2\nu_0\sinh t$ and by (1.6) we have

$$\begin{aligned}\frac{j_0'}{j_0} &> 2\int_0^\infty K_0(2\nu_0\sinh t)\,e^{-2\nu_0\sinh t}\cosh t\,dt = \frac{1}{\nu_0}\int_0^\infty K_0(x)\,e^{-x}\,dx\\ &= \frac{1}{\nu_0},\end{aligned}$$

which completes the proof of Lemma 2.3.

LEMMA 2.4. *If, for any real* $\kappa > 0$ *and* $\nu_0 > 0$, $0 < j_{\nu_0\kappa} \leqq \nu_0$ *then*

$$\frac{d}{d\nu}\,j_{\nu\kappa}\bigg|_{\nu=\nu_0} < 1.$$

Proof. By making use of (1.2) and (1.6) we have

$$j_0' = 2j_0\int_0^\infty K_0(2j_0\sinh t)\,e^{-2\nu_0 t}\,dt < 2j_0\int_0^\infty K_0(2j_0 t)\,e^{-2j_0 t}\,dt = 1,$$

which is the desired result. Thus the proof of Lemma 2.4 is complete.

Now we can enunciate the main result.

THEOREM 2.1. *If for some* $\nu_0 > \varepsilon_0$ *and any given real* $\kappa > 0$, $0 < j_{\nu_0\kappa} \leqq \nu_0 - \varepsilon_0$, *where* ε_0 *is defined in Lemma* 2.1, *then* $j_{\nu\kappa} \leqq \nu - \varepsilon_0$ *and* $(d^2/d\nu^2)\,j_{\nu\kappa} > 0$ *for* $\nu \geqq \nu_0$.

Proof. The first part of Theorem 2.1 is a consequence of Lemma 2.4. In fact this means that the function $j_{\nu\kappa} - \nu$ decreases, i.e.

$$j_{\nu\kappa} - \nu < j_{\nu_0\kappa} - \nu_0 \leqq -\varepsilon_0 \quad \text{for } \nu \geqq \nu_0.$$

To show the second part we use Lemma 2.2 and Lemma 2.3. An application of Lemma 2.3 gives $\nu j'/j > 1$ for $\nu \geqq \nu_0$ and by Lemma 2.2 we obtain $j'' > 0$ for $\nu \geqq \nu_0$.

The proof of Theorem 2.1 is then complete.

Combining (1.4) with Theorem 2.1 we obtain the following.

COROLLARY 2.1. *If* $0<\kappa\leqq(\frac{1}{2}-\varepsilon_0)/\pi$ $(=0.107173\cdots)$, *then the function* $j_{\nu\kappa}$ *is convex for* $\nu\geqq\frac{1}{2}$.

Let us remark that $c_{\nu 1}=j_{\nu\kappa}$ where $\alpha=(1-\kappa)\pi$ hence $c_{\nu 1}$ is convex on the interval $\frac{1}{2}\leqq\nu<\infty$ if $\pi-\frac{1}{2}+\varepsilon_0\leqq\alpha<\pi$.

Acknowledgments. The first author is grateful for the hospitality extended by the Department of Mathematics at the University of Torino, where he was a visitor when this paper was written.

Both authors wish to express their thanks to Prof. H. L. Turrittin, whose constructive criticism and suggestions led to the present improved version of this paper.

REFERENCES

[1] Á. ELBERT, *Concavity of the zeros of Bessel functions*, Studia Sci. Math. Hungar, 12 (1977), pp. 81–88.

[2] Á. ELBERT AND A. LAFORGIA, *On the square of the zeros of Bessel functions*, this Journal, 15 (1984), pp. 206–212.

[3] A. LAFORGIA AND M. E. MULDOON, *Monotonicity and concavity properties of zeros of Bessel functions*, J. Math. Anal. Appl., 98 (1984), pp. 470–477.

[4] L. LORCH AND D. J. NEWMAN, *A supplement to the Sturm separation theorem with applications*, Amer. Math. Monthly, 72 (1965), pp. 359–366.

[5] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge Univ. Press, London 1944.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

SPECIAL FUNCTIONS FOR THE SYMMETRIC SPACE OF POSITIVE MATRICES*

AUDREY TERRAS†

Abstract. Section 1 of the paper gives three applications of a basic principle that eigenfunctions of invariant differential operators are eigenfunctions of invariant integral operators. The first application is a derivation of a noneuclidean analogue of the Poisson summation formula, the second is the evaluation of an integral of Muirhead which has been of interest in multivariate statistics, the third is the evaluation of gamma type integrals arising in the theory of Eisenstein series for the general linear group. Section 2 of the paper concerns K-Bessel functions for the general linear group. It relates such functions with gamma functions for the symplectic group and applies the theory to the estimation of Fourier coefficients of automorphic forms for $GL(3, \mathbb{Z})$.

Key words. general linear group, positive matrices, harmonic analysis on symmetric spaces, fundamental domains for discrete groups, Poisson summation formula, K-Bessel functions of matrix argument, gamma functions, theta functions, Eisenstein series, automorphic forms, spherical functions

Introduction. Here we consider some aspects of analysis on the symmetric space $\mathcal{P}_n$ of positive $n \times n$ matrices. The main motivations for this work are statistical sorts of problems in physics and number theory (see N. Hurt [41] and D. Wallace [42] for some examples).

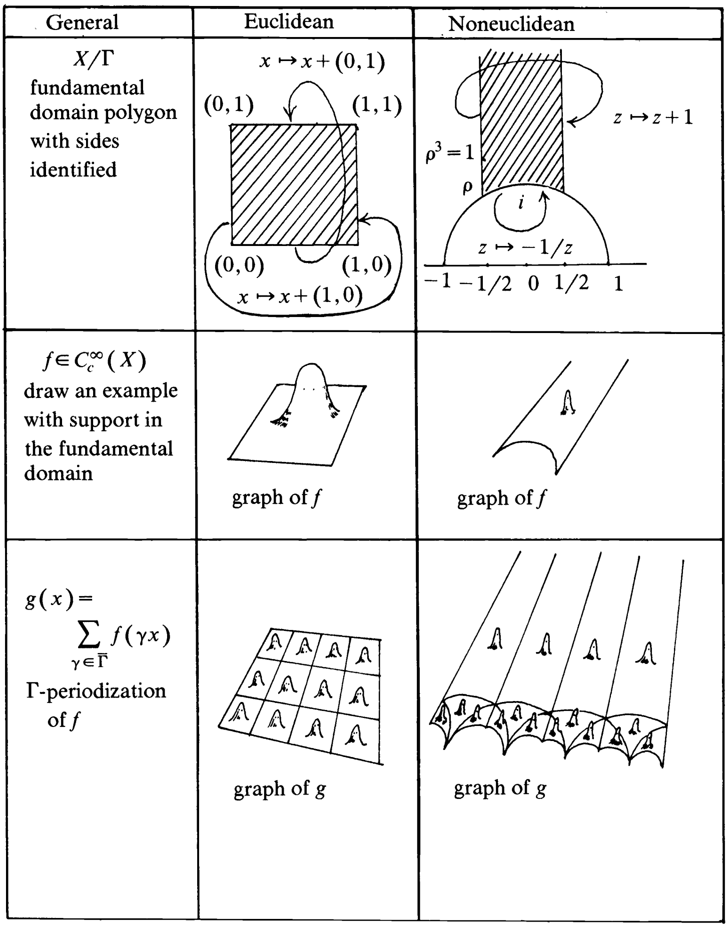
Section 1 concerns applications of the basic principle that eigenfunctions of invariant differential operators are eigenfunctions of invariant integral operators. First, the principle is applied to obtain a noneuclidean analogue of Poisson's summation formula which is simpler than Selberg's trace formula. This result can also be viewed as Mercer's theorem and is thus a prerequisite for the Selberg trace formula (cf. Selberg [22], Terras [27], [28], [32]). The simplicity of this noneuclidean Poisson summation formula suggests that the result may well be a very useful tool in the higher rank situations. There is already much evidence for this (cf. Bartels [2], Hejhal [8], Kudla and Millson [14], Mennicke [18], Patterson [21], Selberg). A second application of the basic principle of §1 is the evaluation of a multiple integral that has been of use in multivariate statistics (cf. Muirhead [19, Thm. 7.2.5]). A third application appears in the study of Eisenstein series for $GL(n, \mathbb{Z})$, the group of integral $n \times n$ matrices of determinant ± 1. Here we generalize a method of Maass [16] and Selberg for obtaining the analytic continuation of Eisenstein series. The method rests on the evaluation of an integral which is an analogue of the gamma functions. The basic principle of this section greatly simplifies the computation.

Section 2 concerns applications of properties of K-Bessel functions for $\mathcal{P}_n$ studied by Bengtson [3]. First, we note that K-Bessel functions are involved in the transformation formula for the noneuclidean analogue of the theta function associated to the Siegel modular group. This occurs because the K-Bessel function for $\mathcal{P}_n$ is the analogue of the gamma function for the symplectic group. One motivation for this study is the hope to obtain higher rank analogues of asymptotic results on the number of lattice points in a circle of radius R as $R \to \infty$ (i.e., noneuclidean circle problems). Bartels [2] has obtained the main term in such asymptotic expansions for any cocompact discrete

*Received by the editors January 31, 1983, and in revised form July 28, 1983.

†Mathematics Department, University of California at San Diego, La Jolla, California 93093. Research partially supported by an NSF Grant.

subgroup Γ of a connected semisimple real Lie group G with finite center. We are interested in the case that the fundamental domain G/Γ is not compact; e.g., $G=\mathrm{GL}(n,\mathbb{R})$ and $\Gamma=\mathrm{GL}(n,\mathbb{Z})$. For this case the spectral decomposition of the G-invariant differential operators on $L^2(K\backslash G/\Gamma)$ has not been completely described (cf. Arthur [1], Jacquet [12], LanGLands [15], Osborne and Warner [20], Venkov [33]). Thus the noneuclidean analogue of Poisson summation associated with Γ cannot be completely described. The second application of K-Bessel functions considered here is a discussion of the growth of Fourier coefficients of automorphic forms for $\mathrm{GL}(n,\mathbb{Z})$. Recently Takhtadzhyan and Vinogradov [26] Imai and Terras [11], and Terras [31] have obtained such expansions. And Takhtadzhyan and Vinogradov [26] have considered applications of such expansions to the extension of work of Kuznetsov on divisor functions. Here we are motivated by the hope to obtain a theory of automorphic forms for $\mathrm{GL}(n,\mathbb{Z})$ which closely parallels the classical theory of modular forms and Maass wave forms when $n=2$. At the end of §2, we include a brief discussion of the relation between K-Bessel functions for $\mathcal{P}_n$ and Whittaker functions of the type studied by Bump [35], Jacquet [13] and others.



General	Euclidean	Noneuclidean
X/Γ fundamental domain polygon with sides identified		
$f\in C_c^\infty(X)$ draw an example with support in the fundamental domain	graph of f	graph of f
$g(x)=\sum_{\gamma\in\Gamma} f(\gamma x)$ Γ-periodization of f	graph of g	graph of g

FIG. 1. *Γ-periodizations of compactly supported functions on X.*

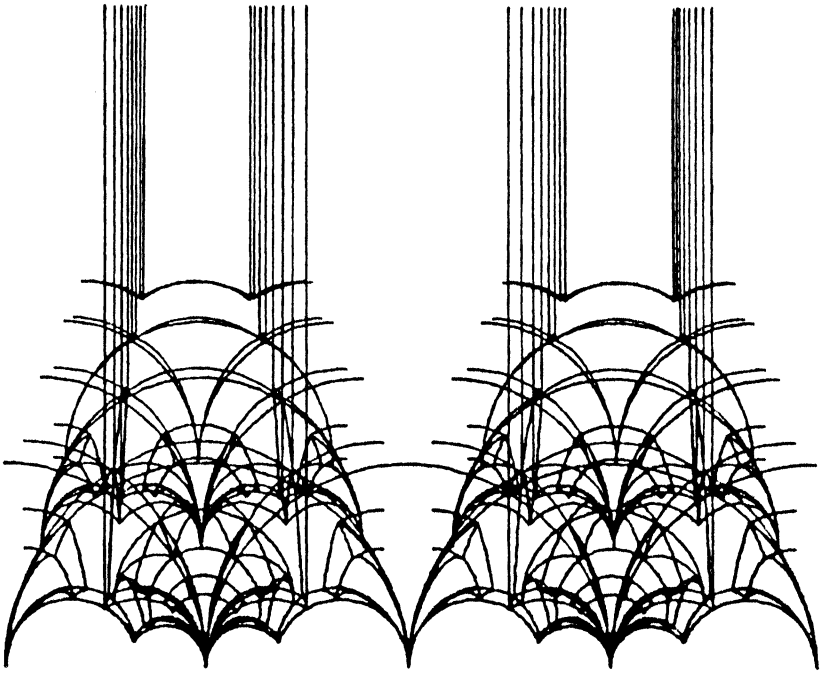


FIG. 2. *Tessellation of the quaternionic upper half plane from* SL$(2,\mathbb{Z}[i])$ *in stereo. Drawn by the* UCSD VAX *computer and Mark Eggert.*

1. Applications of the basic principle on eigenfunctions of invariant integral and differential operators. An introduction with many other references for the problems under consideration appears in Terras [27], [28], [32]. A brief account of the noneuclidean Poisson summation formula for $\Gamma = \mathrm{SL}(2,\mathbb{Z})$ follows. Table 1 summarizes a comparison of euclidean and noneuclidean harmonic analysis for some two-dimensional, rank one symmetric spaces $X = G/K$. Figure 1 pictures the Γ-periodizations of compactly supported functions on X. We are actually interested in higher rank analogues. The simplest example is $\mathscr{P}_3 \subset \mathbb{R}^6$ or the determinant one surface in $\mathscr{P}_3$. To our knowledge, no one has attempted to picture the tessellation of $\mathscr{P}_3$ obtained by letting $\mathrm{GL}(3,\mathbb{Z})$ act on a fundamental domain. To give some idea of the higher dimensional tessellations, we include a stereo picture of a tessellation of hyperbolic 3-space (alias the quaternionic upper half plane) in Fig. 2. Table 2 gives a comparison of spectral expansions of functions in $L^2(X/\Gamma)$ with Γ as in Table 1. For a discussion of these results, see Terras [27].

The relation between Fourier series and Fourier integrals can be expressed in terms of *Poisson's summation formula* for $f \in C_c^\infty(\mathbb{R}^2)$; i.e. f infinitely differentiable with compact support:

$$(1.1)\qquad \text{the } \mathbb{Z}^2\text{-periodization of } f = g(x) = \sum_{n\in\mathbb{Z}^2} f(x+n) = \sum_{n\in\mathbb{Z}^2} \hat{f}(n)e_n(x),$$

TABLE 1

Comparison of euclidean and noneuclidean harmonic analysis in two dimensions.

General	Euclidean	Noneuclidean
symmetric space $x \in X \cong G/K$	$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \cong \mathbb{R}^2/\{0\}$ euclidean plane	$z = x + iy \in H, x \in \mathbb{R}, y > 0$ Poincaré upper half plane
arc length ds	$ds^2 = dx_1^2 + dx_2^2$	$ds^2 = y^{-2}(dx^2 + dy^2)$
Laplacian Δ	$\Delta = \left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} \right)$	$\Delta = y^2 \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)$
G-invariant area $d\mu$	$d\mu = dx_1 \, dx_2$	$d\mu = y^{-2} \, dx \, dy$
$G =$ isometry group	$\mathbb{R}^2$	$g \in G = \mathrm{SL}(2, \mathbb{R})$ $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{matrix} a, b, c, d \in \mathbb{R} \\ ad - bc = 1 \end{matrix}$
$K =$ subgroup of G of elements fixing the origin	$K = \{0\}$	$k \in K = \mathrm{SO}(2), i = \sqrt{-1} =$ origin k orthogonal ${}^t k \cdot k = I$
Action of $g \in G$ on $x \in X$	$x \mapsto x + g$ vector addition	$z \mapsto gz = \frac{az + b}{cz + d}$ fractional linear mapping
$\Gamma =$ discrete subgroup of G	$\Gamma = \mathbb{Z}^2$	$\gamma \in \Gamma = \mathrm{SL}(2, \mathbb{Z})$ $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{matrix} a, b, c, d \in \mathbb{Z} \\ ad - bc = 1 \end{matrix}$ $\overline{\Gamma} = \Gamma / \pm I$
elementary eigenfunctions of Δ	$e_y(x) = \exp(2\pi i \, {}^t x y)$	$p_s(z) = (\mathrm{Im}\, z)^s, \ s \in \mathbb{C}$
eigenvalues	$\Delta e_y = -4\pi^2 \|y\|^2$	$\Delta p_s = s(s-1) p_s$
Helgason–Fourier transform	$\hat{f}(y) = \int_{\mathbb{R}} f(x) \overline{e_y(x)} \, dx$	$\mathscr{H}f(s, k) = \int_H f(z) \overline{p_s(kz)} \, d\mu$ $s \in \mathbb{C}, k \in K$
Fourier inversion or spectral decomposition of Δ (e.g. for $f \in L^2(X)$)	$f(x) = \hat{\hat{f}}(-x)$	$f(z)$ $= \frac{1}{4\pi} \int_{t \in \mathbb{R}} \int_{k \in K} \mathscr{H}f(s, k) p_s(kz)$ $\cdot t \tanh \pi t \, dt$
convolution (defined by convolution on G)	$\widehat{f * g} = \hat{f} \cdot \hat{g}$ (e.g. for $f, g \in L^1(\mathbb{R}^2)$)	$\mathscr{H}(f * g) = (\mathscr{H}f) \cdot (\mathscr{H}g)$ for $f, g \in L^1(H)$ with f or g K-invariant
differentiation	$\Delta f(y) = -4\pi^2 \|y\|^2 \hat{f}(y)$	$\mathscr{H}(\Delta f)(s, k) = \bar{s}(\bar{s} - 1) \mathscr{H}f(s, k)$
heat equation $\begin{Bmatrix} u_t = \Delta u \\ u(x, 0) = f(x) \end{Bmatrix}$ $u(x, t) = f * g_t$ $g_t =$ fundamental solution	$g_t(x)$ $= (4\pi t)^{-1} \exp(-\|x\|^2/4t)$	$g_t(ke^{-r} i)$ $= \sqrt{2} (4\pi t)^{-3/2} e^{-t/4}$ $\cdot \int_r^\infty \frac{b e^{-b^2/4t} \, db}{\sqrt{\cosh b - \cosh r}}$

TABLE 2
Harmonic analysis on the fundamental domain.

General	Euclidean	Noneuclidean
$\{e_\alpha\}_{\alpha\in A}$ complete orthonormal set of eigenfunctions of Δ on X/Γ	$e_a(x)=\exp(2\pi i\,{}^t a x)$ $a\in\mathbb{Z}^2$ purely discrete spectrum	*Continuous Spectrum* of Eisenstein series: $E_s(z)=\sum_{\gamma\in\left(\begin{smallmatrix}* & *\\ 0 & *\end{smallmatrix}\right)\backslash\Gamma}\operatorname{Im}(\gamma z)^s$, $\operatorname{Re}s>1$, with analytic continuation to other values of s *Discrete Spectrum* of cusp forms and constants: $v_0=(3/\pi)^{1/2}$ $\{v_n\}_{n\geqq 1}$, complete orthonormal set of cusp forms (which vanish at ∞ by definition)
$\Delta e_\alpha=\lambda_\alpha e_\alpha$	$\Delta e_a=-4\pi^2\\|a\\|^2 e_a$	$\Delta E_s=s(s-1)E_s$ $\Delta v_0=0(0-1)v_0$ $\Delta v_n=s_n(s_n-1)v_n$
$g(x)=$ $\sum_{\alpha\in A}$ or $\int (g,e_\alpha)e_\alpha(x)\sigma(\alpha)\,d\alpha$ where the integral or sum can be thought of as a Stieltjes integral and the spectral measure $\sigma(\alpha)$ comes from the asymptotics and functional equations of the e_α $(g,h)=\int_{X/\Gamma}g(x)\overline{h(x)}\,dx$	$g(x)=\sum_{a\in\mathbb{Z}^2}(g,e_a)e_a(x)$ ordinary Fourier series	$g(z)=\sum_{n\geqq 0}(g,v_n)v_n(z)$ $+\frac{1}{4\pi i}\int_{\operatorname{Re}s=1/2}(g,E_s)E_s(z)\,ds$ Roelcke–Selberg spectral decomposition of Δ on $L^2(SL(2,\mathbb{Z})\backslash H)$
Poisson summation formula $f\in C_c^\infty(K\backslash X)$ $g(x)=\sum_{\gamma\in\overline{\Gamma}}f(\gamma x)$ $=\sum_{\alpha\in A}\int \hat{f}(\alpha)\overline{e_\alpha(0)}e_\alpha(x)$ $\cdot\sigma(\alpha)\,d\alpha$ (this is also Mercer's theorem)	$\sum_{n\in\mathbb{Z}}f(x+n)$ $=\sum_{n\in\mathbb{Z}}\hat{f}(n)e_n(x)$ Here $e_n(0)=1$	$\sum_{\gamma\in\overline{\Gamma}}f(\gamma z)$ $=\sum_{n\geqq 0}\hat{f}(s_n)\overline{v_n(i)}v_n(z)$ $+\frac{1}{4\pi i}\int_{\operatorname{Re}s=1/2}\hat{f}(s)\overline{E_s(i)}E_s(z)\,ds$ $\hat{f}(x)=\int_H f(z)\overline{p_s(z)}\,d\mu$.
application to the circle problem	$N_{\mathbb{Z}^2}(x)$ $=\#\{n\in\mathbb{Z}^2 \mid d(n,0)\leqq x\}$ $\sim x,\ x\to\infty$ $d(n,0)=\\|n\\|$	$N_\Gamma(x)=$ $\#\{\gamma\in\overline{\Gamma} \mid \cosh d(\gamma i,i)\leqq x\}$ $\sim 6x,\ x\to\infty$ $d(z,w)=$ distance z to w $2\cosh d(\gamma i,i)=\operatorname{Tr}({}^t\gamma\gamma)$
theta function	$\theta_{\mathbb{Z}^2}(a)=\sum_{n\in\mathbb{Z}^2}\exp(-a\,{}^t nn)$ $=\frac{\pi}{a}\theta_{\mathbb{Z}^2}(1/a)$ $\sim\frac{\pi}{a},\ a\to 0+$	$\theta_\Gamma(a)=\sum_{\gamma\in\overline{\Gamma}}\exp\left(-\frac{a}{2}\operatorname{Tr}({}^t\gamma\gamma)\right)$ $f_a(z)=\exp(-a\cosh d(i,z))$ $\hat{f}_a(s)=2\left(\frac{2\pi}{a}\right)^{1/2}K_{\bar{s}-\frac{1}{2}}(a)$ $K_s=K$-Bessel function

with

$$(1.2)\qquad e_n(x)=\exp\{2\pi i(n_1x_1+n_2x_2)\}\quad\text{and}\quad \hat{f}(a)=\int_{\mathbb{R}^2}f(y)\overline{e_a(y)}\,dy.$$

It is possible to weaken the hypotheses on f, but there are examples of functions f for which $f,\hat{f}\in L^1(\mathbb{R})$ but $f(n)=0$ for all $n\in\mathbb{Z}$, while $\hat{f}(0)=1$, $\hat{f}(n)=0$ if $n\in\mathbb{Z}$, $n\neq 0$. Applications of the Poisson summation formula include:

i) a study of the asymptotics of the eigenvalues of the Laplace operator on $L^2(\mathbb{R}^2/\mathbb{Z}^2)$,

ii) proofs of transformation formulas of theta functions and the analytic continuation of zeta functions obtained from theta functions by Mellin transform.

Expositions of these basic facts can be found in many books (cf., for example Terras [27, §§1.3, 1.4]).

Next we will see that it is possible to use a simple lemma about eigenfunctions of integral and differential operators to obtain a noneuclidean analogue of the Poisson summation formula, using the same argument that is usually given to prove (1.1). Suppose that $f\in C_c^\infty(K\backslash H)$; i.e. $f:H\to\mathbb{C}$ is infinitely differentiable with compact support on the Poincaré upper half plane H and $f(kz)=f(z)$ for all $k\in K=SO(2)$ and $z\in H$. The *noneuclidean Poisson sum formula* says:

$$(1.3)\quad \text{the }\Gamma\text{-periodization of } f=g(z) = \sum_{\gamma\in SL(2,\mathbb{Z})/\pm I} f(\gamma z) = \sum_{n\geq 0}\hat{f}(s_n)\overline{v_n(i)}v_n(z)+\frac{1}{4\pi i}\int_{\operatorname{Re} s=1/2}\hat{f}(s)\overline{E_s(i)}E_s(z)\,ds,$$

using the notation of Tables 1 and 2. Here

$$(1.4)\qquad \hat{f}(s)=\int_H f(z)\overline{y^s}\,d\mu$$

$$=\text{the noneuclidean Fourier transform of } f\in C_c^\infty(K\backslash H).$$

This transform is also the Selberg transform (see T. Kubota [43, p. 56]).

The inversion formula for the noneuclidean Fourier transform (4) goes back to Mehler in 1881 and Fock in 1943. Harish-Chandra and Helgason have generalized this to arbitrary symmetric spaces of real semi-simple Lie groups. More information on this subject can be found in references [9], [27], for example. The noneuclidean Poisson summation formula for H/Γ compact goes back to Delsarte (1942). Applications and generalizations have been found by Bartels [2], Elstrodt, Grunewald and Mennicke [4], Hejhal [7], Kudla and Millson [14], Mennicke [18], Patterson [21], and Selberg.

One of the main results needed for a discussion of the noneuclidean Poisson summation formula is Proposition 1 below. First, a few preliminaries are necessary. More details on these preliminaries can be found in the references, particularly Helgason [9], Maass [16], Selberg [22], and Terras [27], [32].

Let

$$\mathcal{P}_n=\{Y\in\mathbb{R}^{n\times n}\,|\,{}^tY=Y,\ Y\text{ positive definite}\}.$$

Here Y positive definite means $Y[x]={}^txYx>0$ for all $x\in\mathbb{R}^n$, $x\neq 0$. One can turn the space $\mathscr{P}_n$ into a Riemannian manifold by providing it with an *arc length*:

$$(1.5)\qquad ds^2=\operatorname{Tr}\left((Y^{-1}dY)^2\right)\quad\text{for } Y=(y_{ij})_{1\leqq i,j\leqq n},\qquad dY=(dy_{ij})_{1\leqq i,j\leqq n}.$$

The *general linear group* $\mathrm{GL}(n,\mathbb{R})=\{g\in\mathbb{R}^{n\times n}\,|\det g\neq 0\}$ acts on $\mathscr{P}_n$ via

$$(1.6)\qquad Y\in\mathscr{P}_n\mapsto Y[g]={}^tgYg\in\mathscr{P}_n.$$

Set $G=\mathrm{GL}(n,\mathbb{R})$ and $K=O(n)$. One can easily show that $\mathscr{P}_n$ can be identified with the quotient space $K\backslash G$ via

$$\begin{array}{ccc} K\backslash G & \to & \mathscr{P}_n \\ Kg & \mapsto & I[g]={}^tgg. \end{array}$$

The action of G on $\mathscr{P}_n$ leaves the arc length (1.5) invariant. The *G-invariant measure* $d\mu$ on $\mathscr{P}_n$ is

$$(1.7)\qquad d\mu_n(Y)=d\mu=|Y|^{-(n+1)/2}\prod_{1\leqq i\leqq j\leqq n}dy_{ij},\qquad |Y|=\det Y.$$

The *Laplacian* on $\mathscr{P}_n$ is

$$(1.8)\qquad \Delta=\operatorname{Tr}\left(\left(Y\frac{\partial}{\partial Y}\right)^2\right)\quad\text{if } \frac{\partial}{\partial Y}=\left(\frac{1}{2}(1+\delta_{ij})\frac{\partial}{\partial y_{ij}}\right)_{1\leqq i,j\leqq n},$$

with $\delta_{ij}=0$ for $i\neq j$ and 1 for $i=j$. There are analogous G-invariant differential operators

$$L_j=\operatorname{Tr}\left(\left(Y\frac{\partial}{\partial Y}\right)^j\right),\qquad j=1,2,\cdots,n,$$

forming an algebraically independent basis for the algebra $D(\mathscr{P}_n)$ of all G-invariant differential operators on $\mathscr{P}_n$. Here a "G-invariant differential operator" is a differential operator that commutes with the action of G on $\mathscr{P}_n$.

The fundamental special functions for $D(\mathscr{P}_n)$ are the *power functions* defined for $s\in\mathbb{C}^n$, $Y\in\mathscr{P}_n$ by

$$(1.9)\qquad p_s(Y)=\prod_{j=1}^n|Y_j|^{s_j},\quad\text{where } Y=\begin{pmatrix} Y_j & * \\ * & * \end{pmatrix},\quad Y_j\in\mathscr{P}_j,$$

$|Y_j|=$ determinant of Y_j. It is easily proved that the power functions are eigenfunctions of all the differential operators in $D(\mathscr{P}_n)$.

The G-invariant integral operators of interest here are the *convolution operators* defined by convolution on the group G itself (Selberg [22] calls the kernel of such a convolution operator a point-pair invariant). In fact, we can define for $g\in C_c(\mathscr{P}_n/K)$, i.e. G continuous and K-invariant with compact support,

$$(1.10)\qquad C_gf(X)=(f*g)(X)=\int_{\mathscr{P}_n}f(Y)g(XY^{-1})\,d\mu.$$

For we can set $\tilde{f}(x)=f({}^txx)$, when $x\in G$. Then

$$\tilde{f}(b)\tilde{g}(ab^{-1})=f({}^tbb)g({}^t(ab^{-1})ab^{-1})=f({}^tbb)g({}^tb^{-1}({}^taa)b^{-1}).$$

When g is K-invariant, we can move the ${}^tb^{-1}$ around in the argument of g and obtain

$$\tilde{f}(b)\tilde{g}(ab^{-1})=f({}^tbb)g(b^{-1}{}^tb^{-1}{}^taa).$$

Properties of convolution operators.

1) The convolution operator C_g defined in (1.10) commutes with the action of G on $\mathscr{P}_n$; i.e., C_g is an invariant integral operator.

2) If $g(Y)=g(Y^{-1})$, then C_g is a self-adjoint integral operator on $L^2(\mathscr{P}_n, d\mu)$.

3) $C_{g*h}f=C_gC_hf$.

4) The convolution operators C_g, for $g\in C_c(\mathscr{P}_n/K)$, form a commutative algebra of operators.

The *Fourier transform* of a function $f\in C_c(\mathscr{P}_n/K)$ as considered by Harish-Chandra and Helgason is:

$$\hat{f}(s)=\int_{Y\in\mathscr{P}_n} f(Y)\overline{p_s(Y)}\,d\mu \quad \text{for } s\in\mathbf{C}^n. \tag{1.11}$$

The inversion formula for K-invariant functions f is due to Harish-Chandra and was extended to $f\in C_c(\mathscr{P}_n)$ by Helgason (cf. Helgason [9] and Terras [27, 32]).

PROPOSITION 1 (eigenfunctions of invariant differential and integral operators).

a) *Let $f\in C^\infty(\mathscr{P}_n)$ be an eigenfunction of all the G-invariant differential operators $L\in D(\mathscr{P}_n)$; i.e., $Lf=\lambda_Lf$, for $\lambda_L\in\mathbf{C}$. Define $s\in\mathbf{C}^n$ by $Lp_s=\lambda_Lp_s$, where p_s denotes the power function* (1.9). *If $g\in C_c^\infty(\mathscr{P}_n/K)$, then f is an eigenfunction of the convolution operator C_g in* (1.10) *with*

$$C_gf=f*g=\hat{g}(\overline{s^*})f,$$

where $\hat{g}$ is defined by (1.11) *and $s^*=(s_{n-1},\cdots,s_1,-(s_1+\cdots+s_n))$. Here $p_s(Y^{-1}[w])=p_{s^*}(Y)$, where*

$$w=\begin{pmatrix} 0 & & 1 \\ & 1 & \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & 0 \end{pmatrix}.$$

b) *Conversely, suppose $f\in C(\mathscr{P}_n)$ is an eigenfunction of all the convolution operators C_g with $g\in C_c^\infty(\mathscr{P}_n/K)$. Then f is also an eigenfunction of all the invariant differential operators.*

For a proof of this proposition see Terras [27]. The result goes back to Selberg [22] and in this context the Fourier transform $\hat{g}$ is called the Selberg transform.

Now we can discuss *the noneuclidean Poisson summation formula* (1.3) by giving a noneuclidean analogue of the usual proof of (1.1). We consider only the case $G=\mathrm{SL}(2,\mathbb{R})$ here because the generalization of the Roelcke–Selberg spectral decomposition to $\Gamma=\mathrm{SL}(n,\mathbb{Z})$ has not been so explicitly worked out, although one expects an analogous result (cf. Arthur [1], Jacquet [12], Venkov [33], and some unpublished notes of Kaori Imai).

The proof of (1.3) proceeds as follows. The idea is simply that the Roelcke–Selberg spectral decomposition of the Γ-periodization of f in (1.3) can be computed using Proposition 1. Thus we must set

$$g(z)=\sum_{\gamma\in\mathrm{SL}(2,\mathbb{Z})/\pm I} f(\gamma z), \qquad z\in H,$$

and compute, for example, the "Fourier coefficient" for g corresponding to the discrete spectrum element v_n (cf. Table 2):

$$(g,v_n)=\int_{H/\Gamma} g(z)\overline{v_n(z)}\,d\mu=(f*\bar{v}_n^{\#})(i),$$

with $v\#(gi)=v(g^{-1}i)$ for $g\in \mathrm{SL}(2,\mathbb{R})$. Using Lemma 1, we see that

$$(g,v_n)=\hat{f}(s_n)\bar{v}_n(i).$$

The calculation proceeds in exactly the same way when v_n is replaced by an Eisenstein series. Since f has compact support, the convolution integral converges absolutely.

One can extend the noneuclidean Poisson summation formula beyond smooth functions with compact support, and that is necessary for most applications.

As an application of the noneuclidean Poisson sum formula let us briefly consider a *noneuclidean analogue of the circle problem* which has been studied by many authors and generalized considerably (cf. Bartels [2], Elstrodt, Grunewald and Mennicke [4], Mennicke [18], Patterson [21], Selberg).

As usual, let $\Gamma=\mathrm{SL}(2,\mathbb{Z})$ and $\Gamma=\Gamma/\pm I$. Consider for $x>0$:

$$N_\Gamma(x)=\#\left\{\gamma\in\bar{\Gamma}\,\middle|\,\cosh d(\gamma i,i)=\frac{1}{2}\mathrm{Tr}({}^t\gamma\gamma)\leqq x\right\}. \tag{1.12}$$

Then Patterson [21] shows that

$$N_\Gamma(x)\sim 6x,\qquad x\to\infty, \tag{1.13}$$

along with an estimate for the error in this asymptotic formula. Mennicke [18] discusses a proof of analogous results involving a *noneuclidean analogue of the theta function*, defined for $a>0$ by:

$$\theta_\Gamma(a)=\sum_{\gamma\in\bar{\Gamma}}\exp\left\{\frac{-a}{2}\mathrm{Tr}({}^t\gamma\gamma)\right\}. \tag{1.14}$$

Formula (1.14) makes sense for $\Gamma=\mathrm{GL}(n,\mathbb{Z})$ or $\mathrm{Sp}(n,\mathbb{Z})$, of course, with $\bar{\Gamma}=\Gamma/\text{center}$.

In order to apply the noneuclidean Poisson sum formula (1.3) to the noneuclidean theta function (1.14), we need to compute the noneuclidean Fourier transform (1.4) of

$$f_a(Y)=\exp\left\{-\frac{a}{2}\mathrm{Tr}(Y)\right\},\qquad Y\in\mathcal{SP}_2=\{Y\in\mathcal{P}_2\,|\,|Y|=1\},$$

since we can identify the Poincaré upper half plane H and $\mathcal{SP}_2$ via

$$\begin{array}{ccc} H & \to & \mathcal{SP}_2 \\ z=x+iy & \mapsto & Y_z=\begin{pmatrix} y^{-1} & 0\\ 0 & y\end{pmatrix}\begin{bmatrix}1 & -x\\ 0 & 1\end{bmatrix}. \end{array}$$

It turns out that if $\hat{f}_a$ denotes the Fourier transform (4), then

$$\hat{f}_a(s)=\int_{z\in H}\exp\left\{\frac{-a}{2}\mathrm{Tr}(Y_z)\right\}y^{\bar{s}-2}\,dx\,dy=2\left(\frac{2\pi}{a}\right)^{1/2}K_{\bar{s}-1/2}(a). \tag{1.15}$$

Here K_s denotes the usual K-Bessel function.

As $a\to 0$, the main term on the right-hand side of the noneuclidean Poisson sum formula for f_a comes from $v_0=(3/\pi)^{1/2}$. A Tauberian theorem completes the proof of

(1.13) (cf. [27, §3.7]) or the references mentioned above for details and alternate discussions).

A second application of Proposition 1 arises in the evaluation of a special integral that occurs in multivariate statistics. First we need some definitions. A *spherical function* on $\mathscr{P}_n$ is a K-invariant eigenfunction $f(Y)$ of the invariant differential operators on $\mathscr{P}_n$ such that $f(I)=1$, $I=$ the identity matrix. Harish-Chandra proved (cf. Helgason [9] or Terras [27]) that if dk denotes Haar measure on $K=O(n)$, with

$$\int_K dk=1,$$

then a spherical function on $\mathscr{P}_n$ has the form

$$h_s(Y)=\int_{k\in K} p_s(Y[k])\,dk. \tag{1.16}$$

Spherical functions which are polynomials in the entries of Y (e.g. for $s\in(\mathbb{Z}^+)^n$) are of interest in multivariate statistics. Such spherical functions have been studied by many authors (cf. Muirhead [19]). Here we wish to indicate a simple way of evaluating the following integral defined for $B\in\mathscr{P}_n$, $r\in\mathbb{C}$, $s\in\mathbb{C}^n$ (with the variables suitably restricted for convergence) by:

$$M(B,r,s)=\int_{Y\in\mathscr{P}_n}\exp\{-\mathrm{Tr}(BY^{-1})\}|BY^{-1}|^r h_s(Y)\,d\mu. \tag{1.17}$$

This integral appears, for example, in the book of Muirhead [19, Thm. 7.2.5]. Since the integral is a convolution, Proposition 1 says that

$$M(B,r,s)=\hat{f}(\bar{s}^*)h_s(B),\quad \text{where } f(Y)=|Y|^r\exp\{-\mathrm{Tr}(Y)\}.$$

It is quite easy to evaluate the Fourier transform (1.11) of f in terms of the *gamma function* for $\mathscr{P}_n$, defined by $s\in\mathbb{C}^n$ by:

$$\Gamma_n(s)=\int_{Y\in\mathscr{P}_n} p_s(Y)\exp\{-\mathrm{Tr}(Y)\}\,d\mu. \tag{1.18}$$

Then, by a result of Ingham [36] and Siegel [37, Vol. I, pp. 326–405] in the 1920's and 30's, we have the factorization:

$$\Gamma_n(x)=\pi^{n(n-1)/4}\prod_{j=1}^{n}\Gamma\left(s_j+\cdots+s_n-\frac{j-1}{2}\right), \tag{1.19}$$

where Γ denotes the ordinary gamma function ($\Gamma=\Gamma_1$). A proof of (1.19) can be found in [27]. The preceding discussion proves:

THEOREM 1 (Muirhead's integral formula).

$$M(B,r,s)=\Gamma_n(s_{n-1},\cdots,s_1,-(s_1,\cdots,s_n+r))h_s(B).$$

Convergence of (1.17) *occurs for* $\mathrm{Re}(s_j+\cdots+s_n+r)<0, j=1,\cdots,n$.

Our final application of Proposition 1 is to the computation of analogues of gamma functions arising in the problem of obtaining analytic continuations of Eisenstein series for $GL(n,\mathbb{Z})$ via analogues of Riemann's method of theta functions—a method used by Riemann to continue the Riemann zeta function and obtain its functional equation. In order to keep our discussion at a simple level, we will only discuss a special case that does not really illustrate the power of the method.

Before describing these results, we need a few more definitions. When generalizing the Roelcke–Selberg spectral decomposition to $\mathrm{GL}(n,\mathbb{Z})$, one must study special functions called *automorphic forms* v for $\Gamma=\mathrm{GL}(n,\mathbb{Z})$ defined to be $v:\mathcal{P}_n\to\mathbb{C}$ having the following three properties:

(1.20)
1) v is an eigenfunction for all $L\in D(\mathcal{P}_n)$; i.e., $Lv=\lambda_L v$, $\lambda_L\in\mathbb{C}$;
2) $v(Y[A])=v(Y)$ for all $A\in\Gamma$;
3) v has at most polynomial growth at infinity; i.e., it grows at most like a power function $p_s(Y)$.

We shall use the notation $A(\Gamma,\lambda)$ for the *space of such automorphic forms* corresponding to a given eigenvalue system λ. Let

$$\mathcal{S}\mathcal{P}_n=\{Y\in\mathcal{P}_n \,|\, |Y|=1\}.$$

Then $\mathcal{S}\mathcal{P}_n$ is the symmetric space corresponding to the Lie group $\mathrm{SL}(n,\mathbb{R})$. We will also consider functions $v:\mathcal{S}\mathcal{P}_n\to\mathbb{C}$ to be automorphic forms for $\mathrm{GL}(n,\mathbb{Z})$ if they satisfy the same 3 conditions with $\mathcal{S}\mathcal{P}_n$ replacing $\mathcal{P}_n$. And we use $A^0(\Gamma,\lambda)$ *to denote the space of automorphic forms for* Γ *on* $\mathcal{S}\mathcal{P}_n$.

Consider the *partial Iwasawa decomposition* of $Y\in\mathcal{P}_n$ corresponding to $n=n_1+n_2$, with $1\leqq n_1,\ n_2\leqq n$, given by

$$(1.21)\qquad \begin{aligned} &Y=a[v],\quad \text{with } a=\begin{pmatrix} a_1(Y) & 0\\ 0 & a_2(Y)\end{pmatrix},\quad a_i(Y)\in\mathcal{P}_{n_i},\\ &v=\begin{pmatrix} I_{n_1} & X\\ 0 & I_{n_2}\end{pmatrix},\quad X\in\mathbb{R}^{n_1\times n_2},\quad I_{n_j}=n_j\times n_j \text{ identity matrix.}\end{aligned}$$

Let $f_j\in A(\mathrm{GL}(n_j,Z),\lambda_j)$ and form the *gamma function* associated to f_1 and f_2:

$$(1.22)\qquad \Gamma(f_1,f_2)=\int_{Y\in\mathcal{P}_n} f_1(a_1(Y))f_2(a_2(Y))\exp\{-\mathrm{Tr}(XY^{-1})\}\,d\mu.$$

The integral will converge for suitably chosen f. Since $f_1(a_1(Y))f_2(a_2(Y))$ is an eigenfunction of the invariant differential operators, it follows from Proposition 1, that there is $s=s(f_1,f_2)\in\mathbb{C}^n$ such that

$$(1.23)\qquad \Gamma(f_1,f_2)=f_1(a_1(Y))f_2(a_2(Y))\Gamma_n(s(f_1,f_2)).$$

This considerably simplifies a proof of a special case of this result given in Maass [16, §7].

It is possible to form many analogues of the Eisenstein series that appears in Table 2. One such *Eisenstein series* for $\mathrm{GL}(n,\mathbb{Z})$ is associated to two automorphic forms $f_i\in A(\mathrm{GL}(n_i,\mathbb{Z}),\lambda_i)$, for $i=1,2$, with $n=n_1+n_2$, via:

$$(1.24)\qquad E_{n_1,n_2}(f_1,f_2;\,Y)=\sum_{A\in\mathrm{GL}(n,\mathbb{Z})/P(n_1,n_2)} f_1(a_1(Y[A]))f_2(a_2(Y[A])).$$

Here $P(n_1,n_2)$ denotes the *parabolic subgroup* of $GL(n,\mathbb{Z})$ consisting of matrices with block form:

$$(1.25)\qquad \begin{pmatrix} A_1 & A_{12}\\ 0 & A_2\end{pmatrix}\quad \text{for } A_i\in\mathrm{GL}(n_i,\mathbb{Z}),\ A_{12}\in\mathbb{Z}^{n_1\times n_2}.$$

We can always decompose $Y \in \mathscr{P}_n$ as

$$Y = |Y|^{1/n} Y^0, \quad \text{where } Y^0 \in \mathscr{S}\mathscr{P}_n. \tag{1.26}$$

Suppose that

$$f_i(Y) = |Y|^{-r_i} f_i^0(Y^0), \quad \text{for some } r_i \in \mathbf{C}, f_i^0 \in A^0(\mathrm{GL}(n_i, \mathbf{Z}), \lambda_i). \tag{1.27}$$

Then the Eisenstein series (1.24) converges when $\mathrm{Re}(r_1 - r_2) > n/2$, assuming that the growth conditions on the f_i^0 imply integrability on the fundamental domains. This can be proved using a generalization of the integral test (cf. Terras [27, §4.4, Exercise 23] or [29]).

When the Hecke operators are introduced, it is possible to relate Eisenstein series (1.24) with zeta functions. Since Hecke operators were discussed in [30] and [27], we will merely recall the definitions. Suppose that

$$f : \mathscr{S}\mathscr{P}_n / \mathrm{GL}(n, \mathbf{Z}) \to \mathbf{C}.$$

Then, for any positive integer m, *the mth Hecke operator* T_m is defined by:

$$T_m f(Y) = \sum_{\substack{A \in \mathbf{Z}^{n \times n} \, \mathrm{rk}\, n / \mathrm{GL}(n, \mathbf{Z}), \\ |A| = m}} f\left((Y[A])^0\right), \tag{1.28}$$

where we use the notation (1.26). Suppose that f is an eigenfunction of all the Hecke operators; i.e., $T_m f = u_m f$, for some $u_m \in \mathbf{C}$, $m \geqq 1$. Form the *Dirichlet series*:

$$L_f(s) = \sum_{m \geqq 1} u_m m^{-s}, \qquad \mathrm{Re}\, s > \frac{n}{4}. \tag{1.29}$$

Define the *zeta function* associated to $f_i \in A(\mathrm{GL}(n_i, \mathbf{Z}), \lambda_i)$, $i = 1, 2$, by:

$$Z(f_1, f_2; Y) = \sum_{A \in \mathbf{Z}^{n \times n} \, \mathrm{rk}\, n / P(n_1, n_2)} f_1(a_1(Y[A])) f_2(a_2(Y[A])). \tag{1.30}$$

Then we have:

THEOREM 2 (The relation between Eisenstein series and zeta functions formed from two lower rank automorphic forms).

$$Z(f_1, f_2; Y) = E(f_1, f_2; Y) L_{f_1}(2r_1) L_{f_2}(2r_2)$$

holds if f_i *are given by* (1.27), $E(f_1, f_2; Y) = E_{n_1, n_2}(f_1, f_2; Y)$ *defined by* (1.24), *and* $Z(f_1, f_2; Y)$ *given by* (1.30).

Proof. Use the same argument as in the proof of Terras [30, Prop. 1]. The main idea is that the same discussion which gives explicit representatives for the quotient summed over in the definition of the mth Hecke operator yields a complete set of representatives for $A \in \mathbf{Z}^{n \times n} \mathrm{rk}\, n / \mathrm{GL}(n, \mathbf{Z})$ of the form $A = BC$, where

$$B \in \mathrm{GL}(n, \mathbf{Z}) / P(n_1, n_2)$$

and

$$C = \begin{pmatrix} C_1 & 0 \\ 0 & C_2 \end{pmatrix}, \qquad C_i \in \mathbf{Z}^{n_i \times n_i} \mathrm{rk}\, n_i / \mathrm{GL}(n_i, \mathbf{Z}).$$

It is now possible to relate the Eisenstein series (1.24) with a *Euclidean theta function* defined for $X, Y \in \mathcal{P}_n$ by:

$$(1.31) \qquad \theta(Y,X) = \sum_{A \in \mathbb{Z}^{n \times n}} \exp\{-\pi \operatorname{Tr}(Y[A]X)\}.$$

Riemann's method of analytic continuation of zeta functions leads us to define the following *matrix Mellin transform*, for r_i suitably restricted so that convergence occurs:

$$(1.32) \qquad J(f_1,f_2) = \int_{\mathcal{P}_n/P(n_1,n_2)} f_1(a_1(X)) f_2(a_2(X)) \theta_n(Y, X^{-1})\, d\mu$$

where

$$\theta_m(Y,X) = \sum_{A \in \mathbb{Z}^{n \times n} \text{ rk } m} \exp\{-\pi \operatorname{Tr}(Y[A]X)\}.$$

It follows from (1.23) that

$$(1.33) \qquad J(f_1,f_2) = \Gamma_n(s(f_1,f_2)) Z(f_1,f_2;\, \pi Y).$$

To see this, you must switch the quotient modulo $V \in P(n_1,n_2)$ from $\mathcal{P}_n$ to $\mathbb{Z}^{n \times n}$. This requires the following calculation:

$$\operatorname{Tr}(Y[AV]X^{-1}) = \operatorname{Tr}(Y[A]X^{-1}[{}^tV]) = \operatorname{Tr}\left(Y[A](X[V^{-1}])^{-1}\right).$$

Since $P(n_1,n_2)$ is a group it follows that $V^{-1} \in P(n_1,n_2)$. So, if $s(f_1,f_2)$ is determined by the eigenvalues of f_1, f_2, (as in (1.23)), then

$$(1.34) \qquad J(f_1,f_2) = \pi^{-(n_1r_1+n_2r_2)} \Gamma_n(s(f_1,f_2)) Z(f_1,f_2;\, Y).$$

It is then possible to use Riemann's method of analytic continuation of the Riemann zeta function, as modified by Selberg (cf. Maass [16], Terras [27], [29] [30]) to study the analytic continuation and functional equations of the Eisenstein series. In this special case, however, the method does not lead to anything new. For we have:

THEOREM 3 (an easy functional equation of the Eisenstein series). *When f_i is as in* (1.27) *and* $\operatorname{Re}(r_1 - r_2) > n/2$, *we have*

$$E_{n_1,n_2}(f_1,f_2;\, Y) = E_{n_2,n_1}(f_2^*,f_1^*;\, Y^{-1}) \quad \textit{if } f^*(Y) = f(Y^{-1}).$$

Proof. Note that if

$$w = \begin{pmatrix} 0 & I_{n_1} \\ I_{n_2} & 0 \end{pmatrix}, \qquad Y = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix} \begin{bmatrix} I_{n_1} & X \\ 0 & I_{n_2} \end{bmatrix},$$

$$Y^{-1}[w] = \begin{pmatrix} a_2^{-1} & 0 \\ 0 & a_1^{-1} \end{pmatrix} \begin{bmatrix} I_{n_2} & -X \\ 0 & I_{n_1} \end{bmatrix},$$

and observe that both series converge for $\operatorname{Re}(r_1 - r_2) > n/2$.

In order to obtain nontrivial functional equations by Riemann's method, one must add an extra dimension, as in Maass [16, p. 267]. We will not discuss all the details here.

Speh [25] has considered analogous results with applications to the determination of the residual spectrum of $\mathrm{GL}(n,\mathbb{Z})$ (cf. also Jacquet [12]).

2. K-Bessel functions for $\mathcal{P}_n$. We consider K-Bessel functions for $\mathcal{P}_n$ which arise in Fourier expansions of automorphic forms for $\mathrm{GL}(n,\mathbb{Z})$. These and related special functions have been studied from many different points of view (see Bengtson [3], Bump [35], Goodman and Wallach [5], Gross, Holman and Kunze [6], Herz [10], Jacquet [13], Muirhead [19], Shalika [24], Terras [27, §4.3]).

For $1 \leqq m \leqq n-1$, define the abelian group:

$$(2.1)\qquad N(m,n-m)=\left\{ U=\begin{pmatrix} I_m & X \\ 0 & I_{n-m}\end{pmatrix} \middle| X\in\mathbb{R}^{m\times(n-m)}\right\}.$$

Then we say that $f:\mathcal{P}_n\to\mathbb{C}$ is a *K-Bessel function* if:

1) $f(Y[\begin{smallmatrix} I & X \\ 0 & I\end{smallmatrix}])=\exp\{2\pi i\mathrm{Tr}({}^tNX)\}f(Y)$, for all $Y\in\mathcal{P}_n$, $X\in\mathbb{R}^{m\times(n-m)}$; i.e., f transforms according to the above character of $N(m,n-m)$;

2) f is an eigenfunction of all the invariant differential operators $L\in D(\mathcal{P}_n)$;

3) f has at most polynomial growth at infinity.

The growth requirement in 3) may appear to be somewhat weak at first GLance. In certain cases, we actually obtain exponential decay (see (2.6) below). However, we are also including singular cases. When $n=2$, for example, we find that

$$f\left(\begin{pmatrix} y & 0 \\ 0 & 1/y\end{pmatrix}\begin{bmatrix} 1 & x \\ 0 & 1\end{bmatrix}\right)=cy^{1/2}K_{s-1/2}(2\pi|R|y)\quad \text{if } R\neq 0,$$

where K denotes the ordinary K-Bessel function. As $y\to\infty$, this function approaches 0 exponentially. But, as $y\to 0$, the function blows up like a polynomial in $1/y$. Moreover, if $R=0$, the function is ay^s+by^{1-s}. The ordinary K-Bessel functions are discussed in detail in Lebedev [44].

We are studying a matrix entry corresponding to a representation of $\mathrm{GL}(n,\mathbb{R})$ induced from a character of $N(m,n-m)$. Kirillov [38] shows that irreducible unitary representations of the nilpotent Lie group N of all upper triangular $n\times n$ matrices with real entries and one's on the diagonal come from inducing characters of this type with $m=[n/2]$. Such representations of N are infinite dimensional. There are also one dimensional representations of N that arise in the theory of Whittaker functions to be considered at the end of this section.

It is easy to give examples of such K-Bessel functions. The *main example of a matrix argument K-Bessel function* is:

$$(2.2)\quad k_{m,n-m}(s|Y,R)=\int_{X\in\mathbb{R}^{m\times(n-m)}}\exp\{2\pi i\mathrm{Tr}({}^tRX)\}p_{-s}\left(Y^{-1}\begin{bmatrix} I & 0 \\ {}^tX & I\end{bmatrix}\right)dX.$$

Here $s\in\mathbb{C}^n$ with s_j restricted to suitable half planes, $Y\in\mathcal{P}_n$, $R\in\mathbb{R}^{m\times(n-m)}$, $1\leqq m\leqq n-1$. The power function p_s is defined in (1.9). Formula (2.2) is useful for demonstrating that $k_{m,n-m}(\cdot|Y,\cdot)$ is indeed an eigenfunction of all the invariant differential operators on $\mathcal{P}_n$.

Another K-Bessel function is obtained by choosing $A\in\mathrm{GL}(m,\mathbb{R})$, $B\in\mathrm{GL}(n-m,\mathbb{R})$, $C\in\mathbb{R}^{m\times(n-m)}$ such that $R=ACB$. Then define:

$$(2.3)\qquad f(Y)=k_{m,n-m}\left(s\middle|Y\begin{bmatrix} {}^tA^{-1} & 0 \\ 0 & {}^tB\end{bmatrix},C\right).$$

It is easily seen that

$$f\left(Y\begin{bmatrix} I_m & X \\ 0 & I_{n-m}\end{bmatrix}\right)=\exp\{2\pi i\operatorname{Tr}({}^t(ACB)X)\}f(X).$$

We shall see that when $n \geqq 3$, it is the functions (2.3) that appear in the Fourier coefficients of Eisenstein series, rather than the simpler functions (2.2). It does not appear to be possible to move the A and B over to the C-variable, in general.

In order to understand the convergence properties of the integrals above, it is useful to define a *matrix argument K-Bessel function*:

$$K_m(s|V,W)=\int_{Y\in\mathscr{P}_m} p_s(Y)\exp\{-\operatorname{Tr}(VY+WY^{-1})\}\,d\mu_n(Y). \tag{2.4}$$

Here $s\in\mathbf{C}^m$, $W,V\in\mathscr{P}_m$. The function can be extended to singular W if the s_j are suitably restricted.

Example 1.

$$k_{1,1}\left((s,0)\middle|\begin{pmatrix} y^{-1} & 0 \\ 0 & y\end{pmatrix},a\right)=\begin{cases} 2\pi^{1/2}|\pi a|^{s-1/2}\Gamma(s)^{-1}y^{1/2}K_{s-1/2}(2\pi|a|y), & a\neq 0,\\ \Gamma(\frac{1}{2})\Gamma(s-\frac{1}{2})\Gamma(s)^{-1}y^{1-s}, & a=0,\end{cases}$$

$$K_1(s|v,w)=2(w/v)^{s/2}K_s(2\sqrt{vw}).$$

Here $K_s(y)$ denotes the ordinary K-Bessel function.

Example 2.

$$\begin{aligned} k_{2,1}((s_1,s_2,0)|I,(a,0)) &= \iint (1+x_1^2)^{-s_1}(1+x_1^2+x_2^2)^{-s_2}\exp(2\pi iax_1)\,dx_1\,dx_2 \\ &= \int (1+x_1^2)^{-s_1-s_2+1/2}\exp(2\pi iax_1)\,dx_1\int(1+y^2)^{-s_2}dy \\ &= k_{1,1}(s_1+s_2-\tfrac{1}{2}|a)B(\tfrac{1}{2},s_2-\tfrac{1}{2}),\end{aligned}$$

where we have used the substitution $x_2=(1+x_1^2)^{1/2}y$ and $B(p,q)=\Gamma(p)\Gamma(q)/\Gamma(p+q)$.

In general, one does not appear to have such a factorization.

Tom Bengtson [3] proves the following formula relating the two Bessel functions:

(2.5)

$$\Gamma_m(-s^*)k_{m,n-m}\left(s_{\text{ext}}\middle|\begin{pmatrix} V & 0 \\ 0 & W\end{pmatrix},R\right)=\pi^{m(n-m)/2}|W|^{m/2}K_m(s\#\,|W[\pi\,{}^tR],V^{-1}),$$

where

$$s\in\mathbf{C}^m,\quad 1\leqq m\leqq n-1, \qquad s_{\text{ext}}=(s,0)\in\mathbf{C}^n,$$

$$s\#=-s+\left(0,\cdots,0,\frac{n-m}{2}\right),\qquad s^*=\left(s_{m-1},\cdots,s_1,-\sum_{j=1}^m s_j\right).$$

Bengtson [3] develops many more properties of these K-Bessel functions. Assuming that V and W lie in $\mathscr{P}_n$, it is shown by Terras [27] that if $\underline{a}$ denotes the smallest element in the set of eigenvalues of V and W, then (for fixed $s\in\mathbf{C}^m$):

$$K_m(s|V,W)=O(a^{-m(m+1)/4}e^{-2ma})\quad\text{as } a\to\infty. \tag{2.6}$$

However, when we study Fourier coefficients of Eisenstein series, we obtain functions of the form (2.3). Then Bengtson's formula (2.5) shows that we will need to study

$$K_m(s\,|\,V,W), \quad \text{with } V \text{ singular.}$$

When $n=3$, $m=2$, the function of interest is

$$K_2(s\,|\,{}^t qq, I), \quad \text{with } q\in\mathbb{R}^2.$$

Bengtson [3] shows that this converges when

$$\begin{aligned} &\operatorname{Re} s_2 \text{ and } \operatorname{Re}(s_1+s_2) < -\tfrac{1}{2}, && \text{if } q=0, \\ &\operatorname{Re} s_2 < 0, && \text{if } q_2=0,\ q_1\neq 0, \\ &\operatorname{Re}(s_1+s_2)<0, && \text{if } q_2\neq 0. \end{aligned} \tag{2.7}$$

Our first concern is the relation between these Bessel functions for $GL(n,\mathbb{R})$ and the gamma function for $\mathrm{Sp}(n,\mathbb{R})$, the *symplectic group* defined by

$$G^* = \mathrm{Sp}(n,\mathbb{R}) = \left\{ g\in \mathrm{SL}(2n,\mathbb{R}) \,\middle|\, {}^t gJg=J, \text{ where } J=\begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}\right\}.$$

The symmetric space for $\mathrm{Sp}(n,\mathbb{R})$ can be identified with the space

$$\mathscr{P}_n^* = \{ W\in \mathscr{P}_{2n} \,|\, W\in \mathrm{Sp}(n,\mathbb{R})\} \tag{2.8}$$

(cf. Terras [27, Ch. 5]). The *partial Iwasawa decomposition* of such a symplectic matrix $W\in\mathscr{P}_n^*$ is:

$$W=\begin{pmatrix} Y^{-1} & 0 \\ 0 & Y\end{pmatrix}\begin{bmatrix} I_n & X \\ 0 & I_n\end{bmatrix} \quad \text{for } Y\in\mathscr{P}_n,\ {}^tX=X\in\mathbb{R}^{n\times n}. \tag{2.9}$$

The basic eigenfunctions of the G^*-invariant differential operators on $\mathscr{P}_n^*$ are the *power functions*:

$$p_s(W)=p_s(Y), \quad \text{for } s\in\mathbb{C}^n, \quad \text{with } W \text{ as in (2.9)}, p_s \text{ as in (1.9)}. \tag{2.10}$$

The *G^*-invariant measure* on $\mathscr{P}_n^*$ is

$$d\mu^*(W)=|Y|^{-(n+1)/2}d\mu_n(Y)\,dX, \tag{2.11}$$

where W is as in (2.9) and $d\mu_n(Y)$ is defined in (1.7). Then the *symplectic analogue of the gamma function* is:

$$\Gamma_n^*(s)=\int_{W\in\mathscr{P}_n^*} p_s(W)\exp\{-\mathrm{Tr}(W)\}\,d\mu_n^*(W). \tag{2.12}$$

THEOREM 4 (the symplectic gamma function is a Bessel function for $\mathrm{GL}(n)$).

$$\Gamma_n^*(s)=\pi^{n(n+1)/4}K_n(s-(0,\cdots,0,\tfrac{1}{2})|I,I).$$

Proof. By definition,

$$\Gamma_n^*(s)=\int_{Y\in\mathscr{P}_n}\int_{\substack{X\in\mathbb{R}^{n\times n},\\ {}^tX=X}} \exp\{-\mathrm{Tr}(Y+Y^{-1}+Y^{-1}[X])\}\,p_s(Y)|Y|^{-(n+1)/2}d\mu_n(Y)\,dX.$$

Perform the integral over X and obtain:

$$\Gamma_n^*(s)=\pi^{n(n+1)/4}\int_{Y\in\mathcal{P}_n}\exp\{-\mathrm{Tr}(Y+Y^{-1})\}p_s(Y)|Y|^{-1/2}d\mu_n(Y).$$

Formula (2.4) completes the proof.

The *noneuclidean analogue of the theta function corresponding to the Siegel modular group* $\mathrm{Sp}(n,\mathbb{Z})=\{A\in\mathrm{Sp}(n,\mathbb{R})|A$ has integral entries$\}=\Gamma$ is

$$(2.13)\qquad \theta_\Gamma(a)=\sum_{\gamma\in\Gamma/\pm I}\exp\{-a\mathrm{Tr}({}^t\gamma\gamma)\}\quad\text{for } a>0.$$

If one can develop the analogue of the Roelcke–Selberg spectral decomposition of the invariant differential operators on $L^2(\mathcal{P}_n^*/\mathrm{Sp}(n,\mathbb{Z}))$, then the noneuclidean Poisson sum formula for $\mathrm{Sp}(n,\mathbb{Z})$ must relate $\theta_\Gamma(a)$ with sums and integrals of functions:

$$(\pi/a)^{n(n+1)/4}K_n(s-(0,\cdots,0,\tfrac{1}{2})|aI,aI).$$

Hopefully, extending ideas of Mennicke [18], it will be possible to use Poisson summation for $\mathrm{Sp}(n,\mathbb{Z})$ and $\mathrm{SL}(n,\mathbb{Z})$ to study nonholomorphic cusp forms for these groups.

As a last application of matrix K-Bessel functions, consider the problem of estimating Fourier coefficients of automorphic forms $f\in A^0(\mathrm{GL}(n,\mathbb{Z}),\lambda)$. Since

$$f\left(Y\begin{bmatrix} I_m & A\\ 0 & I_{n-m}\end{bmatrix}\right)=f(Y),\quad\text{for all } A\in\mathbb{Z}^{m\times(n-m)},\ Y\in\mathcal{P}_n,$$

it follows that $f(Y)$ is a periodic function of period one in each entry of X when Y has the following *partial Iwasawa decomposition*:

$$(2.14)\qquad Y=\begin{pmatrix} V & 0\\ 0 & W\end{pmatrix}\begin{bmatrix} I_m & X\\ 0 & I_{n-m}\end{bmatrix},\quad V\in\mathcal{P}_m,\quad W\in\mathcal{P}_{n-m},\quad X\in\mathbb{R}^{m\times(n-m)}.$$

Evidence from Goodman and Wallach [5], Shalika [24], Imai and Terras [11], Takhtadjan and Vinogradov [26], and Terras [31] leads one to believe that the *Fourier expansion of* f should have the form:

$$(2.15)\qquad f(Y)=\sum_{N\in\mathbb{Z}^{m\times(n-m)}}\ \sum_{\substack{A,B,C\\ N=ACB}}\exp(2\pi i\mathrm{Tr}({}^tNX))a_f(A,B,C)$$
$$\times k_{m,n-m}\left(s\left|\begin{pmatrix} V[{}^tA^{-1}] & 0\\ 0 & W[{}^tB]\end{pmatrix},C\right.\right).$$

Here the matrices $A\in\mathrm{GL}(m,\mathbb{Z})/P$, $B\in\mathrm{GL}(n-m,\mathbb{Z})/P'$, $C\in\mathbb{Z}^{m\times(n-m)}$ with P, $P'=$ parabolic subgroups. When f is an Eisenstein series, the coefficients $a_f(A,B,C)$ are analogues of singular series or divisor functions. Such expansions are the natural analogues of those obtained by Siegel [37, Vol. II, p. 115] and the starting point for the expansions considered by Bump [35], Jacquet, Piatetskii-Shapiro and Shalika [40] and many others. It would be useful to study the Fourier coefficients of other automorphic forms such as those considered by Ash or those corresponding to cubic fields found by Gelbart, Jacquet, and Piatetskii-Shapiro. It would also be interesting to obtain cusp forms from integrals of theta functions using an analogue of the construction of Marie-France Vignéras [34].

We can use the theory of matrix argument K-Bessel functions to estimate the $a_f(A,B,C)$ for bounded $f(Y)$, generalizing an argument of Hecke (cf. Terras [27, §3.5, Exercise 15]).

THEOREM 5 (estimates of Fourier coefficients). *Suppose that* $f \in A^0(\mathrm{GL}(3,\mathbb{Z}),\lambda)$ *is bounded with Fourier expansion* (2.15). *In this case* $B=1$ *and there is a constant* $\kappa>0$ *such that*

$$|a_f(A,1,C)| \leqq \kappa \|C\|^{2\,\mathrm{Re}(s_1+2s_2+1)}.$$

Here we assume that the eigenvalues of f are the same as those of the Bessel function $k_{2,1}(s\,|\,Y,R)$ *under the action of the invariant differential operators.*

Proof. Suppose $|f(Y)| \leqq L$, for all $Y \in \mathscr{SP}_3$. Then by the definition of a Fourier coefficient,

$$\left|a_f(A,1,C)k_{2,1}\left(s\left|\begin{pmatrix} V[{}^tA^{-1}] & 0 \\ 0 & W\end{pmatrix}\right.,C\right)\right|$$
$$\leqq \int_{X\in(\mathbb{R}/\mathbb{Z})^2}\left|f\left(\begin{pmatrix} V & 0 \\ 0 & W\end{pmatrix}\begin{bmatrix} I & X \\ 0 & I\end{bmatrix}\right)\exp\{2\pi i\,\mathrm{Tr}({}^tXN)\}\right|dX \leqq L.$$

So we find that

$$|a_f(A,1,C)| \leqq \frac{L}{\left|k_{2,1}\left(s\left|\begin{pmatrix} V[{}^tA^{-1}] & 0 \\ 0 & W\end{pmatrix}\right.,C\right)\right|}.$$

By (2.5), if $s\# = -s+(0,\tfrac{1}{2})$, $s^* = (s_1, -(s_1+s_2))$,

$$k_{2,1}\left(s\left|\begin{pmatrix} V[{}^tA^{-1}] & 0 \\ 0 & 1/V\end{pmatrix}\right.,C\right) = \pi\Gamma_2(-s^*)^{-1}|V|^{-1}K_2\left(s\#\,\big|\,|V|^{-1}\pi^2C{}^tC, V^{-1}[A]\right).$$

Choose $V \in \mathscr{P}_2$ such that $V = t^{-1}V^0$ with $t = |V|^{-1/2} = \|C\|^{-2}$, $V^0 = A{}^tA$. Then, assuming $s\#$ satisfies the inequalities (2.7):

$$K_2\left(s\#\,|\pi^2|V|^{-1}C{}^tC, V^{-1}[A]\right) = K_2\left(s\#\,|\pi^2t^{-2}C{}^tC, tI\right) = p_{s\#}(tI)K_2\left(s\#\,|\pi^2t^{-1}C{}^tC, I\right).$$

By choosing $t = \|C\|^2$, we insure that the argument $Ct^{-1/2}$ lies in the unit sphere. Since a continuous function takes a minimum on a compact set, we can write:

$$M = \min\left\{K_2\left(s\#\,|\pi^2u{}^tu, I\right)|\ \|u\|=1\right\}.$$

Then

$$|a_f(A,1,C)| \leqq \frac{L|\Gamma(-s^*)|\|C\|^4}{M\pi|p_{-s_1,-s_2+1/2}(\|C\|^2I)|} = \kappa\|C\|^{2\,\mathrm{Re}(s_1+2s_2+1)}.$$

If $-s^* \in \mathbb{C}^2$ is chosen as in the Harish-Chandra–Helgason inversion formula, then we need $\mathrm{Re}\,s_1 = \mathrm{Re}\,s_2 = \tfrac{1}{2}$. Note that

$$K_2\left(s\#\,|q{}^tq, I_2\right) \overset{\text{as } q\to 0}{\to} \Gamma_2(-s_1, s_1+s_2-1/2) = \Gamma(s_2-\tfrac{1}{2})\Gamma(s_1+s_2-1).$$

It follows that by multiplying t by a suitable constant, we can assure that

$$K_2(s\#\,|\pi^2 t^{-1} C'C, I) \neq 0.$$

Note also that $\Gamma_2(-s^*)^{-1} = \Gamma(s_2)^{-1}\Gamma(s_1+s_2-1/2)^{-1} \neq 0$ when $\mathrm{Re}\, s_i = 1/2$, for $i = 1, 2$.

Questions.

1) Can we similarly generalize other arguments from classical automorphic forms that depend on Fourier expansions?

2) Can one compute the Fourier expansion of the Eisenstein series (1.24)?

Remarks on connections with Whittaker functions and Fourier expansions of automorphic forms in series of Whittaker functions. Whittaker functions and Fourier expansions of automorphic forms as sums of these functions are discussed by Bump [35], Jacquet [13], Jacquet, Piatetskii-Shapiro and Shalika [40], Proskurin [39], and Shalika [24]. For $r \in \mathbb{R}^{n-1}$, $Y \in \mathcal{P}_n$, and $s \in \mathbb{C}^n$ with $\mathrm{Re}\, s$ suitably restricted for convergence, the Whittaker function can be defined by:

$$\text{(2.16)} \qquad W(s\,|\,Y, r) = \int_{n \in N} p_{-s}(Y^{-1}[{}^t n]) \exp\left(2\pi i \sum_{i=1}^{n-1} r_i x_{i,i+1}\right) dx,$$

where N is the nilpotent group of matrices n of the form:

$$n = \begin{pmatrix} 1 & & x_{ij} \\ & \ddots & \\ 0 & & 1 \end{pmatrix}.$$

The exponential appearing in the integral is easily seen to be a one-dimensional character of N. The integral itself can be easily shown to converge wherever the numerator in the Harish-Chandra c-function (a function giving the spectral measure for the Fourier transform on $\mathcal{P}_n$) converges (see Jacquet [13] and Terras [27]). One also sees easily that

$$W(s\,|\,Y[n], r) = \exp\left(2\pi i \sum_{i=1}^{n-1} r_i x_{i,i+1}\right).$$

The $W(s\,|\,Y, r)$ are analogous to Eisenstein series with the largest possible number of complex variables ($s \in \mathbb{C}^n$); i.e., the highest dimensional part of the spectrum of the Laplacian. Thus one can use techniques developed by Selberg [23] to obtain $n!$ functional equations of the Whittaker functions (see Bump [35]) by writing them as integrals of "lower rank" Whittaker type functions such as the k-Bessel functions (2.2). This is analogous to writing an Eisenstein series with n complex variables as a sum of Eisenstein series with a smaller number of complex variables (see Terras [27]).

More explicitly, one can write the Whittaker functions as Fourier transforms of k-Bessel functions (2.2). For example, when $n = 3$:

$$W(s\,|\,Y, r) = \int_{x_{12} \in \mathbb{R}} k_{2,1}\left(s \,\middle|\, Y\begin{bmatrix} 1 & -x_{12} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, (0, r_2)\right) \exp(2\pi i r_1 x_{12})\, dx_{12}.$$

Then one can obtain functional equations of the Whittaker functions from those of the lower rank functions, and vice versa, since the k-Bessel function is also a Fourier transform of the Whittaker function. This same sort of idea relates the Fourier expansions (2.15) with those involving Whittaker functions. For example, in the case of a cusp form for $\mathrm{GL}(3,\mathbb{Z})$, Bump [35] obtains the equivalent of the expansion:

$$f(Y)=\sum_{A\in\mathrm{SL}(2,\mathbb{Z})/\left(\begin{smallmatrix}* & *\\ 0 & *\end{smallmatrix}\right)}\ \sum_{r_1,r_2\geq 1} a_{r_1r_2}W\left(s\,\middle|\,Y\begin{bmatrix}A & 0\\ 0 & 1\end{bmatrix},(r_1,r_2)\right).$$

This is obtained by starting with the expansion:

$$f(Y)=\sum_{r_1,r_2\in\mathbb{Z}} f_r(Y),\quad\text{where}$$

$$f_r\left(Y\begin{bmatrix}I & x\\ 0 & 1\end{bmatrix}\right)=\exp(2\pi i\,{}^t rx)f_r(Y),\quad\text{for all } x\in\mathbb{R}^2.$$

Then one notes that if $A\in\mathrm{SL}(2,\mathbb{Z})$, $r\in\mathbb{C}^2$, we have

$$f_r(Y)=f_{{}^tAr}\left(Y\begin{bmatrix}A & 0\\ 0 & 1\end{bmatrix}\right).$$

Furthermore, if $r_2\in\mathbb{Z}$, one has

$$f_{(r_1,r_3)}\left(Y\begin{bmatrix}1 & r_2 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\right)=f_{(r_1,r_3-r_1r_2)}(Y).$$

So, if $r_1=0$, the coefficient must be invariant under

$$Y\mapsto Y\begin{bmatrix}1 & r_2 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}.$$

Thus one can Fourier expand the f_r's with respect to the x_{12} variable in the matrices in the nilpotent group N. That goes from Bessel to Whittaker functions. It is the multiplicity one theorem of Shalika [24] that says the resulting functions must be multiples of Whittaker functions.

REFERENCES

[1] J. Arthur, *The trace formula in invariant form*, Ann. of Math., 114 (1981), pp. 1–74.

[2] H.-J. Bartels, *Nichteuklidische Gitterpunktprobleme und Gleichverteilung in linearen algebraischen Gruppen*, Comment. Math. Helvetici, 57 (1982), pp. 158–172.

[3] T. Bengtson, *Bessel functions on* $\mathcal{P}_n$, Pacific J. Math., 108 (1983), pp. 19–30.

[4] J. Elstrodt, F. Grunewald and J. Mennicke, *Discontinuous groups on three-dimensional hyperbolic space: analytical theory and arithmetic applications*, Russ. Math. Surveys, 38 (1983), pp. 137–168.

[5] R. Goodman and N. Wallach, *Conical vectors and Whittaker vectors*, J. Funct. Anal., 39 (1980), pp. 199–279.

[6] K. I. Gross, W. J. Holman and R. A. Kunze, *A new class of Bessel functions and applications in harmonic analysis*, Proc. Symp. Pure Math., 35, Amer. Math. Soc., Providence, RI, 1979, 407–415.

[7] D. Hejhal, *The Selberg Trace Formula for* $\mathrm{PSL}(2,\mathbb{R})$, *Vols.* I, II, Lecture Notes in Mathematics, 548, 1001, Springer, New York, 1983.

[8] ———, *Some Dirichlet series whose poles are related to cusp forms* (I, II), preprints.

[9] S. Helgason, *Lie groups and symmetric spaces*, in Battelle Rencontres, C. M. DeWitt and J. A. Wheeler, eds., Benjamin, New York, 1968, pp. 1–71.

[10] C. Herz, *Bessel functions of matrix argument*, Ann. of Math., 61 (1955), pp. 474–523.

[11] K. Imai and A. Terras, *Fourier expansions of Eisenstein series for* $\mathrm{GL}(3,\mathbb{Z})$, Trans. Amer. Math. Soc., 273 (1982), pp. 679–694.

[12] H. Jacquet, *Residual spectrum, especially for* GL(n), lecture, Univ. Maryland, Nov., 1982.

[13] ______, *Les fonctions de Whittaker associées aux groupes de Chevalley*, Bull. Soc. Math. France, 95 (1967), pp. 243–309.

[14] S. Kudla and J. Millson, *Harmonic differentials and closed geodesics on Riemann surfaces*, Inv. Math., 54 (1979), pp. 193–211.

[15] R. Langlands, *Eisenstein Series*, Lecture Notes in Mathematics 544, Springer, New York, 1976.

[16] H. Maass, *Siegel's Modular Forms and Dirichlet Series*, Lecture Notes in Mathematics 216, Springer, New York, 1971.

[17] ______, *Über eine neue Art von nichtanalytischen automorphen Funktionen und die Bestimmung Dirichletscher Reihen durch Funktionalgleichung*, Math. Ann., 121 (1949), pp. 141–183.

[18] J. Mennicke, *Vorträge über Selbergs Spurformel* I, U. Bielefeld, W. Germany.

[19] R. Muirhead, *Aspects of Multivariate Statistical Theory*, John Wiley, New York, 1978.

[20] M. S. Osborne and G. Warner, *The Theory of Eisenstein Systems*, Academic Press, New York, 1981.

[21] S. J. Patterson, *A lattice-point problem in hyperbolic space*, Mathematika, 22 (1975), pp. 81–88; 23 (1976), p. 227.

[22] A. Selberg, *Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series*, J. Indian Math. Soc., 20 (1956), pp. 47–87.

[23] ______, *A new type of zeta function connected with quadratic forms*, Report of the Institute in the Theory of Numbers, Univ. Colorado, Boulder, 1959, pp. 207–210.

[24] J. A. Shalika, *The multiplicity one theorem for* GL(n), Ann. Math., 100 (1974), pp. 171–193.

[25] B. Speh, *Unitary representations of* GL(n, $\mathbb{R}$) *with non-trivial* ($\mathfrak{g}$, K) *cohomology*, preprint.

[26] L. A. Takhtadzhyan and A. I. Vinogradov, *Theory of Eisenstein series for the group* SL(3, $\mathbb{R}$) *and its application to a binary problem*, J. Soviet Math., 18 (1982), pp. 293–323.

[27] A. Terras, *Harmonic analysis on symmetric spaces and applications*, to appear.

[28] ______, *Noneuclidean harmonic analysis*, SIAM Rev., 24 (1982), pp. 159–193.

[29] ______, *Integral formulas and integral tests for series of positive matrices*, Pacific J. Math., 89 (1980), pp. 471–490.

[30] ______, *On automorphic forms for the general linear group*, Rocky Mountain J. Math., 12 (1982), pp. 123–143.

[31] ______, *Fourier coefficients of Eisenstein series of one complex variable for the special linear group*, Trans. Amer. Math. Soc., 205 (1975), pp. 97–114.

[32] ______, *Analysis on positive matrices as it might have occurred to Fourier*, Lecture Notes in Mathematics 899, Springer, New York, 1981, pp. 442–478.

[33] A. B. Venkov, *Spectral theory of automorphic functions, the Selberg zeta function, and some problems of analytic number theory*, Russian Math. Surveys, 34 (1979), pp. 79–153.

[34] M. F. Vignéras, *Séries thêta des formes quadratiques indéfinies*, Lecture Notes in Mathematics 627, Springer, New York, 1977, pp. 227–240.

[35] D. Bump, *Automorphic forms on* GL(3, $\mathbb{R}$), Ph.D. Thesis, Univ. of Chicago, Chicago, 1982, to appear in Springer Lecture Notes.

[36] A. E. Ingham, *An integral which occurs in statistics*, Proc. Cambridge Phil. Soc., 29 (1933), pp. 271–276.

[37] C. L. Siegel, *Gesammelte Abhandlungen, Vols.* I–IV, Springer, New York, 1979.

[38] A. A. Kirillov, *Unitary representations of nilpotent Lie groups*, Russian Math. Surveys, 17 (1962), pp. 53–104.

[39] N. V. Proskurin, *Expansions of automorphic functions*, Proc. Steklov Inst., 116 (1982), pp. 119–141.

[40] H. Jacquet, I. I. Piatetski-Shapiro, and J. Shalika, *Automorphic forms on* GL(3), I, II, Ann. of Math., 109 (1979), pp. 169–212, 213–258.

[41] H. Hurt, *Geometric Quantization in Action*, D. Reidel, Amsterdam, 1983.

[42] D. Wallace, *Selberg's trace formula and units in higher degree number fields*, Ph.D. thesis, Univ. California, San Diego, 1982.

[43] T. Kubota, *Elementary Theory of Eisenstein Series*, Wiley, New York, 1973.

[44] N. N. Lebedev, *Special Functions and their Applications*, Dover, New York, 1972.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

THE LIMITING EIGENVALUE DISTRIBUTION OF A MULTIVARIATE F MATRIX*

JACK W. SILVERSTEIN[†]

Abstract. Let X_{ij}, Y_{ij} $i,j=1,2,\cdots$ be i.i.d. $N(0,1)$ random variables and for positive integers p,m,n, let $\overline{X}_p=(X_{ij})$ $i=1,2,\cdots,p$; $j=1,2,\cdots,m$, and $\overline{Y}_p=(Y_{ij})$ $i=1,2,\cdots,p$; $j=1,2,\cdots,n$. Suppose further that $p/m\to y>0$ and $p/n\to y'\in(0,\frac{1}{2})$ as $p\to\infty$. In [5], [6] it is shown that the empirical distribution function of the eigenvalues of $(1/m\,\overline{X}_p\,\overline{X}_p^T)(1/n\,\overline{Y}_p\,\overline{Y}_p^T)^{-1}$ converges i.p. as $p\to\infty$ to a nonrandom d.f.

In the present paper the limiting d.f. is derived.

1. Introduction. Let X_{ij}, $i,j=1,2,\cdots$ be i.i.d. $N(0,1)$ random variables, and for any positive integers p,m, let $W_p=\overline{X}_p\overline{X}_p^T$, $\overline{X}_p=(X_{ij})$ $i=1,2,\cdots,p; j=1,2,\cdots,m$, be the $p\times p$ Wishart matrix $W(I,m)$. It is well known [1], [2], [4] that if $p/m\to y>0$ as $p\to\infty$, then the empirical distribution function F_p of the eigenvalues of $(1/m)W_p$ (i.e. $F_p(x)=(1/p)$ (# of eigenvalues of $(1/m)W_p\leqq x$)) converges a.s. for every $x\geqq 0$ to a nonrandom d.f. F_y, where for $0<y\leqq 1$, F_y has density

(1.1)
$$f_y(x)=\begin{cases}\dfrac{1}{2\pi yx}\sqrt{\left(x-(1-\sqrt{y})^2\right)\left((1+\sqrt{y})^2-x\right)} & \text{for } (1-\sqrt{y})^2<x<(1+\sqrt{y})^2,\\ 0 & \text{otherwise,}\end{cases}$$

and for $1<y<\infty$ F_y has mass $1-1/y$ at zero and density f_y on $((1-\sqrt{y})^2,(1+\sqrt{y})^2)$.

In [6] it is shown that the empirical d.f. of $(1/m)W_pT_p$, under certain conditions on the $p\times p$ matrix T_p, converges in probability to a nonrandom d.f. $\overline{F}$. The specific conditions on T_p are the following:

1) T_p is symmetric positive definite a.s.
2) W_p and T_p are independent.
3) If G_p is the empirical d.f. of the eigenvalues of T_p, then for every positive integer k, $\int x^k dG_p(x)$ converges in L^2 to a nonrandom value H_k, where $\sum_{k=1}^{\infty}H_{2k}^{-1/2k}=\infty$.

The moments $\{E_k\}_{k=1}^{\infty}$ of $\overline{F}$ are also derived. They are given by

$$E_k=\sum_{w=1}^{k}y^{k-w}\sum_{\substack{n_1+\cdots+n_w=k-w+1,\\ n_1+2n_2+\cdots+wn_w=k}}\frac{k!}{n_1!\cdots n_w!w!}H_1^{n_1}\cdots H_w^{n_w}.\tag{1.2}$$

No further information of $\overline{F}$ is given.

* Received by the editors March 4, 1983, and in revised form August 8, 1983. This research was supported by the National Science Foundation under grant MCS-8101703.

[†] Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27650.

In [5] it is shown the conditions are satisfied for $T_p=((1/n)\,\underline{W}_p)^{-1}$ where $\underline{W}_p$ is $W(I,n)$, W_p and $\underline{W}_p$ are independent, and $p/n\to y'\in(0,1/2)$ as $p\to\infty$. In particular, 3) is verified by showing

$$\int x^k dG_p(x)\xrightarrow{L^2}\int_{(1-\sqrt{y'})^2}^{(1+\sqrt{y'})^2}\frac{1}{x^k}\,dF_{y'}(x).$$

The matrix $((1/m)W_p)((1/n)\,\underline{W}_p)^{-1}$ is seen to be a multivariate F matrix, fundamental to statistical work in multivariate analysis.

In this paper we will derive the limiting empirical d.f. of $((1/m)W_p)((1/n)\,\underline{W}_p)^{-1}$. We will show for any $y'\in(0,1)$, if

$$H_k=\int_{(1-\sqrt{y'})^2}^{(1+\sqrt{y'})^2}\frac{1}{x^k}\,dF_{y'}(x),\qquad k=1,2,\cdots,$$

then $\{E_k\}_{k=1}^{\infty}$ are the moments of the d.f. $F_{y,y'}$, where for $0<y\leqq 1$ $F_{y,y'}$ has density

$$f_{y,y'}(x)$$

$$=\begin{cases}\dfrac{(1-y')\sqrt{\left(x-\left(\dfrac{1-\sqrt{1-(1-y)(1-y')}}{1-y'}\right)^2\right)\left(\left(\dfrac{1+\sqrt{1-(1-y)(1-y')}}{1-y'}\right)^2-x\right)}}{2\pi x(xy'+y)}\\[2ex] \qquad\text{for }\left(\dfrac{1-\sqrt{1-(1-y)(1-y')}}{1-y'}\right)^2<x<\left(\dfrac{1+\sqrt{1-(1-y)(1-y')}}{1-y'}\right)^2,\\[2ex] 0\qquad\text{otherwise.}\end{cases}$$

and for $1<y<\infty$ $F_{y,y'}$ has mass $1-1/y$ at zero and density $f_{y,y'}$ on

$$\left(\left(\frac{1-\sqrt{1-(1-y)(1-y')}}{1-y'}\right)^2,\left(\frac{1+\sqrt{1-(1-y)(1-y')}}{1-y'}\right)^2\right).$$

The derivation of $F_{y,y'}$ will be handled in the next section by first evaluating a general expression for $E(e^{sX})$ $s\in\mathbb{C}$, where X is a random variable having moments $\{E_k\}$, and $\{H_k\}$ are the moments of a random variable Y having support on a closed interval on $\mathbb{R}^+$ bounded away from zero. This expression will be seen to involve an integral of a function in the complex plane depending on the generating function of the moments of Y^{-1}. Then $F_{y,y'}$ will be determined by evaluating the integral when Y^{-1} has d.f. $F_{y'}$.

2. Derivation of $F_{y,y'}$. Assume that $\{H_k\}$ are the moments of the random variable Y having support on $[a,b]$ with $0<a<b<\infty$. Let $G(z)=E((1-zY)^{-1})$, $z\in\mathbb{C}$. Then G is analytic on $\mathbb{C}-[1/b,1/a]$ and for $|z|<1/b$, $G(z)=\sum_{k=0}^{\infty}H_kz^k$ $(H_0\equiv 1)$. Let $G_I(z)=E((1-zY^{-1})^{-1})$. Then G_I is analytic on $\mathbb{C}-[a,b]$. Moreover, we have $G_I(z)=1-G(1/z)$, $z\in\mathbb{C}-[a,b]$.

Let X be a random variable having moments $\{E_k\}$ given by (1.2). We may ignore the question of whether $\{E_k\}$ are the moments of a random variable since the following steps will be reversible and we will wind up with $F_{y,y'}$, a proper probability d.f. Expanding $E(e^{sX})$, $s \in \mathbf{C}$, in a formal power series around $s=0$ we have

(1.3)

$$E(e^{sX}) = \sum_{k=0} \frac{E_k s^k}{k!} = 1 + \sum_{k=1}^{\infty} s^k \sum_{w=1}^{k} \frac{y^{k-w}}{w!} \sum_{\substack{n_1+\cdots+n_w=k-w+1,\\ n_1+\cdots+wn_w=k}} \frac{H_1^{n_1}\cdots H_w^{n_w}}{n_1!\cdots n_w!}$$

$$= 1 + \sum_{k=1}^{\infty} s^k \sum_{w=1}^{k} \frac{y^{k-w}}{w!} \sum_{\substack{n_2+2n_3+\cdots+(w-1)n_w=w-1,\\ k-(2n_2+\cdots+wn_w)\geq 0}} \frac{H_2^{n_2}\cdots H_w^{n_w}}{n_2!\cdots n_w!}$$

$$\cdot \frac{H_1^{(k-(2n_2+\cdots+wn_w))}}{(k-(2n_2+\cdots+wn_w))!}$$

$$= 1 - \frac{1}{y} + \frac{1}{y} e^{ysH_1} + \sum_{w=2}^{\infty} \frac{y^{-w}}{w!} \sum_{n_2+\cdots+(w-1)n_w=w-1} \frac{H_2^{n_2}\cdots H_w^{n_w}}{n_2!\cdots n_w!}$$

$$\sum_{k\geq \max(w,2n_2+\cdots+wn_w)} \frac{(sy)^k H_1^{(k-(2n_2+\cdots+wn_w))}}{(k-(2n_2+\cdots+wn_w))!}.$$

Notice when $w \geq 2$ and $n_2+\cdots+(w-1)n_w = w-1$, $2n_2+\cdots+wn_w \geq w$. Therefore

(1.4)

$$E(e^{sX}) = 1 - \frac{1}{y} + \frac{1}{y} e^{ysH_1}$$

$$+ e^{ysH_1} \sum_{w=2}^{\infty} \frac{y^{-w}}{w!} \sum_{n_2+\cdots+(w-1)n_w=w-1} \frac{H_2^{n_2}\cdots H_w^{n_w}}{n_2!\cdots n_w!} (sy)^{2n_2+\cdots+wn_w}$$

$$= 1 - \frac{1}{y} + \frac{1}{y} e^{ysH_1} \sum_{n=0}^{\infty} \frac{s^n}{(n+1)!} \sum_{m_1+2m_2+\cdots+nm_n=n} \frac{(ysH_2)^{m_1}\cdots(ysH_{n+1})^{m_n}}{m_1!\cdots m_n!}.$$

Notice that

$$\sum_{m_1+\cdots nm_n=n} \frac{(ysH_2)^{m_1}\cdots(ysH_{n+1})^{m_n}}{m_1!\cdots m_n!},$$

defined to be 1 when $n=0$, is the coefficient of z^n in the series expansion about $z=0$ of $\exp(ys\sum_{k=1}^{\infty} H_{k+1}z^k) = \exp(ys((G(z)-1)/z - H_1))$. Note also that $1/(n+1)!$ is the coefficient of z^n in the expansion about $z=0$ of $(e^z-1)/z$. Both functions are analytic in a neighborhood of the origin, independent of y and s. Therefore we can write ([3, p. 158])

$$(1.5)\quad E(e^{sX}) = 1 - \frac{1}{y} + \frac{1}{y2\pi i} e^{ysH_1} \oint_{|z|=r<1/b} \frac{(e^{s/z}-1)}{s/z} e^{ys((G(z)-1)/z-H_1)} \left(\frac{1}{z}\right) dz$$

$$= 1 - \frac{1}{y} + \frac{1}{sy2\pi i} \oint_{|z|=r<1/b} e^{s/z} e^{ys((G(z)-1)/z)} dz.$$

Making the substitution $z \to 1/z$ we have

$$(1.6) \qquad E(e^{sX}) = 1 - \frac{1}{y} + \frac{1}{sy2\pi i} \oint_{|z|=r>b} e^{sz} e^{ysz(G(1/z)-1)} z^{-2}\, dz$$

$$= 1 - \frac{1}{y} + \frac{1}{sy2\pi i} \oint_{|z|=r>b} e^{sz - yszG_I(z)} z^{-2}\, dz.$$

Using integration by parts we have

$$(1.7) \qquad E(e^{sX}) = 1 - \frac{1}{y} + \frac{1}{y2\pi i} \oint_{|z|=r>b} \frac{d}{dz}\left(z(1-yG_I(z))e^{sz(1-yG_I(z))}\right)\left(\frac{1}{z}\right) dz.$$

Provided

$$(1.8) \qquad v = z(1 - yG_I(z))$$

is invertible along $|z| = r$, we make the substitution (1.8) and arrive at

$$(1.9) \qquad E(e^{sX}) = 1 - \frac{1}{y} + \frac{1}{y2\pi i} \oint_{|z(v)|=r>b} e^{sv} \frac{1}{z(v)}\, dv.$$

Since $G_I(z) \to 0$ as $|z| \to \infty$, for any $\delta \in (0,1)$ we have for all r sufficiently large

$$(1.10) \qquad (1-\delta)|z| \leqq |v| \leqq (1+\delta)|z|$$

along the contour.

To derive $F_{y,y'}$, $0 < y' < 1$, we apply (1.9) to the case when Y^{-1} has density $f_{y'}$. Using the identity

$$(1.11) \qquad \int_c^d \frac{\sqrt{(x-c)(d-x)}}{x}\, dx = \frac{\pi}{2}(\sqrt{d} - \sqrt{c})^2$$

valid for $0 \leqq c < d$, it is straightforward to show, first for z real, $z > (1 - \sqrt{y'})^{-2}$, and therefore for all $z \in \mathbb{C} - [(1+\sqrt{y'})^{-2}, (1-\sqrt{y'})^{-2}]$,

$$(1.12) \qquad G_I(z) = \frac{1}{2\pi y'} \int_{(1-\sqrt{y'})^2}^{(1+\sqrt{y'})^2} \frac{1}{(1-xz)x} \sqrt{\left(x - (1-\sqrt{y'})^2\right)\left((1+\sqrt{y'})^2 - x\right)}\, dx$$

$$= \frac{1 - z(1-y') + (1-y')\sqrt{\left(z - (1+\sqrt{y'})^{-2}\right)\left(z - (1-\sqrt{y'})^{-2}\right)}}{2y'z}$$

where we will interpret all square roots of the form

$$(1.13) \qquad \sqrt{(z-a_1)(z-a_2)}, \qquad a_1, a_2 \in \mathbb{R}, \quad a_1 < a_2$$

to be positive on (a_2, ∞) and to vary continuously off this interval. Notice then, that the square root will be negative for $z \in (-\infty, a_1)$.

Solving for z in (1.8) we find

$$(1.14)\qquad z=\frac{(2y'/y+(1-y'))v+1-y\pm\sqrt{(v(1-y')+(1-y))^2-4v}}{2(y'/y+1-y')}$$

$$=\frac{(2y'/y+(1-y'))v+1-y\pm(1-y')\sqrt{(v-b_1)(v-b_2)}}{2(y'/y+1-y')}$$

where

$$b_1=\left(\frac{1-\sqrt{1-(1-y)(1-y')}}{1-y'}\right)^2,\qquad b_2=\left(\frac{1+\sqrt{1-(1-y)(1-y')}}{1-y'}\right)^2.$$

Notice in (1.14) if the plus sign in front of the square root is used we would have $z\sim v$ for v large, whereas if the minus sign is used, then $z\sim(y'/y)/((y'/y)+(1-y'))$. Therefore, for r in (1.9) sufficiently large (1.8) is invertible along $|z|=r$ and we have

$$(1.15)\qquad z(v)=\frac{(2y'/y+(1-y'))v+1-y+(1-y')\sqrt{(v-b_1)(v-b_2)}}{2(y'/y+1-y')}$$

and

$$(1.16)\qquad \frac{1}{z(v)}=\frac{(2y'/y+(1-y'))v+1-y-(1-y')\sqrt{(v-b_1)(v-b_2)}}{2v(vy'/y+1)}.$$

Integrating $e^{sv}/z(v)$ along contours as in Fig. 1 when $y\neq 1$, and letting the two horizontal lines approach the real axis, we get (noting the discontinuity of the square root across $[b_1,b_2]$)

$$(1.17)\qquad E(e^{sX})=1-\frac{1}{y}+\frac{1}{y2\pi i}\oint_{|z+y/y'|=r_1<y/y'}e^{sv}\frac{1}{z(v)}dv$$

$$+\frac{1}{y2\pi i}\oint_{|z|=r_2<\min(y/y',b_1)}e^{sv}\frac{1}{z(v)}dv$$

$$+\frac{1}{2\pi}\int_{b_1}^{b_2}e^{sx}\frac{(1-y')\sqrt{(x-b_1)(b_2-x)}}{x(xy'+y)}dx.$$

For $y=1$ the limiting inner contour should not encompass the origin, and we will get (1.17) except the second integral will not appear.

We see that when $v=-y/y'$, the numerator of $1/z(v)$ is zero. Therefore the first integral in (1.17) vanishes. When $v=0$ the numerator of $1/z(v)$ is $2(1-y)$ when $0<y\leqq 1$, and is zero when $y>1$. Therefore, the term involving the second integral in (1.17) is $(1/y-1)I^{(y)}_{(0,1]}$, where I_A is the indicator function on the set A.

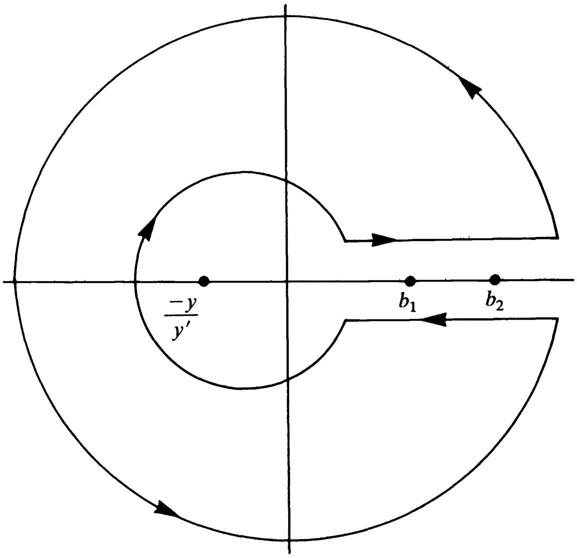


FIG. 1.

We therefore have

$$E(e^{sX}) = \left(1 - \frac{1}{y}\right) I^{(y)}_{(1,\infty)} + \int_{-\infty}^{\infty} e^{sx} f^{(x)}_{y,y'} dx. \tag{1.18}$$

Using the fact that $F_{y,y'}$ is a proper probability d.f. we conclude that (1.18) for $s = it$, $t \in \mathbb{R}$, is the characteristic function of the random variable X with d.f. $F_{y,y'}$, so that the d.f. of X must be $F_{y,y'}$.

REFERENCES

[1] U. GRENANDER AND J. W. SILVERSTEIN, *Spectral analysis of networks with random topologies*, SIAM J. Appl. Math, 32 (1977), pp. 499–519.

[2] D. JONSSON, *Some limit theorems for the eigenvalues of a sample covariance matrix*, J. Multivariate Anal., 12 (1982), pp. 1–38.

[3] E. C. TITCHMARSH, *The Theory of Functions*, 2nd ed., Oxford Univ. Press, London, 1939.

[4] K. W. WACHTER, *The strong limits of random matrix spectra for sample matrices of independent elements*, Ann. Probab., 6 (1978), pp. 1–18.

[5] Y. Q. YIN, Z. D. BAI AND P. R. KRISHNAIAH, *Limiting behavior of the eigenvalues of a multivariate F matrix*, J. Multivariate Anal., 13 (1983), pp. 508–516.

[6] Y. Q. YIN AND P. R. KRISHNAIAH, *A limit theorem for the eigenvalues of product of two random matrices*, J. Multivariate Anal., 13 (1983), pp. 489–507.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

019

SOME SUMMATION FORMULAE FOR NONTERMINATING BASIC HYPERGEOMETRIC SERIES*

A. VERMA† AND V. K. JAIN‡

Abstract. Basic hypergeometric series extensions of some classical results on hypergeometric series are obtained. These include non-terminating extensions of summation formulas of Watson and Whipple and a pair of Cayley–Orr type identities.

AMS-MOS subject classification (1980). Primary 33A35

Key words. basic hypergeometric series, q series identities

1. Introduction. Andrews [2] obtained a q-analogue of a terminating version of Watson and Whipple's summation formulae for ${}_3F_2[1]$:

$$(1.1)\qquad {}_4\Phi_3\left[\begin{matrix} a,b,\sqrt{c},-\sqrt{c}; & q,q \\ \sqrt{abq},-\sqrt{abq},c & \end{matrix}\right]=a^{n/2}\Pi\left[\begin{matrix} aq,bq,\frac{cq}{a},\frac{cq}{b}; & q^2 \\ q,abq,cq,\frac{cq}{ab} & \end{matrix}\right],$$

where $b=q^{-n}$, n a nonnegative integer and

$$(1.2)\qquad {}_4\Phi_3\left[\begin{matrix} a,\frac{q}{a},\sqrt{c},-\sqrt{c}; & q,q \\ -q,e,\frac{c}{e}q & \end{matrix}\right]=q^{n(n+1)/2}\Pi\left[\begin{matrix} ea,\frac{eq}{a},\frac{acq}{e},\frac{cq^2}{ae}; & q^2 \\ e,eq,\frac{cq}{e},\frac{cq^2}{e} & \end{matrix}\right],$$

where $a=q^{-n}$. Later on Jain [9] obtained a q-analogue of terminating Watson and Whipple summation formulae due to Bailey [5] in the form

$$(1.3)\qquad \sum_{r=0}^{n}\frac{[a;q]_r[b;q]_r[q^{-2n};q^2]_rq^r}{[q;q]_r[abq;q^2]_r[q^{-2n};q]_r}=\frac{[aq;q^2]_n[bq;q^2]_n}{[q;q^2]_n[abq;q^2]_n}$$

and

$$(1.4)\qquad {}_4\Phi_3\left[\begin{matrix} a,\frac{q}{a},-q^{-n},q^{-n}; & q,q \\ -q,b,\frac{q^{1-2n}}{b} & \end{matrix}\right]=\frac{[ab;q^2]_n[bq/a;q^2]_n}{[b;q]_{2n}},$$

*Received by the editors March 1, 1982, and in revised form July 18, 1983.

† Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G2H1

‡ Department of Mathematics, Barailley College, Barailley, U. P., India

and gave an alternative proof of (1.1). Andrews [2] showed that the summation formulae (1.1) and (1.2) hold only if the ${}_4\phi_3$ series involved are terminating basic hypergeometric series. In §2 of this note we prove a summation formulae for a nonterminating well-poised ${}_8\phi_7$,

$$(1.5)\quad {}_8W_7\left[\frac{-abc}{\sqrt{q}}; a^2, b^2, c, -c, \frac{-ab\sqrt{q}}{c}; q, \frac{c\sqrt{q}}{ab}\right]$$

$$= \Pi\left[\begin{matrix} -abc\sqrt{q}, -\dfrac{c\sqrt{q}}{ab}; & q \\ -\dfrac{ac\sqrt{q}}{b}, -\dfrac{bc\sqrt{q}}{a} & \end{matrix}\right] \Pi\left[\begin{matrix} a^2q, b^2q, \dfrac{c^2q}{a^2}, \dfrac{c^2q}{b^2}; & q^2 \\ \dfrac{c^2q}{a^2b^2}, q, a^2b^2q, c^2q & \end{matrix}\right]$$

and

$$(1.6)\quad {}_8W_7\left[-c; a, \frac{q}{a}, c, -d, -\frac{q}{d}; q, c\right]$$

$$= \Pi\left[\begin{matrix} -c, -cq; & q \\ cd, -\dfrac{cq}{a}, -ac, \dfrac{cq}{d} & \end{matrix}\right] \cdot \Pi\left[\frac{cdq}{a}, acd, \frac{acq}{d}, \frac{cq^2}{ad}; q^2\right],$$

where ${}_{p+3}W_{p+2}[a; b_1, b_2, \cdots, b_p; q, z]$ denotes the well-poised basic hypergeometric series

$${}_{p+3}\phi_{p+2}\left[\begin{matrix} a, q\sqrt{a}, -q\sqrt{a}, b_1, b_2, \cdots, b_p; & q, z \\ \sqrt{a}, -\sqrt{a}, \dfrac{aq}{b_1}, \dfrac{aq}{b_2}, \cdots, \dfrac{aq}{b_p} & \end{matrix}\right].$$

(1.5) and (1.6) may be regarded as a q-analogue of the nonterminating version of Watson and Whipple summation formulae for nonterminating ${}_3F_2(+1)$ to which they reduce on replacing a, b, c, d by q^a, q^b, q^c and q^d respectively and letting $q \to 1^-$. However (1.5) on setting $b = q^{-n/2}$ and transforming the well-poised ${}_8\phi_7$ by Watson's q-analogue of Whipple's transformation [4, 8.5(2)] $a \to -acq^{-n/2-1/2}$, $c \to -aq^{1/2-n/2}/c$, $g \to q^{-n}$, $d \to -c$, $e \to a^2$, $f \to c^2$) reduces to (1.1). On the other hand setting $a = q^{-n}$ in (1.6) and then once again transforming the well-poised ${}_8\phi_7$ by [4, 8.5(2)] ($a \to -c, c \to -d, d \to -\frac{q}{d}, e \to c, f \to q^{1+n}, g \to q^{-n}$) reduces to (1.2) on replacing c and d by $\sqrt{c}$ and $e/\sqrt{c}$ respectively.

It is of interest to note that (1.6) is a q-analogue of the following summation formula for nonterminating well-poised ${}_6F_5(-1)$ due to Whipple [13, (14.1)] to which it reduces on replacing a, c and d by q^a, $-q^c$ and $-q^d$ respectively and letting $q \to 1^-$

$$(1.7)$$

$${}_6F_5\left[\begin{matrix} a, 1+\dfrac{a}{2}, \dfrac{1}{2}+x, \dfrac{1}{2}-x, \dfrac{1}{2}+y, \dfrac{1}{2}-y; & -1 \\ \dfrac{a}{2}, \dfrac{1}{2}+a-x, \dfrac{1}{2}+a+x, \dfrac{1}{2}+a-y, \dfrac{1}{2}+a+y & \end{matrix}\right]$$

$$= \frac{\pi}{2^{2a-1}}\Gamma\left[\begin{matrix} \dfrac{1}{2}+a+x, \dfrac{1}{2}+a-x, \dfrac{1}{2}+a+y, \dfrac{1}{2}+a-y; \\ a, 1+a, \dfrac{1+a+x+y}{2}, \dfrac{1+a+x-y}{2}, \dfrac{1+a-x+y}{2}, \dfrac{1+a-x-y}{2} \end{matrix}\right].$$

On the other hand replacing a,b and c by q^a, $-q^{1/2+x-a}$ and $-q^y$ respectively in (1.5) and letting $q\to 1^-$ yields a known summation formula for nonterminating very well-poised ${}_6F_5(-1)$ due to Whipple [13; (15.73)]:

$$(1.8)\qquad {}_6F_5\left[\begin{matrix} x+y,\dfrac{2+x+y}{2},2a,1+2x-2a,y,1+x-y; & -1 \\ \dfrac{x+y}{2},1+x+y-2a,y-x+2a,1+x,2y & \end{matrix}\right]$$

$$=\Gamma\left[\begin{matrix} 1+x,\dfrac{1}{2},y+\dfrac{1}{2},2a+y-x,1+x+y-2a; \\ a+y-x,a+\dfrac{1}{2},1+x-a,y-a+\dfrac{1}{2},1+x+y \end{matrix}\right].$$

In §3 using (1.5) and (1.6) the sum of a bilateral ${}_8\psi_8$ is obtained which is a q-analogue of M. Jackson's [6] bilateral analogue of Watson and Whipple summation formulae.

The note is concluded by obtaining a q-analogue of two transformations of Bailey [3] and applying them to find q-analogues of Cayley–Orr type identities due to Bailey [3].

2. Proof of (1.5) and (1.6). In the transformation [12,(5.1)]:

$$(2.1)\qquad {}_{10}W_9\left[-a;a,b,-b,-c,c,-e^2q^n,q^{-n};q,\frac{a^2q^3}{(bce)^2}\right]$$

$$=\frac{[-aq;q]_n[e^2;q]_n}{a^n[-q;q]_n[e^2/a;q]_n}\,{}_5\Phi_4\left[\begin{matrix} a,aq,e^4q^{2n},\dfrac{a^2q^2}{b^2c^2},q^{-2n}; & q^2,q^2 \\ e^2,e^2q,\dfrac{a^2q^2}{b^2},\dfrac{a^2q^2}{c^2} & \end{matrix}\right],$$

setting $e=aq/bc$ and then replacing a,b,c by a^2, $a\sqrt{q}/b$ and $ab\sqrt{q}/c$ respectively and transforming the resulting well-poised ${}_{10}\phi_9$ on the left-hand side by [4, 8.5(1)] and the ${}_4\phi_3$ on the right-hand side by [10; 8.3]

$$(2.2)\qquad {}_4\Phi_3\left[\begin{matrix} a,b,c,q^{-n}; & q,q \\ e,g,h & \end{matrix}\right]=\frac{\left[\dfrac{g}{c};q\right]_n\left[\dfrac{eg}{ab};q\right]_n}{[g;q]_n\left[\dfrac{eg}{abc};q\right]_n}$$

$$\cdot{}_4\Phi_3\left[\begin{matrix} \dfrac{e}{a},\dfrac{e}{b},c,q^{-n}; & q,q \\ e,\dfrac{cq^{1-n}}{g},\dfrac{cq^{1-n}}{h} & \end{matrix}\right],$$

where $egh = abcq^{1-n}$, we obtain

(2.3)

$$ {}_{10}W_9\left[-\frac{abc}{\sqrt{q}}; c, -c, b^2, a^2, -\frac{ab\sqrt{q}}{c}, -c^2q^n, q^{-n}; q, q\right] $$

$$ = \frac{\left[\frac{c^2q}{a^2}; q^2\right]_n \left[\frac{c^2q}{b^2}; q^2\right]_n \left[-abc\sqrt{q}; q\right]_n}{\left[c^2q; q^2\right]_n \left[\frac{c\sqrt{q}}{ab}; q\right]_n \left[-\frac{bc\sqrt{q}}{a}; q\right]_n \left[-\frac{ac\sqrt{q}}{b}; q\right]_n} $$

$$ \cdot {}_4\Phi_3\left[\begin{matrix} a^2, b^2, \frac{a^2b^2}{c^4}q^{1-2n}, q^{-2n}; & q^2, q^2 \\ a^2b^2q, \frac{a^2q^{1-2n}}{c^2}, \frac{b^2q^{1-2n}}{c^2} & \end{matrix}\right]. $$

In (2.3) let $n \to \infty$ and sum the resulting ${}_2\Phi_1$ on the right-hand side by the q-analogue of Gauss' summation theorem [4, 8.4(3)]. We get (1.5).

To prove (1.6) we transform the well-poised ${}_{10}\Phi_9$ on the left-hand side of (2.3) by [4, 8.5(1)] to obtain

(2.4)

$$ {}_{10}W_9\left[-c; \frac{a\sqrt{q}}{b}, \frac{b\sqrt{q}}{a}, -\frac{c\sqrt{q}}{ab}, c, -\frac{ab\sqrt{q}}{c}, -c^2q^n, q^{-n}; q, q\right] $$

$$ = \frac{\left[\frac{c^2q}{a^2}; q^2\right]_n \left[\frac{c^2q}{b^2}; q^2\right]_n \left[-cq; q\right]_n \left[-ab\sqrt{q}; q\right]_n \left[c^2; q\right]_n}{\left[c^2q; q^2\right]_n \left[c; q\right]_n \left[\frac{c^2\sqrt{q}}{ab}; q\right]_n \left[-q; q\right]_n \left[-\frac{bc\sqrt{q}}{a}; q\right]_n \left[-\frac{ac\sqrt{q}}{b}; q\right]_n} $$

$$ \cdot {}_4\Phi_3\left[\begin{matrix} a^2, b^2, \frac{a^2b^2}{c^4}q^{1-2n}, q^{-2n}; & q^2, q^2 \\ a^2b^2q, \frac{a^2}{c^2}q^{1-2n}, \frac{b^2}{c^2}q^{1-2n} & \end{matrix}\right]. $$

In (2.4) letting $n \to \infty$ and summing the resulting ${}_2\Phi_1$ by q-analogue of Gauss' summation theorem [4, 8.4(3)], we get (1.6) after replacing a by $ab/\sqrt{q}$ and then setting $b^2 = cq/ad$.

Note that (1.5) for $c \to 0$ reduces to the q-analogues of Gauss' second summation theorem due to Andrews [1]. In (1.6) replace d by d/c and let $c \to 0$ to get the q-analogues of Bailey's summation theorem due to Andrews [1].

3. Jackson [6] obtained a summation formula for a bilateral hypergeometric series which contains the Watson and Whipple summation formulas as special cases. We begin this section by obtaining the q-analogues of Jackson's formula in the form:

(3.1)

$$\Pi\left[\begin{array}{l} \frac{aq}{b},\frac{aq}{c},\frac{b^2c}{a},\frac{-aq}{b},\frac{q}{b},\frac{q}{c},-\frac{q}{b},\frac{b^2c}{a^2}; \quad q \\ aq,y,\frac{a^2q}{b^2y},\frac{q}{a},\frac{y}{a},\frac{aq}{b^2y} \end{array}\right]$$

$$\cdot {}_8\psi_8\left[\begin{array}{l} b,-b,c,y,\frac{a^2q}{b^2c},\frac{a^2q}{b^2y},q\sqrt{a},-q\sqrt{a}; \quad q,-\frac{b^2}{a} \\ \frac{aq}{b},-\frac{aq}{b},\frac{aq}{c},\frac{aq}{y},\frac{b^2c}{a},\frac{b^2y}{a},\sqrt{a},-\sqrt{a} \end{array}\right]$$

$$=-\left(\frac{a}{y}\right)\Pi\left[\begin{array}{l} \frac{aq}{by},\frac{aq}{cy},-\frac{aq}{by},\frac{b^2c}{ay}; \quad q \\ \frac{a}{y},\frac{q}{y},y,\frac{a^2q}{b^2y^2},\frac{aq}{b^2},\frac{yq}{a},-\frac{b^2}{a} \end{array}\right]$$

$$\cdot\Pi\left[\begin{array}{l} q^2,\frac{cqy}{a},\frac{ayq^2}{cb^2},\frac{yb^2q}{ac},\frac{ycb^4}{a^3}; \quad q^2 \\ \frac{b^2y^2q}{a^2} \end{array}\right]$$

$$+\frac{a^2q}{b^2y}\Pi\left[\begin{array}{l} q^2,\frac{acq^2}{b^2y},\frac{a^3q^3}{b^4yc},\frac{aq^2}{cy},\frac{b^2cq}{ay},\frac{b^2y^2}{a^2}; \quad q^2 \\ \frac{a^2q^3}{b^2y^2} \end{array}\right]$$

$$\cdot\Pi\left[\begin{array}{l} \frac{b^2y}{ac},\frac{b^4y^2}{a^3q^2},\frac{b^4cy}{a^3q},\frac{a^3q^3}{b^4y^2}; \quad q \\ \frac{b^2y}{aq},\frac{b^2y}{a^2},\frac{a^2q}{b^2y},\frac{b^2y^2}{a^2q},\frac{aq}{b^2},\frac{aq^2}{b^2y},\frac{a^3q^2}{b^4y^2},\frac{b^4y^2}{a^3q},-\frac{b^2}{a} \end{array}\right].$$

To prove (3.1) set $a_1=a$, $a_3=y$, $a_4=a^2q/b^2y$, $a_7=b$, $a_8=c$, $a_9=a^2q/b^2c$, $a_{10}=-b$ in Jackson's [7, 2.2] transformation connecting a ${}_8\psi_8$ with two ${}_8\phi_7$ series (which can also be obtained by setting $N=4$ and specializing the parameters in a transformation of Slater [11, 2(7)]) to obtain

(3.2)

$$\Pi\left[\begin{matrix}\frac{aq}{b},-\frac{aq}{b},\frac{aq}{c},\frac{b^2c}{a},\frac{q}{b},-\frac{q}{b},\frac{q}{c},\frac{b^2c}{a^2}; & q\\ aq,y,\frac{a^2q}{b^2y},\frac{q}{a},\frac{y}{a},\frac{aq}{b^2y} & \end{matrix}\right]$$

$$\cdot{}_8\psi_8\left[\begin{matrix}b,-b,c,y,\frac{a^2q}{b^2c},\frac{a^2q}{b^2y},q\sqrt{a},-q\sqrt{a}; & q,-\frac{b^2}{a}\\ \frac{aq}{b},-\frac{aq}{b},\frac{aq}{c},\frac{aq}{y},\frac{b^2c}{a},\frac{b^2y}{a},\sqrt{a},-\sqrt{a} & \end{matrix}\right]$$

$$=y\Pi\left[\begin{matrix}\frac{yq}{b},\frac{yq}{c},\frac{b^2cy}{a^2},-\frac{yq}{b},q,\frac{aq}{by},\frac{aq}{cy},-\frac{aq}{by},\frac{b^2c}{ay},\frac{a}{y^2}; & q\\ \frac{a}{y},\frac{q}{y},y,\frac{a^2q}{b^2y^2},\frac{y^2}{a},\frac{aq}{b^2},\frac{yq}{a},\frac{aq}{y^2} & \end{matrix}\right]$$

$$\cdot{}_8W_7\left[\frac{y^2}{a};\frac{by}{a},-\frac{by}{a},\frac{cy}{a},\frac{ayq}{b^2c},\frac{aq}{b^2};q;-\frac{b^2}{a}\right]$$

$$+\frac{a^2q}{b^2y}\Pi\left[\begin{matrix}\frac{a^2q^2}{b^3y},\frac{a^2q^2}{b^2cy},\frac{cq}{y},-\frac{a^2q^2}{b^3y},q-\frac{by}{a},\frac{b^2y}{ac},\frac{b^4yc}{a^3q},\frac{by}{a},\frac{b^4y^2}{a^3q^2}; & q\\ \frac{b^2y}{aq},\frac{b^2y}{a^2},\frac{a^2q}{b^2y},\frac{b^2y^2}{a^2q},\frac{aq}{b^2},\frac{aq^2}{b^2y},\frac{a^3q^2}{b^4y^2},\frac{b^4y^2}{a^3q} & \end{matrix}\right]$$

$$\cdot{}_8W_7\left[\frac{a^3q^2}{b^4y^2};\frac{aq}{by},-\frac{aq}{by},\frac{acq}{b^2y},\frac{a^3q^2}{b^4yc},\frac{aq}{b^2};q,-\frac{b^2}{a}\right].$$

Summing both ${}_8W_7$'s on the right-hand side of (3.2) by (1.5), we get (3.1). It may be noted that (3.1) for $b=a$ yield a q-analogue of (1.7) whereas in (3.1) first setting $b=a$ and then letting $a=\sqrt{q}, y=q^{3/4}$ we get a q-analogue of yet another summation formula of Whipple [2, 3(i), p.97]. On the other hand (3.1) may also be regarded as a q-analogue of another summation formula for ${}_6H_6(-1)$ due to Jackson [8, 1.2] to which it reduces on letting $q\to 1^-$ in the usual way.

4. We begin this section by obtaining the following q-analogue of two transformations due to Bailey [3]:

(4.1)

$$
{}_{10}W_9\left[-abq^{-1/2-n};\ -q^{-n},\ -\frac{abq^{1/2-n}}{c^2},\ -c,c,a^2,b^2,q^{-n};q,q\right]
$$

$$
=\frac{\left[-ab\sqrt{q}\,;q\right]_n\left[-\frac{\sqrt{q}}{ab};q\right]_n\left[\frac{c^2q}{a^2};q^2\right]_n\left[\frac{c^2q}{b^2};q^2\right]_n}{\left[-\frac{a\sqrt{q}}{b};q\right]_n\left[-\frac{b\sqrt{q}}{a};q\right]_n\left[c^2q;q^2\right]_n\left[\frac{c^2q}{a^2b^2};q^2\right]_n}
$$

$$
\cdot{}_4\phi_3\left[\begin{matrix} a^2,b^2,\dfrac{a^2b^2q^{1-2n}}{c^4},q^{-2n}; & q^2,q^2 \\ a^2b^2q,\dfrac{a^2q^{1-2n}}{c^2},\dfrac{b^2q^{1-2n}}{c^2} & \end{matrix}\right]
$$

(4.2)

$$
=\frac{\left[-\frac{\sqrt{q}}{ab};q\right]_n\left[\frac{c^2q}{a^2};q^2\right]_n\left[a^2q;q^2\right]_n}{\left[-\frac{a\sqrt{q}}{b};q\right]_n\left[-\frac{b\sqrt{q}}{a};q\right]_n\left[ab\sqrt{q}\,;q\right]_n\left[\frac{c^2q}{a^2b^2};q^2\right]_n}
$$

$$
\cdot{}_4\phi_3\left[\begin{matrix} b^2,\dfrac{q^{1-2n}}{c^2},\dfrac{c^2}{b^2},q^{-2n}; & q^2,q^2 \\ \dfrac{a^2q^{1-2n}}{c^2},\dfrac{q^{1-2n}}{a^2},c^2q & \end{matrix}\right].
$$

The transformation (4.1) has been written by transforming the well-poised ${}_{10}\phi_9$ in the left-hand side of (2.3) by [4; 8.5(1)], whereas (4.2) is obtained by transforming the ${}_4\phi_3$ in the right-hand side of (4.1) by (2.2). The transformation (4.1) for $c=ab/\sqrt{q}$ yields the summation formula

$$
{}_4\phi_3\left[\begin{matrix} a^2,b^2,\dfrac{q^{3-2n}}{a^2b^2},q^{-2n}; & q^2,q^2 \\ a^2b^2q,\dfrac{q^{2-2n}}{a^2},\dfrac{q^{2-2n}}{b^2} & \end{matrix}\right]=\frac{\left[a^2;q\right]_n\left[b^2;q\right]_n}{\left[\frac{a^2b^2}{q};q\right]_n\left[a^2b^2q;q^2\right]_n}\cdot\frac{\left[\frac{a^2b^2}{q};q\right]_{2n}\left[-q;q\right]_n}{\left[a^2;q^2\right]_n\left[b^2;q^2\right]_n},
$$

whereas for $c=a/\sqrt{q}$, (4.1) yields the interesting summation formula

$$
{}_{10}W_9\left[abq^{-1/2-n};a^2,b^2,\frac{a}{\sqrt{q}},-\frac{a}{\sqrt{q}},\frac{bq^{3/2-n}}{a},-q^{-n},q^{-n};q,q\right]
$$

$$
=\frac{\left[b^2;q^2\right]_n\left[-\frac{\sqrt{q}}{ab};q\right]_n\left[\frac{a^2}{b^2q};q^2\right]_n q^n}{\left[\frac{1}{b^2};q^2\right]_n\left[ab\sqrt{q}\,;q\right]_n\left[-\frac{a\sqrt{q}}{b};q\right]_n\left[-\frac{b\sqrt{q}}{a};q\right]_n}.
$$

On the other hand using (4.1) and (4.2) the following identities of the Cayler–Orr type can be written.

THEOREM I. *If*

$$
{}_4\Phi_3\left[\begin{matrix} a^2,b^2,c,-c; & q,z \\ c^2,ab\sqrt{q},-ab\sqrt{q} & \end{matrix}\right]
{}_4\Phi_3\left[\begin{matrix} \dfrac{c\sqrt{q}}{ab},-\dfrac{c\sqrt{q}}{ab},-\dfrac{a\sqrt{q}}{b},-\dfrac{b\sqrt{q}}{a}; & q,qz \\ -q,-abq^{3/2},-\dfrac{c^2\sqrt{q}}{ab} & \end{matrix}\right]
$$

$$
+\frac{ab}{\sqrt{q}}\,{}_4\Phi_3\left[\begin{matrix} a^2,b^2,c,-c; & q,qz \\ c^2,ab\sqrt{q},-ab\sqrt{q} & \end{matrix}\right]
{}_4\Phi_3\left[\begin{matrix} \dfrac{c\sqrt{q}}{ab},-\dfrac{c\sqrt{q}}{ab},-\dfrac{a\sqrt{q}}{b},-\dfrac{b\sqrt{q}}{a}; & q,z \\ -q,-abq^{3/2},-\dfrac{c^2\sqrt{q}}{ab} & \end{matrix}\right]
$$

$$
=\frac{ab}{\sqrt{q}}\sum_{n=0}^{\infty}\frac{\left(1+ab\sqrt{q}\right)\left[-\dfrac{\sqrt{q}}{ab};q\right]_{n+1}}{\left(1+abq^{n+1/2}\right)\left[-\dfrac{c^2\sqrt{q}}{ab};q\right]_n}a_nz^n
$$

then

$$
{}_2\Phi_1\left[\begin{matrix} a^2,b^2; & q^2,qz \\ a^2b^2q & \end{matrix}\right]
{}_2\Phi_1\left[\begin{matrix} \dfrac{c^2q}{a^2},\dfrac{c^2q}{b^2}; & q^2,z \\ \dfrac{c^4q}{a^2q^2} & \end{matrix}\right]
=\sum_{n=0}^{\infty}\frac{\left[c^2q;q^2\right]_n}{\left[c^4q/a^2b^2;q^2\right]_n}a_nz^n.
$$

THEOREM II. *If*

$$
{}_4\Phi_3\left[\begin{matrix} a^2,b^2,c,-c; & q,z \\ c^2,ab\sqrt{q},-ab\sqrt{q} & \end{matrix}\right]
{}_4\Phi_3\left[\begin{matrix} \dfrac{c\sqrt{q}}{ab},-\dfrac{c\sqrt{q}}{ab},-\dfrac{a\sqrt{q}}{b},-\dfrac{b\sqrt{q}}{a}; & q,qz \\ -q,-abq^{3/2},-\dfrac{c^2\sqrt{q}}{ab} & \end{matrix}\right]
$$

$$
+\frac{ab}{\sqrt{q}}\,{}_4\Phi_3\left[\begin{matrix} a^2,b^2,c,-c,; & q,qz \\ c^2,ab\sqrt{q},-ab\sqrt{q} & \end{matrix}\right]
{}_4\Phi_3\left[\begin{matrix} \dfrac{c\sqrt{q}}{ab},-\dfrac{c\sqrt{q}}{ab},-\dfrac{a\sqrt{q}}{b},-\dfrac{b\sqrt{q}}{a}; & q,z \\ -q,-abq^{3/2},-\dfrac{c^2\sqrt{q}}{ab} & \end{matrix}\right]
$$

$$
=\sum_{n=0}^{\infty}\frac{\left[-\dfrac{\sqrt{q}}{ab};q\right]_{n+1}}{\left[-\dfrac{c^2\sqrt{q}}{ab};q\right]_n}\,\frac{ab\left(1+ab\sqrt{q}\right)}{\left(1+abq^{n+1/2}\right)}a_nz^n
$$

then

$$ {}_2\phi_1\left[\begin{matrix} b^2, \frac{c^2}{b^2}; & q^2, qz \\ c^2q & \end{matrix}\right] {}_2\phi_1\left[\begin{matrix} a^2q, \frac{c^2q}{a^2}; & q^2, z \\ c^2q & \end{matrix}\right] = \sum_{n=0}^{\infty} \frac{[a^2b^2q; q^2]_n}{[c^2q; q^2]_n} a_n z^n. $$

Theorems I and II are q-analogues of known results of Bailey [3, Thms. I and II].

REFERENCES

[1] G. E. ANDREWS, *On the q-analogue of Kummers' theorem and applications*, Duke Math. J., 40 (1973), pp. 525–528.

[2] ______, *On the q-analogue of Watson and Whipple summations*, this Journal, 7 (1976), pp. 332–336.

[3] W. N. BAILEY, *Some theorems connecting products of hypergeometric series*, Proc. London Math. Soc., (2) 38 (1935), pp. 377–384.

[4] ______, *Generalized Hypergeometric Series*, Cambridge Tract, 1935.

[5] ______, *On sums of terminating* ${}_3F_2(1)$, Quart J. Math. (Oxford), (2) 4 (1953), pp. 237–240.

[6] M. JACKSON, *A generalization of the theorems of Watson and Whipple on the sum of the series* ${}_3F_2$, J. London Math. Soc., 24 (1949), pp. 238–240.

[7] ______, *On well-poised bilateral hypergeometric series of the type* ${}_8\psi_8$, Quart. J. Math. (Oxford), (2) 1(1950), pp. 63–68.

[8] ______, *A note on the summations of a particular* ${}_6H_6$ *with argument* -1, J. London Math. Soc., 27 (1952), pp. 124–126.

[9] V. K. JAIN, *Some transformations of basic hypergeometric functions*, this Journal, 12(1981), pp. 957–961.

[10] D. B. SEARS, *On the transformation theory of basic hypergeometric functions*, Proc. London Math. Soc., (2) 53(1951), pp. 158–180.

[11] L. J. SLATER, *General transformations of bilateral series*. Quart. J. Math. (Oxford), (2) 3(1952), pp. 72–80.

[12] A. VERMA AND V. K. JAIN, *Transformations between basic hypergeometric series on different bases and identities of Rogers-Ramanujan type*, J. Math. Anal. Appl., 76 (1980), pp. 230–269.

[13] F. J. W. WHIPPLE, *On well-poised series, generalized hypergeometric series having parameters in pair, each pair with the same sum*, Proc. London Math. Soc., (2) 25 (1926), pp. 525–544.

SIAM J. MATH. ANAL.
Vol. 16, No. 3, May 1985

020

RATIONAL INTERPOLATION TO e^x, II.*

PETER B. BORWEIN[†]

Abstract. The following estimate is derived for the error in approximating e^x by rational functions. Let π_n denote the polynomials of degree at most n.

THEOREM. *Let* $\gamma_1, \gamma_2, \cdots, \gamma_{2n+1}$ *be points* (*not necessarily distinct*) *in* $[0,\alpha]$, $\alpha<2$. *Choose* P_n, $Q_n \in \pi_n$ *so that*

$$P_n(\gamma_i)-Q_n(\gamma_i)e^{-\gamma_i}=0 \quad \text{for } i=1,2,\cdots,2n+1.$$

Then for $x\in[0,\alpha]$

$$|P_n(x)/Q_n(x)-e^{-x}|\leq C_\alpha \frac{n!n!}{(2n)!(2n+1)!}\left|\prod_{i=1}^{2n+1}(x-\gamma_i)\right|$$

and

$$|P_n(x)/Q_n(x)-e^{-x}|\geq D_\alpha \frac{n!n!}{(2n)!(2n+1)!}\left|\prod_{i=1}^{2n+1}(x-\gamma_i)\right|,$$

where C_α *and* D_α *depend only on* α.

1. Introduction. We derive precise estimates for the error in interpolating e^{-x} on $[0,\alpha]$, $\alpha<2$, by rational functions whose numerators and denominators have the same degree. These estimates show that, up to a constant, the optimal choice of interpolation points are the zeros of the Chebyshev polynomials shifted to the interval $[0,\alpha]$. The estimates provide another proof of the main diagonal case of the Meinardus conjecture concerning the error in best approximation to e^x, at least, up to a constant and on a smaller interval. (See [1], [2], [3, p. 168], [4] and [5].)

Let π_n denote the real algebraic polynomials of degree at most n.

THEOREM. *Let* $\gamma_1, \gamma_1, \cdots, \gamma_{2n+1}$ *be points* (*not necessarily distinct*) *in* $[0,\alpha]$, *where* $\alpha<2$. *Choose* $P_n, Q_n \in \pi_n$ *so that*

$$P_n(\gamma_i)-Q_n(\gamma_i)e^{-\gamma_i}=0 \quad \textit{for } i=1,2,\cdots,2n+1.$$

Then, for $x\in[0,\alpha]$,

$$|P_n(x)/Q_n(x)-e^{-x}|\leq C_\alpha \frac{n!n!}{(2n)!(2n+1)!}\left|\prod_{i=1}^{2n+1}(x-\gamma_i)\right|$$

and

$$|P_n(x)/Q_n(x)-e^{-x}|\geq D_\alpha \frac{n!n!}{(2n)!(2n+1)!}\left|\prod_{i=1}^{2n+1}(x-\gamma_i)\right|,$$

where

$$\left(\frac{2-\alpha}{163}\right)^2\leq D_\alpha\leq C_\alpha\leq\frac{9}{(2-\alpha)^3}.$$

* Received by the editors February 22, 1983, and in revised form October 3, 1983. This research was supported in part by the Natural Science and Engineering Research Council of Canada.

[†] Department of Mathematics, Statistics and Computing Science, Dalhousie University, Halifax, Nova Scotia B3H 4H8, Canada.

If we set all the γ_i to zero in the above theorem then we get bounds for the error in main diagonal Padé approximation.

The theorem is a refinement of a similar result in [1].

2. Preliminaries. We proceed, initially, exactly as in [1, p. 143]. Suppose that $P_n, Q_n \in \pi_n$ and suppose that $P_n(x)-Q_n(x)e^{-x}$ has $2n+1$ zeros on the interval $[0,\alpha]$. If $Q_n(x)=q_0+q_1x+\cdots+q_nx^n$ then on taking $n+1$ derivatives

$$(1)\quad (P_n(x)-Q_n(x)e^{-x})^{(n+1)}=(Q_n(x)e^{-x})^{(n+1)}=\sum_{k=0}^{n}\binom{n+1}{k}Q_n^{(k)}e^{-x}(-1)^{(n+1-k)}$$

$$=(-1)^{n+1}e^{-x}\sum_{k=0}^{n}\frac{x^k}{k!}\sum_{j=0}^{n-k}\binom{n+1}{j}(-1)^j(k+j)!q_{k+j}.$$

Since $(Q_n(x)e^{-x})^{(n+1)}$ has n zeros on $[0,\alpha]$, we deduce that there exist $\beta_1,\cdots,\beta_n\in[0,\alpha]$ so that

$$\sum_{k=0}^{n}\frac{x^k}{k!}\sum_{j=0}^{n-k}\binom{n+1}{j}(-1)^j(k+j)!q_{k+j}=q_n\prod_{i=1}^{n}(x-\beta_i).$$

Thus, if $q_n\prod_{i=1}^{n}(x-\beta_i)=b_0+b_1x+\cdots+b_nx^n$, we have
(2)

$$\begin{bmatrix}\binom{n+1}{0} & -\binom{n+1}{1} & +\binom{n+1}{2} & \cdots & (-1)^n & \binom{n+1}{n}\\ 0 & \binom{n+1}{0} & -\binom{n+1}{1} & \cdots & (-1)^{n-1} & \binom{n+1}{n-1}\\ 0 & 0 & \binom{n+1}{0} & \cdots & (-1)^{n-2} & \binom{n+1}{n-2}\\ \vdots & \vdots & \vdots & \cdots & & \vdots\\ 0 & 0 & 0 & \cdots & & \binom{n+1}{0}\end{bmatrix}\begin{bmatrix}q_0 0!\\ q_1 1!\\ q_2 2!\\ \vdots\\ q_n n!\end{bmatrix}=\begin{bmatrix}b_0 0!\\ b_1 1!\\ b_2 2!\\ \vdots\\ b_n n!\end{bmatrix}.$$

We can invert (2) to obtain

$$(3)\quad\begin{bmatrix}\binom{n}{n} & \binom{n+1}{n} & \binom{n+2}{n} & \cdots & \binom{2n}{n}\\ 0 & \binom{n}{n} & \binom{n+1}{n} & \cdots & \binom{2n-1}{n}\\ 0 & 0 & \binom{n}{n} & \cdots & \binom{2n-2}{n}\\ \vdots & \vdots & \vdots & & \vdots\\ 0 & 0 & 0 & \cdots & \binom{n}{n}\end{bmatrix}\begin{bmatrix}b_0 0!\\ b_1 1!\\ b_2 2!\\ \vdots\\ b_n n!\end{bmatrix}=\begin{bmatrix}q_0 0!\\ q_1 1!\\ q_2 2!\\ \vdots\\ q_n n!\end{bmatrix}.$$

We observe that (3) can be easily derived from (2) combined with the fact that the (m,n) Padé approximant (the case where $b_0=b_1=\cdots=b_{n-1}=0$) to e^{-x} is given by

$$\sum_{v=0}^{m}\frac{\binom{m}{v}}{\binom{m+n}{v}}\frac{(-x)^v}{v!}\Big/\sum_{v=0}^{n}\frac{\binom{n}{v}}{\binom{n+m}{v}}\frac{x^v}{v!}.$$

We now consider $e^xP_n(x)-Q_n(x)$ and perform similar calculations to those above. We write $P_n(x)=p_0+\cdots+p_nx^n$ and we deduce the existence of $\alpha_1,\cdots,\alpha_n\in[0,\alpha]$ so that

$$(e^xP_n(x))^{(n+1)}=e^xp_n\prod_{i=1}^{n}(x-\alpha_i),$$

where

$$p_n\prod_{i=0}^{n}(x-\alpha_i)=a_0+\cdots+a_nx^n,$$

(4)
$$\begin{bmatrix} \binom{n+1}{0} & \binom{n+1}{1} & \binom{n+1}{2} & \cdots & \binom{n+1}{n} \\ 0 & \binom{n+1}{0} & \binom{n+1}{1} & \cdots & \binom{n+1}{n-1} \\ 0 & 0 & \binom{n+1}{0} & \cdots & \binom{n+1}{n-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \binom{n+1}{0} \end{bmatrix} \begin{bmatrix} p_00! \\ p_11! \\ p_22! \\ \vdots \\ p_nn! \end{bmatrix} = \begin{bmatrix} a_00! \\ a_11! \\ a_22! \\ \vdots \\ a_nn! \end{bmatrix},$$

and

(5)
$$\begin{bmatrix} \binom{n}{n} & -\binom{n+1}{n} & \binom{n+2}{n} & \cdots & (-1)^n\binom{2n}{n} \\ 0 & \binom{n}{n} & -\binom{n+1}{n} & \cdots & (-1)^{n-1}\binom{2n-1}{n} \\ 0 & 0 & \binom{n}{n} & \cdots & (-1)^{n-2}\binom{2n-2}{n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \binom{n}{n} \end{bmatrix} \begin{bmatrix} a_00! \\ a_11! \\ a_22! \\ \vdots \\ a_nn! \end{bmatrix} = \begin{bmatrix} p_00! \\ p_11! \\ p_22! \\ \vdots \\ p_nn! \end{bmatrix}.$$

The information about P_n and Q_n that allows us to analyse the error in interpolating e^x is contained in the following lemma.

LEMMA. *Suppose that $P_n(x)=p_0+p_1x+\cdots+p_nx^n$ and suppose that $Q_n=q_0+q_1x+\cdots+q_nx^n$ where $q_0>0$. Suppose also that $P_n(x)-Q_n(x)e^{-x}$ has $2n+1$ zeros at $\gamma_1,\cdots,\gamma_{2n+1}\in[0,\alpha]$.*

Then:

a) *P_n has alternating coefficients*;

b) $|p_n|\le(n!/(2n)!)|p_0|$;

c) *if $\alpha\le2$, then*

$$|p_n|\ge\left(\frac{4}{27}\right)\frac{n!}{(2n)!}|p_0| \quad \textit{and} \quad |q_n|\ge\left(\frac{4}{27}\right)\frac{n!}{(2n)!}|q_0|;$$

d) *if $\alpha<2$ then Q_n has positive coefficients,*

$$q_n\le\left(\frac{2}{2-\alpha}\right)\frac{n!}{(2n)!}q_0 \quad \textit{and} \quad |Q_n(\alpha)|\le\frac{q_02e^\alpha}{(2-\alpha)},$$

e) *if* $\alpha<2$ *then*

$$\frac{1}{e^{3\alpha/2}}\le\frac{P_n(0)}{Q_n(0)}\le\frac{4}{(2-\alpha)^2}.$$

Proof. That P_n and Q_n can be found with the desired interpolation properties is a consequence of results in [3, pp. 16 and 165].

Part a) is a direct consequence of (5) and the observation that the a_i in (4) alternate in sign.

Part b) follows from (5) and the above, that is,

$$|p_0|=\sum_{i=0}^{n}|i!a_i|\binom{n+i}{n}\ge\frac{(2n)!}{n!}|a_n|=\frac{(2n)!}{n!}|p_n|.$$

To prove part c) we see that, if $0\le\alpha_1,\cdots,\alpha_n\le\alpha<2$ and

$$a_n\prod_{i=1}^{n}(x-\alpha_i)=a_0+\cdots+a_nx^n,$$

then

$$|a_k|\le\binom{n}{k}\alpha^{n-k}|a_n|.$$

Thus, from (5) (or (3) for the second part),

$$|p_0|\le\sum_{k=0}^{n}\binom{n+k}{n}k!|a_k|\le|a_n|\sum_{k=0}^{n}\frac{(n+k)!\alpha^{n-k}}{k!(n-k)!}.$$

Since $2^{n-k}/(n-k)!\le(\frac{2}{3})^{n-k}\cdot\frac{9}{2}$,

$$|p_0|\le\frac{9|a_n|}{2}\sum_{k=0}^{n}\frac{(n+k)!}{k!}\left(\frac{2}{3}\right)^{n-k}\le\frac{9|a_n|}{2}\cdot\frac{(2n)!}{n!}\sum_{k=0}^{n}\left(\frac{1}{2}\right)^{n-k}\left(\frac{2}{3}\right)^{n-k}\le\frac{27}{4}|p_n|\frac{(2n)!}{n!}.$$

The first part of d) follows from an examination of (3) using the facts that, for $i\le n$,

$$(i-1)!|b_{i-1}|\le\alpha(i!)|b_i|\quad\text{and}\quad\binom{n+i-1}{n}\le\frac{1}{2}\binom{n+i}{n}.$$

The second part of d) is proved by noting that

$$q_0\ge n!|b_n|\binom{2n}{n}-(n-1)!|b_{n-1}|\binom{2n-1}{n}\ge\left(1-\frac{\alpha}{2}\right)\frac{(2n)!}{n!}q_n.$$

To see the final part of d), note that

$$i!q_i=\sum_{k=0}^{n-i}\binom{n+k}{n}(k+i)!b_{k+i}$$

and

$$\binom{n+k-1}{n}(k+i-1)!|b_{k+i-1}|\le\binom{n+k}{n}(k+i)!|b_{k+i}|.$$

Hence, since the b_k alternate in sign,

$$i!q_i \leq \binom{2n-i}{n}(n!)|b_n| \leq \binom{2n}{n} n!|b_n| = \frac{(2n)!}{n!} q_n$$

or

$$q_i \leq \left(\frac{2}{2-\alpha}\right)\frac{q_0}{i!}.$$

Thus

$$Q_n(x) \leq \frac{2q_0}{2-\alpha} e^x.$$

Finally, from (5), one can show that

$$|(i+1)!p_{i+1}| \leq \frac{1}{2}|i!p_i|.$$

Since,

$$e^{-\gamma_1} = \frac{p_0}{q_0} \frac{\sum_{i=0}^n (p_i/p_0)(\gamma_1)^i}{\sum_{i=0}^n (q_i/q_0)(\gamma_1)^i}.$$

It follows that

$$1 - \frac{\gamma_1}{2} \leq \sum_{i=0}^n \frac{p_i}{p_0}(\gamma_1)^i \leq e^{\gamma_1/2},$$

$$1 \leq \sum_{i=0}^n \frac{q_i}{q_0}(\gamma_1)^i \leq \frac{2e^{\gamma_1}}{2-\alpha}$$

and

$$\frac{1}{e^{3\gamma_1/2}} \leq \frac{p_0}{q_0} \leq \frac{4}{(2-\gamma_1)(2-\alpha)}.$$

3. Proof of the theorem. Let $P_{n+k}, Q_{n+k} \in \pi_{n+k}$ be such that

$$\frac{P_{n+k}(x)}{Q_{n+k}(x)} - e^{-x}, \qquad k = 0, 1, \cdots$$

has $2k$ zeros at zero and a single zero at each of the γ_i. Then, for $x \in [0, \alpha]$

$$R_k(x) = \frac{P_{n+k+1}(x)}{Q_{n+k+1}(x)} - \frac{P_{n+k}(x)}{Q_{n+k}(x)} = \frac{\alpha_{k+1} x^{2k} \prod_{i=1}^{2n+1}(x-\gamma_i)}{Q_{n+k+1}(x) Q_{n+k}(x)}. \tag{6}$$

Also, if $P_{n+k}(x) = 1 + \cdots + p_{n+k,k}x^{n+k}$ and $Q_{n+k} = q_{0,k} + \cdots + q_{n+k,k}x^n$, then

$$\alpha_{k+1} = p_{n+k+1,k+1} \cdot q_{n+k,k} - p_{n+k,k} \cdot q_{n+k+1,k+1}$$

and by parts a) and d) of the lemma

$$|\alpha_{k+1}| = |p_{n+k+1,k+1} \cdot q_{n+k,k}| + |p_{n+k,k} \cdot q_{n+k+1,k+1}|.$$

Parts b), c) and d) of the lemma yield the following bounds for a_{k+1}:

$$|\alpha_{k+1}|\le\left(\frac{2}{2-\alpha}\right)\frac{(n+k+1)!(n+k)!}{(2n+2k+2)!(2n+2k)!}\left(|q_{0,k}\cdot p_{0,k+1}|+|q_{0,k+1}\cdot p_{0,k}|\right)$$

and

$$|\alpha_{k+1}|\ge\left(\frac{4}{27}\right)^2\frac{(n+k+1)!(n+k)!}{(2n+2k+2)!(2n+2k)!}\left(|q_{0,k}\cdot p_{0,k+1}|+|q_{0,k+1}\cdot p_{0,k}|\right).$$

From part d) of the lemma

$$q_{0,k}\le Q_{n+k+1}(x)\le(q_{0,k})\frac{2e^{\alpha}}{2-\alpha}.$$

For $k\ge0$ we note that $q_{0,k+1}=p_{0,k}=1$. Thus, for $k\ge1$

$$|R_k(x)|\le\frac{4}{2-\alpha}\left(x^{2k}\prod_{i=1}^{2n+1}|x-\gamma_i|\right)\frac{(n+k+1)!(n+k)!}{(2n+2k+2)!(2n+2k)!}$$

and

$$|R_k(x)|\ge2\left(\frac{4-2\alpha}{27e^{\alpha}}\right)^2\left(x^{2k}\prod_{i=1}^{2n+1}|x-\gamma_i|\right)\frac{(n+k+1)!(n+k)!}{(2n+2k+2)!(2n+2k)!}.$$

For $k=0$

$$|R_0(x)|\le\left(\frac{2}{2-\alpha}\right)\left(\prod_{i=1}^{2n+1}|x-\gamma_i|\right)\frac{(n+1)!n!}{(2n+2)!(2n)!}\left(1+\frac{P_n(0)}{Q_n(0)}\right)$$

and

$$|R_0(x)|\ge\left(\frac{4-2\alpha}{27e^{\alpha}}\right)^2\left(\prod_{i=1}^{2n+1}|x-\gamma_i|\right)\frac{(n+1)!n!}{(2n+2)!(2n)!}\left(1+\frac{P_n(0)}{Q_n(0)}\right).$$

Note that

$$\frac{(n+k+2)!(n+k+1)!}{(2n+2k+4)!(2n+2k+2)!}\Big/\frac{(n+k+1)!(n+k)!}{(2n+2k+2)!(2n+k)!}\le\frac{1}{16(n+k)^2}.$$

Thus,

$$\left|e^{-x}-\frac{P_n(x)}{Q_n(x)}\right|=\sum_{k=0}^{\infty}|R_k(x)|$$
$$\le\left(\frac{2}{2-\alpha}\right)\left(\prod_{i=1}^{2n+1}|x-\gamma_i|\right)\frac{(n+1)!n!}{(2n+2)!2n!}\left[1+\frac{4}{(2-\alpha)^2}+2\sum_{k=1}^{\infty}\frac{1}{16(n+k)^2}\right]$$

and

$$\left|e^{-x}-\frac{P_n(x)}{Q_n(x)}\right|$$
$$\ge\left(\frac{4-2\alpha}{27e^{\alpha}}\right)^2\left(\prod_{i=1}^{2n+1}|x-\gamma_i|\right)\frac{(n+1)!n!}{(2n+2)!2n!}\left[1+\frac{1}{e^3}-2\sum_{n=1}^{\infty}\frac{1}{16(n+k)^2}\right].\qquad\square$$

REFERENCES

[1] P. Borwein, *Rational Interpolation to e^x*, J. Approx. Theory, 35 (1982), pp. 142–147.

[2] D. Braess, *On the conjecture of Meinardus on rational approximation of e^x*, J. Approx. Theory, 36 (1982), pp. 317–320.

[3] G. Meinardus, *Approximation of Functions: Theory and Numerical Methods*, Springer-Verlag, New York/Berlin, 1967.

[4] G. Nemeth, *Relative rational approximation of the function e^x*, Math. Notes, 21 (1977), pp. 325–328.

[5] D. J. Newman, *Approximation with Rational Functions*, CBMS Regional Conference Series in Applied Mathematics 41, American Mathematical Society, Providence, RI, 1979.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

001

ON THE ASYMPTOTIC BEHAVIOR OF SOLUTIONS TO KINETIC EQUATIONS*

M. J. LEITMAN†

Abstract. A "kinetic equation" here refers to a Cauchy initial value problem in $L^2((-\infty,\infty))$

$$\frac{dx_t}{dt} = -Ax_t, \qquad x_0 = g,$$

where A is a bounded, symmetric, positive semidefinite, linear operator in $L^2((-\infty,\infty))$ with a mean-value property: $\int Ax = 0$. The specific operators considered typify a class which arises in the study of certain stochastic differential equations. This work addresses the following question: For which initial functions g in $L^2((-\infty,\infty))$ do the solutions decay in the L^2-norm more slowly than any exponential function? (The L^1-norms are constant.) It is shown that a sufficient, but not necessary, condition that g induce this slow decay is that g have nonzero mean: $\int g \neq 0$.

1. Formulation of the problem and principal results. By a "kinetic equation" we mean a Cauchy problem in L^2 of the following form:[1]

$$\text{(K)} \qquad \begin{aligned} &\frac{d}{dt}x_t = -Ax_t, \qquad t \geqq 0, \\ &x_0 = g, \end{aligned}$$

where A is a bounded symmetric linear operator in L^2 which is positive semidefinite and has zero mean:[2] $Ax \in L^1 \Rightarrow \int Ax = 0$.

Roughly speaking, solutions x_t of (K) have the following properties:[3]

(i) if $g \geqq 0$ then $x_t \geqq 0$, for all $t \geqq 0$;
(ii) if $g \in L^1$, then $x_t \in L^1$ and $\|g\|_1 = \|x_t\|_1$, for all $t \geqq 0$;
(iii) if P denotes the projection onto the null space of A, then $\|x_t - Pg\|_2$ decreases monotonically to zero as $t \to \infty$.

We are concerned here with the rate at which x_t approaches Pg. Specifically, we will describe those initial conditions g for which this decay is exponential and those for which solutions decay more slowly than any exponential.

The significance of this problem is discussed in [3], primarily with reference to the propagation of plane waves in a random medium. It is worth pointing out, however, that $-A$ generates a Markov process whose values for measurable sets $\mathscr{B}$ are generated by taking the initial function g to be the characteristic function of $\mathscr{B}$. The "kinetic equation" may thus be construed as the Kolmogorov equation associated with the Markov process. Here we concentrate on the mathematical question of decay rate, referring to [3] as needed.

*Received by the editors February 28, 1983. This work was supported by the National Science Foundation under grant MCS 80-02852.

†Department of Mathematics and Statistics, Case Western Reserve University, Cleveland, Ohio 44106

[1] Here, and in the sequel, L^p always means $L^p((-\infty,\infty))$, for $p \geqq 1$.

[2] For functions $f \in L^1$ we write $\int f$ for $\int_{-\infty}^{\infty} f(x)\,dx$ when no confusion is likely.

[3] For details not explicitly presented here, we refer to our earlier paper [3]. Note that we have replaced A by $-A$, for convenience.

The operator A is an integral operator in L^2, with kernel ψ of the following form:

$$\text{(A)} \qquad \mu \mapsto (Ax)(\mu) = \int_{-\infty}^{\infty} \psi(\mu,\nu)[x(\mu)-x(\nu)]\,d\nu, \qquad \mu \in (\infty,\infty).$$

Its kernel ψ is required to satisfy:

(A1) $\psi(\mu,\nu) \geqq 0$;
(A2) $\psi(\mu,\nu) = \psi(\nu,\mu)$;
(A3) $\psi(\mu,\nu) \leqq \bar{\psi} < \infty$;
(A4) $\phi(\mu) \overset{\text{def'n.}}{=} \int_{-\infty}^{\infty} \psi(\mu,\mu)\,d\nu \leqq \bar{\phi} < \infty$;
(A5) $\int_{-\infty}^{\infty} \phi^2(\mu)\,d\mu < \infty$;
(A6) for $\lambda > 0$, the set $\phi^{-1}((0,\lambda))$ has positive Lebesgue measure.

In [3] it was shown that (A1)–(A4) together imply that A is a symmetric, positive semidefinite,[4] bounded linear operator in $L^2(\|A\|_2 \leqq 2\bar{\phi})$. Moreover, $x \in L^1$ implies $Ax \in L^1$, and $\int Ax = 0$. The inclusion of (A5) guarantees that A takes L^2 boundedly into $L^1(\|A\|_{2,1} \leqq 2\|\phi\|_2)$, in which case $\int Ax = 0$ for all $x \in L^2$, and that zero is in the essential spectrum of A. The further inclusion of (A6) suffices to guarantee that zero is not an isolated point of the spectrum of A.

Remark 1. A is not compact as an operator in L^2. In fact, it may have no eigenvalues at all of finite multiplicity, say if the interval $(0,\bar{\phi})$ is in the range of ϕ. Moreover, if g is an eigenfunction corresponding to a positive eigenvalue, then necessarily $g \in L^1$ and $\int g = 0$.

Remark 2. Technically, condition (A5) guarantees that the integral operator $x \mapsto \Psi x$ given by

$$\mu \mapsto (\Psi x)(\mu) = \int_{-\infty}^{\infty} \psi(\mu,\nu)x(\nu)\,d\nu$$

is compact in L^2. It then follows that the operators $x \mapsto Ax$ and the L^2-multiplier $x \to \phi x$ have the same essential spectrum, the essential range of ϕ. Clearly zero is in the essential spectrum of $x \mapsto \phi x$, and condition (A6) guarantees that it is not an isolated point.

Remark 3. The "acoustic kernels" which originally motivated this study (see [3] and §2); are of the type $\psi(\mu,\nu) = f(\mu^2 - \nu^2)$, where $f \in L^1 \cap L^\infty$, $f \geqq 0$, and f is even; e.g. $f(s) = 1/(1+s^2)$. Such kernels satisfy (A1)–(A5); moreover, (A6) is redundant, since A is then positive definite, and so zero is in the continuous spectrum of A rather than being an eigenvalue.

Another example, discussed in [3] and in the next section, is that of convolution; that is, $\psi(\mu,\nu) = f(\mu - \nu)$, where $f \in L^1 \cap L^\infty$, $f \geqq 0$, and f is even. In this case (A1)–(A4) and (A6) are satisfied, but (A5) is not. It will turn out that A is positive definite, and that zero is in the continuous spectrum of A. Also, A does not map L^2 into L^1 but merely $L^1 \cap L^2 \to L^1 \cap L^2$.

Separable kernels $\psi(\mu,\nu) = f(\mu)f(\nu)$ which satisfy (A1)–(A4) always satisfy (A5), but not necessarily (A6). For example, if $f = \chi_{[-1,1]}$, the characteristic function of the interval $[-1,1]$, the spectrum of A consists of exactly two points $\{0, 2\}$, each of which is an eigenvalue of infinite multiplicity.

[4] In [3] ψ was assumed positive rather than nonnegative. In the latter case, A may or may not be positive definite.

The solution of the Cauchy problem (K) is given through the solution semigroup $\{T_t : t \geqq 0\}$ generated by $-A$; that is, $x_t = T_t g \equiv \exp(-tA)g$, $t \geqq 0$. In [3] it is shown that this is an analytic semigroup of bounded, symmetric, positive definite, positive, linear operators in L^2. If $g \in L^1 \cap L^2$, then $T_t g \in L^1 \cap L^2$ and $\int T_t g = \int g$, for $t \geqq 0$. Finally, if P denotes the projection onto the null space $N(A)$ of A, and $Q = 1 - P$ is its complement, the function $t \mapsto \|T_t Q g\|_2$ decreases monotonically to zero as $t \to \infty$.[5]

Our main result concerns the rate at which $t \mapsto \|T_t Q g\|_2$ approaches zero. For each $h \in L^2$, $h \neq 0$, we will see that $t \mapsto (d/dt)\ln\|T_t h\|_2$ is a negative, nondecreasing function for $t \geqq 0$. Hence $\lambda_h \geqq 0$ is well defined through

$$\lambda_h = -\lim_{t \to \infty} \frac{d}{dt} \ln\|T_t h\|_2.$$

For $h = 0$ we formally write $\lambda_0 = \infty$, so that $h \to \lambda_h$ makes sense for all $h \in L^2$.

If $g \in L^2$ is such that $\lambda_{Qg} > 0$, we say that *g decays to Pg exponentially fast with decay rate* λ_{Qg} or, simply $g \in ED$. For in this case it is easily seen that

$$\|T_t Q g\|_2 \leqq \|Qg\|_2 e^{-\lambda_{Qg} t}, \qquad t \geqq 0,$$

and, for each $\varepsilon > 0$, there is a $t_\varepsilon \geqq 0$, such that

$$\|T_t Q g\|_2 \geqq \|T_{t_\varepsilon} Q g\|_2 e^{-(\lambda_{Qg} + \varepsilon)(t - t_\varepsilon)}, \qquad t \geqq t_\varepsilon.$$

If $g \in L^2$ is such that $\lambda_{Qg} = 0$, we say that *g decays to Pg more slowly than any exponential* or, simply, $g \in SED$.

To formulate our main result succinctly it is convenient to define a subset $\mathscr{C}$ in L^2 by

$$\mathscr{C} = \left\{ x \in L^2 : \text{either } x \notin L^1 \text{ or both } x \in L^1 \text{ and } \int x \neq 0 \right\}.$$

Its complement $\mathscr{D}$ in L^2 is the subspace of zero means:

$$\mathscr{D} = \left\{ x \in L^2 : x \in L^1 \text{ and } \int x = 0 \right\}.$$

We may now state our *main result*: If (A1)–(A6) hold, then

$$Qg \in \mathscr{C} \Rightarrow g \in SED \quad \text{or, equivalently,} \quad g \in ED \Rightarrow Qg \in \mathscr{D}.$$

Remark 4. In [3], where $P \equiv 0$, it was conjectured that g decayed to zero more slowly than any exponential whenever g was nonnegative and nonzero. Since all such functions are in $\mathscr{C}$, the conjecture is certainly true.

Remark 5. The main result is not quite as sharp as one would like, for our method does not seem to tell us which functions g such that $Qg \in \mathscr{D}$ are also in SED. That there may be such functions will be shown by explicit example in the next section.

We examine the solution of (K) by means of Hille's famous formulae [1]:

(H1)
$$T_t g = \int e^{-\lambda t} d_\lambda E_\lambda g$$

and

(H2)
$$\|T_t g\|_2^2 = \int e^{-2\lambda t} d\|E_\lambda g\|_2^2,$$

[5] These assertions were verified in [3] for A positive definite, in which case $P \equiv 0$.

where $\lambda \mapsto E_\lambda$ is the resolution of the identity of the operator A. From (H2) we obtain the explicit formula

$$\text{(H3)} \qquad -\frac{d}{dt}\ln\|T_t Qg\|_2 = \frac{\int \lambda e^{-2\lambda t} d_\lambda \|E_\lambda Qg\|_2^2}{\int e^{-2\lambda t} d_\lambda \|E_\lambda Qg\|_2^2}.$$

The heart of the argument lies in the analysis of the resolution of the identity of the operator A. To this end we abstract just those properties of A which are needed. Henceforth, A is assumed to be a bounded, symmetric, positive semidefinite, linear operator in L^2 such that

(B1) $g \in L^2 \Rightarrow Ag \in \mathscr{D}$;

(B2) zero is a nonisolated point in the spectrum of A.

We will prove our main result as the following.

THEOREM. *Let A be a bounded, symmetric, positive semidefinite, linear operator in L^2 which satisfies* (B1) *and* (B2). *Then*

$$Qg \in \mathscr{C} \Rightarrow g \in SED \quad \textit{or, equivalently,} \quad g \in ED \Rightarrow Qg \in \mathscr{D}.$$

Remark 6. In Remark 3 we observed that convolution kernels do not satisfy (B1). However, they may, as in the first example in §2, satisfy:

(B1)′ $Ag \in L^1 \cap L^2 \Rightarrow \int Ag = 0$.

In this case we have a modification of the theorem.

THEOREM (alternate version). *Let A be a bounded, symmetric, positive semidefinite, linear operator in L^2 which satisfies* (B1)′ *and* (B2). *Then, for those $g \in L^2$ such that $Qg \in L^1 \cap L^2$,*

$$g \in ED \Rightarrow \int Qg = 0$$

or equivalently

$$\int Qg \neq 0 \Rightarrow g \in SED.$$

Thus, under these weaker hypotheses, we do not conclude that all those functions $g \in L^2$ such that $Qg \notin L^1$ are in SED.

The hypotheses of these two theorems may be motivated by the following heuristic argument. Suppose that zero is an accumulation point of the spectrum of A, but not itself an eigenvalue. If g is an initial function with components in infinitely many of the corresponding "eigenspaces", then $E_\lambda g$ cannot be zero near $\lambda = 0$. The Hille formula (H1) then shows that the induced decay cannot be exponentially fast. Clearly such a g cannot be an eigenfunction, since every eigenfunction must induce exponentially fast decay. Now the properties of A guarantee that the L^1-eigenvalues all have zero mean, whence the hypothesis that the functions without zero mean ought to be the ones which induce decay slower than any exponential. The examples in the next section show that this heuristic argument, while essentially correct, is somewhat simplistic.

2. Some examples. The case of convolution is simple and physically less interesting [3]; nevertheless it provides insight. We suppose $\psi(\mu, \nu) = f(\mu - \nu)$, where $f \in L^1 \cap L^\infty$ is nonnegative and even. This is sufficient to guarantee that (A1)–(A4) and (A6) hold. The formulae which follow will show that A is positive definite so that zero is in its continuous spectrum.

If we denote the Fourier transform in L^2 by $(\hat{\ })$, and use Plancherel's theorem on solutions to (K), we get

$$\|T_t g\|_2^2 = \int_{-\infty}^{\infty} e^{-2\check{f}(\omega)t} |\hat{g}(\omega)|^2 d\omega$$

where

$$\omega \mapsto \check{f}(\omega) = \sqrt{2\pi}\,[\hat{f}(0) - \hat{f}(\omega)] = 2\int_0^{\infty} [1-\cos(\omega s)]\, f(s)\, ds.$$

Alternatively, the resolution of the identity can be computed directly, as in [1], to yield

$$(E_\lambda g)(\mu) = \int_{-\infty}^{\infty} \left\{ \frac{1}{2\pi} \int_{\check{f}^{-1}([0,\lambda])} e^{i\omega(\mu-\nu)} d\omega \right\} g(\nu)\, d\nu,$$

and hence

$$\langle E_\lambda g, g \rangle = \int_{\check{f}^{-1}([0,\lambda])} |\hat{g}(\omega)|^2 d\omega.$$

From Hille's formula (H2) we get

$$\|T_t g\|_2^2 = \int_0^{\infty} e^{-2\lambda t} d_\lambda \left(\int_{\check{f}^{-1}([0,\lambda])} |\hat{g}(\omega)|^2 d\omega \right).$$

From its definition we see that $\check{f}$ is continuous, $\check{f}(0)=0$, $\check{f}(\omega)>0$ for $\omega \neq 0$, and $\lim_{|\omega|\to\infty} \check{f}(\omega) = \hat{f}(0) = \int f$. The set $\check{f}^{-1}(0,\lambda)$ has nonzero Lebesgue measure for $\lambda > 0$. As a consequence, the function $\lambda \mapsto \langle E_\lambda g, g\rangle$ is continuous at zero for every $g \in L^2$, so that zero cannot be an eigenvalue of A. If $g \in L^1 \cap L^2$ and $\int g \neq 0$, then $|\hat{g}(\omega)| > 0$ for ω near zero. In this case $\langle E_\lambda g, g\rangle$ is positive for all $\lambda > 0$. Hence, zero is in the continuous spectrum of A ($P \equiv 0$). Of course A is positive definite, since $E_\lambda \equiv 0$ for $\lambda \leqq 0$. We have thus shown that $g \in L^1 \cap L^2$ and $\int g \neq 0$ together imply that $g \in SED$. Now there are certainly functions $g \in \mathscr{D}$ for which $\hat{g}$ vanishes at zero but not in any neighborhood of zero.[6] Such functions, of course, are also in SED. It follows from a result of Polya that there are also functions in $\mathscr{D}$ with arbitrary positive decay rate.[7]

We have seen that the conclusion of the Theorem, and its alternate in Remark 6, applies to this example. But, as pointed out earlier, (B1) does not hold. Furthermore the conditions on f do not seem to guarantee that (B1)$'$ holds either. To guarantee that (B1)$'$ does hold, add to the assumption the requirement that[8] $\int_0^\infty s^\alpha f(s)\, ds < \infty$, for some $\alpha \geqq \frac{1}{2}$.

[6] The function $g(\mu) = \chi_{[0,1)}(|\mu|) - \chi_{[1,2)}(|\mu|)$ is certainly in $\mathscr{D}$. In this case $\hat{g}(\omega) = 2(\sin\omega)/\omega - (\sin 2\omega)/2\omega$ so that $\hat{g}(0)=0$ but $\hat{g}(\omega)>0$ for ω near zero.

[7] To construct a function in $\mathscr{D}$ with arbitrary positive decay rate, use the method in Chung [2, Thm. 6.5.3] and the discussion which follows.

[8] For $x \in L^2$, $x \notin L^1$, and $y \in L^2$ defined by

$$y = \left(\int f \right) x - f * x,$$

we would like to conclude that $y \in L^1 \Rightarrow \int y = 0$. This is equivalent to $|\hat{y}(0)| < \infty \Rightarrow \hat{y}(0) = 0$, where $\hat{y} = \hat{x}\check{f}$. Now if f has finite (fractional) moment $\alpha \geqq \frac{1}{2}$, then $\check{f}(\omega) = O(\omega^\alpha)$ as $\omega \to 0$. If $\hat{y}(0) \neq 0$, then $\hat{x}$ cannot be in L^2. Hence $y(0) = 0$.

Observe that if f has finite second moment: $\int_0^\infty s^2 f(s)\,ds<\infty$, then $\check{f}(\omega)=O(\omega^2)$ as $|\omega|\to 0$. In this case $\|T_t g\|_2^2=O(1/\sqrt{t})$ as $t\to\infty$, a case of algebraic decay as in the heat equation.

The second example is a modified form of the first: $\psi(\mu,\nu)=f(\mu-\nu)+f(\mu+\nu)$, where again $f\in L^1\cap L^\infty$ is nonnegative and even. This kernel has the same symmetry as nonnegative even functions of $(\mu^2-\nu^2)$; that is, ψ is symmetric and even in each variable.

We can proceed as in the convolution example above to obtain

$$\|T_t g\|_2^2=\int_{-\infty}^{\infty}\exp(-4\check{f}(\omega)t)|\hat{g}_{\text{even}}(\omega)|^2 d\omega+\exp\left(-4\int f\right)\|g_{\text{odd}}\|_2^2$$

where g_{even} and g_{odd} are the even and odd parts of g. In this example, all the odd functions in L^2 are eigenfunctions of A corresponding to the eigenvalue $2\int f$; and zero is still in the continuous spectrum ($P=0$). All the comments of the preceding example apply here as well to the even functions in L^2.

The third example to be considered has the same symmetry as the second, but satisfies (A1)–(A6) and, hence, (B1), (B2).

For $n=1,2,3,\cdots$ set $\delta_n=\sqrt{n}-\sqrt{n-1}$, $I_n=[\sqrt{n-1},\sqrt{n})$, and write χ_n for the characteristic function of the interval I_n. Now define ψ by

$$\psi(\mu,\nu)=\sum_{n=1}^{\infty}\chi_n(|\mu|)\chi_n(|\nu|).$$

Then ϕ is given by

$$\phi(\mu)=2\sum_{n=1}^{\infty}\delta_n\chi_n(|\mu|).$$

Since $\sqrt{n}\,\delta_n\to\frac{1}{2}$ as $n\to\infty$, it follows that $\phi\in L^2\cap L^\infty$, but $\phi\notin L^1$. Clearly (A1)–(A6) hold.

The calculation of A is straightforward, yielding

$$(Ag)(\mu)=2\sum_{n=1}^{\infty}\delta_n[g(\mu)-\bar{g}_n]\chi_n(|\mu|),$$

where $\bar{g}_n$ is the mean value of the even part of g on I_n;

$$\bar{g}_n=\frac{1}{\delta_n}\int_{I_n}g_{\text{even}}(\nu)\,d\nu.$$

The null space $N(A)$ of A consists of all functions of the form $\sum_{n=1}^{\infty}c_n\chi_n(|\mu|)$ for which $\sum_{n=1}^{\infty}\delta_n c_n^2$ is finite. The projection P onto $N(A)$ is

$$(Pg)(\mu)=\sum_{n=1}^{\infty}\bar{g}_n\chi_n(|\mu|).$$

The solution to (K) can now be given explicitly by

$$(T_t g)(\mu)=(Pg)(\mu)+\sum_{n=1}^{\infty}e^{-2\delta_n t}[g(\mu)-\bar{g}_n]^2\chi_n(|\mu|).$$

Thus

$$\|T_tQg\|_2^2=\sum_{n=1}^{\infty}e^{-2\delta_n t}\int_{-I_n\cup I_n}[g(\nu)-\bar{g}_n]^2 d\nu$$

where $-I_n$ is the interval $(-\sqrt{n}\,,-\sqrt{n-1}\,]$.

From this last formula we see that $g\in L^2$ is in ED if and only if, for some $N\geqq 1$, $g_{\text{odd}}=0$ and $g_{\text{even}}=\bar{g}_n$ on each I_n, $n\geqq N$. In this case $\lambda_{Qg}=2\delta_N>0$, for the least such N. By direct use of this formula we see that if $Qg\in\mathscr{C}$ then $g\in SED$. However, all the odd functions $g\in L^1\cap L^2$ with unbounded support also lie in SED. For such functions $Qg=g\in\mathscr{D}$. If $g\in L^1\cap L^2$, g is even, and $g\neq\bar{g}_n$ on an infinite sequence of I_n's, then $Qg\in\mathscr{D}$ but $g\in SED$.

Our final example is the most interesting. Let $\psi(\mu,\nu)=f(\mu^2-\nu^2)$, where once more we take $f\in L^1\cap L^\infty$ to be nonnegative and even. As pointed out in Remark 3, the "acoustic kernel" $f(s)=1/(1+s^2)$ is of this type [3]. Under these conditions (A1)–(A6) are satisfied. In fact, ϕ is even and

$$\phi(\mu)=\begin{cases}\displaystyle\int_0^{\infty}f(s)\frac{ds}{\sqrt{s}}, & \mu=0,\\[2ex] \displaystyle\frac{1}{|\mu|}\int_{-\mu^2}^{\infty}f(s)\frac{ds}{\sqrt{1+s/\mu^2}}, & |\mu|\neq 0\end{cases}$$

so that $|\mu|\phi(\mu)\to\int f$ as $|\mu|\to\infty$.

As before, symmetry allows us to consider the even and odd functions separately. Since $x_{\text{odd}}\mapsto Ax_{\text{odd}}=\phi x_{\text{odd}}$, the L^2-multiplier, we have from (H2)

$$\|T_tg\|_2^2=\int e^{-2\lambda t}d_\lambda\langle E_\lambda g_{\text{even}},g_{\text{even}}\rangle+\int e^{-2\lambda t}d_\lambda\left(\int_{\phi^{-1}([0,\lambda])}|g_{\text{odd}}(\mu)|^2d\mu\right).$$

Hence the odd functions $g\in L^2$ with unbounded support are in SED, while those with bounded support are in ED.

It is not hard to show that A is positive definite [3], so that zero is in the continuous spectrum. If $g\in L^2$ is even $(g\neq 0)$ and $g\notin L^1$, then, $g\in\mathscr{C}$, and, by the Theorem, to SED. On the other hand, if $g\in L^1\cap L^2$ is even $(g\neq 0)$ and $\int g=0$, then $g\in\mathscr{D}$. At this point we cannot tell if such a g is in SED. Thus SED certainly contains the odd functions with unbounded support, and the even functions which have nonzero mean or which are not integrable at all.

3. The spectral resolution of the identity and the asymptotic behavior of solutions. As a preliminary we specify some notation. For $f\in L^1$ we write $\int f$ for the Lebesgue integral $\int_{-\infty}^{\infty}f(s)\,ds$. For functions f and g such that $fg\in L^1$ we write $\langle f,g\rangle$ for $\int fg$. Then for $f,g\in L^2$, $\langle f,g\rangle$ is the standard inner product and $\|f\|_2\equiv\langle f,f\rangle^{1/2}$ is the L^2-norm of f.

Since A is a symmetric linear operator in L^2, it has a real spectral resolution of the identity $\lambda\mapsto E_\lambda$. Recall that $\{E_\lambda:\ -\infty<\lambda<\infty\}$ is a family of bounded symmetric linear operators in L^2 such that:

(i) $\lambda\mapsto E_\lambda$ is (normalized) left-continuous in the uniform operator topology;

(ii) $E_{\lambda'}E_{\lambda''}=E_{\min\{\lambda',\lambda''\}}$.

Properties ((B1), (B2)), or ((B1)′, (B2)), further imply:

$$
\begin{aligned}
&E_\lambda = 0, && -\infty < \lambda \leqq 0,\\
\text{(iii)}\ &E_\lambda \text{ is a nonzero projection}, && 0 < \lambda \leqq \|A\|_2,\\
&E_\lambda = 1, \text{ the identity}, && \|A\|_2 < \lambda < \infty .
\end{aligned}
$$

$$
\text{(iv)}\ \lim_{\lambda \downarrow 0} E_\lambda = \begin{cases} 0 & \text{if } \lambda = 0 \text{ is in the continuous spectrum of } A,\\ P \neq 0 & \text{if } \lambda = 0 \text{ is an eigenvalue of } A. \end{cases}
$$

The spectral theorem in L^2 can be stated as follows: for each pair $g,h \in L^2$, $\lambda \to \langle E_\lambda g, h\rangle$ is of bounded variation on $(-\infty, \infty)$, and

$$
\langle g,h\rangle = \int d_\lambda \langle E_\lambda g,h\rangle,
$$

$$
\langle Ag,h\rangle = \int \lambda d_\lambda \langle E_\lambda g,h\rangle .
$$

Our main technical result is embodied in the following lemma.

LEMMA. *Let A be a symmetric, positive semidefinite linear operator in L^2. Then for each $g \in L^2$, and $0 < a < b < \infty$, there is an $f_{a,b} \in L^2$, depending on f, such that*

$$
Af_{a,b} = E_{[a,b)}\, g .
$$

Moreover,

$$
\|f_{a,b}\|_2 \leqq \frac{1}{a} \|E_{[a,b)}\, g\|_2 .
$$

The proof of this lemma will be deferred until the next section. In this section its consequences will be explored. To facilitate this we define $\hat{\lambda}_h$, for each $h \in L^2$ by

$$
\hat{\lambda}_h = \inf\{\lambda : \langle E_\lambda h, h\rangle > 0\}.
$$

COROLLARY 1. *Let* (B1) *hold. If $g \in L^2$ is such that $\hat{\lambda}_g > 0$, then $g \in \mathscr{D}$; that is $g \in L^1 \cap L^2$ and $\int g = 0$.*

Proof. For $b > \|A\|_2$, $E_{[a,b)}g = g - E_a g$. Now $E_a g = 0$ for $0 < a \leqq \hat{\lambda}_g$. For this choice of a and b, use the lemma to conclude that $Af_{a,b} = g$. By (B1) we have $g \in \mathscr{D}$. □

COROLLARY 2. *Let A satisfy* (B1). *For each $g \in L^2$, and $0 < a < b < \infty$, we have $E_{[a,b)}g \in \mathscr{D}$. If, in addition, $\|A\|_{2,1} < \infty$, then*

$$
\|E_{[a,b]}g\|_1 \leqq \frac{\|A\|_{2,1}}{a} \|E_{[a,b)}\, g\|_2 .
$$

The boundedness of A as an L^2 operator is not required in the inequality.

Proof. From the lemma we have that

$$
Af_{a,b} = E_{[a,b)}\, g .
$$

Hence (B1) implies $E_{[a,b)}g \in \mathscr{D}$.

Next, if $\|A\|_{2,1} < \infty$ we have

$$
\|E_{[a,b)}\, g\|_1 = \|Af_{a,b}\|_1 \leqq \|A\|_{2,1} \|f_{a,b}\|_2 \leqq \frac{\|A\|_{2,1}}{a} \|E_{[a,b)}\, g\|_2 . \qquad \square
$$

Observe that if (B1) holds Corollary 2 implies that $E_a g \in L^1 \cap L^2$ if and only if $E_b g \in L^1 \cap L^2$. Hence, the boundedness of A guarantees that $g \in L^1 \cap L^2$ if and only if $E_\lambda g \in L^1 \cap L^2$ for some, and hence all, $\lambda > 0$. In this case $\int g = \int E_\lambda g$, $\lambda > 0$.

Proof of the theorem. From Corollary 2 we see that

(i) if $g \in L^2$ and $g \notin L^1$, then $E_\lambda g \notin L^1$ for $\lambda > 0$;

(ii) if $g \in L^1 \cap L^2$ and $\int g \neq 0$, then $\int E_\lambda g \neq 0$ for $\lambda > 0$;

that is, if $g \in \mathscr{C}$, then $\hat{\lambda}_g = 0$.

From (H3) it follows easily that $\lambda_{Qg} = \hat{\lambda}_{Qg}$; that is, $\hat{\lambda}_{qg}$ is the decay rate of g. Since we have assumed that (B1), (B2) hold, the theorem is proved. □

Remark 7. Suppose (B1)′ holds instead of (B1). Then, from the lemma, we see that $E_{[a,b)} g \in L^1 \cap L^2$ implies $\int E_{[a,b)} g = 0$. We then obtain a modified version of Corollary 1.

COROLLARY 1′. *Let* (B1)′ *hold. If* $g \in L^1 \cap L^2$ *is such that* $\hat{\lambda}_g > 0$, *then* $g \in \mathscr{D}$; *that is,* $\int g = 0$.

From this we see that if $Qg \in L^1 \cap L^2$ and $g \in ED$, then $\int Qg = 0$. This proves the alternate version of the theorem in Remark 6.

4. Proof of the lemma. For each $g \in L^2$ the function $\lambda \to \gamma_g(\lambda)$ induced by g through

$$\lambda \mapsto \gamma_g(\lambda) = \langle E_\lambda g, g \rangle$$

is nondecreasing and left-continuous. Moreover

$$\text{(i)} \quad \gamma_g(\lambda) = 0, \qquad \lambda \leqq 0,$$

$$\text{(ii)} \quad \gamma_g(\lambda) = \|g\|_2^2, \qquad \lambda > \|A\|_2,$$

$$\text{(iii)} \quad \lim_{\lambda \downarrow 0} \gamma_g(\lambda) = \|Pg\|_2^2.$$

Let $\lambda \mapsto \tilde{f}(\lambda)$ be measurable with respect to the Lebesgue-Stieltjes measure $d\gamma_g$ induced on $(-\infty, \infty)$ by the function γ_g. Denote by L_g^2 the Hilbert Space of such functions $\tilde{f}$ for which

$$\int \tilde{f}^2(\lambda)\, d_\lambda \gamma_g(\lambda) < \infty.$$

The inner product of $\tilde{f}, \tilde{h} \in L_g^2$ is given by $\langle \tilde{f}, \tilde{h} \rangle_g = \int \tilde{f}(\lambda) \tilde{h}(\lambda)\, d_\lambda \gamma_g(\lambda)$, and the L_g^2 norm of $\tilde{f}$ is denoted by $\|\tilde{f}\|_{g,2}$. The elements of L_g^2, while still called "functions", must be understood as equivalence classes with respect to $d\gamma_g$ in the usual way.

As mentioned in the Spectral Theorem, the function $\lambda \mapsto \langle E_\lambda g, h \rangle$ is of bounded variation on $(-\infty, \infty)$. Indeed, from the polar equality

$$\langle E_\lambda g, h \rangle = \frac{1}{4} [\langle E_\lambda(g+h), (g+h) \rangle - \langle E_\lambda(g-h), (g-h) \rangle]$$

$$= \frac{1}{4} [\gamma_{(g+h)}(\lambda) - \gamma_{(g-h)}(\lambda)],$$

which is the difference of two bounded nondecreasing functions.

CLAIM 1. *There is an isometry* $\tilde{f} \xrightarrow{\mathscr{V}} f$ *taking* L_g^2 *onto* L^2 *realized by*

$$\text{(V)} \qquad \langle f,h\rangle = \int \tilde{f}(\lambda)\, d_\lambda \langle E_\lambda g,h\rangle$$

for all $h \in L^2$.

Proof of Claim 1. We will show that the right-hand side of (V) defines a continuous linear functional on L^2 for each $\tilde{f} \in L_g^2$. Then by the representation theorem of F. Riesz, there is a unique $f \in L^2$ such that (V) holds for all $h \in L^2$. Write $f = \mathscr{V}\tilde{f}$.

The map $\tilde{f} \xrightarrow{\mathscr{V}} f$ is clearly linear. To see that $\|f\|_2 = \|\tilde{f}\|_{g,2}$, proceed as follows. By (V),

$$\begin{aligned}\langle E_\lambda g, f\rangle &= \int \tilde{f}(\mu)\, d_\mu \langle E_\mu g, E_\lambda g\rangle \\ &= \int \tilde{f}(\mu)\, d_\mu \langle E_\mu E_\lambda g, g\rangle \\ &= \int \tilde{f}(\mu)\, d_\mu \langle E_{\min\{\lambda,\mu\}} g, g\rangle \\ &= \int_{-\infty}^{\lambda} \tilde{f}(\mu)\, d_\mu \langle E_\mu g, g\rangle.\end{aligned}$$

Thus

$$\langle E_\lambda g, f\rangle = \int_{-\infty}^{\lambda} \tilde{f}(\mu)\, d_\mu \gamma_g(\mu).$$

But then

$$\begin{aligned}\|f\|_2^2 = \langle f,f\rangle &= \int \tilde{f}(\lambda)\, d_\lambda \langle E_\lambda g, f\rangle \\ &= \int \tilde{f}(\lambda)\, d_\lambda\left(\int_{-\infty}^{\lambda} \tilde{f}(\mu)\, d_\mu \gamma_g(\mu)\right) \\ &= \int |\tilde{f}(\lambda)|^2 d_\lambda \gamma_g(\lambda) \\ &= \|\tilde{f}\|_{g,2}.\end{aligned}$$

To complete the proof we must show that the right-hand side of (V) defines a continuous linear functional on L^2 for each $\tilde{f} \in L_g^2$. Let $\{I_k\}$ denote any partition of $(-\infty, \infty)$, and let $\lambda_k \in I_k$.

First suppose $\tilde{f}$ is continuous. Since the support of $d\gamma_g$ is bounded, $\tilde{f}$ is certainly in L_g^2. Consider sums of the form

$$\sum_k \tilde{f}(\lambda_k) \langle E_{I_k} g, h\rangle.$$

Note that $E_{I_k}g=0$ if $I_k\subset(-\infty,0]\cup(\|A\|_2,\infty)$, so that the sums are finite. Then

$$\left|\sum_k \tilde{f}(\lambda_k)\langle E_{I_k}g,h\rangle\right| \leqq \sum_k |\tilde{f}(\lambda_k)||\langle E_{I_k}g,E_{I_k}h\rangle|$$

$$\leqq \left(\sum_k |\tilde{f}(\lambda_k)|^2\langle E_{I_k}g,g\rangle\right)^{1/2}\left(\sum_k\langle E_{I_k}h,h\rangle\right)^{1/2}$$

$$\leqq \left(\sum_k |\tilde{f}(\lambda_k)|^2\langle E_{I_k}g,g\rangle\right)^{1/2}\|h\|_2.$$

From the definition of the Lebesgue Stieltjes integral we have

$$\left|\int\tilde{f}(\lambda)\,d_\lambda\langle E_\lambda g,h\rangle\right| \leqq \left(\int\tilde{f}(\lambda)^2 d_\lambda\gamma_g(\lambda)\right)^{1/2}\|h\|_2\leqq\|\tilde{f}\|_{g,2}\|h\|_2$$

for all continuous f. Since the continuous functions are dense in L_g^2, Claim 1 is verified. □

Remark 8. Denote the range of $\mathcal{V}$ by M_g. Let $\hat{M}_g$ be the subspace of L^2 generated as follows:

$$\hat{M}_g=\operatorname{cl}\operatorname{spn}\{E_bg-E_ag:-\infty<a<b<\infty\}.$$

Then $\hat{M}_g=M_g$. To see this, note that

$$E_bg-E_ag=E_{[a,b)}g=\int\chi_{[a,b)}(\lambda)\,d_\lambda E_\lambda g.$$

Hence $\hat{M}_g\subset M_g$. But $\operatorname{spn}\{\chi_{[a,b)}:-\infty<a<b<\infty\}$ is dense in L_g^2, so that $\hat{M}_g=M_g$.

Remark 9. $g\in M_g$. Indeed, since $d_\gamma g$ is bounded, it follows that 1^*, the constant function with value 1, is in L_g^2. Thus $g=\mathcal{V}1^*$.

On the strength of Claim 1 we write

$$f=\int\tilde{f}(\lambda)\,d_\lambda E_\lambda g$$

to denote $\mathcal{V}\tilde{f}$.

We have shown that L_g^2 is isometrically isomorphic to the (closed) subspace M_g of L^2. The next result shows that A is, in effect, multiplication in L_g^2.

CLAIM 2. *The restriction of A to M_g is equivalent to multiplication in L_g^2:*

$$((\mathcal{V}^{-1}A\mathcal{V})\tilde{f})(\lambda)=\lambda\tilde{f}(\lambda).$$

Proof. We first show that, for $f\in M_g$ and $h\in L^2$, $d_\lambda\langle E_\lambda f,h\rangle=\tilde{f}(\lambda)d_\lambda\langle E_\lambda g,h\rangle$. Indeed,

$$\langle E_\lambda f,h\rangle=\langle f,E_\lambda h\rangle=\int\tilde{f}(\mu)\,d_\mu\langle E_\mu g,E_\lambda h\rangle$$

$$=\int\tilde{f}(\mu)\,d_\mu\langle E_\lambda E_\mu g,h\rangle$$

$$=\int\tilde{f}(\mu)\,d_\mu\langle E_{\min\{\lambda,\mu\}}g,h\rangle$$

$$=\int_{-\infty}^{\lambda}\tilde{f}(\mu)\,d_\mu\langle E_\mu g,h\rangle.$$

Then from the spectral theorem we have

$$\langle Af,h\rangle=\int\lambda\, d_\lambda\langle E_\lambda f,h\rangle=\int\lambda\tilde{f}(\lambda)\, d_\lambda\langle E_\lambda g,h\rangle.$$

Finally, since $\lambda\mapsto\lambda\tilde{f}(\lambda)$ is in L_g^2 whenever $\tilde{f}\in L_g^2$, we see that $(\mathscr{V}^{-1}Af)(\lambda)=\lambda\tilde{f}(\lambda)$, where $f=\mathscr{V}\tilde{f}$. This proves Claim 2. □

At last we turn to the

Proof of the lemma. For $0<a<b<\infty$ the function

$$\mu\mapsto\tilde{f}_{a,b}(\mu)=\frac{1}{\mu}\chi_{[a,b)}(\mu)$$

is in L_g^2. Then for $f_{a,b}=\mathscr{V}\tilde{f}_{a,b}$ we have, by Claim 1,

$$\|f_{a,b}\|_2^2=\|\tilde{f}_{a,b}\|_{g,2}^2=\int\frac{1}{\mu^2}\chi_{[a,b)}(\mu)\, d_\mu\gamma_g(\mu)\leqq\frac{1}{a^2}\|E_{[a,b)}g\|_2^2.$$

Next, by Claim 2,

$$\langle Af_{a,b},h\rangle=\int\mu\frac{1}{\mu}\chi_{[a,b]}(\mu)\, d_\mu\langle E_\mu g,h\rangle$$
$$=\int_a^b d_\mu\langle E_\mu g,h\rangle=\langle E_{[a,b)}g,h\rangle.$$

The lemma is proved. □

Acknowledgments. The author is grateful to Professors J. A. Langsam and V. J. Mizel for many valuable discussions.

REFERENCES

[1] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[2] K. L. Chung, *A Course in Probability Theory*, Academic Press, New York, 1974.

[3] M. J. Leitman, *An integrodifferential equation for plane waves propagating into a random fluid: asymptotic behavior*, this Journal, 12 (1981), pp. 560–571.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

GENERALIZED OPERATOR RICCATI EQUATIONS*

HENDRIK J. KUIPER†

Abstract. The Riccati equation

$$\frac{dU}{dt}=AU+UB^*+UCU+D$$

on the space $\mathscr{L}(X)$ of bounded linear operators on a reflexive Banach space X arises in control theory and transport theory. A more general problem is the following. Let A and B be closed linear operators in X and X^* respectively and let $\mathscr{F}$ be a map from $[0,T)\times\mathscr{L}(X)$ into $\mathscr{L}(X)$ and consider the initial value problem

$$\frac{dU}{dt}=\mathscr{Cl}[AU+UB^*]+\mathscr{F}(t,U),\qquad U(0)=U_0,$$

on $\mathscr{L}(X)$. It is shown that for a certain class of initial conditions, determined by A, B and the geometry of X, there exist continuously differentiable solutions with respect to the uniform operator topology. It is also shown that if A and B have compact inverses, then there exist solutions with respect to the strong operator topology for arbitrary initial conditions.

1. Introduction. Let $\mathscr{L}(X,Y)$ be the linear space of all bounded linear operators from the Banach space X into the Banach space Y and let $\mathscr{L}(X)=\mathscr{L}(X,X)$. When we endow this space with the strong operator topology we obtain a topological vector space which we will denote by $\mathscr{L}_s(X,Y)$. Similarly we can obtain a Banach space $\mathscr{L}_u(X,Y)$ by imposing the uniform operator topology. In either one of the two spaces $\mathscr{L}_s(X)$ or $\mathscr{L}_u(X)$ we can look at the infinite-dimensional Riccati equation

$$\frac{dS}{dt}=\mathscr{Cl}[AS+SB]+SCS+D,\qquad S(0)=S_0, \tag{1.1}$$

where $\mathscr{Cl}$ denotes operator closure and where A and B are closed linear operators in X. This equation arises in optimal control theory (see e.g. [14]) as well as in transport theory (see e.g. [11], [18]).

Several results have been obtained since Lions proved the existence of distributional solutions. In particular we mention the work of Curtain and Pritchard [3], DaPrato [4], [5], Lukes and Russell [15], Tartar [21], and Temam [22]. All of these results deal with either distributional solutions or solutions which are differentiable in the weak operator topology. The question arises whether there exist solutions in $\mathscr{L}_s(X)$ (i.e., strongly differentiable) or even in $\mathscr{L}_u(X)$. Generally, when the coefficients A and B are unbounded, one can not obtain a solution in $\mathscr{L}_u(X)$ unless the initial condition is such that the orbit will be restricted to some smaller subspace of operators such as the compact linear operators. Such a restriction is sometimes natural. For example (1.1) arises in linear filtering theory where S represents a covariance operator which indeed, typically is compact. An existence theorem for strong solutions for equation (1.1), with noncompact initial values, was obtained in [11] in case X is a Hilbert space. The proof

*Received by the editors March 1, 1983, and in revised form February 9, 1984. This research was partly supported by the National Science Foundation under grant MCS 81-03441.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85287.

is very much dependent on the quadratic structure of the right-hand side of (1.1) and can not be extended to more general equations such as

$$\frac{dS}{dt}=AS+SB+\Phi(S), \qquad S(0)=S_0. \tag{1.2}$$

Here we have deleted the operator closure which is, however, implied. We will do so in the future when there is no risk of ambiguity. Tartar studied this equation [23] and obtained existence of distributional solutions as well as many qualitative results such as strong continuity from the right and a priori estimates. Generalized Riccati equations such as (1.2) arise in some applications such as optimal control of linear systems with state dependent white noise and quadratic cost. The finite-dimensional version of this problem is studied, e.g., in [23] and [24]. The infinite-dimensional version is similar and, of course, also of practical interest.

Let $\mathscr{C}(X)$ denote the closed linear operators and let $\mathscr{K}_0(X)$ denote the compact linear operators in X. For the sake of simplcity we shall assume that X is a reflexive space and we shall identify X^{**} with X. Using Sobolevskii's results on nonlinear semigroups [20] we shall obtain in §4 the existence of a continuously differentiable solution of

$$\frac{dS}{dt}=\mathscr{C}\ell[A(t,S)S+SB(t,S)^*]+\mathscr{F}(t,S), \qquad S(0)=S_0\in\mathscr{K}, \tag{1.3}$$

in a subspace $\mathscr{K}$ of $\mathscr{L}_u(X)$, where $\mathscr{K}_0(x)\subset\mathscr{K}$. Here A: $[0,t_0)\times\mathscr{L}_u(X)\rightarrow\mathscr{C}(X)$, B: $[0,t_0)\times\mathscr{L}_u(X)\rightarrow\mathscr{C}(X^*)$ and $\mathscr{F}$: $[0,t_0)\times\mathscr{L}_u(X)\rightarrow\mathscr{L}_u(X)$. The definition of $\mathscr{K}$ depends on A and B. For example, if A and B are bounded linear operators then $\mathscr{K}=\mathscr{L}_u(X)$, as we should expect.

We also prove the existence of strong solutions for arbitrary initial values in $\mathscr{L}_u(X)$. Although our results allow X to be a Banach space, it should be pointed out that they are new results even if X is a Hilbert space. The major difficulty in the proofs of these results lies in the fact that before we can apply Sobolevskii's results we must show that the operator

$$\mathbf{A}\colon S\rightarrow\mathscr{C}\ell[AS+SB^*]$$

can be extended to a densely defined closed linear operator which is accretive. Of course, this is not in general true and our approach depends heavily on the relation between A,B and the geometry of X. First of all, we need the norm attaining operators to be dense in $\mathscr{L}_u(X)$. Lindenstrauss [13] showed that this is the case when X is reflexive, although that is not a necessary condition. For example, Iwanik [9] proved that the norm attaining operators are dense in $\mathscr{L}_u(L^1[0,1])$. Secondly, we need (very loosely speaking) the geometry of X (resp. X^*) to be "nice" on that part where A (resp. B) becomes unbounded. To make this statement rigorous, we employ the concept of a π-space which was introduced by Lindestrauss. These concepts will be discussed in §2. The results from the theory of semigroups of operators which we will need will be given in §3.

Although the results obtained here are for the case where the underlying space X is reflexive, it is possible to extend them to cases where this is not true. For example, $L^1[0,1]$ is a π-space and the norm attaining operators are dense in $\mathscr{L}_u(L^1[0,1])$. These two main requirements having been met, we can obtain existence of solution to Riccati equations on $\mathscr{L}(L^1[0,1])$. At the end of this article we will point out how to alter the hypotheses in order to handle the situation where X is not reflexive.

2. Preliminaries. We first define what we shall mean by a π-space.

DEFINITION 1. A Banach space Y is called a π_λ-space if there exists a collection $\{F_\tau\}_{\tau\in T}$ of finite-dimensional subspaces, directed by inclusion, such that their union is dense in Y and such that for each τ there exists a linear projection operator π_τ, with $\pi_\tau Y = F_\tau$ and $\|\pi_\tau\| \leqq \lambda$. The space Y is said to be a π-space if it is a π_λ-space for some $\lambda \geqq 1$.

When this concept was introduced by Lindenstrauss, he referred to it instead as a "space with the projection-approximation property". Browder and deFigueiredo [1] introduced the notation "π_1-space", which was subsequently used by several other authors. Similarly Singer in his book on the theory of bases [18, p. 611] refers to these spaces as "spaces which have an extended π_λ-basis". Before we give a few examples of π-spaces, it should be noted that a space Y is a π-space if there exists a uniformly bounded net $\{u_\lambda\}$ of finite rank projection operators which converges to the identity, I, in $\mathscr{L}_s(Y)$ ([18, Thm. 18.4]). Also if Y^* is a π-space then so is Y ([18, Thm. 18.8]).

Clearly any Hilbert space is a π-space, as well as any Banach space which has a Schauder basis: Let $\{e_1, e_2, \cdots\}$ be a Schauder basis, and for any $x = \sum_{i=1}^{\infty}\alpha_i e_i$ let $\pi_n x = \sum_{i=1}^{n}\alpha_i e_i$. Since $\lim_{n\to\infty}\pi_n x = x$ for all x we conclude, using the uniform boundedness theorem, that there exists a $k>0$ such that $\|\pi_n\| \leqq k$ for all n. As another example suppose μ is a σ-finite measure on a set Ω, then $L^p(\mu)$, $1 \leqq p \leqq \infty$, is a π-space. Indeed, consider the set of all partitions τ of Ω into finitely many sets of finite measure Ω_1, $\Omega_2, \cdots, \Omega_m$ and (unless μ is a finite) a set Ω_0 of infinite measure. These form a directed set, T, in the obvious manner. Let $\tau = \{\Omega_0, \Omega_1, \Omega_2, \cdots, \Omega_m\} \in T$ and set F_τ equal to the linear space of all functions of the form $\sum_{i=1}^{m}\alpha_i\chi_i$, where χ_i is the characteristic function of Ω_i. Finally define

$$\pi_\tau f = \sum_{i=1}^{m} \mu(\Omega_i)^{-1}\left\{\int_{\Omega_i} f\,d\mu\right\}_{\chi_i}.$$

One easily verifies the requirements of a π_i-space are satisfied.

Other Banach spaces which are π-spaces include c_0, and $C(S)$, S any compact metric space [17].

If Y is a π_k-space with associated projection operators $\{\pi_F\}$, indexed here by their ranges, then the net $\pi_F y$ tends to y for each $y \in Y$. To see this suppose ε is an arbitrary positive number. Then there exists an element z which lies in the range of one of these projections, say π_G, such that $\|y - z\| \leqq \varepsilon(k+1)^{-1}$. Then

$$\|y - \pi_F y\| \leqq \|y - z\| + \|z - \pi_F z\| + \|\pi_F z - \pi_F y\| \leqq \varepsilon$$

for any $F \supset G$.

LEMMA 2. *Let $\mathscr{D}$ be a dense linear subspace of a π-space Y. Then there exists a constant k_1 and a net $\{\pi_\alpha\}$ of linear projection operators such that, for each α, $\pi_\alpha Y$ is a finite-dimensional subspace of $\mathscr{D}$, $\|\pi_\alpha\| \leqq k_1$, and such that $\pi_\alpha \to I$ in the strong topology. Similarly if $\mathscr{D}^*$ is a dense linear subspace of Y^*, then there exists a net $\{\tilde{\pi}_\beta\}$ of linear projection operators in Y such that $\tilde{\pi}_\beta^* Y^*$ is a finite-dimensional subspace of $\mathscr{D}^*$, $\|\tilde{\pi}_\beta\| \leqq k_1$ for all β, and $\tilde{\pi}_\beta \to I$ in the strong topology.*

Proof. Let B_ρ be the ball of radius ρ centered at 0 in Y and let ∂B_ρ be its surface. If S is any subspace of Y then we let $S_\rho = S \cap B_\rho$ and $\partial S_\rho = S \cap \partial B_\rho$. Since Y is an π-space there exists a net $\{\tilde{\pi}_F\}$ of linear projection operators, $\|\tilde{\pi}_F\| \leqq k$, indexed by their ranges (i.e., $\pi_F Y = F$ and $\cup F$ is dense in Y). We note that if p is any bounded linear projection operator in Y with range R and kernel K then $\|p\|\operatorname{dist}(K,\partial R_1)=1$. This follows from the definition of the norm:

$$\|p\|=\sup\left\{\left.\frac{\|r\|}{\|r-\kappa\|}\right| r\in\partial R_1,\kappa\in K\right\}$$

$$=\left[\inf\left\{\ \|r-\kappa\|\,\middle|\, r\in\partial R_1,\kappa\in K\right\}\right]^{-1}.$$

In particular if K^F denotes the kernel of $\tilde{\pi}_F$ then $Y=F+K^F$ and $\operatorname{dist}(K^F,\partial F_1)=\|\tilde{\pi}_F\|^{-1}\geqq k^{-1}$. Since $\mathscr{D}$ is dense in Y we can find a linear subspace G^F such that $G^F\subset\mathscr{D}$, $G_1^F\subset F_1+\delta F_1+\delta K_1^F$, and $\partial G_1^F\subset\partial F_1+\delta F_1+\delta K_1$ with $0<\delta<1/4kd$, $d=\dim F$. Hence

$$\operatorname{dist}\left(\partial G_1^F,K^F\right)>k^{-1}-\frac{2}{4kd}\geqq\frac{1}{2k}.$$

Clearly $Y=G^F\oplus K^F$ and the associated projection operator π_F which projects onto G^F along K^F satisfies $\|\pi_F\|\leqq 2k$. Next we estimate $\|\pi_F-\tilde{\pi}_F\|$. Suppose $x\in Y$, $\|x\|=1$, then $\pi_F x+(I-\pi_F)x=\tilde{\pi}_F x+(I-\tilde{\pi}_F)x$. Therefore $\pi_F x-\tilde{\pi}_F x\in(G_{2k}^F-F_k)\cap K^F\subset 2k(G_1^F-F_1)\subset 2k(F_{2+\delta}-\delta K_1^F)\cap K^F\subset 2k\delta K_1^F$. Hence $\|\pi_F-\tilde{\pi}_F\|\leqq 2k\delta<1/\dim F$. But this means that the net $\pi_F x$ converges to x. We next consider the second case. Since π_F^* is a projection operator on Y^* whose range is finite-dimensional, the same argument as above shows there exists a projection operator p_F whose range is contained in $\mathscr{D}^*$ and which satisfies $\|p_F-\pi_F^*\|\leqq 1/\dim\pi_F^*Y^*$. Letting $\hat{\pi}_F=p_F^*$ we are done.

The concept of numerical range of an operator in a Hilbert space naturally carries over to operators in Banach spaces by means of the duality map:

$$J\colon X\to 2^{X^*},$$
$$Jx=\left\{x^*\in X^*\,\middle|\,\|x^*\|^2=\|x\|^2=\left\langle x,x^*\right\rangle\right\}.$$

We now define the numerical range, $\theta(L)$, of an operator $L\in\mathscr{C}(X)$ as the closure of

$$\left\{z\in\mathbb{C}\,|\,z=\langle Lx,x^*\rangle,x\in\mathscr{D}(L),\|x\|=1,x^*\in Jx\right\}.$$

Definition 3. We say that L is of type (ω,δ), $\omega\in\mathbb{R}$, $0<\delta<\pi/4$, if $L\in\mathscr{C}(X)$ and:

(i) $\sigma(L)\cup\Theta(L)\subset\Sigma_{\omega,\delta}\equiv\{z\in\mathbb{C}\,|\,|\arg(\omega-z)|\leqq\pi/2-\delta\}$;

(ii) there exists a closed linear subspace $X_L\subset X$ such that $\mathscr{D}(L)=X_L+\mathscr{E}_L$ where $\bar{\mathscr{E}}_L$ is a π-space, and $X=X_L\oplus\bar{\mathscr{E}}_L$.

Suppose F is a map from a set S into $\mathscr{C}(X)$ and that there exist ω,δ and a decomposition $X=X_L\oplus\bar{\mathscr{E}}_L$ as above such that $F(s)$ satisfies (i) and (ii) for each $s\in S$. Then we will say that F is of class (ω,δ).

The first part of the definition is like the usual restriction put on a linear operator in order to ensure that it generates a holomorphic semigroup. The last part of the definition is satisfied if, for example, X is a π-space, or if L is bounded.

We shall assume

$$A\colon\mathscr{D}(A)\subset X\to X \text{ and } B\colon\mathscr{D}(B)\subset X^*\to X^* \text{ are of type } (\omega,\delta).$$

This means that $\mathscr{D}(A)=X_A+\mathscr{E}_A$ and that there exists a bounded linear projection operator π onto $\bar{\mathscr{E}}_a$ along X_A (i.e., $\operatorname{Ker}\pi=X_A$). Similarly $\mathscr{D}(B)=X_B^*+\mathscr{E}_B^*$ and there exists a projection operator ρ^* onto $\bar{\mathscr{E}}_B^*$ along X_B^*. Clearly $X^{**}=(X_B^*)^*+(\mathscr{E}_B^*)^*$, so that by means of the canonical identification of X with X^{**} we have $X=X_B+\bar{\mathscr{E}}_B$ where we have used the notation $(X_B^*)^*=X_B$ and $(\mathscr{E}_B^*)^*=\bar{\mathscr{E}}_B$. Let ρ denote the projection onto $\bar{\mathscr{E}}_B$ along X_B. By Lemma 2 there exist nets $\{\tilde{\pi}_\alpha\}$ and $\{\tilde{\rho}_\beta\}$ of projection operators on $\bar{\mathscr{E}}_A$ and $\bar{\mathscr{E}}_B$ respectively such that $\tilde{\pi}_\alpha$ tends to the identity on $\bar{\mathscr{E}}_A$ while $\tilde{\rho}_\beta$ tends to the identity on $\bar{\mathscr{E}}_B$ and such that $\text{Range}(\tilde{\pi}_\alpha)\subset\mathscr{E}_A$ and $\text{Range}(\tilde{\rho}_\beta)^*\subset\mathscr{E}_B^*$. We set $\pi_\alpha=\tilde{\pi}_\alpha\pi$ and $\rho_\beta=\tilde{\rho}_\beta\rho$, so that $\pi_\alpha\to\pi$ and $\rho_\beta\to\rho$ where, for some constant k and all α and β, $\|\pi_\alpha\|\leqq k$ and $\|\rho_\beta\|\leqq k$.

DEFINITION 4. Let $P=I-\pi$, $R=I-\rho$, $P_\alpha=P+\pi_\alpha$, $R_\beta=R+\rho_\beta$ and

(i) $\mathscr{D}_0(\mathbf{A})=\{S\in\mathscr{L}_u(X)|P_\alpha SR_\beta$ for some α and some $\beta\}$;

(ii) $\mathscr{K}=\overline{\mathscr{D}_0(\mathbf{A})}$, closure in $\mathscr{L}_u(X)$.

Clearly $\mathbf{A}S=\mathscr{C}\ell[AS+SB^*]$ is a well-defined linear operator on $\mathscr{L}(X)$ with domain $\mathscr{D}_0(\mathbf{A})$ since $AS=AP_\alpha S$ and $\mathscr{C}\ell[SB^*]=[BR_\beta^*S^*]^*$ are bounded linear linear operators on X.

One also sees that the members of $\mathscr{K}$ are the operators which are of the form $C+L$ where C is compact and $L=(I-\pi)L(I-\rho)$ is bounded. If X is a π-space we may simply take $X=\bar{\mathscr{E}}_A=\bar{\mathscr{E}}_B$ and $\mathscr{K}$ the collection of all compact linear operators, although this is rather restrictive. Instead one might check the initial value S_0 and see if $AS_0+S_0B^*$ is a densely defined bounded linear operator. If so one might try to find a collection $\mathscr{K}$ as defined above which contains S_0. As we will see, what this implies is that there exists a uniform solution locally (i.e. a $C^1([0,t_0),\mathscr{L}_u(X))$ solution for some $t_0>0$).

It should be observed that A defined on $\mathscr{D}_0(\mathbf{A})$ is a closable operator in $\mathscr{K}$ whose closure is given by $S\to\mathscr{C}\ell[AS+SB^*]$. To see this suppose $\{S_n\}$ is a sequence in $\mathscr{D}_0(\mathbf{A})$ with $\lim_{n\to\infty}\|S_n\|=0$ and with $\mathbf{A}(S_n)$ tending to $T\in\mathscr{K}$. We must show $T=0$. Suppose x is an arbitrary element in $\mathscr{D}(B^*)$, then $\lim_{n\to\infty}\mathscr{C}\ell[S_nB^*]x=\lim_{n\to\infty}S_n(B^*x)=0$. Hence $\lim_{n\to\infty}AS_nx=Tx$. But S_nx tends to zero and A is closed. Therefore $Tx=0$. Since $\mathscr{D}(B^*)$ is dense, we have $T=0$. In the future we will let $\mathbf{A}$ denote the closure in $\mathscr{K}$ of the previously defined operator $\mathbf{A}$, and $\mathscr{D}(\mathbf{A})$ will denote its domain. Suppose $S_n\in\mathscr{D}_0(\mathbf{A})$ and $S_n\to S\in\mathscr{D}(\mathbf{A})$ and $\mathbf{A}S_n\to\mathbf{A}S$ in $\mathscr{L}(X)$. If $x\in\mathscr{D}(B^*)$ then $\mathbf{A}S_nx$ converges to $(\mathbf{A}S)x-SB^*x$ while $S_nx\to Sx$. Since A is a closed operator $Sx\in\mathscr{D}(A)$ and so $(\mathbf{A}S)x=ASx+SB^*x$ for all x in the dense subspace $\mathscr{D}(B^*)$. This proves $\mathbf{A}S=\mathscr{C}\ell[AS+SB^*]$.

We let $\mathbf{I}$ denote the identity map on $\mathscr{L}(X)$ and let

$$\mathbf{A}_\lambda=\mathbf{A}-(\lambda+2\omega+1)\mathbf{I}. \tag{2.1}$$

LEMMA 5. *For any* $S\in\mathscr{K}$ *we have*

(i) $\|S-\lambda\mathbf{A}S\|\geqq(1-2\lambda\omega)\|S\|$, *for all* $\lambda\in(0,1/2\omega)$;

(ii) $\|\mathbf{A}_\lambda S\|\geqq(1+|\lambda|)\|S\|/C$, *for some constant* C *and all* λ *with* $\text{Re}\lambda\geqq0$.

Proof. We only need to prove these inequalities for a dense collection of operators in $\mathscr{K}$. We will only consider the norm attaining members of $\mathscr{D}_0(\mathbf{A})$. These are the operators S such that there exists an x, $\|x\|=1$, satisfying $\|Sx\|=\|S\|$. Let $S\in\mathscr{K}$, then we know we can find an operator S' such that $\pi S'\rho=\pi_\alpha S'\rho_\beta$ for some α and some β and such that S' is arbitrarily close to S in norm. A theorem of Lindenstrauss [13] tells us that the operators L in $\mathscr{L}(Y,Z)$ whose second dual, L^{**}, are norm attaining are dense in $\mathscr{L}(Y,Z)$. The proof of this theorem actually gives us a little bit more. It shows that given $T\in\mathscr{L}(Y,Z)$ and $\varepsilon>0$, there exists a compact linear C_ε such that $\|C_\varepsilon\|<\varepsilon$, $\text{Range}\,C_\varepsilon\subset\text{Range}\,T$, $\text{Ker}\,C_\varepsilon\supset\text{Ker}\,T$ and such that $(T+C_\varepsilon)^{**}$ is norm-attaining. This means that we can find a norm-attaining operator S'', arbitrarily close in norm to S', and satisfying $\pi S''\rho=\pi_\alpha S''\rho_\beta$. Therefore let us assume that in fact $S=P_\alpha SR_\beta\in\mathscr{D}(\mathbf{A})$ and attains its norm at the unit vector x_0. Let $Sx_0=y_0$. We claim $S^*Jy_0\subset\|S\|^2Jx_0$. To see this, suppose $y_0^*\in Jy_0$; then $\|S\|^{-2}S^*y_0^*\leqq1$, and

$$\left\langle\|S\|^{-2}S^*y_0^*,x_0\right\rangle\in\|S\|^{-2}\langle JSx_0,Sx_0\rangle=\{1\}.$$

Therefore we have, letting $z_0^* = \|S\|^{-1} JSx_0$:

$$\begin{aligned}\|S - \lambda \mathbf{A} S\| &\geqq \left| \left\langle z_0^*, Sx_0 - \lambda ASx_0 - \lambda SB^* x_0 \right\rangle \right| \\ &= \left| \left\langle \|S\|^{-2} JSx_0, Sx_0 \right\rangle \right| - \lambda \left\langle \|S\|^{-2} JSx_0, ASx_0 \right\rangle \\ &\quad - \lambda \left\langle \|S\|^{-2} S^* JSx_0, B^* x_0 \right\rangle \Big| \|S\| \\ &\subset \left\{ |1 - \lambda\sigma_1 - \lambda\sigma_2| \|S\| : \sigma_1, \sigma_2 \in \Sigma_{\omega,\delta} \right\}.\end{aligned}$$

Clearly $|1 - \lambda\sigma_1 - \lambda\sigma_2| \geqq 1 - 2\lambda\omega$. The proof of (ii) proceeds similarly, showing that $\|\mathbf{A}_\lambda S\| \geqq |\lambda + 2\omega + 1 - \sigma_1 - \sigma_2| \|S\|$ where σ_1, $\sigma_2 \in \Sigma_{\omega,\delta}$. But

$$|\lambda + 2\omega + 1 - \sigma_1 - \sigma_2| \geqq \operatorname{dist}(\lambda, \Sigma_{2\omega,\delta} - 2\omega - 1) \geqq (1 + |\lambda|)\sin\delta.$$

This means (ii) is satisfied with $C = (\sin\delta)^{-1}$.

This lemma shows that $2\omega\mathbf{I} - \mathbf{A}$ is an accretive operator on $\mathscr{K}$.

LEMMA 6. *Whenever* $\operatorname{Re}\lambda > 0$, $\mathbf{A}_\lambda^{-1}$ *exists as a bounded linear operator in* $\mathscr{K}$. *Moreover* $\|\mathbf{A}_\lambda^{-1}\| < C(1 + |\lambda|)$ *for some constant C.*

Proof. Consider the operator $K \to AK - \omega K$ with domain $\mathscr{D}_0$ defined earlier. Denote the closure of this operator by $\mathbf{A}_A$. Its domain is dense in $\mathscr{K}$ and $\lambda\mathbf{I} - \mathbf{A}_A$ is invertible on $\mathscr{K}$ whenever $\lambda > 0$. Indeed, using the inequality

$$|\langle \{(\lambda + \omega)I - A\}x, x^* \rangle^*| \geqq \lambda \langle x, x^* \rangle$$

which holds for all $\lambda > 0$ and $x^* \in Jx$, we see that for $L \in \mathscr{K}$ and $\lambda > 0$ we have

$$\left\| (\lambda\mathbf{I} - \mathbf{A}_A)^{-1}(L) \right\| = \left\| [(\lambda + \omega)I - A]^{-1} L \right\| \leqq (1/\lambda)\|L\|.$$

Applying the Hille–Yosida theorem (see next section) we conclude that $\mathbf{A}_A$ generates a strongly continuous (with respect to the uniform operator topology) semigroup of contraction operators on $\mathscr{K}$. We denote this semigroup by $\mathbf{T}_A(s)$. Of course $\mathbf{A} - \omega\mathbf{I}$ generates a semigroup $T_A(s)$ of contraction operators on X. Consider $K \in \mathscr{K}$, $x \in X$ and

$$\psi(s) = [\mathbf{T}_A(s)(K)]x - T_A(s)(Kx).$$

Clearly $\psi(0) = 0$ and $\psi'(s) = (A - \omega I)\psi(s)$ for $s > 0$. By uniqueness of the solution to this initial value problem it follows that $\mathbf{T}_A(s)(K) = T_A(s)K$. Similarly we can extend the operator $K \to \mathscr{Cl}(KB^*)$-ωK to a closed, densely defined operator $\mathbf{A}_{B^*}$ on $\mathscr{K}$ which generates a strongly continuous semigroup $\mathbf{T}_{B^*}(s)$ on $\mathscr{K}$. Let $x^* \in X^*$ and let $T_B(s)$ be the strongly continuous semigroup on X^* generated by $B - \omega I$. Consider

$$\phi(s) = [\mathbf{T}_{B^*}(s)(K)]^* x^* - T_B(s) K^* x^*.$$

It satisfies $\phi(0) = 0$. Using the fact that taking adjoints is continuous with respect to the uniform topology, we see that $d\phi/ds = (B - \omega I)\phi(s)$ and hence, again by uniqueness,

$$\mathbf{T}_{B^*}(s)(K) = KT_B(s)^*.$$

Hence defining $\mathbf{T}(s)(K) = T_A(s)KT_B(s)$ we see that $\mathbf{T}(= \mathbf{T}_A \circ \mathbf{T}_B = \mathbf{T}_B \circ \mathbf{T}_A)$ is a strongly continuous semigroup of contraction operators on $\mathscr{K}$. Taking the derivative of $\mathbf{T}(s)$ in the weak operator topology we see that its infinitesimal generator must be $\mathbf{A} - 2\omega\mathbf{I}$.

Applying the Hille–Yosida theorem again, but this time in the other direction, we see that $\mathbf{A}_\lambda^{-1}$ exists for all $\lambda > 0$. The rest follows from the previous lemma.

3. Some results from the theory of semigroups. In this section we will state several results whose proofs can be found in [6], [7] and [20]. Included are several estimates which we will need. These will be stated more carefully than they usually are. More specifically, the constants which appear in these estimates, and which are usually treated generically, can be though of as functions of ω and δ where ω and δ are chosen such that the numerical range of the generator is contained in $\Sigma_{\omega,\delta}$.

Let A be a densely defined closed linear operator on a Banach space Z and let $\mathscr{D}(A)$ denote its domain. In the previous section we indicated the use of the Hille–Yosida theorem. This theorem says that A is the infinitesimal generator of a strongly continuous semigroup of contraction operators if $\lambda I - A$ is invertible and $\|(\lambda I - A)\| < 1/\lambda$ for all $\lambda > 0$.

A somewhat similar result is the following. Suppose that the resolvent set of A contains the wedge

$$S_{\zeta,\eta} = \{ z \in \mathbb{C} \,|\, |\arg(z-\zeta)| \leqq \pi/2 + \eta \} \tag{3.1}$$

for some $\zeta \in \mathbb{R}$ and some $0 < \eta < \pi/2$, and that

$$\|(\lambda I - A)^{-1}\| \leqq \frac{M}{|\lambda - \zeta|}, \qquad \lambda \in S_{\zeta,\eta}, \tag{3.2}$$

where M is some constant. Then A generates an analytic semigroup, denoted e^{At}. A slightly stronger hypothesis is to assume that the spectrum and numerical range of A are contained in the wedge $\Sigma_{\omega,\delta}$ (= closure of the complement of $S_{\omega,\delta}$) for some $\omega \in \mathbb{R}$ and $0 < \delta < \pi/2$. To see this, let $0 < \eta < \delta$, $\zeta > \omega$ and x a unit vector in Z. Then for any $x^* \in Jx$, $|\langle (\lambda I - A)x, x^* \rangle| = |\lambda - \langle Ax, x^* \rangle| \geqq \operatorname{dist}(\lambda, \Theta(A)) \geqq \operatorname{dist}(\lambda, \Sigma_{\omega,\delta})$. One may easily verify that if δ, and hence also η, is taken to be less than $\pi/4$ then for $\lambda \in S_{\zeta,\eta}$, with $\phi = |\arg(\lambda - \zeta)|$ and $\psi = |\arg(\lambda - \omega)|$, we have

$$\operatorname{dist}(\lambda, \Sigma_{\omega,\delta}) = \begin{cases} |\lambda-\zeta|\cos(\phi-\delta) + (\zeta-\omega)\cos\delta & \text{if } \psi \geqq \delta, \\ |\lambda-\omega| = \left[|\lambda-\zeta|^2 + (\zeta-\omega)^2 + 2|\lambda-\zeta|(\zeta-\omega)\cos\phi \right]^{1/2} & \text{if } \psi \leqq \delta, \end{cases}$$

$$\geqq \begin{cases} |\lambda-\zeta|\sin(\delta-\eta) + (\zeta-\omega)\cos\delta & \text{if } \psi \geqq \delta, \\ |\sqrt{\cos\delta}\,\left[|\lambda-\zeta| + \zeta - \omega \right] & \text{if } \psi \leqq \delta. \end{cases}$$

Since $\sin(\delta-\eta) \leqq \cos\delta \leqq \sqrt{\cos\delta}$ we have

$$\|(\lambda - A)^{-1}\| \leqq \frac{[\sin(\delta-\eta)]^{-1}}{|\lambda-\zeta| + \zeta - \omega}, \qquad \lambda \in S_{\zeta,\eta}. \tag{3.3}$$

We note that the above calculation actually proves that the continuous and point spectra of A are contained in the closure of its numerical range. The residual spectrum of A is contained in the closure of $\Theta(A^*)$ and hence the hypothesis on A may be changed from $\sigma(A) \cup \Theta(A) \subset \Sigma_{\omega,\delta}$ to $\Theta(A) \cup \Theta(A^*) \subset \Sigma_{\omega,\delta}$. For bounded operators $\Theta(A) \subset \Theta(A^*)$, however this might not be true for certain unbounded closed operators. If it were, then at least on reflexive spaces, the hypothesis would just need to be

$\Theta(A)\subset\Sigma_{\omega,\delta}$. If $\omega<0$, then letting $M=[\sin(\delta-\eta)\min(1,|\omega|)]^{-1}$ and $\zeta=0$ we obtain

$$\|(\lambda-A)^{-1}\|\leqq\frac{M}{|\lambda|+1}\quad\text{if }\operatorname{Re}\lambda\leqq 0,$$

which is the inequality assumed in [7] and [20] in order to ensure that A in fact generates an analytic semigroup.

The crucial inequality on which most other estimates depend can be easily obtained from the identity

$$A^m e^{tA}=\frac{1}{2\pi i}\int_\Gamma \lambda^m(\lambda I-A)^{-1}e^{\lambda t}d\lambda$$

where Γ can, for example, be taken to be the boundary of $S_{\tilde\omega,\delta/2}$, $0<\eta<\delta$, $\omega<\tilde\omega$ oriented in the upward direction, i.e. $\lambda=\tilde\omega+r\exp\pm(\frac{\pi}{2}+\frac{\delta}{2})i$, $0\leqq r<\infty$. Using (3.3) we then obtain after some manipulation and letting, for example, $\tilde\omega=\omega+|\omega|/2$:

$$\|A^m e^{tA}\|\leqq C(\omega,\delta,m)e^{\tilde\omega t}t^{-m}. \tag{3.4}$$

The value of $\tilde\omega$ is important only insofar as that it has to exceed ω unless $\omega=0$.

For the rest of this section let us assume that $\omega<0$, $\tilde\omega=\omega/2$. Then we may also define fractional powers:

$$A^{-\alpha}=\frac{e^{-i\pi\alpha}}{\Gamma(\alpha)}\int_0^\infty e^{sA}s^{\alpha-1}ds\qquad(\alpha>0),$$
$$A^\alpha=(A^{-\alpha})^{-1}.$$

For any real α the domain $\mathscr{D}(A^\alpha)$ is also dense in Z and $\mathscr{D}(A^\alpha)\subset\mathscr{D}(A^\beta)$ if $\beta\leqq\alpha$. Moreover

$$A^\alpha A^\beta v=A^\beta A^\alpha v=A^{\alpha+\beta}v\quad\text{for all }v\in\mathscr{D}(A^\gamma),$$

where $\gamma=\max(\alpha,\beta,\alpha+\beta)$, and

$$\|A^\beta v\|\leqq C(\alpha,\beta,\gamma,\omega,\delta)|A^\gamma v|^{(\beta-\alpha)/(\gamma-\alpha)}\|A^\alpha v\|^{(\gamma-\beta)/(\gamma-\alpha)} \tag{3.5}$$

for all $v\in\mathscr{D}(A^\gamma)$, $\alpha<\beta<\gamma$. This inequality together with (3.4), using $\tilde\omega=\omega/2$, yields

$$\|A^\beta e^{tA}\|\leqq C(\omega,\delta,\beta)e^{-\omega t/2}t^{-\beta}. \tag{3.6}$$

Next we allow A to depend on the parameter t, requiring however that $\mathscr{D}(A(t))\cap\mathscr{D}(A(s)$ is dense for any s, $t\geqq 0$ and that there exists a $0<a<1$ and a constant C_T such that

$$\|(A(\tau)-A(\tau))A^{-1}(s)\|\leqq C_T|t-\tau|^a$$

uniformly for all $0\leqq t$, τ, $s\leqq T$. We also assume that the numerical range $\Theta(A(t))\subset\Sigma_{\omega,\delta}$ for all t so that the above inequalities are satisfied uniformly for all t. It is known that there exists a *fundamental solution* or *propagation operator* $U(t,s)\in\mathscr{L}(Z)$ for $0\leqq s\leqq t$ i.e. $(t,s)\to U(t,s)$ is strongly continuous in t and s, the derivative $\partial U(t,s)/\partial t$ exists in the strong topology and belongs to $\mathscr{L}(Z)$ and is also strongly continuous in t for $t>s\geqq 0$. Moreover the range of $U(t,s)$ contains the domain $\mathscr{D}(A(t))$ of $A(t)$ for $t>s$

and

$$\frac{\partial U(t,\tau)}{\partial t}=A(t)U(t,\tau) \qquad (t<\tau),$$
$$U(\tau,\tau)=I.$$

We can verify that in fact the domain $\mathscr{D}(A(t))$ is independent of t (see Lemma 7).

Let f: $[0,\infty)\to Z$ be uniformly Hölder continuous. Then

$$u(t)=U(t,0)u_0+\int_0^t U(t,s)f(s)\,ds \tag{3.7}$$

is the unique solution to the Cauchy problem

$$\frac{du}{dt}=A(t)u+f(t), \qquad 0<t<T,$$
$$u(0)=u_0.$$

Actually (and we shall make use of this fact) we only need f to be uniformly Hölder continuous on compact subintervals of $(0,\infty)$. This can be seen by splitting the integral in (3.7) into two parts (one from 0 to $t/2$, the other from $t/2$ to t) before proceeding with the usual differentiability proof such as given, e.g., in [7, p. 129]. The following a priori estimates can be obtained using the fact that

$$U(t,\tau)=\exp(t-\tau)A(\tau)+\int_\tau^t \exp(t-s)A(s)\Phi(s,\tau)\,ds,$$

where Φ solves the Volterra equation

$$\Phi(t,\tau)=\phi_1(t,\tau)+\int_\tau^t \Phi(t,s)\phi_1(s,\tau)\,ds$$

where

$$\phi_1(t,\tau)=[A(\tau)-A(t)]\exp(t-\tau)A(\tau).$$

This means that again, as before, the various constants can be thought of as depending upon the operators $A(t)$ only via the parameters ω and δ. In other words the estimates are uniformly valid for all $A(t)$ provided their numerical ranges $\Theta(A(t))\subset\Sigma_{\omega,\delta}$. It should also be noted that even though the constants C appearing in the estimates below will also depend on a, T and C_T, we will not explicitly indicate this.

$$\|A^\gamma(s)[U(t,\tau)-\exp(\tau-t)A(t)]A^{-\beta}(\tau)\|\leqq C(\omega,\delta,\gamma,\beta)|t-\tau|^{\alpha+\beta-\gamma} \tag{3.8}$$
$$\text{for } 0\leqq s,\ 0<\tau\leqq t\leqq T,\ 0\leqq\beta\leqq 1,\ 0\leqq\gamma\leqq 1.$$

$$\|A^\gamma(t)U(t,\tau)A^{-\beta}(\tau)\|\leqq C(\omega,\delta,\gamma,\beta)|t-\tau|^{\beta-\gamma} \tag{3.9}$$
$$\text{for } 0\leqq\tau<t\leqq T \text{ and } 0\leqq\beta\leqq\gamma<1+\alpha.$$

$$\|A^\gamma(t)A^{-\beta}(\tau)\|\leqq C(\omega,\delta,\gamma,\beta)\|A(t)A(\tau)^{-1}\|^\gamma \qquad \text{for } 0\leqq\gamma<\beta\leqq 1. \tag{3.10}$$

$$\|A^\gamma(s)A(t)[U(t,\tau)-\exp(t-\tau)A(t)]A(t)A^{-\beta}(\tau)\| \tag{3.11}$$
$$\leqq C(\omega,\delta,\gamma,\beta,\varepsilon)|t-\tau|^{\beta-\gamma-1+\alpha-\varepsilon} \quad \text{for any } \varepsilon>0 \text{ and } 0\leqq\beta\leqq 1,\ 0\leqq\gamma<\alpha.$$

If $\beta>0$ or if $\beta=0$ and $s=t$ then $\varepsilon=0$ is allowed.

$$(3.12)\quad \left\| A^{\gamma}(s)\left[\int_{\tau}^{t+h} U(t+h,s)f(s)\,ds-\int_{\tau}^{t} U(t,s)f(s)\,ds\right]\right\| \leqq C(\omega,\delta,\gamma)h^{1-\gamma}(|\log h|+1)\max_{t\leqq r\leqq t+h}\|f(r)\|$$

for $0\leqq\tau\leqq t\leqq t+h\leqq T,\ 0\leqq\gamma<1.$

Finally we consider the nonlinear Cauchy problem

$$(3.13)\quad \begin{aligned}&\frac{du}{dt}-A(t,u)u=f(t,u)\qquad(0<t<t_0),\\ &u(0)=u_0.\end{aligned}$$

It has a unique solution on $[0,t^*]$ for some $0<t^*\leqq t_0$, which is continuously differentiable (in the norm topology) on $(0,t^*)$ and is continuous on $[0,t^*]$ provided the following hypotheses are satisfied.

(I). $A(0,u_0)$ is a closed operator with dense domain D_0,

$$(3.14)\quad \|[\lambda I-A(0,u_0)]^{-1}\|\leqq\frac{C}{1+|\lambda|}\qquad(\mathrm{Re}\,\gamma\leqq 0).$$

(II). For all $v\in\mathscr{B}_R(0)=\{w\,|\,\|w\|<R\}$, $A(t,v)$ is well defined on D_0 for all $0\leqq t\leqq t_0$. Furthermore for any $\tau\in[0,t_0]$ and w, $v\in\mathscr{B}_R(0)$

$$(3.15)\quad \|[A(t,v)-A(\tau,w)]A^{-1}(\tau,w)\|\leqq C(R)\big(|t-\tau|^{\sigma}+\|v-w\|\big)\quad\text{where } 0<\sigma<1.$$

(III). For all t, $\tau\in[0,t_0]$ and w, $v\in\mathscr{B}_R(0)$

$$(3.16)\quad \|f(t,v)-f(\tau,w)\|\leqq C(R)\big(|t-\tau|^{\sigma}+\|v-w\|\big).$$

(IV). $u_0\in\mathscr{B}_R(0)\cap D(A^{\beta}(0,u_0))$ for some $\beta>0$.

Actually the hypotheses as stated above are more restrictive than they need to be (see e.g.. [7] or [20]).

We conclude this section with three lemmas which will give us some qualitative results for the nonlinear problem which we will need in order to prove existence of strongly differentiable solutions for (1.3).

Let Y be a Banach space and let $\mathscr{B}_R$ denote the open ball of radius R in Y. We have

LEMMA 7. *Suppose that, for each* $t\in[0,T)$ *and* $u\in Y$, $A(t,u)$ *is a closed linear operator with dense domain and with numerical range contained in* $\Sigma_{\omega,\delta}$, $\omega<0$. *Let* $A_0=A(0,0)$ *and suppose that for each* $R>0$ *there exists a constant* $C(R)$ *such that* $\mathscr{D}(A(t,u))\cap\mathscr{D}(A(s,v))$ *is dense in* Z *for all* (t,u) *and* (s,v) *in* $Q_R\equiv[0,T-R^{-1})\times\mathscr{B}_R$ *and*

$$(3.17)\quad \|[A(t,u)-A(s,v)]A_0^{-1}\|\leqq C(R)\big(|t-s|^{\sigma}+\|u-v\|\big).$$

Then $\mathscr{D}(A(t,u))$ *is independent of t and u. Suppose in addition that there exist numbers* $k>0$ *and* $0<\gamma<1$ *such that for each t the map* $\Phi(t,u)$ *defined by*

$$\Phi(t,u)=A_0^{-\gamma}A(t,u)A_0^{-\gamma}$$

is continuous. Let $\{u_{\gamma}\}$ *be bounded net of Hölder continuous maps from* $[0,T)$ *into Y such that* $\|u_{\gamma}(t)-u(t)\|\to 0$ *uniformly on compact intervals in* $(0,T)$. *Denote by* $U(t,s;u_{\gamma})$ *the* *fundamental solution for the operator* $A(t,u_\gamma(t))$. *Then*

$$\left\|U(t,s;u_\gamma)-U(t,s;u)\right\|\to 0$$

for any $0\leqq s\leqq t<T$.

Proof. First we endow $[0,T)\times \mathrm{Y}$ with the topology induced by the metric

$$\rho((t,u),(s,v))=|t-s|+\|u-v\|.$$

Let $\mathscr{D}=\mathscr{D}(A_0)$ and let $\mathscr{S}$ be all $(t,u)\in[0,t_0)\times Y$, for which $\mathscr{D}(A(t,u))=\mathscr{D}$ and let $\mathscr{S}_0$ be the component of $\mathscr{S}$ which contains $(0,0)$. It will suffice to show that $\mathscr{S}_0$ is both open and closed in $[0,T)\times Y$. Let $(t',u')\in\mathscr{S}_0\cap Q_R$ and choose s and v such that $(s,v)\in Q_R$ and $C(R)(|t'-s|+\|u'-v\|)<1$. If $x\in\mathscr{D}(A(t',u'))\cap\mathscr{D}(A(s,v))=\mathscr{D}_0$ then

$$\begin{aligned}\|(A(s,v)-A(t',u'))x\|&\leqq\left\|(A(s,v)-A(t',u'))A(t',u')A(t',u')^{-1}A(t',u')x\right\|\\&\leqq b\|A(t',u')x\| \qquad (b\leqq 1).\end{aligned}$$

Hence $A(s,v)-A(t',u')$, restricted to $\mathscr{D}_0$, is an $A(t',u')$-bounded operator (see e.g. [10, p. 190]) and hence $A(t',u')$ and $A(s,v)$ have the same domain. This proves that $\mathscr{S}_0$ is open, but the same technique shows that the complement of $\mathscr{S}_0$ must also be open, and hence $\mathscr{S}_0$ is also closed. Before we prove the second part of the lemma, we note that since u_γ is uniformly Hölder continuous on compact subintervals of $[0,T)$, $A(t,u_\gamma(t))$ indeed generates a fundamental solution. We have

$$\frac{\partial}{\partial t}U(t,s;u_\gamma)=A(t,u_\gamma(t))U(t,s;u_\gamma). \tag{3.18}$$

Hence letting $\Delta_{\gamma\delta}(t,s)=U(t,s;u_\gamma)-U(t,s;u_\delta)$ we have $\Delta_{\gamma\delta}(0,0)=0$ and for $T>t>s\geqq 0$:

$$\frac{\partial}{\partial t}\Delta_{\gamma\delta}(t,s)=A(t,u_\gamma(t))\Delta_{\gamma\delta}(t,s)+\left(A(t,u_\gamma(t))-A(t,u_\delta(t))\right)U(t,s;u_\delta)$$

which means

$$\Delta_{\gamma\delta}(t,s)=\int_s^t U(t,\tau;u_\gamma)\left[A(\tau,u_\gamma(\tau))-A(\tau,u_\delta(\tau))\right]U(\tau,s;u_\delta)\,d\tau.$$

Choosing $0<\varepsilon<1-\gamma$ we have

(3.19)

$$\begin{aligned}\|\Delta_{\gamma\delta}(t,s)\|\leqq\int_s^t&\left\|U(t,\tau;u_\gamma)A^{\gamma+\varepsilon}(\tau,u_\gamma(\tau))\right\|\\&\times\left\|A^{-\gamma-\varepsilon}(\tau,u_\gamma(\tau))A_0^\gamma\right\|\left\|A_0^{-\gamma}\left[A(\tau,u_\gamma(\tau))-A(\tau,u_\delta(\tau))\right]A_0^{-\gamma}\right\|\\&\times\left\|A_0^\gamma A^{-\gamma-\varepsilon}(\tau,u_\delta(\tau))\right\|\left\|A^{\gamma+\varepsilon}(\tau,u_\delta(\tau))U(\tau,s;u_\delta)\right\|d\tau\\\leqq\int_s^t& C(\omega,\delta,\gamma,\varepsilon)(t-\tau)^{-\gamma-\varepsilon}\left\|A_0^{-\gamma}\left[A(\tau,u_\gamma(\tau))-A(\tau,u_\delta(\tau))\right]A_0^{-\gamma}\right\|(\tau-s)^{-\gamma-\varepsilon}d\tau.\end{aligned}$$

The constant $C(\omega,\delta,\gamma,\varepsilon)$ also depends on T (which causes no problem at all) and on

$$\sup_\gamma\ \sup_{0\leqq s<t\leqq T}\left\|U(t,s;u_\gamma)\right\|.$$

However, in the next lemma we shall show that that this quantity is finite (in fact $=\exp T$). We can therefore now apply the dominated convergence theorem to show that the net $\|\Delta_{\gamma\delta}(t,s)\|$ converges to zero. Here we regard $\{\gamma\delta\}$ as a directed set in the obvious manner: $\gamma\delta>\gamma'\delta'$ if $\gamma>\gamma'$ and $\delta>\delta'$.

The a priori estimates which we have for the fundamental solution $U(t,s)$ depend on the modulus of continuity to $A(t)A(0)^{-1}$. In order to prove a global existence theorem for the nonlinear problem (1.3) we will need the following estimate which depends only on ω.

LEMMA 8. *Suppose $U(t,s)$ is a fundamental solution for the closed linear operators $A(t)$ on the Banach space Z and suppose that the numerical range $\Theta(A(t))\subset\Sigma_{\omega,\delta}$ for all t. Then*

$$\|U(t,s)\|\leqq\exp\omega(t-s). \tag{3.20}$$

Proof. We will use some results and notation found in [16, pp. 31–45]. Let $\phi(z)=\frac{1}{2}\|z\|^2$ and $\psi(z)=\|z\|$. These are convex functionals and therefore possess one-sided Gateaux derivatives as well as subdifferentials. We define these as

$$\delta_{\pm}\psi(z)y=\lim_{h\to0\pm}\frac{\psi(z+hy)-\psi(z)}{h}$$

and

$$\partial\psi(z)=\{x^*\in Z^*|\psi(z+x)\geq\psi(z)+\operatorname{Re}\langle x^*,x\rangle\text{ for all }x\text{ in }Z\}.$$

First we note three facts (see [16]). First, that if $x^*\in\partial\psi(z)$ then $\|z\|x^*\in\partial\phi(z)$. This follows immediately from the definition of subdifferentials. Secondly, $\partial\phi=J$, the duality map. Thirdly that

$$\{\langle x^*,y\rangle|x^*\in\partial\psi(z)\}=[\delta_-\psi(z)y,\delta_+\psi(z)y]. \tag{3.21}$$

Such an equality does not hold for ϕ since we need the additional properties $\psi(x+y)\leqq\psi(x)+\psi(y)$ and $\psi(rx)=r\psi(x)$ for $r>0$, in order to obtain (3.21) and clearly ϕ fails to satisfy the second of these. Let us use the notation

$$x(t)=U(t,s)x(0)\quad\text{and}\quad x'(t)=A(t)U(t,s)x(0).$$

Since $x(t+h)=x(t)+hx'(t)+o(h)$ we see that $\theta(t)\equiv\|x(t)\|$ satisfies

$$\frac{d^+}{dt}\theta(t)=\delta^+\psi(x(t))x'(t)=\langle x^*,x'(t)\rangle$$

for some $x^*\in\partial\psi(x(t))\subset\|x(t)\|^{-1}J(x(t))$. This means

$$\frac{d^+}{dt}\theta(t)\in\theta(t)\Sigma_{\omega,\delta}$$

which in turn implies $\theta(t)\leqq\exp\omega(t-s)$.

LEMMA 9. *Suppose that*

$$\frac{du}{dt}=A(t,u)u+f(t,u)\qquad(0<t<T)$$

where $A(t,u)$ is of type (ω,δ) and where

$$\|f(t,u)\|\leqq a(t)+b(t)\|u\|$$

for some positive, continuous functions a and b. Then for each $0 \leqq \gamma < 1$ *there exists a continuous function* ψ_γ*, determined by* $\|u(0)\|$, $\gamma, \omega, \delta, a$ *and* b*, such that*

$$\|A^\gamma(t,u(t))u(t)\| \leqq \psi_\gamma(t) \quad \text{on } (0,T).$$

Proof. Lemma 9 and (3.9) show us that if we let $y(s) = \|u(s)\|$ then

$$y(t) \leqq e^{\omega t}\left[y(0) + \int_0^t a(s)e^{-\omega s}\,ds + \int_0^t b(s)e^{-\omega s}y(s)\,ds\right].$$

Defining the first two terms in the brackets to be $\tilde{a}(t)$ we have

$$e^{-\omega t}y(t) \leqq \tilde{a}(t) + \int_0^t b(s)e^{-\omega s}y(s)\,ds.$$

We can now apply Gronwall's inequality to find a function ψ_0 (in terms of $\tilde{a}$ and b) such that

$$y(t) \leqq \psi_0(t) \quad \text{on } [0,T). \tag{3.22}$$

Now letting $z(t) = \|A^\gamma(t,u(t))u(t)\|$ we have

$$z(t) \leqq \|A^\gamma(t,u(t))U(t,0;u)u(0)\| + \int_0^t \|A^\gamma(t,u(t))U(t,s;u)\|\|f(s,u(s))\|\,ds.$$

By using (3.7), (3.9), (3.16) and (3.22) we get

$$\begin{aligned} z(t) \leqq{}& C(\omega,\delta,\gamma,\hat{C}(\psi_0(t)))t^{-\gamma} \\ &+ \int_0^t C(\omega,\varepsilon,\gamma,\hat{C}(\psi_0(t)))(t-s)^{-\gamma}(a(s)+b(s)\psi_0(s))\,ds. \end{aligned}$$

We now define $\psi_\gamma(t)$ to be the right-hand side of this inequality:

$$\|A^\gamma(t,u(t))u(t)\| \leqq \psi_\gamma(t) \quad \text{on } (0,T). \tag{3.23}$$

4. Existence theorems. In this section we will first prove the existence of solutions in $C^1([0,t_0),\mathscr{K})$ to the problem

$$\frac{dU}{dt} = A(t,U(t))U(t) + U(t)B(t,U(t))^* + \mathscr{F}(t,U(t)) \tag{4.1}$$

for the initial conditions

$$U(0) = U_0 \in \mathscr{D}_0, \tag{4.2}$$

where $\mathscr{D}_0$ is the domain of $\mathbf{A}_0$, the closure of the operation on $\mathscr{L}_u(X)$ defined by

$$L \to \mathscr{C}\!\ell[A(0,0)L + LB(0,0)^*],$$

as described in §2. We conclude by proving the existence of strong solutions in (4.1) with initial conditions

$$U(0) = U_0 \in \mathscr{L}(X) \tag{4.3}$$

under more restrictive hypotheses on $A(t,U)$ and $B(t,U)$. We will always assume $\mathscr{F}$ to be a well-defined map from $[0,T) \times \mathscr{L}(X)$ into $\mathscr{L}(X)$. However in order to get our first existence theorem we must of course assume that $\mathscr{F}(t,K) \in \mathscr{K}$ whenever $t \in [0,T)$

and $K \in \mathscr{K}$. We list the hypotheses needed to ensure the existence of a solution in $C^1([0,t^*),\mathscr{K})$ for some $0 < t^* < T$.

(H.0) : $\mathscr{F}(t,K) \in \mathscr{K}$ for $0 \leqq t < T$ and $K \in \mathscr{K}$.

(H.1) : For each $t \in [0,T)$ and $U \in \mathscr{L}(X)$ with $\|U\| < r$, the linear operators $A(t,U)$ and $B(t,U)$ are of type (ω,δ).

(H.2) : There exist a constant $\sigma > 0$ and a positive function h such that for each t_1, $t_2 \in [0,T)$, $t_1 < t_2$, and K_1, $K_2 \in \mathscr{B}_r(0)$ we have

$$\|\mathscr{F}(t_1,K_1) - \mathscr{F}(t_2,K_2)\| < h(t_2)\left(|t_1 - t_2|^\sigma + \|K_1 - K_2\|\right).$$

(H.3) : A, B and $\mathbf{A}$ satisfy property (P_r) on $[0,T)$.

DEFINITION. A map M: $[0,T) \times \mathscr{L}(Z) \to \mathscr{L}(Z)$ will be said to satisfy property (P_r) on $[0,T)$ if there exist constants $\sigma > 0$, $\tau > 0$, and a positive function H such that for each t_1, $t_2 \in [0,T)$, $t_1 \leqq t_2$ and K_1, $K_2 \in \mathscr{L}(Y)$ with $\|K_1\| < r$, $\|K_2\| < r$ we have

$$\left\|(M(t_1,K_1) - M(t_2,K_2))(M(0,0) + \tau \cdot \mathrm{id})^{-1}\right\| \leqq H(t_2)\left(|t_1 - t_2|^\sigma + \|K_1 - K_2\|\right).$$

It can easily be seen that (H.3) is satisfied if, for example, $A(t,U) = A(0,0) + \alpha(t,U)$ and $B(t,U) = B(0,0) + \beta(t,U)$ where α and β satisfy continuity conditions of the type imposed on $\mathscr{F}$ by (H.2).

We may assume, without loss of generality, that for $t \leqq T$ the hypotheses are satisfied with $\omega \leqq -\frac{1}{2}$, because if this were not the case then the change of variables $V = e^{-(2\omega+1)t}U$ would transform the problem into one where the hypotheses are satisfied with $\omega = -\frac{1}{2}$. It should be noted however that this also produces a change in the value of r, changing it to $re^{-(2\omega+1)T}$.

Next we note that by Lemma 7 the domains of $A(t,U)$, $B(t,U)$ and $\mathbf{A}(t,U)$ are independent of t and U. We can apply Sobolevskii's existence theorem to conclude the existence of a local solution. Let $t^* \leqq T$ be the largest value so that (4.1)–(4.2) has a solution U: $[0,t^*) \to \mathscr{K}$. Let

$$\gamma_1(t) \equiv \|A(t,U(t))U(t) + U(t)B(t,U(t))^*\|.$$

Now, applying [7, Thm. 16.5, p. 175] we see that if $\gamma_1(t)$ is bounded on $[0,t^*)$ then the solution can be extended to $[0,t^*]$. If the hypotheses are satisfied for all $r > 0$ then it follows we can extend to an even integer interval, a contradiction. Actually if we can show $\gamma_\beta \equiv \|\mathbf{A}^\beta U\|$ remains bounded on $[0,T)$ then we can deduce that the solution U exists on all of $[0,T)$, provided $U_0 \in \mathscr{D}(\mathbf{A}_0^\beta)$. We have therefore proven the next theorem.

THEOREM 10. *Suppose* (H.0)–(H.3) *are satisfied. Then the domain* $\mathscr{D}$ *of* $\mathbf{A}(t,U)$ *is independent of t and U. For each* $U_0 \in \mathscr{D}$ *(or more generally* $\in \mathscr{D}(\mathbf{A}_0^\beta)$, $\beta > 0$) *with* $\|U_0\| < r$ *there exists* $0 < t_1 \leqq T$ *and a unique solution* $U \in C^1([0,t_1),\mathscr{L}_u(X))$ *to problem* (4.1)–(4.2). *If the hypotheses are satisfied for arbitrarily large r, then there exists* $0 < t^* \leqq T \leqq \infty$ *and a solution* $U \in C^1([0,t^*),\mathscr{L}_u(X)$ *such that either* $t^* = T$ *or* $\limsup \gamma_1(t) = \infty$. *If there also exist continuous functions a and b on* $[0,T)$ *such that*

$$\|\mathscr{F}(t,U)\| \leqq a(t) + b(t)\|U\|,$$

then $t^* = T$.

We note that since the fundamental solution $\mathbf{U}(t,s;U)$ associated with the map $\mathbf{A}(t,U)$ is of the form

$$\mathbf{U}(t,s;U)K = U_A(t,s;U)KU_B(t,s;U)^*$$

the solution U, whose existence was proven in the above lemma, must satisfy the integral equation

$$\begin{aligned} U(t) = {} & U_A(t,0;U)U(0)U_B(t,0;U)^* \\ & + \int_0^t U_A(t,s;U)\mathcal{F}(s,U(s))U_B(t,s;U)^*\,ds. \end{aligned}$$

We can use this to obtain estimates on t^*. Consider, for example the case where

$$\mathcal{F}(t,U) = U(t)C(t)U(t) + D(t),$$

where $C, D\colon [0,\infty) \to \mathcal{L}(X)$ are sufficiently well behaved maps. Then

$$\|U(t)\| \leqq e^{2\omega t}\|U(0)\| + \int_0^t e^{2\omega(t-s)}\left[c(s)\|U(s)\|^2 + d(s)\right] ds$$

where $\|C(t)\| < c(t)$ and $\|D(t)\| \leqq d(t)$, $c \in C^2[0,\infty)$, $d \in C[0,\infty)$. Then $\|U(t)\| \leqq y(t)$ where $y(t)$ is the solution of the corresponding equality, i.e.

$$y'(t) = 2\omega y(t) + c(t)y(t)^2 + d(t).$$

One now easily shows that, letting $\|U(0)\| = y_0$, we have

$$\|U(t)\| \leqq c(t)^{-1}[z(t) - g(t)]$$

where

$$\begin{aligned} & z' = z^2 + h(t), \qquad z(0) = c(0)y_0 + g(0), \\ & g(t) = \omega + \frac{1}{2}(\ln c(t))'; \\ & h(t) = c(t)d(t) - \left[\omega + \frac{1}{2}(\ln c(t))'\right]^2 + \frac{1}{2}(\ln c(t))'', \end{aligned}$$

and hence t^* is the largest value such that $z(t)$ is defined on $[0,t^*)$. In particular if we assume c and d are constant then $h \equiv cd - \omega^2$ and we can find t^* explicitly. Let $\mu = |cd - \omega^2|^{-1/2}$; then

$$t^* = \begin{cases} \frac{1}{2}\mu\left[\pi - 2\tan^{-1}\mu(cy_0 + \omega)\right] & \text{if } cd > \omega^2, \\ [cy_0 + \omega]^{-1} & \text{if } cd = \omega^2, y_0 > -\dfrac{\omega}{c}, \\ \frac{1}{2}\mu\ln|(1 + \mu cy_0 + \mu\omega)/(1 - \mu cy_0 - \mu\omega)| & \text{if } cd < \omega^2, y_0 > \dfrac{-\omega + 1/\mu}{c}, \\ \infty & \text{if } cd \leqq \omega^2, y_0 \leqq \dfrac{-\omega + 1/\mu}{c}, \end{cases}$$

where we interpret $1/\mu = 0$ when $cd = \omega^2$.

In order to obtain existence of strong solution in $\mathcal{L}(X)$ for equation (4.1) with the general initial condition (4.3) we also assume

(H.4) : $A(0, U_0) + (\omega + \sigma)I$ and $B(0, U_0) + (\omega + \sigma)I$ have compact inverses for $\sigma > 0$.

(H.5) : There exists a $0 < \gamma < 1$ such that $A(0,0)^{-\gamma}A(t,U)A(0,0)^{-\gamma}$ is continuous from $[0,T) \times \mathcal{L}_u(X)$ into $\mathcal{L}_u(X)$.

THEOREM 11. *Suppose* (H.1)–(H.5) *are satisfied with r arbitrarily large. Then there exists a* $t_1 \in (0,T]$ *such that* (4.1), *with initial condition* (4.3), *has a strongly differentiable solution U on* $[0,t_1)$. *U is uniformly Hölder continuous (with arbitrary exponent in* (0,1)) *on any compact subinterval of* $(0,t_1)$ *and* t_1 *may be chosen such that* $t_1 = T$ *or*

$$\limsup_{t \uparrow t_1} \|\mathscr{F}(t,U(t))\| = \infty. \tag{4.4}$$

Proof. Let us assume for the moment that there exists a positive number R such that

$$\|\mathscr{F}(t,U)\| \leqq R \tag{4.5}$$

for all $t \in [0,T)$ and all $U \in \mathscr{L}(X)$. Let us also consider the problems

$$\frac{dU}{dt} = A(t,U(t))U(t) + U(t)B^*(t,U(t)) + P_\alpha \mathscr{F}(t,U(t))R_\beta \tag{4.6}$$

with initial condition

$$U(0) = P_\alpha U_0 R_\beta. \tag{4.7}$$

We know that there exist uniform solutions on all of $[0,T)$ (see (3.13), (3.14)). We denote these solutions by $U_{\alpha\beta}$. We can consider the indices $\alpha\beta$ as forming a directed set in the obvious manner and hence we have a net $U_{\alpha\beta}$ which we wish to show has a subnet which converges in the compact-open topology of $C((0,T), \mathscr{L}(X))$. By the Arzela–Ascoli theorem this is true provided the $U_{\alpha\beta}$'s are equicontinuous on compact intervals and the operators $\{U_{\alpha\beta}(t)\}$ form a precompact set in $\mathscr{L}_u(X)$ for each $t \in (0,T)$. First we note that

$$\begin{aligned} U_{\alpha\beta}(t) = {} & U_A(t,0;U_{\alpha\beta})P_\alpha U_0 R_\beta U_B(t,0;U_{\alpha\beta})^* \\ & + \int_0^t U_A(t,s;U_{\alpha\beta})P_\alpha \mathscr{F}(s,U_{\alpha\beta}(s))R_\beta U_B(t,s;U_{\alpha\beta})^*\,ds \end{aligned} \tag{4.8}$$

where

$$\begin{aligned} & \|U_A(t+h,s;U_{\alpha\beta}) - U_A(t,s;U_{\alpha\beta})\| \\ & = \left\| \int_t^{t+h} A(\tau,U_{\alpha\beta}(\tau))U_A(\tau,s;U_{\alpha\beta})\,d\tau \right\| \leqq \frac{Ch}{|s-t|} \end{aligned} \tag{4.9}$$

and a similar inequality for U_B^*. Therefore the first term on the right-hand side of (4.8) satisfies the necessary equicontinuity condition. Applying [7, Lemma 14.4, p. 163] we see that

$$\begin{aligned} & \left\| \int_0^{t+h} U_A(t+h,s;U_{\alpha\beta})P_\alpha \mathscr{F}(s,U_{\alpha\beta}(s))R_\beta U_B^*(t+h,s;U_{\alpha\beta})\,ds \right. \\ & \left. - \int_0^t U_A(t,s;U_{\alpha\beta})P_\alpha \mathscr{F}(s,U_{\alpha\beta}(s))R_\beta U_B^*(t,s;U_{\alpha\beta})\,ds \right\| \leqq CRh(|\log h| + 1), \end{aligned} \tag{4.10}$$

thus proving that the $U_{\alpha\beta}$'s are equicontinuous on compact subintervals of $(0,T)$. This, by the way, also shows that the $U_{\alpha\beta}$'s are uniformly bounded on compact intervals and hence that we may apply Lemma 7. In order to prove the compactness requirement, we write:

$$
\begin{aligned}
&A^{\rho}(0,U_0)U_{\alpha\beta}(t)B^{*\rho}(0,U_0)\\
&=\left[A^{\rho}(0,U_0)A^{-2\rho}(t,U_{\alpha\beta}(t))\right]\left[A^{2\rho}(t,U_{\alpha\beta}(t))U_A(t,0;U_{\alpha\beta})\right]P_\alpha U_0R_\beta\\
&\qquad\times\left[B^{2\rho}(t,U_{\alpha\beta}(t))U_B(t,0;U_{\alpha\beta})\right]^*\left[B^{\rho}(0,U_0)B^{-2\rho}(t,U_{\alpha\beta}(t))\right]^*\\
&\quad+\int_0^t\left[A^{\rho}(0,U_0)A^{-2\rho}(t,U_{\alpha\beta}(t))\right]\left[A^{2\rho}(t,U_{\alpha\beta}(t))U_A(t,s;U_{\alpha\beta})\right]P_\alpha\mathscr{F}(s,U_{\alpha\beta}(s))R_\beta\\
&\qquad\qquad\times\left[B^{2\rho}(t,U_{\alpha\beta}(t))U_B(t,0;U_{\alpha\beta})\right]^*\left[B^{\rho}(0,U_0)B^{-2\rho}(t,U_{\alpha\beta}(t))\right]^*ds.
\end{aligned}
$$

Using inequalities (3.9) and (3.10) we see that the first term is uniformly bounded for each fixed $t>0$ and all α and β and that the second term can similarly be bounded by

$$k\int_0^t(t-s)^{-4\rho}ds<\infty$$

provided $\rho<1/4$. Therefore, for each $t>0$ there exists a number r such that

$$U_{\alpha\beta}(t)\in A^{-\rho}(0,U_0)\mathscr{B}_r(0)B^{*-\rho}(0,U_0)\quad\text{for all }\alpha\text{ and }\beta.$$

However this set is compact in $\mathscr{L}_u(X)$. By the Arzela–Ascoli theorem the $\{U_{\alpha\beta}\}$ are compact with respect to the compact-open topology of $C((0,T),\mathscr{L}_u(X))$ and hence there exists a subnet $U_{\alpha'\beta'}$ which tends to an element $U\in C((0,T),\ \mathscr{L}_u(X))$ in that topology. Moreover we note that by (4.9) and (4.10)

$$\|U(t+h)-U(t)\|\leqq CRh(|\log h|+1)+Ch/t$$

for all $t>0$ and $h>0$. By Lemma 7 we note that $U_A(t,s;\,U_{\alpha\beta})\to U_A(t,s;\,U)$ and similarly for $U_B^*(t,s;\,U_{\alpha\beta})$. We may therefore apply the Lebesgue dominated convergence theorem for Banach space valued integrals and obtain

$$U(t)=U_A(t,0;U)U_0U_B^*(t,0;U)+\int_0^t U_A(t,s;U)\mathscr{F}(s,U(s))U_B^*(t,s;U)\,ds. \tag{4.11}$$

We denote $P_\alpha\mathscr{F}(t,U)R_\beta$ by $\mathscr{F}_{\alpha\beta}(t,U)$ and $P_\alpha U_0R_\beta$ by $U_{0\alpha\beta}$. The integrals

$$
\begin{aligned}
\int_0^t&\left\{A(t,U_{\alpha\beta})\exp[A(t,U_{\alpha\beta})(t-s)]\mathscr{F}_{\alpha\beta}(t,U_{\alpha\beta})\exp[B^*(t,U_{\alpha\beta})(t-s)]\right.\\
&\qquad\left.+\exp[A(t,U_{\alpha\beta})(t-s)]\mathscr{F}_{\alpha\beta}(t,U_{\alpha\beta})\mathscr{C}\ell\left[\exp[B^*(t,U_{\alpha\beta})(t-s)]B^*(t,U_{\alpha\beta})\right]\right\}ds
\end{aligned}
$$

exist as improper Riemann integrals and as Lebesgue integrals in the strong topology and are equal to

$$\mathscr{I}_{\alpha\beta}(t)=-\mathscr{F}_{\alpha\beta}(t,U_{\alpha\beta})+\exp[A(t,U_{\alpha\beta})t]\mathscr{F}_{\alpha\beta}(t,U_{\alpha\beta})\exp[B^*(t,U_{\alpha\beta})t].$$

We define

$$\begin{aligned}
J_{\alpha\beta}(t)= & \int_0^t A(t,U_{\alpha\beta})\left\{U_A(t,s;U_{\alpha\beta})-\exp\left[A(t,U_{\alpha\beta})(t-s)\right]\right\} \\
& \qquad \times\mathscr{F}_{\alpha\beta}(s,U_{\alpha\beta})U_B^*(t,s;U_{\alpha\beta})\,ds \\
& + \int_0^t A(t,U_{\alpha\beta})\exp\left[A(t,U_{\alpha\beta})(t-s)\right]\mathscr{F}_{\alpha\beta}(s,U_{\alpha\beta}) \\
& \qquad \times\left\{U_B^*(t,s;U_{\alpha\beta})-\exp\left[B^*(t,U_{\alpha\beta})(t-s)\right]\right\}ds \\
& + \int_0^t A(t,U_{\alpha\beta})\exp\left[A(t,U_{\alpha\beta})(t-s)\right] \\
& \qquad \times\left\{\mathscr{F}_{\alpha\beta}(s,U_{\alpha\beta})-\mathscr{F}_{\alpha\beta}(t,U_{\alpha\beta})\right\}\exp\left[B^*(t,U_{\alpha\beta})(t-s)\right]ds \\
& + \int_0^t U_A(t,s;U_{\alpha\beta})\mathscr{F}_{\alpha\beta}(s,U_{\alpha\beta}) \\
& \qquad \times\mathscr{C}\ell\left[\left\{U_B^*(t,s;U_{\alpha\beta})-\exp\left[B^*(t,U_{\alpha\beta})(t-s)\right]B^*(t,U_{\alpha\beta})\right\}\right]ds \\
& + \int_0^t \left\{U_A(t,s;U_{\alpha\beta})-\exp\left[A(t,U_{\alpha\beta})(t-s)\right]\right\}\mathscr{F}_{\alpha\beta}(s,U_{\alpha\beta}) \\
& \qquad \times\mathscr{C}\ell\left\{\exp\left[B^*(t,U_{\alpha\beta})(t-s)\right]B^*(t,U_{\alpha\beta})\right\}ds \\
& + \int_0^t \exp\left[A(t,U_{\alpha\beta})(t-s)\right]\left\{\mathscr{F}_{\alpha\beta}(s,U_{\alpha\beta})-\mathscr{F}_{\alpha\beta}(t,U_{\alpha\beta})\right\} \\
& \qquad \times\mathscr{C}\ell\left\{\exp\left[B^*(t,U_{\alpha\beta})(t-s)\right]B^*(t,U_{\alpha\beta})\right\}ds.
\end{aligned}$$

All the terms occurring in the definition of $\mathscr{J}_{\alpha\beta}$ are well defined as can be seen by using the various estimates in §3. We next define

$$\begin{aligned}
\mathscr{H}_{\alpha\beta}(t)= & A(t,U_{\alpha\beta})U_A(t,0;U_{\alpha\beta})U_{0\alpha\beta}U_B^*(t,0;U_{\alpha\beta}) \\
& + U_A(t,0;U_{\alpha\beta})U_{0\alpha\beta}\mathscr{C}\ell\left\{U_B^*(t,0;U_{\alpha\beta})B^*(t,U_{\alpha\beta})\right\}+\mathscr{F}_{\alpha\beta}(t,U_{\alpha\beta}).
\end{aligned}$$

We now note that $\mathscr{I}_{\alpha\beta}+\mathscr{J}_{\alpha\beta}+\mathscr{H}_{\alpha\beta}$ is just another way of writing the right hand side of (4.6) so that

$$\frac{dU_{\alpha\beta}}{dt}=\mathscr{I}_{\alpha\beta}+\mathscr{J}_{\alpha\beta}+\mathscr{H}_{\alpha\beta}.$$

Let $x^*\in\mathscr{D}(A^*)$ and $y\in\mathscr{D}(B^*)$; then using inequalities (3.9), (3.8) and (3.6) we can apply Lemma 7 and the dominated convergence theorem to the expression

$$\left\langle x^*,\left(\mathscr{I}_{\alpha\beta}(t)+\mathscr{J}_{\alpha\beta}(t)+\mathscr{H}_{\alpha\beta}(t)\right)y\right\rangle.$$

Letting $\mathscr{I}$, $\mathscr{J}$ and $\mathscr{H}$ denote the expressions corresponding to $\mathscr{I}_{\alpha\beta}$, $\mathscr{J}_{\alpha\beta}$ and $\mathscr{H}_{\alpha\beta}$ respectively but with $U_{\alpha\beta}$ replaced by U, and using the fact that $\mathscr{D}(A^*)$ and $\mathscr{D}(B^*)$ are dense we have

$$\mathscr{I}(t)+\mathscr{J}(t)+\mathscr{H}(t)=\mathscr{C}\ell\{A(t,U)U(t)+U(t)B^*(t,U)\}+\mathscr{F}(t,U).$$

We can also apply Lemma 7 and the dominated convergence theorem to the expressions

$$U_{\alpha\beta}(t)x = U_{0\alpha\beta}x + \int_0^t \left[\mathscr{I}_{\alpha\beta}(s) + \mathscr{J}_{\alpha\beta}(s) + \mathscr{H}_{\alpha\beta}(s)\right] x\, ds$$

to conclude that

$$U(t)x = U_0 x + \int_0^t [\mathscr{I}(s) + \mathscr{J}(s) + \mathscr{H}(s)]\, x\, ds \tag{4.12}$$

for all $x \in x$. Hence we can differentiate U and we have

$$\begin{aligned}\frac{dUx}{dt} &= [\mathscr{I}(t) + \mathscr{J}(t) + \mathscr{H}(t)]\\ &= \mathscr{C}\ell\{A(t,U)U(t) + U(t)B^*(t,U)\}x + \mathscr{F}(t,U)x.\end{aligned}$$

This completes the proof for the case where $\mathscr{F}$ is uniformly bounded. If $\mathscr{F}$ does not satisfy (4.5) for any R then we define

$$\mathscr{F}_R(t,U) = \max\left(1, \|\mathscr{F}(t,U)\| R^{-1}\right)^{-1} \mathscr{F}(t,U).$$

Using what we have just proven we can deduce the existence of a solution U_R on an interval $[0, t_R)$ for the problem with $\mathscr{F}$ replaced by $\mathscr{F}_R$. Here t_R is the first value of t where $\|\mathscr{F}(t, U(t))\|$ attains the value R. In fact we have a net of uniform solutions which converge to a strong solution U_R on $[0, t_R)$ of the problem (4.1) with initial condition (4.3). Now we can take the corresponding net of solutions for $\mathscr{F}_{2R}$ (i.e. with the same initial conditions) and find a subnet which converges to a strong solution U_{2R} which perforce agrees with U_R on $[0, t_R)$. This implies that there exists a value $0 < t_1 \leqq \infty$ and a strong solution U on $[0, t_1)$ such that $t_1 = \infty$ or such that

$$\limsup_{t \uparrow t_1} \|\mathscr{F}(t, U(t))\| = \infty.$$

We have not been able to show uniqueness of strong solutions to generalized Riccati equations. However if A and B^* do not depend on U, i.e. ($A = A(t)$, $B^* = B^*(t)$) then the Lipschitz continuity of $\mathscr{F}(t, U)$ with respect to U easily is seen to imply uniqueness within the class of strong solutions which also satisfy the integral equation (4.11). In fact by imposing Lipschitz continuity on $A_0^{-\gamma}A(t,U)A_0^{-\gamma}$ for some $0<\gamma<1$ we can use the ideas in Lemma 7 (more specifically equation (3.19)) to extend this uniqueness result to the case where A and B do depend on U. In particular we have existence and uniqueness in case $A(t,U) = A_0(t) + \alpha(t,U)$ and $B(t,U) = B_0(t) + \beta(t,U)$ where α and β satisfy the hypothesis (H.2) imposed on $\mathscr{F}$.

One may obtain similar existence theorems when the space X is not reflexive. Suppose, for example, that X is the dual space of some other Banach space: $X = Z^*$. We may then suppose that B is densely defined in Z and that its domain is $Z_B + \mathscr{E}_B^*$ where $Z = Z_B \oplus \mathscr{E}_B^*$ and $\mathscr{E}_B^{**}$ is a π-space. This induces a similar spitting on X, $X = X_B \oplus \mathscr{E}_B$ and we can proceed as before up to Lemma 5. It is in the proof of this Lemma that only significant use was made of reflexivity in deducing that the norm attaining operators were dense in $\mathscr{L}_u(X)$. Therefore we could instead of reflexivity assume that the norm attaining operators are dense in $\mathscr{L}_u(X)$. Another approach is to consider the map $L \to \mathscr{C}\ell[A^{**}L + LB^{***}]$ provided A^{**} and B^{***} are well defined with dense domain and with numerical ranges contained in $\Sigma_{\omega,\delta}$. If this is so then, using the fact that the members of $\mathscr{L}_u(X)$ whose second conjugates are norm attaining are dense, we can show that Lemma 5 still holds.

REFERENCES

[1] F. E. BROWDER AND D. G. DEFIGUEIREDO, *J-monotone nonlinear operators in Banach spaces*, Nederl. Akad. Wet. Proc., Ser A, 69 (1966), pp. 412–420.

[2] M. G. CRANDALL AND A. PAZY, *Nonlinear evolution equations in Banach spaces*, Israel. J. Math., 11 (1972), pp. 57–94.

[3] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite-dimensional Riccati equation*, J. Math. Anal. Appl., 47 (1974), pp. 43–57.

[4] 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.

[5] ______, *Quelques resultats d'existence, unicité et régularité pour un problème de la théorie du contrôle*, J. Math. Pures. Appl., 51 (1973), pp. 353–375.

[6] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Vol.* I, Academic Press, New York, 1958.

[7] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[8] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Colloquium Publications, Vol. 31, American Mathematical Society, Providence, RI, 1957.

[9] A. IWANIK, *Norm attaining operators on Lebesgue spaces*, Pacific J. Math., 83 (1979), pp. 381–386.

[10] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[11] H. J. KUIPER AND S. M. SHEW, *Strong solutions for infinite-dimensional Riccati equations arising in transport theory*, this Journal, 11 (1980), pp. 211–222.

[12] J. LINDENSTRAUSS, *Extension of compact operators*, Mem. AMS 48, American Mathematical Society, Providence, RI, 1963.

[13] ______, *On operators which attain their norm*, Israel J. Math., 1 (1963), pp. 139–148.

[14] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1970.

[15] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, SIAM J. Control Optim., 7 (1963), pp. 101–121.

[16] R. H. MARTIN, *Nonlinear Operators and Differential Equations in Banach Spaces*, John Wiley, New York, 1976.

[17] E. MICHAEL AND A. PELCZYNSKI, *Separable Banach spaces which admit l_n^∞ approximations*, Israel Math. J., 4 (1966), pp. 189–198.

[18] R. REDHEFFER, *On the relation of transmission line theory to scattering and transfers*, J. Math. Phys., 41 (1962), pp. 1–41.

[19] I. SINGER, *Bases in Banach Spaces, Vol.* II, Springer, New York, 1981.

[20] P. E. SOBOLEVSKII, *Equations of parabolic type in Banach space*, Trans. Amer. Math. Soc., 1 (1965), pp. 1–62.

[21] L. TARTAR, *Sur l'étude directe d'équations non linéaires intervenant en théorie du contrôle optimal*, J. Funct. Anal., 7 (1971), pp. 85–115.

[22] R. TEMAM, *Sur l'équation de Riccati associée à des opérateurs non bornés en dimension infinite*, J. Funct. Anal., 8 (1971), pp. 85–115.

[23] W. M. WONHAM, *Optimal stationary control of linear systems with state-dependent noise*, SIAM J. Control, 5 (1967), pp. 468–500.

[24] ______, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681–697.

[25] K. YOSIDA, *Functional Analysis*, Springer-Verlag, New York, 1966.

003

AN ABSTRACT MODEL FOR RADIATIVE TRANSFER IN AN ATMOSPHERE WITH REFLECTION BY THE PLANETARY SURFACE*

W. GREENBERG[†] AND C. V. M. VAN DER MEE[‡]

Abstract. A Hilbert space model is developed that applies to radiative transfer in a homogeneous, plane-parallel planetary atmosphere. Reflection and absorption by the planetary surface are taken into account by imposing a reflective boundary condition. The existence and uniqueness of the solution of this boundary value problem are established by proving the invertibility of a scattering operator using the Fredholm alternative.

1. Introduction. It is well known (cf. [10], [1], [13], [8]) that on neglecting polarization and thermal emission the transfer of radiation through a plane-parallel, vertically homogeneous planetary atmosphere of finite optical thickness τ can be described by the abstract differential equation

$$(1.1)\qquad (Tg)'(x) = -Ag(x), \qquad 0<x<\tau,$$

where T is a bounded injective self-adjoint operator and A a positive self-adjoint compact perturbation of the identity, both of them acting on a complex Hilbert space H. For the mth Fourier component problem in radiative transfer one has $H=L_2[-1,1]$, while T and A are given by

(1.2)

$$(Th)(\mu) = \mu h(\mu),$$

$$(Ah)(\mu) = h(\mu) - \frac{c}{4\pi}\int_{-1}^{1}\int_{0}^{2\pi} p\left(\mu\mu' + \sqrt{1-\mu^2}\sqrt{1-(\mu')^2}\cos\alpha\right)\cos m\alpha\, h(\mu')\,d\alpha\,d\mu'.$$

Here the phase function p is nonnegative and $\int_{-1}^{1} p(t)\,dt = 2$, while $0 \leqq c \leqq 1$ is the albedo of single scattering (cf. [5], [15], [11]). Equation (1.2) also appears in neutron transport theory (see [3], [7]).

In the mathematical literature (1.1) usually is endowed with partial-range boundary conditions of the form

$$(1.3)\qquad Q_+g(0) = f_+ \in \operatorname{Ran} Q_+, \qquad Q_-g(\tau) = f_- \in \operatorname{Ran} Q_-,$$

where $Q_\pm$ is the $(\cdot,\cdot)$-orthogonal projection onto the maximal T-positive/negative T-invariant subspace of H. For the specific T in (1.2) one in fact has

$$(Q_+h)(\mu) = \begin{cases} h(\mu), & \mu>0, \\ 0, & \mu<0, \end{cases} \qquad (Q_-h)(\mu) = \begin{cases} 0, & \mu>0, \\ h(\mu), & \mu<0. \end{cases}$$

*Received by the editors March 22, 1983 and in revised form October 26, 1983.

[†] Department of Mathematics, University of Florence, Viale Morgagni 67A, Florence, Italy. Permanent Address, Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. The research of this author was supported by the U.S. Department of Energy under grant DE-AS05 80ER10711.

[‡] Department of Physics and Astronomy, Free University, De Boelelaan 1081, 1081 HV Amsterdam, the Netherlands.

Although in neutron physics equations (1.3) are a realistic set of boundary conditions (because neutrons typically do not reflect), in planetary physics equations (1.3) are satisfied only on neglecting reflection by the planetary surface. In order to formulate abstract boundary conditions to (1.1) that describe reflection and absorption by the planetary surface for the example (1.2), we assume the existence of a signature operator J (i.e., $J=J^*=J^{-1}$) such that

$$(1.4) \qquad JT=-TJ, \qquad JA=AJ.$$

For T and A in (1.2) one may, indeed, take

$$(1.5) \qquad (Jh)(\mu)=h(-\mu).$$

Now let $\mathscr{R}$ be an operator on $\operatorname{Ran} Q_+$ such that $0 \leqq T\mathscr{R} \leqq T$ on $\operatorname{Ran} Q_+$. Then on (1.1) we impose the boundary conditions

$$(1.6) \qquad Q_+g(0)=f_+, \qquad Q_-g(\tau)=J\mathscr{R}Q_+g(\tau).$$

We call J an *inversion symmetry*, $\mathscr{R}$ the *surface reflection operator* and (1.1) with boundary conditions (1.6) an (*abstract*) *planetary problem*. Equation (1.1) with boundary conditions (1.3) we shall call an (*abstract*) *finite-slab problem*, which is the name prevalent in neutron physics. By a solution of the planetary problem shall be meant a continuous function g: $[0,\tau]\to H$ such that Tg is differentiable on $(0,\tau)$ in the strong sense and (1.1) and (1.6) are satisfied.

Equations (1.6) are so-called reflective boundary conditions. In rarefied gas dynamics [4] and radiative transfer [6] they are common practice. It has only been recently that Beals and Protopopescu [2] have given a rigorous treatment of such problems for the generalized Fokker–Planck equation. However, their boundary conditions differ from (1.6) and do not show a general abstract form. In the present article we shall draw on some results of van der Mee on the abstract finite-slab problem [13] and reflection and transmission operators [14] as well as on an inner product of Beals [1].

Under the weak assumption that $\operatorname{Ran}(I-A)\subseteq \operatorname{Ran}|T|^{\alpha}$ for some $0<\alpha<1$, the finite-slab problem (1.1)–(1.3) has a unique solution given by

$$(1.7) \quad g(x)=\Big[e^{-xT^{-1}A}PP_+ + e^{(\tau-x)T^{-1}A}PP_- + (I-xT^{-1}A)(I-P)\Big]V_\tau^{-1}(f_+ + f_-).$$

As we shall point out in §2, $T^{-1}A$ is self-adjoint with respect to an equivalent inner product, except possibly for an isolated eigenvalue at zero. Then P, $I-P$, PP_+ and PP_- are the spectral projections of $T^{-1}A$ corresponding to the nonzero, zero, positive and negative part of the spectrum, respectively, while

$$(1.8)$$
$$V_\tau = Q_+\Big[PP_+ + e^{\tau T^{-1}A}PP_- + (I-P)\Big] + Q_-\Big[PP_- + e^{-\tau T^{-1}A}PP_+ + (I-\tau T^{-1}A)(I-P)\Big]$$

is an invertible operator. The result is due to van der Mee [13], a parallel proof of the invertibility of V_τ (but for strictly positive A) was found by Hangelbroek, and a related result, with the solution in some extension of the Hilbert space H but for more general A, was proved by Beals [1]. As in [14], we write

$$(1.9) \qquad g(0)=R_{+\tau}f_+ + T_{-\tau}f_-, \qquad g(\tau)=T_{+\tau}f_+ + R_{-\tau}f_-,$$

where $R_{\pm\tau}$ are *reflection operators* and $T_{\pm\tau}$ *transmission operators*. These operators are uniquely specified by (1.9) and the requirement $R_{\pm\tau}Q_{\mp}=T_{\pm\tau}Q_{\mp}=0$, and their closed form can be found using (1.7).

Let us combine (1.9) with the boundary conditions (1.6) and apply Q_+ to the left. We obtain

$$(1.10)\qquad (Q_+-Q_+R_{-\tau}J\mathscr{R})Q_+g(\tau)=Q_+T_{+\tau}f_+.$$

Once $Q_+g(\tau)$ has been found from (1.10), one computes $Q_-g(\tau)$ from (1.6) and gets the solution in the form (1.7) with $f_-=Q_-g(\tau)$. Hence, the vector equation (1.10) is equivalent to the abstract planetary problem. From (1.4) one finds $JQ_{\pm}=Q_{\mp}J$ and $JR_{\pm\tau}=R_{\mp\tau}J$ (cf. [14]), whence

$$J(Q_+-Q_+R_{-\tau}J\mathscr{R})=(Q_--Q_-R_{+\tau}\mathscr{R}J)J.$$

In order to solve (1.10) we thus have to investigate the invertibility of the *$\mathscr{R}$-scattering operator*

$$(1.11)\qquad S_{\mathscr{R}}=I-Q_+R_{-\tau}J\mathscr{R}-Q_-R_{+\tau}\mathscr{R}J.$$

We state the main results of this article.

THEOREM 1.1. *Let* $\operatorname{Ran}(I-A)\subseteq\operatorname{Ran}|T|^{\alpha}$ *for some* $0<\alpha<1$, *and let* $0\leqq T\mathscr{R}\leqq T$ *on* $\operatorname{Ran}Q_+$. *Then the $\mathscr{R}$-scattering operator in* (1.11) *is invertible.*

Using standard semigroup theory we then have as a consequence the next theorem.

THEOREM 1.2. *Let* $\operatorname{Ran}(I-A)\subset\operatorname{Ran}|T|^{\alpha}$ *for some* $0<\alpha<1$, *and let* $0\leqq T\mathscr{R}\leqq T$ *on* $\operatorname{Ran}Q_+$. *Then for every* $f_+\in\operatorname{Ran}Q_+$ *the planetary problem* (1.1) *and* (1.6) *has a unique solution, which is given by* (1.7) *where*

$$f_-=J\mathscr{R}S_{\mathscr{R}}^{-1}Q_+T_{+\tau}f_+.$$

We have required that $0\leqq T\mathscr{R}\leqq T$ on $\operatorname{Ran}Q_+$. For the example of radiative transfer this implies that the radiative flux returning from the planetary surface does not exceed the flux incident to the surface. For $\mathscr{R}=0$ one has total absorption, for $\mathscr{R}=I$ specular reflection and for

$$(\mathscr{R}h)(\mu)=2\int_0^1\mu'h(\mu')\,d\mu'$$

diffuse reflection. In [2] and [4] the only surface reflection operators studied are $\mathscr{R}=(1-\alpha)I$ where, in rarefied gas dynamics terminology, $0\leqq\alpha\leqq1$ is the accommodation coefficient. In [6] the more general surface reflection operator

$$(\mathscr{R}h)(\mu)=\rho_sh(\mu)+2\rho_d\int_0^1\mu'h(\mu')\,d\mu'$$

is used, where $\rho_s+\rho_d\leqq1$, $\rho_s\geqq0$ and $\rho_d\geqq0$. In all cases the hypothesis $0\leqq T\mathscr{R}\leqq T$ on $\operatorname{Ran}Q_+$ is fulfilled. If the phase function $p\in L_r[-1,1]$ with $r>1$, then $\operatorname{Ran}(I-A)\subseteq\operatorname{Ran}|T|^{\alpha}$ for every $0<\alpha<(r-1)/2r$ [13, §VI.1].

In §2 we shall review some properties of reflection and transmission operators, partly from [14] and partly hitherto unknown. In §3 we prove the invertibility of the $\mathscr{R}$-scattering operator for $\mathscr{R}=I$ (specular reflection). Finally, in §4 we prove Theorem 1.1.

2. Reflection and transmission operators. Throughout the present and the next section T is a bounded injective self-adjoint operator and A a positive operator, defined on the complex Hilbert space H. We assume that $I-A$ is compact and $\mathrm{Ran}(I-A)\subseteq \mathrm{Ran}|T|^{\alpha}$ for some $0<\alpha<1$. By $Q_{\pm}$ we denote the orthogonal projection onto the maximal positive/negative T-invariant subspace of H. If $\mathrm{Ker}A=\{0\}$, then, as Hangelbroek [9] observed, H is a Hilbert space with respect to the inner product

$$(2.1)\qquad (x,y)_A=(Ax,y)$$

and $T^{-1}A$ is self-adjoint with respect to (2.1). The $(\cdot,\cdot)_A$-orthogonal projection onto the maximal positive/negative $T^{-1}A$-invariant subspace of H is denoted by $P_{\pm}$. If $\mathrm{Ker}A\neq\{0\}$, then $T^{-1}A$ has a nonzero and finite-dimensional zero root linear manifold

$$Z_0(T^{-1}A)=\{x\in H/\exists n\geqq 0:(T^{-1}A)^n x=0\},$$

while the $(\cdot,\cdot)$-orthogonal complement $Z_1=\{T[Z_0(T^{-1}A)]\}^{\perp}$ of the subspace $T[Z_0(T^{-1}A)]$ is $T^{-1}A$-invariant and a Hilbert space with respect to (2.1) and the restriction of $T^{-1}A$ to Z_1 is $(\cdot,\cdot)_A$-selfadjoint. The projection of H onto Z_1 along $Z_0(T^{-1}A)$ we denote by P, whence the $(\cdot,\cdot)_A$-orthogonal projection onto the maximal positive/negative $T^{-1}A$-invariant subspace of H is given by $PP_{\pm}$, where $Z_0(T^{-1}A)\subseteq \mathrm{Ker}\,PP_{\pm}$. The idea to study $T^{-1}A$ on the finite-codimensional subspace Z_1 was first exploited by Lekkerkerker [12] for neutron transport with isotropic scattering.

For every $f\in H$ the abstract finite-slab problem (1.1)–(1.3), where $f_{\pm}=Q_{\pm}f$, has a unique solution g, which is given by (1.7) (see [13]). In terms of the solution g one may specify in a unique way reflection operators $R_{\pm\tau}$ and transmission operators $T_{\pm\tau}$ such that $R_{\pm\tau}Q_{\mp}=T_{\pm\tau}Q_{\mp}=0$ (see (1.9)). More precisely, if $f_{+}=0$ (resp. $f_{-}=0$), then $g(0)=T_{-\tau}f_{-}$ (resp. $g(0)=R_{+\tau}f_{+}$) and $g(\tau)=R_{-\tau}f_{-}$ (resp. $g(\tau)=T_{+\tau}f_{+}$). The expression (1.7) can now be used to find the following explicit formulas:

$$(2.2)\qquad R_{+\tau}=\left[PP_{+}+e^{\tau T^{-1}A}PP_{-}+(I-P)\right]V_{\tau}^{-1}Q_{+},$$

$$(2.3)\qquad T_{+\tau}=\left[PP_{-}+e^{-\tau T^{-1}A}PP_{+}+(I-\tau T^{-1}A)(I-P)\right]V_{\tau}^{-1}Q_{+},$$

$$(2.4)\qquad R_{-\tau}=\left[PP_{-}+e^{-\tau T^{-1}A}PP_{+}+(I-\tau T^{-1}A)(I-P)\right]V_{\tau}^{-1}Q_{-},$$

$$(2.5)\qquad T_{-\tau}=\left[PP_{+}+e^{\tau T^{-1}A}PP_{-}+(I-P)\right]V_{\tau}^{-1}Q_{-}.$$

Using (1.8) one easily finds

$$(2.6)\qquad Q_{\pm}R_{\pm\tau}=Q_{\pm},\qquad Q_{\mp}T_{\pm\tau}=0,$$

$$(2.7)\qquad R_{\pm\tau}Q_{\pm}=R_{\pm\tau},\qquad T_{\pm\tau}Q_{\pm}=T_{\pm\tau}.$$

We also find that $R_{\pm\tau}$ is a projection operator such that $R_{\pm\tau}-Q_{\pm}$ is compact (cf. [14]). In a less elementary way (see [14]) one derives the intertwining properties

$$(2.8)\qquad TR_{\pm\tau}=(I-R_{\mp\tau}^{*})T,\qquad TT_{\pm\tau}=T_{\pm\tau}^{*}T.$$

PROPOSITION 2.1. *One has the decompositions*

$$\mathrm{Ran}\,R_{\pm\tau}\oplus\mathrm{Ran}\,Q_{\mp}=H.$$

Proof. Put

$$U_1 = Q_+ R_{+\tau} + Q_-(I - R_{+\tau}).$$

Then the invertibility of U_1 is easily proved equivalent to the decomposition

$$\operatorname{Ran} R_{+\tau} \oplus \operatorname{Ran} Q_- = H.$$

Using (2.6) one computes that

$$U_1 = Q_+ + Q_- - (I - Q_+) R_{+\tau} = I - (R_{+\tau} - Q_+),$$

whence $I - U_1$ is compact. If $U_1 h = 0$, then the vector $Q_+ h = -(I - R_{+\tau})h \in \operatorname{Ran} Q_+ \cap \operatorname{Ran} Q_- = \{0\}$, which implies $Q_+ h = (I - R_{+\tau})h = 0$ and therefore

$$h = R_{+\tau} h + (I - R_{+\tau})h = R_{+\tau} h = R_{+\tau} Q_+ h = 0.$$

Thus $\operatorname{Ker} U_1 = \{0\}$ and the invertibility of U_1 is clear. Q.E.D.

PROPOSITION 2.2. *One has the decomposition*

$$\operatorname{Ran} R_{+\tau} \oplus \operatorname{Ran} R_{-\tau} = H.$$

Proof. Assume that, for some $k, l \in H$,

$$R_{+\tau} k = R_{-\tau} l.$$

Putting $x_\pm = \pm Q_\pm(k - l)$ one finds $h = k + x_- = l + x_+$ and

$$R_{+\tau} h = R_{-\tau} h.$$

On premultiplying this equality by Q_+ and Q_- one gets

$$h = R_{+\tau} h = R_{-\tau} h,$$

which implies (see (2.2)–(2.5))

$$T_{+\tau} h = T_{+\tau} R_{-\tau} h = 0, \qquad T_{-\tau} h = T_{-\tau} R_{+\tau} h = 0.$$

Hence,

$$h = R_{+\tau} h + T_{-\tau} h = \left[PP_+ + e^{\tau T^{-1}A} PP_- + (I - P) \right] V_\tau^{-1} h,$$

$$h = R_{-\tau} h + T_{+\tau} h = \left[e^{-\tau T^{-1}A} PP_+ + PP_- + (I - \tau T^{-1}A)(I - P) \right] V_\tau^{-1} h.$$

From these equations one finds

$$0 = \pm \int_0^{\pm\infty} (1 - e^{\mp \tau/z}) F(dz) V_\tau^{-1} h, \qquad (I - P) V_\tau^{-1} h = (I - P) h \in \operatorname{Ker} A,$$

where F is the resolution of the identity of $(T^{-1}A|Z_1)^{-1}$ (as a self-adjoint operator with respect to (2.1)). Hence, $PP_\pm V_\tau^{-1} h = PP_\pm h = 0$, while

$$h = V_\tau h \in \operatorname{Ker} A.$$

Then, since $T_{\pm\tau} h = 0$, we have

$$PP_\pm V_\tau^{-1} Q_\pm h = 0, \qquad PP_\pm V_\tau^{-1} Q_\mp h = 0,$$
$$(I - P) V_\tau^{-1} Q_- h = 0, \qquad (I - \tau T^{-1}A)(I - P) V_\tau^{-1} Q_+ h = 0.$$

All this implies

$$V_\tau^{-1}Q_\pm h \in \operatorname{Ker} A.$$

Using that $V_\tau x = x$ for $x \in \operatorname{Ker} A$ (cf. (1.8)), one obtains

$$Q_\pm h \in \operatorname{Ker} A.$$

However, we also have $V_\tau^{-1}y = y$ for $y \in \operatorname{Ker} A$. Thus, in view of $T_{\pm\tau}h = 0$,

$$Q_+h = (I-P)V_\tau^{-1}Q_+h = (I-\tau T^{-1}A)(I-P)V_\tau^{-1}Q_+h = 0,$$
$$Q_-h = (I-P)V_\tau^{-1}Q_-h = 0,$$

which implies $h = 0$. From this we find the injectivity of the operator

$$U_2 = R_{+\tau}(I-R_{-\tau}) + (I-R_{+\tau})R_{-\tau}.$$

However, the compactness of $R_{\pm\tau} - Q_\pm$ implies that

$$U_2 - I = (R_{+\tau}-Q_+) - R_{+\tau}(R_{-\tau}-Q_-) + (I-R_{+\tau})(R_{-\tau}-Q_-)$$

is a compact operator. As $\operatorname{Ker} U_2 \subseteq \operatorname{Ran} R_{+\tau} \cap \operatorname{Ran} R_{-\tau} = \{0\}$, we conclude that U_2 is invertible. From the invertibility of U_2 we easily derive this proposition. Q.E.D.

We note that neither of the proofs of the propositions required the existence of an inversion symmetry J satisfying (1.4). In case there exists such an inversion symmetry, one may conclude that

$$JQ_\pm = Q_\mp J, \qquad JPP_\pm = PP_\mp J, \tag{2.9}$$
$$JR_{\pm\tau} = R_{\mp\tau}J, \quad JT_{\pm\tau} = T_{\mp\tau}J. \tag{2.10}$$

3. Invertibility of the scattering operators. First we prove Theorem 1.1 for $\mathscr{R} = I$ (specular reflection).

PROPOSITION 3.1. *The $\mathscr{R}$-scattering operator for $\mathscr{R} = I$ is invertible.*

Proof. We have

$$S_I = I - Q_+R_{-\tau}J - Q_-R_{+\tau}J.$$

Clearly this operator is reduced by the orthogonal decomposition

$$\{x \in H / Jx = x\} \oplus \{x \in H / Jx = -x\} = H \tag{3.1}$$

(see (2.9)–(2.10)) and therefore it suffices to prove the invertibility of the operators $I \pm (Q_+R_{-\tau} + Q_-R_{+\tau})$. As a result of (2.6) one has to prove the invertibility of the operators $R_{+\tau} + R_{-\tau}$ and $2I - (R_{+\tau} + R_{-\tau})$. Since $R_{\pm\tau} - Q_\pm$ is compact, both of these operators are compact perturbations of the identity and therefore it is sufficient to prove that neither $\lambda = 0$ nor $\lambda = 2$ is an eigenvalue of $R_{+\tau} + R_{-\tau}$.

If $(R_{+\tau} + R_{-\tau})h = 0$, then

$$R_{+\tau}h = -R_{-\tau}h \in \operatorname{Ran} R_{+\tau} \cap \operatorname{Ran} R_{-\tau}.$$

Using Proposition 2.2 one finds $R_{\pm\tau}h = 0$ and therefore $Q_\pm h = Q_\pm R_{\pm\tau}h = 0$. So we may exclude $\lambda = 0$ as an eigenvalue of $R_{+\tau} + R_{-\tau}$.

If $(R_{+\tau} + R_{-\tau})k = 2k$, then

$$(I-R_{+\tau})k = -(I-R_{-\tau})k \in \operatorname{Ran} Q_+ \cap \operatorname{Ran} Q_- = \{0\},$$

whence $k=R_{+\tau}k=R_{-\tau}k$. Proposition 2.2 implies $k=0$, which excludes $\lambda=2$ as an eigenvalue of $R_{+\tau}+R_{-\tau}$. Q.E.D.

The next result will play an important role in the proof of Theorem 1.1 but is also interesting for its own sake.

PROPOSITION 3.2. *For* $0<\tau<\infty$ *the operator* $|T|(R_{+\tau}+R_{-\tau})$ *is self-adjoint and satisfies*

$$0\leqq|T|(R_{+\tau}+R_{-\tau})\leqq 2|T|.$$

If Q_τ *denotes the projection of H onto* $\operatorname{Ran}R_{+\tau}$ *along* $\operatorname{Ran}R_{-\tau}$, *we have*

$$[R_{+\tau}+R_{-\tau}]^{-1}=Q_+Q_\tau+Q_-(I-Q_\tau), \tag{3.2}$$

$$[2I-(R_{+\tau}+R_{-\tau})]^{-1}=Q_\tau Q_++(I-Q_\tau)Q_-. \tag{3.3}$$

Proof. With the help of (2.7) and the elementary identities

$$R_{+\tau}Q_\tau=Q_\tau, \qquad R_{-\tau}(I-Q_\tau)=I-Q_\tau$$

one easily proves (3.2) and (3.3).

Using (2.8) one computes that

$$\{|T|(R_{+\tau}+R_{-\tau})\}^*=T\{(I-R_{-\tau})+(I-R_{+\tau})\}(Q_+-Q_-).$$

Next one exploits (2.7) and subsequently (2.6) and derives

$$\{|T|(R_{+\tau}+R_{-\tau})\}^*=2|T|-T(R_{+\tau}-R_{-\tau})=|T|(R_{+\tau}+R_{-\tau}),$$

which establishes the self-adjointness of $|T|(R_{+\tau}+R_{-\tau})$. Hence, the eigenvalues of $R_{+\tau}+R_{-\tau}$ are situated on the real line. It suffices to prove that $\sigma(R_{+\tau}+R_{-\tau})=\{\lambda\in\mathbb{C}/\lambda I-(R_{+\tau}+R_{-\tau})$ is not invertible$\}\subseteq(0,2)$.

Using (1.7) and (1.8) one concludes that

$$\lim_{\tau\downarrow 0}\|I-(R_{+\tau}+R_{-\tau})\|=0. \tag{3.4}$$

However, for every $0<\tau<\infty$ the operator $R_{+\tau}+R_{-\tau}$ is a compact perturbation of the identity. If $\sigma(R_{+\tau_0}+R_{-\tau_0})\not\subseteq(0,2)$, either the smallest eigenvalue of $R_{+\tau_0}+R_{-\tau_0}$ is negative or the largest eigenvalue exceeds 2. Because both the infimum and supremum of $\sigma(R_{+\tau}+R_{-\tau})$ depend continuously on τ and (3.4) holds true, there must exist $0<\tau_1<\tau_0$ such that either 0 or 2 is an eigenvalue of $R_{+\tau_1}+R_{-\tau_1}$, which is a contradiction. Hence, $\sigma(R_{+\tau}+R_{-\tau})\subseteq(0,2)$ for all $0<\tau<\infty$. Q.E.D.

We remark that

$$(Q_+-Q_-)[R_{+\tau}+R_{-\tau}]=[2I-(R_{+\tau}+R_{-\tau})](Q_+-Q_-),$$

so that the (real) spectrum of $R_{+\tau}+R_{-\tau}$ is symmetric with respect to $\lambda=1$.

Proof of Theorem 1.1. Let us first extend the surface reflection operator $\mathscr{R}$ from $\operatorname{Ran}Q_+$ to H by putting

$$\mathscr{R}h=\mathscr{R}Q_+h+J\mathscr{R}JQ_-h, \qquad h\in H.$$

Then the $\mathscr{R}$-scattering operator is given by

$$S_{\mathscr{R}}=I-[Q_-R_{+\tau}+Q_+R_{-\tau}]\mathscr{R}J.$$

Since $J\mathscr{R}=\mathscr{R}J$, this operator is reduced by the decomposition (3.1). Thus it suffices to establish the invertibility of the two operators

$$I+[Q_-R_{+\tau}+Q_+R_{-\tau}]\mathscr{R}=I+(R_{+\tau}+R_{-\tau}-I)\mathscr{R},$$
$$I-[Q_-R_{+\tau}+Q_+R_{-\tau}]\mathscr{R}=I-(R_{+\tau}+R_{-\tau}-I)\mathscr{R},$$

both of which are compact perturbations of the identity.

Following Beals [1] we introduce the completion H_T of H with respect to the inner product

$$(x,y)_T=(|T|x,y).$$

As the (extended) operator $\mathscr{R}$ satisfies $0\leqq|T|\mathscr{R}\leqq|T|$, one has

$$0\leqq(\mathscr{R}x,x)_T=(|T|\mathscr{R}x,x)\leqq(x,x)_T, \qquad x\in H\subseteq H_T,$$

and therefore $\mathscr{R}$ extends to a positive contraction on H_T, also denoted by $\mathscr{R}$. Proposition 3.2 implies that $R_{+\tau}+R_{-\tau}-I$ extends to a strict contraction on H_T. Hence,

$$(R_{+\tau}+R_{-\tau}-I)\mathscr{R}$$

has H_T-norm strictly less than unity. We thus find the invertibility of the operators $I\pm(R_{+\tau}+R_{-\tau}-I)\mathscr{R}$ on H_T. On the original Hilbert space H these operators have zero null space and are compact perturbations of the identity and therefore invertible too. Q.E.D.

REFERENCES

[1] R. Beals, *An abstract treatment of some forward-backward problems of transport and scattering*, J. Funct. Anal., 34 (1979), pp. 1–20.

[2] R. Beals and V. Protopopescu, *Half-range completeness for the Fokker–Planck equation*, J. Stat. Phys., 32 (1983), pp. 565–584.

[3] K. M. Case and P. F. Zweifel, *Linear Transport Theory*, Addison-Wesley, Reading, MA, 1967.

[4] C. Cercignani, *The Kramers problem for a not completely diffusing wall*, J. Math. Anal. Appl., 10 (1965), pp. 568–586.

[5] S. Chandrasekhar, *Radiative Transfer*, 2nd rev. ed., Dover, New York, 1960.

[6] C. Devaux, C. E. Siewert and Y. L. Yuan, *The complete solution for radiative transfer problems with reflecting boundaries and internal sources*, Astrophys. J., 253 (1982), pp. 773–784.

[7] J. J. Duderstadt and W. R. Martin, *Transport Theory*, Wiley-Interscience, New York, 1979.

[8] W. Greenberg, C. V. M. van der Mee and P. F. Zweifel, *Generalized kinetic equations*, Int. Eqs. Oper. Theor., 7 (1984), pp. 60–95.

[9] R. J. Hangelbroek, *Linear analysis and solution of neutron transport equations*, Transp. Theor. Stat. Phys., 5 (1976), pp. 1–85.

[10] ______, *The dispersion function in neutron transport theory*, in Differential Equations and Applications, W. Eckhaus and E. M. de Jager, eds., North-Holland, Amsterdam, 1977.

[11] H. C. van de Hulst, *Multiple Light Scattering*, Pergamon Press, New York, 1980.

[12] C. G. Lekkerkerker, *The linear transport equation. The degenerate case $c=1$. I. Full-range theory*; II. *Half-range theory*, Proc. Royal Soc. Edinburgh, 75A (1976), pp. 259–282 and 283–295.

[13] C. V. M. van der Mee, *Semigroup and Factorization Methods in Transport Theory*, Math. Center Tract no. 146, Amsterdam, 1981.

[14] ______, *Transport equation on a finite domain. I. Reflection and transmission operators and diagonalization*, Int. Eqs. Oper. Theor., 6 (1983), pp. 572–601.

[15] V. V. Sobolev, *Light Scattering in Planetary Atmospheres*, Pergamon Press, New York, 1975.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

LINEAR OPERATORS RELATED TO TIME-INVARIANT DISCRETE FILTERS*

JOSÉ M. MORAL MEDINA†

Abstract. Discrete filters are studied as continuous translation-invariant linear operators in some fixed, but otherwise rather general, sequence space. The central questions raised is whether or not it is true that every filter L acts upon any input sequence f in a "convolution form", that is, the output Lf can be expressed as $f * h$, where h is a fixed sequence completely determined by L. For the most important input space considered, namely l^∞, the conjecture above is found to be false if the usual normed topology is adopted. However, the conjecture becomes true when l^∞ is endowed with the weak* topology, which also appears to be a natural assumption from a physical viewpoint.

1. Introduction. In electrical engineering, a discrete time-invariant linear system L is described as a linear physical device with constant features, which assigns to every input signal f (a bilateral sequence) an output signal Lf. We shall often refer to such a system as a filter.

Simple and well-known heuristic reasoning permits us to conjecture that the behaviour of the filter is governed by the formula $Lf = f * h$, where $h = L\delta$ is the so-called impulse-response of the filter, that is, the output corresponding to the "impulse signal" δ defined by $\delta(0)=1$, $\delta(k)=0$ if $k \neq 0$. A large part of this paper is devoted to an examination of the validity of such a conjecture.

1.1. General considerations. Let us consider the translation operator T, which assigns to each sequence f the sequence Tf whose nth term is $(Tf)(n) = f(n-1)$. The constancy of the features of the device is expressed mathematically by imposing the condition $L(Tf) = T(Lf)$ for any input sequence f, which means that a delay in the input only causes an identical delay in the output. Thus, to clarify the question of whether or not the outputs of a filter can be represented as the convolution of the inputs with a fixed signal, it is necessary to study the general form of continuous translation-invariant linear (c.t.i.l.) operators acting on different sequence spaces. The results we shall state here can be interpreted as confirmation of the validity of the conjectured representation, conditional on the replacement of the normed topology on l^∞ by a more suitable one. We think the weak* topology reflects better than the normed one the continuity of the physical phenomenon under consideration.

In order to introduce some necessary mathematical preliminaries, let us call $\omega = \omega(Z)$ the F-space [3, p. 51] of all complex bilateral sequences endowed with the topology of pointwise convergence. If $\lambda \subset \omega$ and $\chi \subset \omega$ are topological vector spaces we denote by $B(\lambda, \chi)$ the space of continuous linear operators from λ to χ, and by $\mathrm{TI}(\lambda, \chi)$ the subspace of those which are translation-invariant. As is well known, when λ and χ are Banach spaces, so is $B(\lambda, \chi)$ with the usual definition of the norm [10, p. 234]. Moreover, the same is easily seen to be true for $\mathrm{TI}(\lambda, \chi)$, when T is continuous on χ.

1.2. A preliminary result. In this subsection we will show that there is an isomorphism between $\mathrm{TI}(\lambda, \omega)$ and the space of continuous linear functionals on λ (i.e., the

*Received by the editors August 12, 1980, and in revised form December 14, 1982.

†E. T. S. Ing. Caminos, Universidad Politécnica, Barcelona, Spain. Now in E.T.S. Ing. Agrónomos, Universidad Politécnica, Madrid, Spain.

conjugate space λ^* of λ). This fact will enable us to replace every argument about translation-invariant operators by one about linear functionals.

To this effect we assume the normed topology in l^∞, given by $\|f\|_\infty = \sup\{|f(n)|: n \in Z\}$ and we state the following:

LEMMA 1. *Let* λ, $\lambda \subset \omega$, *be a topological vector space such that* T *is a topological isomorphism from* λ *onto itself. Then* $\mathrm{TI}(\lambda, \omega)$ *is isomorphic to* λ^*. *Furthermore*, $\mathrm{TI}(\lambda, \omega) = \mathrm{TI}(\lambda, l^\infty)$ *whenever* λ *is an F-space and* T *an isometry, and* $\mathrm{TI}(\lambda, l^\infty)$ *and* λ^* *are norm-isomorphic if in addition* λ *is a Banach space.*

Proof. For each $n \in Z$, let us call π_n the nth projection from ω onto the space C of complex numbers. Let us define $\psi(L) = \gamma_L$ for each $L \in \mathrm{TI}(\lambda, \omega)$, where $\gamma_L(f) = (Lf)(0)$ for every $f \in \lambda$. It is clear that ψ is a linear and injective mapping from $\mathrm{TI}(\lambda, \omega)$ to λ^*. Suppose $\gamma \in \lambda^*$ and let L be the translation-invariant linear mapping from λ to ω given by $(Lf)(n) = \gamma(T^{-n}f)$ for all $f \in \lambda$ and $n \in Z$. All mappings $\pi_n L = \gamma T^{-n}$ are continuous, and then $L \in \mathrm{TI}(\lambda, \omega)$. Obviously $\gamma_L = \gamma$, which proves the first part of the lemma. As to the second, if $L \in \mathrm{TI}(\lambda, \omega)$ let $\varepsilon > 0$ be such that $|(Lg)(0)| < 1$ whenever $g \in \lambda$ and $d(g, 0) < \varepsilon$. Given f in λ we can find $\alpha > 0$ such that $d(\alpha f, 0) = d(T^{-n}(\alpha f), 0) < \varepsilon$ for each integer n. Therefore,

$$|(LT^{-n}(\alpha f))(0)| = \alpha |(Lf)(n)| < 1$$

and $Lf \in l^\infty$. The continuity of L as an operator from λ to l^∞ follows [3, p. 58] from the continuity of all mappings $\pi_n L$. Thus $L \in \mathrm{TI}(\lambda, l^\infty)$, and $\mathrm{TI}(\lambda, \omega) \subset \mathrm{TI}(\lambda, l^\infty)$. The reciprocal inclusion is evident, as well as the third part of the lemma.

Remark 1. Lemma 1 is not always valid for spaces of nonbilateral (ordinary) sequences, as ψ may fail to be surjective in this case, see [11].

2. All spaces χ considered in this paper are linear subspaces of ω with topologies stronger than the one induced by ω, so that $\mathrm{TI}(\lambda, \chi) \subset \mathrm{TI}(\lambda, \omega)$.

2. Operators representable as convolutions with sequences.

2.1. Operators defined on the space ω. We start this subsection with a result which serves to justify the necessity of considering proper subspaces of ω in the mathematical formulation of filtering theory. We denote by φ the linear space of all sequences with a finite number of nonnull coordinates.

THEOREM 1. *Let* L *be an operator from* ω *to* ω *and* $h = L\delta$. *Then* L *belongs to* $\mathrm{TI}(\omega, \omega)$ *if and only if* $h \in \varphi$ *and* $Lf = f * h$ *for all* f *in* ω. *Moreover, the collection* $\mathcal{T} = \{T^k: k \in Z\}$ *is an algebraic basis of* $\mathrm{TI}(\omega, \omega)$.

Proof. The theorem follows easily from the first part of Lemma 1, since ω^* is isomorphic to φ in the following sense [1, p. 50]. If $\gamma \in \omega^*$ there is a sequence $h \in \varphi$ such that $\gamma(f) = \sum_k h(k) f(-k)$ for every $f \in \omega$, and, conversely, each $h \in \varphi$ defines, via this formula, a functional γ from ω^*. If we let $\delta_k = T^k \delta$ for each integer k, it is obvious that $h(k) = \gamma(\delta_{-k})$. As to the last assertion of the theorem it suffices to observe that the equality $L = \sum_{k=-m}^{n} h(k) T^k$ for fixed m, $n \in Z^+$ is equivalent to $(Lf)(j) = \sum_{k=-m}^{n} h(k) f(j-k)$ for all $f \in \omega$ and $j \in Z$.

COROLLARY 1. *If* $L \in \mathrm{TI}(\omega, \omega)$ *and the sequence* $h = L\delta$ *has more than one nonnull coordinate*, L *is not injective.*

In fact, if α is a nonnull root of $\sum h(k) z^k$ the image under L of the sequence (α^{-k}) is the null sequence, and so $\operatorname{Ker} L \neq \{0\}$.

In other words, the filter destroys information: the input cannot be known from the output, in contrast to the situation for ordinary sequences (11, p. 362).

Remark 3. The importance of Theorem 1 lies in the fact that it identifies the class of all c.t.i.l. filters which admit arbitrary inputs, with the class of those filters that may be decomposed into a finite number of delay (translation), multiplier and adding elements operating in a nonrecursive way, i.e., without processing the coordinates $(Lf)(k)$, $k<n$, to get $(Lf)(n)$. To enlarge the class of c.t.i.l. filters we must restrict the inputs by choosing smaller spaces, such as φ, l^p $(p \geqq 1)$ and c_0. As usual φ is topologized as the inductive limit of the family of spaces $\{\varphi_{mn}\colon m,n \in Z^+\}$. Each φ_{mn} is the space of those sequences vanishing outside of the interval $[-m,n]$ of Z, endowed with the only possible topological vector space structure. The topology on φ is thereby the strongest which induces on each φ_{mn} its own topology, and a mapping from φ to another topological space is continuous if and only if its restriction to each φ_{mn} is continuous.

2.2. Operators defined on the space φ. The arguments put forward in Remark 3 lead us to consider filters whose inputs belong to φ.

THEOREM 2. *Let L be an operator from φ to ω and $h=L\delta$. Then L is linear and translation-invariant if and only if $Lf=f*h$ for all f in φ, and then $L \in \mathrm{TI}(\varphi,\omega)$. Moreover, $L \in \mathrm{TI}(\varphi,\varphi)$ if and only if in addition $h \in \varphi$.*

Proof. If L is linear and translation-invariant and $f \in \varphi$, then from $f=\Sigma_k f(k)T^k\delta$ we get $Lf=\Sigma_k f(k)T^k h$, since the summation is over a finite number of nonnull terms. Thus $Lf=f*h$. Also, L is continuous, its restriction to each φ_{mn} being a linear operator on a finite-dimensional topological vector space. The remaining assertions are evident.

Remark 4. In contrast with the above result it is worth noting that φ is dense in ω if we consider the pointwise topology, and so we only obtain as c.t.i.l. filters those which consist of a finite number of delay, multiplier and adding elements.

2.3. Operators defined on the spaces l^p and c_0. In order to gain theoretical insight into filter operations, we will now consider spaces larger than φ. We begin with the spaces l^p, $p \geqq 1$, and the space c_0 of the sequences f for which $\lim_{|n| \to \infty} f(n)=0$. The norm on each l^p is defined as usual by $\|f\|_p=(\Sigma_k|f(k)|^p)^{1/p}$, and c_0 is regarded as a subspace of l^∞, so that both of them are Banach spaces. We denote the norm of L as an element of $\mathrm{TI}(l^p,l^r)$ by $\|L\|_{p,r}$ $(1 \leqq p,r \leqq \infty)$. The conjugate index of p shall be noted by p' $(1/p+1/p'=1)$.

THEOREM 3. *Let $p \geqq 1$ be a real number, L an operator from l^p to l^∞, and $h=L\delta$. Then L belongs to $\mathrm{TI}(l^p,l^\infty)$ if and only if $h \in l^{p'}$ and $Lf=f*h$ for all f in l^p. Furthermore, $\mathcal{T}=\{T^k\colon k \in Z\}$ is a Banach basis for $\mathrm{TI}(l^p,l^\infty)$ and $\|L\|_{p,\infty}=\|h\|_{p'}$. If $L \in \mathrm{TI}(l^p,l^\infty)$ and $h \in L^q$, $1 \leq q \leq p'$, then $L \in \mathrm{TI}(l^p,l^r)$, where $1/r=1/p+1/q-1$, and $\lim_{m,n\to\infty}\|L-\Sigma_{k=-m}^n h(k)T^k\|_{p,r}=0$.*

Proof. The first assertion and the equality $\|L\|_{p,\infty}=\|h\|_{p'}$ may be deduced from Lemma 1 and the well-known representation of the conjugate of l^p [1, p. 67]. Let p,q,r be as above and $h \in l^q$. Then Young's inequality [4, p. 199] gives us $\|f*h\|_r \leqq \|f\|_p\|h\|_q$ for all $f \in l^p$, which proves that $L \in \mathrm{TI}(l^p,l^r)$. Let us set $h_{mn}=\Sigma_{k=-m}^n h(k)T^k\delta$, so that $\|h_{mn}-h\|_q \to 0$ as m, $n \to \infty$. For every f in l^p we have:

$$\left\|\left(L-\sum_{k=-m}^{n} h(k)T^k\right)f\right\|_r=\|f*h-f*h_{mn}\|_r \leqq \|h-h_{mn}\|_q\|f\|_p.$$

Therefore, $\lim_{m,n\to\infty}\|L-\Sigma_{k=-m}^n h(k)T^k\|_{p,r}=0$, i.e., $L=\Sigma_{-\infty}^{\infty}h(k)T^k$ in $\mathrm{TI}(l^p,l^r)$. This expression of L is unique, since $\lim_{m,n\to\infty}\|\Sigma_{k=-m}^n a_kT^k\|_{p,r}=0$ implies

$$\lim_{m,n\to\infty}\left\|\sum_{k=-m}^{n} a_kT^k\delta\right\|_r=\left(\sum_{-\infty}^{\infty}|a_k|^r\right)^{1/r}=0$$

which in turn implies $a_k = 0$ for all integer k. To prove that $\mathscr{T}$is a basis for $\mathrm{TI}(l^p, l^\infty)$ it suffices to observe that $h \in l^{p'}$ for each L in this space, and therefore $r = \infty$.

Remark 5. When r is finite we can not assert that $\mathscr{T}$is a basis for $\mathrm{TI}(l^p, l^r)$, since there are operators L for which h does not belong to l^q (and so Young's inequality does not hold). For example, if $p > 1$ and $h(k) = 2/(2k+1)\pi$ for each k, then the corresponding operator L belongs to $\mathrm{TI}(l^p, l^p)$ [12, p. 321]. However, h does not belong to l^1. For operators from l^2 to l^2 there is a very precise result relying upon the Fourier transform $\hat{h}$ of h, i.e., the function in $L^2[0, 2\pi]$ whose kth Fourier coefficient is $h(k)$. It asserts that *an operator L in* $\mathrm{TI}(l^2, l^\infty)$ *belongs to* $\mathrm{TI}(l^2, l^2)$ *if and only if* $\hat{h} \in L^\infty[0, 2\pi]$–i.e., if $\hat{h}$ is essentially bounded—*in which case* $\|L\|_{2,2} = \|\hat{h}\|_\infty$; *$L$ is injective if and only if $\hat{h}$ vanishes in no subset of* $[0, 2\pi]$ *of positive measure*; *and for every $g \in l^2$ there is a unique $f \in l^2$ such that $Lf = g$ if and only if* $1/\hat{h} \in L^\infty[0, 2\pi]$. Most of this result is proved in [14, p. 168]. In particular, when the "transfer function" $\hat{h}$ of the filter is continuous and never zero the filter is bijective from l^2 onto itself and the input can be recovered from the output using the known relation between the convolution of two functions and the product of their Fourier transforms.

For completeness, we next state without proof a theorem about filters with inputs from c_0, whose intersect stems from the fact that c_0 is the smallest closed translation-invariant subspace of l^∞ containing the sequence δ.

THEOREM 4. *Let L be an operator from c_0 to l^∞ and $h = L\delta$. Then L belongs to* $\mathrm{TI}(c_0, l^\infty)$ *if and only if $h \in l'$ and $Lf = f * h$ for all f in c_0. Furthermore, $\mathscr{T} = \{T^k\colon k \in Z\}$ is a Banach basis for* $\mathrm{TI}(c_0, l^\infty)$ *and* $\|L\| = \|h\|_1$.

The proof follows almost exactly the lines of Theorem 4, taking into account that the conjugate space c_0^* of c_0 is isometrically isomorphic to l^1 [10, p. 201]. It is also worthy of note that $\mathrm{TI}(c_0, l^\infty) = \mathrm{TI}(c_0, c_0)$ and that $\mathrm{TI}(l^p, l^\infty) = \mathrm{TI}(l^p, c_0)$, whenever $p > 1$ [14, p. 331], [6, p. 295].

3. Operators not representable as a convolution with a sequence. Right at the beginning of this paper we posed the question of whether or not a filter would always be of "convolution form". We now show that this question gets a negative answer if we use the normed topology for l^∞.

3.1. Preliminary counterexamples. The simplest extension of c_0 is the space $c = c_0 \oplus (v)$, where (v) is the linear manifold spanned by v, i.e., the constant unit sequence (each coordinate equals 1). The space c may also be extended to the space $cc = c \oplus (u)$, where u is the unit-step sequence ($u(n) = 0$ *if* $n < 0$, $u(n) = 1$ *if* $n \geqq 0$). Thus, cc is the space of all sequences f for which $\lim_{k\to\infty} f(k) = \eta$ and $\lim_{k\to-\infty} f(k) = \zeta$ exist (and are finite), and c is the subspace for which $\eta = \zeta$. For both the spaces c and cc (which are closed within l^∞ normed in the usual way) the question of filter representation has an easy answer, which we summarize in the following theorems.

THEOREM 5. *Let L be an operator from c to l^∞ and $h = L\delta$. Then L belongs to* $\mathrm{TI}(c, l^\infty)$ *if and only if $h \in l^1$ and there exists a number A such that*

$$\forall f \in c, \quad Lf = A \cdot \lim_{|n|\to\infty} f(n) \cdot v + f * h.$$

For any operator of this form $\|L\| = \|h\|_1 + |A|$, *and* $L \in \mathrm{TI}(c, c)$.

Proof. We can define an isometrical isomorphism between c^* and l^1 by assigning to each γ in c^* a sequence $h \in l^1$ and a number A such that [1, p. 66]

$$\forall f \in c, \quad \gamma(f) = A \cdot \lim_{|n|\to\infty} f(n) + \sum_{-\infty}^{\infty} h(k) f(-k)$$

and then $\|\gamma\|=\|h\|_1+|A|$. Now, an application of Lemma 1 proves the first two assertions of Theorem 5, and it only remains to prove that $f*h\in c$, whenever $f\in c$ and $h\in l^1$. If $\lim_{|n|\to\infty}f(n)=\eta$ the sequence $f-\eta v$ belongs to c_0. By putting $\alpha=\eta\Sigma_{-\infty}^{\infty}h(k)$ we get

$$(f-\eta v)*h=f*h-\eta(v*h)=f*h-\alpha v\in c_0.$$

From here it is clear that $f*h\in c$.

THEOREM 6. *Let L be an operator from cc to l^∞ and $h=L\delta$. Then L belongs to* $\mathrm{TI}(cc,l^\infty)$ *if and only if $h\in l^1$ and there exists two numbers A,B such that*

$$\forall f\in cc,\quad Lf=\left(A\cdot\lim_{n\to\infty}f(n)+B\cdot\lim_{n\to-\infty}f(n)\right)\cdot v+f*h.$$

For any operator of this form $\|L\|=\|h\|_1+|A|+|B|$ and $L\in\mathrm{TI}(cc,cc)$.

Proof. It can be easily seen that, as in the preceding proof, we can define an isometrical isomorphism between cc^* and l^1 by assigning to each γ in cc^* a sequence $h\in l^1$ and two numbers A,B such that $\|\gamma\|=\|h\|_1+|A|+|B|$ and

$$\forall f\in cc,\quad \gamma(f)=A\cdot\lim_{n\to\infty}f(n)+B\cdot\lim_{n\to-\infty}f(n)+\sum_{-\infty}^{\infty}h(k)f(-k).$$

By applying Lemma 1, we get .ne first two assertions. To prove that $f*h\in cc$ if $f\in cc$ and $h\in l^1$, let us set $\lim_{k\to\infty}f(k)=\eta$ and $\lim_{k\to-\infty}f(k)=\zeta$, so that $g=f-\zeta v-(\eta-\zeta)u\in c_0$ and $g*h\in c_0$. Then $(f*h)(n)=(g*h)(n)+\zeta\Sigma_{-\infty}^{\infty}h(k)+(\eta-\zeta)\Sigma_{-\infty}^{n}h(k)$ for every $n\in Z$, and $(f*h)(n)$ approaches $\eta\Sigma_{-\infty}^{\infty}h(k)$ as $n\to\infty$ and $\zeta\Sigma_{-\infty}^{\infty}h(k)$ as $n\to-\infty$. Therefore $f*h\in cc$.

Remark 6. These two theorems show us that any operator belonging to $\mathrm{TI}(c_0,l^\infty)$ can be extended to c in infinitely many ways. This surprising result is due to the fact that the restriction of any L from $\mathrm{TI}(c,l^\infty)$ to both c_0 and to its closed complement (v) in c are not related by continuity, for both spaces are closed and translation-invariant and therefore we can get a c.t.i.l. operator on c by combining only two c.t.i.l. operators, one defined on c_0 and the other on (v). Thus, the sole constraint on extensions of $L\in\mathrm{TI}(c_0,l^\infty)$ to (v) stems from the equality $Tv=v$. This implies $T(Lv)=Lv$, and so $Lv=Kv$ for some constant K.

The extension of an operator L on c to an operator $\bar{L}$ on cc can be viewed in a similar fashion. Equality $u-Tu=\delta$ holds, and we must have $\bar{L}u-T(\bar{L}u)=h$, whence $\bar{L}u=Kv+u*h$ for some constant K. Again the continuity has not played any role whatsoever.

The preceding considerations could also be used to get direct proofs of Theorems 5 and 6.

3.2. Operators defined on the normed space l^∞. We now come to our final goal in studying the form of c.t.i.l. operators on l^∞. These operators appear important to us if we assume that any filtering theory that intends to be complete should pay some attention to such a standard signals as the (sampled) sinusoids and, by extension, to the class of almost periodic signals, which contains all those signals resulting from the discretization of periodic analogical signals.

The conjugate space of the normed space l^∞ is isometrically isomorphic to $ba(Z)$ [3, p. 296], the Banach space of all bounded finitely additive measures on the algebra $\mathscr{P}(Z)$ of all subsets of Z. The norm of an element μ in $ba(Z)$ is given by $\|\mu\|=\sup\Sigma_{i=1}^{n}|\mu(E_i)|$, where the supremum is taken over all partitions $\{E_i\}_{1\leqq i\leqq n}$ of Z. The isomorphism can be defined by mapping each γ in $(l^{\infty})^{*}$ to the element μ from $ba(Z)$ given by

$$\forall E \subset Z, \quad \mu(E) = \gamma(\chi_{-E}),$$

where $-E = \{n \in Z\colon\ -n \in E\}$ and χ is the characteristic function. If we call $\tilde{f}$ the sequence symmetric to f, the above relation is equivalent to

$$\forall f \in l^{\infty}, \quad \gamma(f) = \int_{Z} \tilde{f}\, d\mu$$

the integral being understood in a generalized Riemann–Stieltjes or Lebesgue sense [5, p. 869], [10, p. 401] or, equivalently, as a limit of finite sums corresponding to linear combinations of characteristic functions approximating f in l^{∞} [8, p. 425]. As customary, we call the convolution of $f \in l^{\infty}$ and $\mu \in ba(Z)$ the sequence given by $(f * \mu)(n) = \int_{Z} T^{n}\tilde{f}\, d\mu$ for all $n \in Z$.

THEOREM 7. *Let L be an operator from l^{∞} to l^{∞}, $h = L\delta$ and μ be the set function on $\mathscr{P}(Z)$ defined by $\mu(E) = (L\chi_{-E})(0)$ for every $E \subset Z$. Then L belongs to* $\mathrm{TI}(l^{\infty}, l^{\infty})$ *if and only if $\mu \in ba(Z)$ and $Lf = f * \mu$ for all f in l^{∞}. Furthermore, in this case $\|L\|_{\infty,\infty} = \|\mu\|$, $h \in l^{1}$ and $h(n) = \mu(\{n\})$ for all $n \in Z$.*

Proof. We can define an isometrical isomorphism between $(l^{\infty})^{*}$ and $ba(Z)$ as we have just described. Given L in $\mathrm{TI}(l^{\infty}, l^{\infty})$, we set $\gamma_{L}(f) = (Lf)(0)$ for every $f \in l^{\infty}$ as in Lemma 1 so that $(Lf)(n) = \gamma_{L}(T^{-n}f)$ for all $n \in Z$. The corresponding measure μ_{L} is defined by

$$\forall E \subset Z, \quad \mu_{L}(E) = (L\chi_{-E})(0)$$

or, equivalently,

$$\forall f \in l^{\infty}, \forall n \in Z, \quad (Lf)(n) = \int_{Z} (T^{-n}f)^{\tilde{}}\, d\mu_{L} = (f * \mu_{L})(n).$$

Then $Lf = f * \mu_{L}$ as desired, and also $\|L\|_{\infty,\infty} = \|\gamma_{L}\| = \|\mu_{L}\|$. Since the restriction of L to c_{0} belongs to $\mathrm{TI}(c_{0}, l^{\infty})$, Theorem 4 shows that $h \in l^{1}$. Finally, $\chi_{-\{n\}} = \delta_{-n}$ implies

$$\forall n \in Z, \quad \mu_{L}(\{n\}) = (L\delta_{-n})(0) = h(n).$$

This completes the proof.

Remark 7. Engineers are primarily concerned with causal systems, which are defined as those which yield a causal output –i.e., are null on negative integers—for a causal input. If L is a translation-invariant linear operator from λ to χ and $T(\lambda) = \lambda$, $T(\chi) = \chi$, it can easily be seen that L is causal if and only if, given any integer m, $f \in \lambda$ and $f(n) = 0$ for $n < m$ imply $(Lf)(n) = 0$ for $n < m$. This property is adopted as a definition when we turn to consider operators L whose inputs and outputs are causal sequences. If $\delta \in \lambda$ and $Lf = f * h$ for all f in λ, then L is causal if and only if the impulse-response h is causal. When L can not be represented by a convolution with a (fixed) sequence this condition may not be sufficient for causality. In particular, an operator L in $\mathrm{TI}(l^{\infty}, l^{\infty})$ is causal if and only if its associated measure μ vanishes on all subsets of the set Z^{-} of negative integers.

Any translation-invariant linear operator from λ_{N} to χ_{N} is causal—in the sense defined above—when λ_{N} and χ_{N} are linear subspaces of the space ω_{N} of all causal sequences and $T(\lambda_{N}) \subset \lambda_{N}$, $T(\chi_{N}) \subset \chi_{N}$. Theorems 1–4 remain valid with a few unessential changes if we substitute for each space its intersection with ω_{N}. Moreover, when

λ and χ are linear subspaces of ω such that $T(\lambda)=\lambda$, $T(\chi)=\chi$, any translation-invariant linear operator L from $\lambda\cap\omega_N$ to $\chi\cap\omega_N$ verifies $Lf=f*h$ for every $f\in\lambda\cap\omega_N$, where h is a sequence in ω_N uniquely determined by L. For the particular cases corresponding to Theorems 1–7 L is continuous only if h is a causal signal belonging in each case to the space referred to in the theorem. This does not contradict Theorem 7, since $\int_Z T^n\tilde{f}d\mu=(f*h)(n)$ for each $n\in Z$ when f and μ vanish on Z^-. Operators on spaces of causal signals are treated in (11).

Any sequence h in l^1 defines, through the formula $Lf=f*h$, an operator L in $\mathrm{TI}(l^\infty,l^\infty)$, as is obvious from the inequality $\|f*h\|_\infty\leqq\|f\|_\infty\|h\|_1$. The corresponding measure $\mu\in ba(Z)$, defined by $\mu(E)=(\chi_{-E}*h)(0)=\sum_{k\in E}h(k)$ for every $E\subset Z$, is countably additive. In fact, since h is an absolutely convergent series, if E is the union of countably many pairwise disjoint subsets E_i of Z we have:

$$\mu(E)=\sum_{k\in E}h(k)=\sum_i\sum_{k\in E_i}h(k)=\sum_i\mu(E_i).$$

Conversely, the result which follows points out that, if μ is a countably additive bounded measure on $\mathscr{P}(Z)$, the corresponding operator L can be expressed by means of the convolution with h.

THEOREM 8. *Let L be an operator from l^∞ to l^∞, $h=L\delta$ and μ be the set function on $\mathscr{P}(Z)$ defined by $\mu(E)=(L\chi_{-E})(0)$ for every $E\subset Z$. Then $h\in l^1$ and $Lf=f*h$ for all f in l^∞, if and only if $L\in\mathrm{TI}(l^\infty,l^\infty)$ and μ is countably additive. Moreover, in this case* $\lim_{m,n\to\infty}\|L-\sum_{k=-m}^n h(k)T^k\|_{\infty,\infty}=0$.

Proof. Let us suppose that $L\in\mathrm{TI}(l^\infty,l^\infty)$ and μ is countably additive. According to Theorem 7, we can write, for every $E\subset Z$

$$(L\chi_E)(0)=\mu(-E)=\sum_{k\in -E}\mu(\{k\})=\sum_{k\in -E}h(k)=\sum_{-\infty}^{\infty}\chi_E(-k)h(k).$$

We consider now the functional γ_L defined in Lemma 1 and the functional $\gamma\in(l^\infty)^*$ given by $\gamma(f)=\sum_{-\infty}^{\infty}f(-k)h(k)$ for every f in l^∞. It is clear from the above equality that $\gamma_L(g)=\gamma(g)$ for any g which is a linear combination of characteristic functions of subsets of Z. Now, the set of all these linear combinations is norm-dense in l^∞ [8, p. 425], so that $\gamma_L=\gamma$. Therefore,

$$(Lf)(n)=\gamma(T^{-n}f)=\sum_{k=-\infty}^{\infty}f(n-k)h(k)$$

for all $f\in l^\infty$ and $n\in Z$, which concludes the proof of the "if" part of the theorem. The "only if" part was proved before. An argument analogous to that used to prove Theorem 3, proves the last assertion.

Remark 8. Theorem 8 reveals the difference between operators in $\mathrm{TI}(l^\infty,l^\infty)$ expressible as a convolution with a sequence (operators that we will say are of type L_h) and operators which are not (which we call of type L_μ). The behaviour of any operator of type L_h out of the norm-closed space c_0 is determined by the behaviour on c_0 (represented by the "impulse-response" $L_h\delta$), and all sequences, whether or not they belong to c_0, are "translated" by the operator in the same way. This does not happen for operators of type L_μ, whose associated measures behave on some subsets of Z independently of how they behave in the one-point sets. In this case we can assert that, for any f in l^∞,

$$Lf = \lim_{\Pi} \sum_{i=1}^{p} \mu(E_i) T^{r_i} f,$$

where $\Pi = \{E_i\}_{1 \leqq i \leqq p}$ is a partition of Z, $r_i \in E_i$, and the limit is taken in ω following the well-known ordering of partitions of Z. We can say that $L = \int_Z T^s \, d\mu(s)$ in the sense of the strong operator topology of $B(l^\infty, \omega)$, while the integration can be interpreted in $\|\cdot\|_{\infty,\infty}$ for an operator L of type L_h.

Remark 9. Unfortunately, no finitely but noncountably additive measure on $\mathscr{P}(Z)$ can be defined in a constructive way. Indeed, such a measure can be extended to a Radon measure on the Stone compactification βZ of Z, and this Radon measure does not vanish on some subset of $\beta Z \sim Z$, each of whose points requires Zorn's lemma to be defined [2, p. 38].

Remark 10. The result expressed by Theorem 7 is partially contained in a very general and abstract theorem of multiplier theory [9, p. 147]. This latter result is stated within the framework of a general locally compact group, for which reason it is, nevertheless, less clearly defined than the result here obtained for the particular group Z.

3.3. A significant example. In spite of Remark 9 above we have a large number of c.t.i.l. operators, not representable as a convolution with a sequence, that can be defined in a constructive way on subspaces of l^∞ wide enough to allow the development of a quite complete filtering theory within them, and of course wide enough for any practical purpose. To be precise, let c_m consist of all sequences f in l^∞ for which $\lim_{n\to\infty}(1/(2n+1))\Sigma_{k=-n}^{n} f(k)$ exists. This subspace c_m is closed within l^∞, and contains the nonseparable space $ap \oplus c_0$, where $ap = ap(Z)$ stands for the space of almost periodic bilateral sequences.

Now define L as follows:

$$\forall f \in c_m, \quad Lf = \lim_{n\to\infty} \frac{1}{2n+1} \sum_{k=-n}^{n} f(k) \cdot v.$$

Obviously, $L \in \mathrm{TI}(c_m, c_m)$. However, $Lf = 0$ for every $f \in c_0$ —and so $L\delta = 0$—which proves that L is not representable as a convolution with a sequence.

The algebra R_d of subsets of Z on which the corresponding measure μ is defined consists of all $E \subset Z$ for which there exists the so-called density of E, $d(E) = \lim_{n\to\infty}(\mathrm{card}([-n,n] \cap E)/(2n+1))$. Then $\mu(E) = (L\chi_{-E})(0) = d(E)$.

In a forthcoming paper about periodic signals, some other examples of nonconvolution operators will be presented and studied.

4. A natural filtering topology on l^∞. We now turn back to the arguments in Remark 8. From an intuitive viewpoint it appears that the continuity of any operator of type L_h is in some heuristic sense stronger than that of any operator of type L_μ. The behaviour of the latter on c_0 does not completely determine its behaviour on the remainder of the space l^∞, although there exists some interdependence due to the nonexistence of a closed complement of c_0 in l^∞ [8, p. 426], and also to the translation-invariance of L. When we define Lf for $f \notin c_0$, we also define $L(T^n f)$, and we can not do it in a free way if the linear manifold generated by $\{T^n f \colon n \in Z\}$ contains a sequence for which L has already been defined. Anyhow it seems natural to search for an adequate topology in l^∞ for which the only translation-invariant linear operators that will remain continuous will be those corresponding to countably additive measures. A desirable property of this topology is that it be strictly stronger than the pointwise convergence topology, a condition certainly fulfilled by the weak and by the weak* topologies on l^∞. As to the first, no operator in $\mathrm{TI}(l^\infty, l^\infty)$ fails to be continuous if we consider l^∞ endowed with it [3, p. 422]. On the other hand, the weak* topology is adequate for our purposes, as the following theorem expresses:

THEOREM 9. *Let L be an operator from l^∞ to ω and $h=L\delta$. If we consider l^∞ endowed with the weak* topology, L is a c.t.i.l. operator if and only if $h\in l^1$ and $Lf=f*h$ for all f in l^∞. In this case L is also a c.t.i.l. operator from l^∞ to l^∞.*

Proof. A collection of seminorms generating the weak* topology on l^∞ is $\{p_\alpha: \alpha\in l^1\}$, where $p_\alpha(f)=|\Sigma_{-\infty}^{\infty}f(k)\alpha(k)|=|(f*\tilde{\alpha})(0)|$ for all f in l^∞. If $h\in l^1$ and $Lf=f*h$ for every f in l^∞, given $\alpha\in l^1$ we can write:

$$p_\alpha(Lf)=|((f*h)*\tilde{\alpha})(0)|=|(f*(h*\tilde{\alpha}))(0)|=p_\beta(f),$$

where $\beta=\tilde{h}*\alpha$. Therefore L is continuous from l^∞ to l^∞, and *a fortiori* to ω. To prove the converse, let us observe that, for all $f\in l^\infty$ and $\alpha\in l^1$

$$\sum_{-\infty}^{\infty} f(k)\alpha(k)=\lim_{m,n\to\infty}\sum_{j=-m}^{n} f(j)\alpha(j)=\lim_{m,n\to\infty}\sum_{k=-\infty}^{\infty}\alpha(k)\sum_{j=-m}^{n} f(j)\delta_j(k)$$

and, therefore, $f=\lim_{m,n\to\infty}\Sigma_{j=-m}^{n}f(j)T^j\delta$ in the weak* topology. If L is a c.t.i.l. operator from l^∞ to ω,

$$Lf=\lim_{m,n\to\infty}\sum_{j=-m}^{n} f(j)(LT^j\delta)=\lim_{m,n\to\infty}\sum_{j=-m}^{n} f(j)T^jh$$

in ω. Then $(Lf)(k)=(f*h)(k)$ for every $k\in Z$. A simple argument shows that this implies $h\in l^1$.

Within the framework of multiplier theory an abstract result is proved [9, p. 74] which is not difficult to interpret as yielding Theorem 9 except for inessential details.

Remark 11. Note that l^∞ with the weak* topology is not an F-space, and so does not fulfill the requirements of the second part of Lemma 1 for the input space. This is the reason why we have chosen ω as the output space in the statement of Theorem 9.

Remark 12. To get deeper insight into the physical meaning of the weak* topology, we may note that any f in l^∞ is, as we have already written, the limit of its "sections" $\Sigma_{j=-m}^{n}f(j)T^j\delta$. This seems to make this topology better reflect physical reality, since infinite signals are not actually observable, but only conceived of as the limits of finite signals, which does not imply uniform convergence.

Remark 13. To underline the arguments of Remark 12, consider the operator L on the space c_m defined in §3.3. For each $\alpha\in[0,2\pi)$ let $e_\alpha(k)=e^{i\alpha k}$ for every $k\in Z$. Then $Le_\alpha=0$ if $\alpha\neq 0$ and $Le_0=v$, so that a little change in the value of α yields a great change in the corresponding output. This does not contradict the normed continuity of L, as this small variation in α deeply changes the norm of $e_\alpha-e_0$. However, in practical situations we only have finite registers of signals, and it is not possible to exactly determine the value of α. When we consider signals belonging to the space p of periodic sequences, this fact may be particulary important, because the collection of the e_α for α rational is an algebraic basis of p, and then a small error in the determination of the coordinate of an $f\in p$ with respect to e_0 may yield a large error in the determination of the output. Thus, it does not seem convenient to consider such a system L continuous. This is another reason why the weak* topology is a natural topology for filtering.

Acknowledgment. The author wishes to express his gratitude to Professor E. Garbayo, of the Politechnical University of Madrid, for his advice, constructive suggestions and criticism throughout the realization of this work, which has been possible with his encouragement.

REFERENCES

[1] S. BANACH, *Opérations linéaires*, Chelsea, New York, 1955.

[2] R. E. CHANDLER, *Hausdorff Compactifications*, Marcel Dekker, New York and Basel, 1965.

[3] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part* I, Interscience, New York, 1958.

[4] G. H. HARDY, J. E. LITTLEWOOD AND G. POLYA, *Inequalities*, Cambridge Univ. Press, Cambridge, 1952.

[5] T. H. HILDEBRANDT, *On bounded linear functional operations*, Trans. Amer. Math. Soc., 36 (1934), pp. 868–875.

[6] E. HEWITT AND K. A. ROSS, *Abstract Harmonic Analysis*, Springer, Berlin-Heidelberg-New York, 1963.

[7] R. HOLMES, *Mathematical foundations of signal processing*, SIAM Rev., 21 (1979), pp. 361–388.

[8] G. KÖTHE, *Topological Vector Spaces* I, Springer, New York, 1969.

[9] R. LARSEN, *An Introduction to the Theory of Multipliers*, Springer, Berlin-Heidelberg-New York, 1971.

[10] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1958.

[11] B. L. D. THORP, *Operators which commute with translations*, J. London Math. Soc., 39 (1964), pp. 359–369.

[12] E. C. TITCHMARSH, *Reciprocal formulae involving series and integrals*, Math. Z., 25 (1926), pp. 321–347.

[13] K. YOSIDA AND E. HEWITT, *Finitely additive measures*, Trans. Amer. Math. Soc., 72 (1952), pp. 46–66.

[14] A. ZYGMUND, *Trigonometrical Series*, Vol. I, Cambridge Univ. Press, Cambridge, 1968.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

005

STRONG RESOLVENT CONVERGENCE OF DIFFUSION OPERATORS*

HAROLD E. BENZINGER[†]

Abstract. It is shown that differential operators arising from boundary value problems with eigenvalue parameter in the boundary condition occur as the limits, in the sense of a generalized notion of strong resolvent convergence, of families of Sturm–Liouville operators modeling heat flow in a rod, where the diffusion coefficient becomes arbitrarily large in half of the rod, thus modeling a mixing-diffusion problem. The generalized notion of strong resolvent convergence is defined, and a development is given in the setting of the abstract theory of self-adjoint operators in a Hilbert space and the theory of semigroups.

Key words. differential operators, eigenvalue parameter in boundary condition, semigroups, strong resolvent convergence

1. Introduction. In this paper it is shown that the differential operator modeling a problem involving both mixing and diffusion in a composite rod is the limit, in the sense of a suitably generalized notion of strong resolvent convergence, of a directed set of differential operators modeling pure diffusion problems. Regarding these differential operators as the infinitesimal generators of strongly continuous semigroups which describe the time evolution of the physical systems, we then show how the time evolution of the mixing-diffusion problem approximates (or, conversely, is approximated by) the time evolution of a pure diffusion problem.

Consider a rod occupying the interval $-1 \leqq x \leqq 1$. If heat energy is subject to ordinary diffusion in the portion $-1 \leqq x \leqq 0$, and is instantaneously mixed in the portion $0 \leqq x \leqq 1$, so that the temperature in this interval is independent of position, then the resulting differential operator (which will be described in detail in §2) has the eigenvalue parameter appearing in the boundary conditions. A method for reformulating such a problem as a self-adjoint operator was given by J. Walter in [5]. We show that this formulation arises naturally as the limit of a directed set of standard diffusion problems, where the diffusion coefficient in $0 \leqq x \leqq 1$ becomes large without bound.

In §2 we give precise formulations of the differential operators modeling the diffusion and mixing-diffusion problems, and we show that these operators are semibounded. In §3 we present expressions for the resolvent operators, eigenvalues and eigenfunctions of the differential operators, and in §4 we examine the behavior of these objects, for the diffusion operators, as the diffusion coefficient in $0 \leqq x \leqq 1$ becomes arbitrarily large, showing how they approximate the related objects for the mixing-diffusion problem. Strong resolvent convergence of a family of self-adjoint operators is defined in [6, §9.3]. See also [4, p. 206]. In §5 we are concerned with abstract spectral theory, giving a generalized definition of strong resolvent convergence and developing related theorems. In §6 we consider the semigroups arising from operators converging in the strong resolvent sense, and interpret these results for the diffusion and mixing-diffusion problems.

*Received by the editors August 2, 1983, and in revised form March 19, 1984.

[†] Department of Mathematics, University of Illinois, 1409 West Green Street, Urbana, Illinois 61801.

2. Formulation of the differential operators. Let $I_0=[-1,0]$, $I_1=[0,1]$. We first consider a pure diffusion problem, described by

$$(2.1\text{a})\qquad \left(p_0^2(x)u_0'\right)'+q_0(x)u_0=\lambda w_0^2(x)u_0+w_0^2(x)f_0(x),\qquad x \text{ in } I_0,$$

$$(2.1\text{b})\qquad \delta^{-1}\left(p_1^2(x)u_1'\right)'+q_1(x)u_1=\lambda w_1^2(x)u_1+w_1^2(x)f_1(x),\qquad x \text{ in } I_1,$$

where $p_j^2\in C(I_j)$, $w_j^2\in C(I_j)$, $q_j\in L^\infty(I_j)$, $f_j\in L^2(I_j;w_j^2)$, $\lambda\in\mathbf{C}$, $\delta>0$, $p_j^2(x)>0$, $w_j^2(x)>0$, and q_j is real. The notation $L^2(I_j;w_j^2)$ refers to the L^2 space on I_j with weight function w_j^2. The parameter $\delta>0$ can be used to control the diffusivity of the portion of the rod in I_1, in particular for δ close to zero, temperature gradients in I_1 will decay more rapidly than those in I_0. The boundary conditions associated with this problem are

$$(2.2\text{a})\qquad u_0(-1)=0,\qquad p_2^2(1)u_1'(1)=0,$$

$$(2.2\text{b})\qquad u_0(0)=u_1(0),\qquad p_0^2(0)u_0'(0)=\delta^{-2}p_1^2(0)u_1'(0).$$

For the mixing-diffusion problem, we give a more detailed discussion of the formulation of the boundary value problem. Let $v_j(x,t)$ denote the temperature in I_j, where t denotes the time. Since heat flow is still controlled by diffusion in I_0, we have

$$(2.3)\qquad w_0^2(x)\frac{\partial v_0(x,t)}{\partial t}=\frac{\partial}{\partial x}\left(p_0^2(x)\frac{\partial v_0}{\partial x}\right)+q_0(x)v_0(x,t),$$

$$(2.4)\qquad v_0(-1,t)=0.$$

In I_1, the assumption of complete mixing means that $v_1(x,t)$ is independent of x, and in fact equals $v_0(1,t)$:

$$(2.5)\qquad v_1(x,t)=v_0(1,t),\qquad x\in I_1.$$

Such an assumption is consistent with the physical nature of the problem only if the end point $x=1$ is perfectly insulated. Thus the heat energy $h_1(t)$ in I_1 can change only through transfers between I_0 and I_1, and leakage across the sides of the rod, $0<x<1$. We have

$$(2.6)\qquad h_1(t)=\int_0^1 w_1^2(x)v_1(x,t)\,dx=v_0(0,t)W_1^2,$$

where $W_1^2=\int_0^1 w_1^2(x)\,dx$. Let $Q_1=\int_0^1 q_1(x)\,dx$. Then

$$\frac{dh_1(t)}{dt}=-p_0^2(0)\frac{\partial v_0(0,t)}{\partial x}+Q_1v_0(0,t),$$

so, using (2.6), we have

$$(2.7)\qquad W_1^2\frac{\partial v_0(0,t)}{\partial t}=-p_0^2(0)\frac{\partial v_0(0,t)}{\partial x}+Q_1v_0(0,t).$$

Equations (2.3), (2.4), (2.7) define the time dependent problem. Using separation of variables: $v_0(x,t)=T(t)u_0(x)$, we are led to

$$(2.8\text{a})\qquad \left(p_0^2(x)u_0'\right)'+q_0(x)u_0=\lambda w_0^2(x)u_0+w_0^2(x)f_0(x),\qquad x \text{ in } I_0,$$

$$(2.8\text{b})\qquad -p_0^2(0)u_0'(0)+Q_iu_0(0)=\lambda W_1^2u_0(0)+W_1^2z,\qquad z \text{ in } \mathbf{C},$$

$$(2.8\text{c})\qquad u_0(-1)=0.$$

Let $H = L^2(I_0; w_0^2) \oplus L^2(I_1, w_1^2)$, and let $V = L^2(I_0; w_0^2) \oplus \mathbb{C}(W_1^2)$, where $\mathbb{C}(W_1^2)$ denotes the complex plane with norm $\|z\| = W_1|z|$. The diffusion problem (2.1), (2.2) determines a self-adjoint operator T_δ on H, with domain $\mathscr{D}(\delta)$ consisting of all (u_0, u_1) in H such that u_j' exists, $p_j^2 u_j'$ is in $AC(I_j)$, $(p_1^2 u_j')'$ is in $L^2(I_j, w_j^2)$, and the boundary conditions are satisfied. For such functions, T_δ is defined by

$$T_\delta \begin{bmatrix} u_0 \\ u_1 \end{bmatrix} = \begin{bmatrix} \dfrac{1}{w_0^2}\left[\left(p_0^2 u_0'\right)' + q_0 u_0\right] \\ \dfrac{1}{w_1^2}\left[\delta^{-2}\left(p_1^2 u_1'\right)' + q_1 u_1\right] \end{bmatrix}, \tag{2.9}$$

and the problem (2.1), (2.2) can be reformulated as

$$T_\delta \begin{bmatrix} u_0 \\ u_1 \end{bmatrix} = \lambda \begin{bmatrix} u_0 \\ u_1 \end{bmatrix} + \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} \tag{2.10}$$

for (f_0, f_1) in H.

The mixing-diffusion problem (2.8) defines a self-adjoint operator T_0 on V with domain $\mathscr{D}(0)$ consisting of all vectors (u_0, z) such that u_0 and $p_0^2 u_0'$ are in $AC(I_0)$, $(p_0^2 u_0')'$ is in $L^2(I_0; w_0^2)$, $z = \lim_{x \to 0^-} u_0(x)$, and $u_0(-1) = 0$. See [5]. Note that the absolute continuity of u_0 and $p_0^2 u_0'$ implies that $u_0(x)$ and $p_0^2(x) u_0'(x)$ have limits as $x \to 0^-$, and since p_0^2 is continuous, $\lim_{x \to 0^-} p_0^2(x) u_0'(x) = p_0^2(0) u_0'(0)$. For such (u_0, z), we have

$$T_0 \begin{bmatrix} u_0 \\ u_0(0) \end{bmatrix} = \begin{bmatrix} \dfrac{1}{w_0^2}\left[\left(p_0^2(x) u_0'(x)\right)' + q_0(x) u_0(x)\right] \\ \dfrac{1}{W_1^2}\left[-p_0^2(0) u_0'(0) + Q_1 u_0(0)\right] \end{bmatrix}, \tag{2.11}$$

and the problem (2.8) becomes

$$T_0 \begin{bmatrix} u_0 \\ u_0(0) \end{bmatrix} = \lambda \begin{bmatrix} u_0 \\ u_0(0) \end{bmatrix} + \begin{bmatrix} f_0 \\ z \end{bmatrix} \tag{2.12}$$

for (f_0, z) in V.

These self-adjoint operators are semibounded. Let

$$Q(x) = \begin{bmatrix} q_0(x) \\ q_1(x) \end{bmatrix}, \qquad w^2(x) = \begin{bmatrix} w_0^2(x) \\ w_1^2(x) \end{bmatrix}.$$

THEOREM 2.13. *For u in $\mathscr{D}(\delta)$,*

$$(T_\delta u, u) \leq \left[\max_{-1 \leq x \leq 1} \frac{Q(x)}{w^2(x)}\right] \|u\|^2, \qquad \delta \geqq 0.$$

Proof. For $\delta > 0$, using integration by parts and the boundary conditions, we have

$$(T_\delta u, u) = -\int_{-1}^{0} p_0^2 |u_0'|^2 - \frac{1}{\delta^2}\int_0^1 p_1^2 |u_1'|^2 + \int_{-1}^{1} Q|u|^2.$$

Since $p_0^2 > 0$, $p_1^2 > 0$, we can eliminate them, obtaining

$$(T_\delta u, u) \leq \int_{-1}^{1} \frac{Q(x)}{w^2(x)} |u(x)|^2 w^2(x)\, dx,$$

from which the result follows. The case that $\delta = 0$ is similar.

3. Resolvent operators, eigenvalues and eigenfunctions. Consider the problem (2.1), (2.2). Let $\lambda = -\rho^2$, and let $\phi_1(x,\rho)$, $\phi_2(x,\rho)$ be a fundamental set of solutions for the homogeneous form of (2.1a), and let $\psi_1(x,\rho,\delta)$, $\psi_2(x,\rho,\delta)$ be a fundamental set of solutions for

$$(p_1^2(x)u_1')' + \delta^2 q_1(x) u_1 = -\rho^2\delta^2 w_1^2(x) u_1, \tag{3.1}$$

which is equivalent to the homogeneous form of (2.1b). We assume these solutions satisfy

$$\begin{bmatrix} \phi_1 & \phi_2 \\ p_0^2\phi_1' & p_0^2\phi_2' \end{bmatrix}(-1,\rho) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\begin{bmatrix} \psi_1 & \psi_2 \\ p_1^2\psi_1' & p_1^2\psi_2' \end{bmatrix}(1,\rho,\delta) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Since ρ, δ appear in the differential equations as entire functions, these fundamental sets are entire functions of ρ, and the fundamental set for (3.1) is an entire function of δ. See [1, p. 37]. It is easily seen that each of these fundamental sets has constant Wronskian ($=1$). Let

$$\Phi(x,y,\rho) = -\phi_1(x,\rho)\phi_2(y,\rho) + \phi_2(x,\rho)\phi_1(y,\rho),$$
$$\Psi(x,y,\rho,\delta) = -\psi_1(x,\rho,\delta)\psi_2(y,\rho,\delta) + \psi_2(x,\rho,\delta)\psi_1(y,\rho,\delta).$$

Using the method of variation of parameters, we see that the general solutions of the nonhomogeneous equations (2.1) are

$$u_0(x,\rho) = A\phi_1(x,\rho) + B\phi_2(x,\rho) + \int_{-1}^{x} \Phi(x,y,\rho) w_0^2(y) f_0(y)\, dy, \tag{3.2}$$

$$u_1(x,\rho,\delta) = C\psi_1(x,\rho,\delta) + D\psi_2(x,\rho,\delta) - \delta^2 \int_{x}^{1} \Psi(x,y,\rho,\delta) w_1^2(y) f_1(y)\, dy. \tag{3.3}$$

It is easily seen that the conditions (2.2) are satisfied if $A = D = 0$ and if B, C satisfy

(3.4)

$$\begin{bmatrix} \phi_2(0,\rho) & -\psi_1(0,\rho,\delta) \\ p_0^2(0)\phi_2'(0,\rho) & -\delta^{-2}p_1^2(0)\psi_1'(0,\rho,\delta) \end{bmatrix} \begin{bmatrix} B \\ C \end{bmatrix}$$
$$= -\begin{bmatrix} \int_{-1}^{0} \Phi(0,y,\rho) w_0^2(y) f_0(y)\, dy + \delta^2 \int_{0}^{1} \Psi(0,y,\rho,\delta) w_1^2(y) f_1(y)\, dy \\ \int_{-1}^{0} p^2(0)\Phi_x(0,y,\rho) w_0^2(y) f_0(y)\, dy + \int_{0}^{1} p_1^2(0)\Psi_x(0,y,\rho,\delta) w_1^2(y) f_1(y)\, dy \end{bmatrix}.$$

The determinant of the coefficient matrix is

$$D(\rho,\delta)=-\delta^{-2}p_1^2(0)\psi_1'(0,\rho,\delta)\phi_2(0,\rho)+p_0^2(0)\phi_2'(0,\rho)\psi_1(0,\rho,\delta). \tag{3.5}$$

For fixed $\delta>0$, the zeros of D give the eigenvalues of T_δ.

For the problem (2.8), we use (3.2), and note that conditions (2.8b, c) are satisfied if $A=0$ and B satisfies

$$\begin{aligned}[t] &\left[-p_0^2(0)\phi_2'(0,\rho)+\left(Q_1+\rho^2W_1^2\right)\phi_2(0,\rho)\right]B\\ &\quad=W_1^2z-\left(Q_1+\rho^2W_1^2\right)\int_{-1}^{0}\Phi(0,y,\rho)w_0^2(y)f_0(y)\,dy\\ &\qquad+\int_{-1}^{0}p_0^2(0)\Phi_x(0,y,\rho)w_0^2(y)f_0(y)\,dy. \end{aligned} \tag{3.6}$$

Let $D_0(\rho)$ denote the coefficient of B in (3.6). The zeros of this function give the eigenvalues of T_0.

For $\delta>0$, the eigenfunctions of T_δ arise, when $D_\delta(\rho)=0$, as the nontrivial solutions of (3.2), (3.3) when $f_0=0, f_1=0$. Thus

$$B\phi_2(0,\rho)=C\psi_1(0,\rho,\delta),$$

so

$$B=E\psi_1(0,\rho,\delta),\qquad C=E\phi_2(0,\rho),$$

and the eigenfunctions are

$$u(x,\rho,\delta)=E\begin{bmatrix}\psi_1(0,\rho,\delta)\phi_2(x,\rho)\\ \phi_2(0,\rho)\psi_1(x,\rho,\delta)\end{bmatrix}, \tag{3.7}$$

where $\|u\|=1$ provided

(3.8)

$$E=\left[|\psi_1(0,\rho,\delta)|^2\int_{-1}^{0}|\phi_2(x,\rho)|^2w_0^2(x)\,dx+|\phi_2(0,\rho)|^2\int_0^1|\psi_1(x,\rho,\delta)|^2w_1^2(x)\,dx\right]^{-1/2}.$$

For $\delta=0$ and ρ a zero of $D_0(\rho)$, we have

$$u(x,\rho,0)=E\begin{bmatrix}\phi_2(x,\rho)\\ \phi_2(0,\rho)\end{bmatrix}, \tag{3.9}$$

with normalization constant

$$E=\left[\int_{-1}^{0}|\phi_2(x,\rho)|^2w_0^2(x)\,dx+W_1^2|\phi_2(0,\rho)|^2\right]^{-1/2}. \tag{3.10}$$

4. Behavior as $\delta\to 0$.

LEMMA 4.1. *The following limits hold unformly for* $0\leqq x\leqq 1$ *and* ρ *in any compact set*:

$$\lim_{\delta\to 0}\begin{bmatrix}\psi_1(x,\rho,\delta) & \psi_2(x,\rho,\delta)\\ \psi_1'(x,\rho,\delta) & p_1^2(x)\psi_2'(x,\rho,\delta)\end{bmatrix}=\begin{bmatrix}1 & -\int_x^1 p_1^{-2}(y)\,dy\\ 0 & 1\end{bmatrix}.$$

Proof. Since ψ_1, ψ_2 are analytic as functions of δ, they converge, as $\delta \to 0$, to the corresponding fundamental sets for $\delta = 0$. For the uniformity with respect to the other parameters, see [1, p. 37].

LEMMA 4.2.

$$\lim_{\delta \to 0} \delta^{-2} p_1^2(0) \psi_1'(0,\rho,\delta) = Q_1 + \rho^2 W_1^2,$$

uniformly for ρ *in any compact set.*

Proof. Using the differential equation (3.1) and the condition $p_1^2(1)\psi_1'(1,\rho,\delta) = 0$, we have

$$\delta^{-2} p_1^2(0) \psi_1'(0,\rho,\delta) = \int_0^1 \left[q_1(y) + \rho^2 w_1^2(y) \right] \psi_1(y,\rho,\delta)\, dy.$$

The result then follows from Lemma 4.1.

LEMMA 4.3.

$$\lim_{\delta \to 0} -D(\rho,\delta) = D_0(\rho),$$

uniformly in each disc $|\rho| \leqq R$.

Proof. This is a direct consequence of the defining formulas and Lemma 4.2.

LEMMA 4.4. *Let* $R > 0$ *be fixed, and let* $\{\rho_k^0\}$, $k = 1, \cdots, N(R)$ *denote the zeros of* $D_0(\rho)$ *in* $|\rho| < R$. *Let* $\varepsilon > 0$ (*but sufficiently small*) *be given. There exists a positive number* $\Delta = \Delta(\varepsilon, R)$ *such that if* $\delta < \Delta$, *then* $D(\rho,\delta)$ *has* $N(R)$ *zeros* $\{\rho_k^\delta\}$ *in* $|\rho| < R$, *and* $|\rho_k^\delta - \rho_k^0| < \varepsilon$.

Proof. Let $\varepsilon > 0$ be small enough so that the finitely many circles $|\rho - \rho_k^0| = \varepsilon$, $k = 1, \cdots, N(R)$ do not intersect the ρ_k^0. Then for some $\alpha > 0$, $|D_0(\rho)| \leqq \alpha$ on these circles. Using Lemma 4.3, we can select δ small enough so that $|D_0(\rho) + D(\rho,\delta)| < \alpha$ in $|\rho| \leqq R$. Thus by Rouché's theorem, the zeros of $-D(\rho,\delta)$ in $|\rho| < R$ lie within the discs $|\rho - \rho_k^0| < \varepsilon$, and multiplicities are preserved.

Next we consider the resolvent operators

$$R_\delta(\lambda) = (\lambda I - T_\delta)^{-1}, \qquad \delta \geqq 0.$$

From (3.2), (3.3), we see that for $\delta > 0$.

$$(4.5) \qquad R_\delta(\lambda) \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = - \begin{bmatrix} B\phi_2(x,\rho) + \int_1^x \Phi(x,y,\rho) w_0^2(y) f_0(y)\, dy \\ C\psi_1(x,\rho,\delta) - \delta^2 \int_x^1 \Psi(x,y,\rho,\delta) w_1^2(y) f_1(y)\, dy \end{bmatrix},$$

where B, C satisfy (3.4). From (3.2),

$$(4.6) \qquad R_0(\lambda) \begin{bmatrix} f_0 \\ z \end{bmatrix} = - \begin{bmatrix} B\phi_2(x,\rho) + \int_{-1}^x \Phi(x,y,\rho) w_0^2(y) f_0(y)\, dy \\ B\phi_2(0,\rho) + \int_{-1}^0 \Phi(0,y,\rho) w_0^2(y) f_0(y)\, dy \end{bmatrix},$$

where B satisfies (3.6).

Let P_1 denote the projection of $L^2(I_1; w_1^2)$ onto $\mathbb{C}(W_1^2)$ given by

$$P_1 f_1 = \frac{1}{W_1^2} \int_0^1 f_1(y) w_1^2(y)\, dy.$$

THEOREM 4.7. *For each* (f_0,f_1) *in* H,

$$\lim_{\delta\to 0} R_\delta(\lambda)\begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = R_0(\lambda)\begin{bmatrix} f_0 \\ P_1 f_1 \end{bmatrix},$$

uniformly for $|\lambda| \leqq R$, $|\lambda - \lambda_k^0| \geqq \varepsilon > 0$, *where convergence is in the norm of* H.

Proof. For $\delta > 0$, let $B_\delta(\rho;f_0,f_0)$, $C_\delta(\rho;f_0,f_1)$ denote the solutions of (3.4), and let $B_0(\rho;f_0,z)$ denote the solution of (3.6). Using (3.4) along with Lemmas 4.1, 4.2, 4.3, we see that

$$\lim_{\delta\to 0} B_\delta(\rho;f_0,f_1) = B_0(\rho;f_0,P_1 f),$$

$$\lim_{\delta\to 0} C_\delta(\rho;f_0,f_1) = B_0(\rho;f_0,P_1 f)\phi_2(0,\rho) + \int_{-1}^{0} \Phi(0,y;\rho) w_0^2(y) f_0(y)\,dy,$$

where the convergence is uniform with respect to ρ in any compact set bounded away from the zeros of $D_0(\rho)$ (since reciprocals of D_0, D_δ are used). Then using (4.5), (4.6), the result is established (along with the stronger result that convergence holds uniformly on $[-1,1]$).

Let $u_{k\delta}$, $k = 1,\cdots,N(R)$, $\delta \geqq 0$, denote the normalized eigenfunctions of T_δ.

THEOREM 4.8.

$$\lim_{\delta\to 0} u_{k\delta} = u_{k0},$$

uniformly for $k = 1,\cdots,N(R)$, *where convergence is in the norm of* H (*and also in the uniform norm on* $[-1,1]$).

Proof. This is a direct consequence of the expressions (3.7)–(3.9), and Lemma 4.1.

5. Strong resolvent convergence. The situation described in Theorem 4.7 is now considered in an abstract setting. Let H be a complex Hilbert space and let V be a closed subspace. Let P denote the orthogonal projection of H onto V. Let $\{T_\delta\}$, $\delta > 0$, denote a family of self-adjoint operators defined in H, and let T_0 denote a self-adjoint operator defined in V.

DEFINITION 5.1. $\{T_\delta\}$ converges to T_0 in the sense of strong resolvent convergence if for some z in $\mathbb{C} - \mathbb{R}$, and each f in H,

$$\lim_{\delta\to 0} R_\delta(z) f = R_0(z) P f,$$

in the norm of H.

For the special case $V = H$, $P = I_H$, this definition and consequences are discussed in [6, §9.3], [4, p. 206].

LEMMA 5.2. *If* $R_\delta(z)f \to R_0(z)Pf$ *for each* f *in* H, *then*

$$R_\delta^k(z) f \to R_0^k(z) P f, \qquad k \geqq 1.$$

Proof. This can be proved by induction on k, using

$$R_0^k(z) P = (R_0(z) P)^k$$

and the uniform boundedness of the resolvent operators:

$$\|R_\delta(z)\| \leqq |\operatorname{Im} z|^{-1}. \tag{5.3}$$

THEOREM 5.4. *If Definition* 5.1 *holds for one nonreal* z_0, *then it holds for all nonreal* z. *Further, the convergence in Definition* 5.1 *holds uniformly with respect to* z *in each disc*

$$|z-z_0| \leqq |\operatorname{Im} z_0| - \alpha, \qquad \alpha > 0.$$

Proof. Using the analyticity of resolvent operators and (5.3), we obtain (modifying [6, p. 294]),

$$\|R_\delta(z)f - R_0(z)Pf\| \leqq \sum_{k=0}^{N} |z_0 - z|^k \|R_\delta^{k+1}(z_0)f - R_0^{k+1}(z_0)Pf\| + 2\sum_{k=N+1}^{\infty} |z_0 - z|^k |\operatorname{Im} z_0|^{-k-1}. \tag{5.5}$$

Given $\varepsilon > 0$ and $|z-z_0| \leqq |\operatorname{Im} z_0| - \alpha$, the second term on the right of (5.5) can be made smaller than ε by making N large. For all such z, the first term can be made smaller than ε by making δ sufficiently small, using Lemma 5.2. This establishes the uniform convergence. The remainder of the proof follows [6, Theorem 9.15].

Remark 5.6. Since for any self-adjoint operator T,

$$\|R(z,T)\| \leqq [\operatorname{dist}(z, \sigma(T))]^{-1},$$

we can modify the proof of Theorem 5.4 to see that if x_0 is a real number in $\rho(T_0)$, and if for $\delta > 0$ sufficiently small, some fixed interval $|x - x_0| < \alpha$ lies in $\rho(T_0)$, then $R_\delta(x_0)f \to R_\delta(x_0)Pf$.

For the case $V = H$, it is proved in [6, Thm. 9.17] that if Definition 5.1 holds, then for any bounded, continuous function u: $\mathbb{R} \to \mathbb{C}$, $u(T)f \to u(T_0)f$. If $V \neq H$, the result $u(T_0)f \to u(T_0)Pf$ is obviously false for all such u, since if $u(x) \equiv 1$, then

$$u(T_\delta) = I_H \neq I_V = u(T_0).$$

For the same reason, Lemma 5.2 does not hold for $k = 0$.

THEOREM 5.7. *Let* u: $\mathbb{R} \to \mathbb{C}$ *be a continuous function such that*

$$\lim_{|x| \to \infty} u(x) = 0. \tag{5.8}$$

Assume Definition 5.1 *is satisfied. Then for each* f *in* H,

$$\lim_{\delta \to 0} u(T_\delta)f = u(T_0)Pf.$$

Proof. As explained in the proof of [6, Thm. 9.17], the set $\mathscr{P}$ of polynomials in $(\pm i - x)^{-1}$ is dense (with respect to the supremum norm) in the set $C(\hat{\mathbb{R}})$ of bounded continuous functions $u(x)$ on $\mathbb{R}$ such that $u(+\infty) = u(-\infty)$. Now for u in $C(\hat{\mathbb{R}})$, v in $\mathscr{P}$, f in H, we have

(5.9)

$$\|u(T_\delta)f - u(T_0)Pf\| \leqq \|u(T_\delta)f - v(T_\delta)f\| + \|v(T_\delta f) - v(T_0)Pf\| + \|v(T_0)Pf - u(T_0)Pf\|.$$

Let $\varepsilon > 0$ be given, and select v such that

$$|u(x) - v(x)| < \varepsilon, \quad \text{all real } x. \tag{5.10}$$

Then by the functional calculus for self-adjoint operators [6, Theorem 7.14],

$$\|u(T_\delta)-v(T_\delta)\|_H<\varepsilon, \qquad \|u(T_0)-v(T_0)\|_V<\varepsilon.$$

Thus (5.9) becomes

$$\|u(T_\delta)f-u(T_0)Pf\|<2\varepsilon\|f\|+\|v(T_\delta)f-v(T_0)Pf\|. \tag{5.11}$$

Since $(\pm i-x)^{-1}\to 0$ as $|x|\to\infty$, we see that $\alpha:=\lim_{x\to\infty}v(x)$ is the constant term in the polynomial v. If $u(\pm\infty)=0$, then from (5.10), $|\alpha|<\varepsilon$, so if $w=v-\alpha$, we see that w is a polynomial in $(\pm i-x)^{-1}$, the constant term is equal to zero, and

$$|u(x)-w(x)|<2\varepsilon.$$

Using w in place of v, we have

$$\|u(T_\delta)f-u(T_0)Pf\|<4\varepsilon\|f\|+\|w(T_\delta)f-w(T_0)Pf\|. \tag{5.12}$$

Using Lemma 5.2, this final term can be made smaller than $\varepsilon\|f\|$ by making δ sufficiently small.

COROLLARY 5.13. *If f is in V and u is in $C(\hat{\mathbb{R}})$ then*

$$\lim_{\delta\to 0}u(T_\delta)f=u(T_0)f.$$

Proof. Since $f=Pf$, we can go from (5.11) to (5.12) without assuming $u(\pm\infty)=0$.

6. Semigroups. The self-adjoint operator T_δ ($\delta\geqq 0$) generates a strongly continuous semigroup $\{U_\delta(t)\}$, $t\geqq 0$ if it is bounded from above. We assume there exists a real number K, and positive number δ_0 such that for $0\leqq\delta\leqq\delta_0$, and all u in the domain of T_δ,

$$(T_\delta u,u)\leqq K\|u\|^2. \tag{6.1}$$

By Theorem 2.13, this holds for the differential operators considered earlier. We note the easily proved identity

$$R(\pm i,tT_\delta)=t^{-1}R(\pm it^{-1},T_\delta). \tag{6.2}$$

THEOREM 6.3. *Let f be in H and let $0<t_1<t_2<\infty$. Then*

$$\lim_{\delta\to 0}U_\delta(t)f=U_0(t)Pf,$$

uniformly for $t_1\leqq t\leqq t_2$. If $K<0$, then convergence holds uniformly for $t_1\leqq t<\infty$.

Proof. Let

$$u(x)=\begin{cases} e^x, & x\leqq K,\\ e^{2K-x}, & x>K.\end{cases}$$

Then $U_\delta(t)=u^t(T_\delta)$. Let $y=x-K$. Then for $t\geqq 0$,

$$u^t(x)=e^{Kt}\begin{cases} e^{yt}, & y\leqq 0\\ \\ e^{-yt}, & y>0\end{cases} \quad := e^{Kt}u_0^t(y).$$

Note that

$$u_0^t(y)=u_0(ty).$$

Let $v_0(y)$ be a polynomial in $(\pm i-y)^{-1}$ such that, for some preassigned $\varepsilon>0$, $|u_0(y)-v_0(y)|<\varepsilon$ for all real y. Note that v_0 can be chosen with constant term equal to zero, and for each $t\geqq 0$,

$$\text{(6.4)}\qquad |u_0^t(y)-v_0(ty)|<\varepsilon.$$

If $v(x)=e^K v_0(y)$, then for all real x,

$$\text{(6.5)}\qquad |u^t(x)-v(tx)|=e^{Kt}|u_0(ty)-v_0(ty)|<\varepsilon e^{Kt}\leqq \varepsilon e^{Kt_2}.$$

Now

$$\text{(6.6)}\qquad U_\delta(t)=e^{Kt}u_0^t(T_\delta),$$

so

$$\begin{aligned}\text{(6.7)}\quad \|U_\delta(t)f-U_0(t)Pf\| &= e^{Kt}\|u_0^t(T_\delta)f-u_0^t(T_0)Pf\|\\ &\leqq e^{Kt_2}\big\{\|u_0^t(T_\delta)f-v_0(tT_\delta)f\|+\|v_0(tT_\delta)f-v_0(tT_0)Pf\|\\ &\qquad +\|v_0(tT_\delta)Pf-u_0^t(T_0)Pf\|\big\}.\end{aligned}$$

Using (6.4) in the first and last terms on the right of (6.8) gives

$$\text{(6.8}'\text{)}\qquad \|U_\delta(t)f-U_0(t)Pf\|\leqq e^{Kt_2}\big\{2\varepsilon\|f\|+\|v_0(tT_\delta)f-v_0(tT_0)Pf\|\big\}.$$

Since $v_0(tT_\delta)$ is a polynomial in $R(\pm i,tT_\delta)$, one can use (6.2) and Theorem 5.4 to see that for $t_1\leqq t\leqq t_2$, we can select δ sufficiently small so that

$$\|v_0(tT_\delta)f-v_0(tT_0)Pf\|<\varepsilon\|f\|.$$

This establishes the first part of the theorem. If $K<0$, we return to (6.7) and note that since $|u_0(y)|\leqq 1$,

$$\|u_0^t(T_\delta)f-u_0^t(T_0)Pf\|\leqq\varepsilon\|f\|,\qquad \delta\geqq 0.$$

Thus given $\varepsilon>0$, we select t_2 large enough so that $t\geqq t_2$ implies

$$e^{Kt}\|u_0^t(T_\delta)f-u_0^t(T_0)Pf\|<\varepsilon\|f\|,\qquad \delta\geqq 0.$$

For $t_1\leqq t\leqq t_2$, we use the first part of the proof.

For f in v, the interval of uniform convergence for the semigroups can contain $t=0$, provided we make the further assumption that any finite interval is contained in a larger interval (a,b) such that for each f in V,

$$\text{(6.9)}\qquad \lim_{\delta\to 0}\int_a^b dE_\delta(x)f=\int_a^b dE_0(x)f.$$

In the case of the differential operators considered earlier, this is certainly true, either by direct computation, or more abstractly, by exploiting the gaps in the spectra, the strong resolvent convergence, and the representation of the integrals in (6.9) as contour integrals.

THEOREM 6.10. *If* (6.9) *holds, and if f is in V, then*

$$\lim_{\delta\to 0} U_\delta(t)f = U_0(t)f,$$

uniformly on each interval $[0,t_1]$. *If $K<0$, convergence holds uniformly for $t\geqq 0$.*

Proof. Let $\varepsilon>0$ be given and let f in V be given. Then

$$f=\int_{-\infty}^{\infty} dE_0(x)f,$$

so there exists an interval (a,b) such that if

$$h_0 := \int_a^b dE_0(x)f,$$

then $\|f-h_0\|<\varepsilon$. Assume (a,b) is as in the assumption, and let

$$h_\delta := \int_a^b dE_\delta(x)f, \qquad \delta\geqq 0.$$

Then there exists $\Delta=\Delta(\varepsilon,\mathrm{f})>0$ such that $0\leqq\delta\leqq\Delta$ implies

$$\|h_\delta - h_0\|<\varepsilon. \tag{6.11}$$

It suffices to consider the last term in (6.8.1) (with $f=Pf$). We have

(6.12)

$$\begin{aligned}&\|v_0(tT_\delta)f - v_0(tT_0)f\|\\ &\quad\leqq \|v_0(tT_\delta)f - v_0(tT_\delta)h_\delta\|\\ &\qquad + \|v_0(tT_0)h_\delta - h_\delta\| + \|h_\delta - h_0\| + \|h_0 - v_0(tT_0)h_0\| + \|v_0(tT_0)h_0 - v_0(tT_0)f\|.\end{aligned}$$

Since $|v_0(y)|\leqq 1+\varepsilon$ and $\|f-h_\delta\|<2\varepsilon$ for $0\leqq\delta\leqq\Delta$, we have

$$\|v_0(tT_\delta)f - v_0(tT_\delta)h_\delta\|<2(1+\varepsilon)\varepsilon, \qquad \delta\geqq 0.$$

This takes care of the first and last terms on the right of (6.12). For the middle term we use (6.11). For the remaining two terms, we note that

$$v_0(tT)h_\delta - h_\delta = \int_a^b (v_0(tx)-1)\,dE_\delta(x)h_\delta.$$

For $a<x<b$, there exists $t(\varepsilon)>0$ so that $0\leqq t\leqq t(\varepsilon)$ implies

$$|v(tx)-1|<\varepsilon.$$

Thus

$$\|v_0(tT_\delta)h_\delta - h_\delta\|<\varepsilon\|h_\delta\|<\varepsilon(\|f\|+\varepsilon),$$

for $0\leqq t\leqq t(\varepsilon)$ and $0\leqq\delta\leqq\Delta$. If $t(\varepsilon)<t_1$, we use the previous theorem on $[t(\varepsilon),t_1]$, with δ possibly still smaller. If $K<0$, then we again use the previous proof to obtain uniform convergence for all $t\geqq 0$.

In the case of the diffusion operators T_δ ($\delta > 0$) and the mixing-diffusion operator T_0, given $f=(f_0,f_1)$ in $H=L^2(I_0,w_0^2)\oplus L^2(I_1,w_1^2)$, we have $V=L^2(I_0,w_0^2)\oplus \mathbf{C}(W_1^2)$ and

$$P\begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = \begin{bmatrix} f_0 \\ P_1 f_1 \end{bmatrix}, \tag{6.13}$$

where P_1 is defined just above Theorem 4.7.

Let $v_\delta(x,t)$ denote the solution of

$$\frac{\partial}{\partial t} v_\delta(x,t) = T_\delta v_\delta(x,t), \qquad v_\delta(x,0)=f, \qquad \delta>0. \tag{6.14}$$

Then

$$v_\delta(x,t) = (U_\delta(t)f)(x). \tag{6.15}$$

Also let

$$\frac{\partial}{\partial t} v_0(x,t) = T_0 v_0(x,t), \qquad v_0(x,0)=Pf. \tag{6.16}$$

Then

$$v_0(x,t) = (U_0(t)Pf)(x). \tag{6.17}$$

Thus the temperature v_δ on any time interval $0<t_1\leqq t\leqq t_2<\infty$ will approach the mixing temperature as $\delta\to 0$, we can select $t_2=\infty$ if the rods are poorly insulated ($K<0$), and we can select $t_1=0$ if $f=(f_0,f_1)$ is already in V, i.e., f_1 is constant.

REFERENCES

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

[2] C. T. Fulton, *Two-point boundary value problems with eigenvalue parameter contained in the boundary conditions*, Proc. Royal Soc. Edinburgh, 77A (1977), pp. 293–308.

[3] D. B. Hinton, *An expansion theorem for an eigenvalue problem with eigenvalue parameter in the boundary condition*, Quart J. Math. Oxford (2), 30 (1979), pp. 33–42.

[4] T. Kato, *Perturbation Theory for Linear Operators*, 2nd edition, Springer-Verlag, Berlin-Heidelberg-New York, 1976.

[5] J. Walter, *Regular eigenvalue problems with an eigenvalue parameter in the boundary condition*, Math. Z., 133 (1973), pp. 301–312.

[6] J. Weidmann, *Lineare Operatoren in Hilberträumen*, B. G. Teubner, Stuttgart, 1976.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

006

ON THE SINGULARITIES OF SINGULAR STURM–LIOUVILLE EXPANSIONS AND AN ASSOCIATED CLASS OF ELLIPTIC P.D.E.'S*

AHMED I. ZAYED[†] AND GILBERT G. WALTER[‡]

Abstract. We consider elliptic equations of the form $u_{xx}+u_{yy}+a(x,y)u_x+b(x,y)u_y+c(x,y)u=0$ that are separable in the polar coordinates (r,θ). Upon separating the variables one obtains an angular equation of the form $d^2w/d\theta^2+(\lambda-p(\theta))w=0$; $0<\theta<2\pi$ where $p(\theta)$ may have a singularity at one of the end points but is otherwise analytic in $0<\operatorname{Re}\theta<2\pi$. This leads to the singular Sturm–Liouville problem $d^2y/dx^2+(\lambda-q(x))y=0$; $0<x<\infty$, $y(0,\lambda)\cos\beta+y'(0,\lambda)\sin\beta=0$ and $|y(\infty,\lambda)|<\infty$. Let $\phi(x,\lambda)$ be the solution of this system with $\phi(0,\lambda)=\sin\beta$, $\phi'(0,\lambda)=\cos\beta$ and for $f(x)\in L^2(0,\infty)$ put $F(\lambda)=\int_0^\infty f(x)\phi(x,\lambda)\,dx$. We show that if $F(\lambda)=O(e^{-c\sqrt{\lambda}})$ as $\lambda\to\infty$, then the singularities of the analytic function $f(t)=\int_{-\infty}^{\infty}F(\lambda)\phi(t,\lambda)\,d\rho(\lambda)$, where $\rho(\lambda)$ is the spectral function, can be located by relating them to the singularities of the associated Laplace transform $g(z)=\int_0^\infty F(s^2)e^{isz}\,ds$, $\lambda=s^2$.

AMS-MOS subject classifications (1980). Primary 34B25; secondary 35B50

1. Introduction. The study of the analytic properties of solutions of partial differential equations as well as the location of their singular points is of great importance in both modern and classical physics. An extensive literature on this subject has been established in the areas of quantum mechanics, quantum field theory and in particular in the theory of potential scattering [13], [18], [20], [21].

Not surprisingly, the first attempt in that direction was to locate the singularities of solutions of the Laplace equation i.e. the singularities of harmonic functions. In his celebrated paper [19], Nehari devised a technique based on what is known as Hadamard's argument to locate the singular points of a harmonic function in the unit disc. On the other hand, Bergman [1], [2] was the first to use integral operators to map holomorphic functions of one or several complex variables onto harmonic functions. Putting Nehari's and Bergman's ideas together, Gilbert in a series of paper [7], [8], [9], and [10] was able to study the singularities of harmonic functions in n variables as well as the singularities of solutions of more general elliptic equations.

Gilbert's technique has also been applied to solutions of certain meta-parabolic and pseudo-parabolic systems [11]. Different approaches to the study of the analytic properties of solutions of partial differential equations have been considered by other people, e.g., Vekua [22] and Garabedian [6].

In this paper we will give a procedure that will enable one to find the singularities of the solutions of the elliptic equations

$$u_{xx}+u_{yy}+a(x,y)u_x+b(x,y)u_y+c(x,y)u=0 \tag{1.1}$$

under the assumption that it can be solved by the separation of variables technique. We assume that (1.1) may be separated in polar coordinates and to this end we require that

* Received by the editors August 9, 1983.

[†] Department of Mathematics, California Polytechnic State University, San Luis Obispo, California 93407.

[‡] Department of Mathematics, University of Wisconsin, Milwaukee, Wisconsin 53201.

the coefficients have the simplified form

$$a(x,y)=\alpha(r)\cos\theta-\frac{\beta(\theta)}{r}\sin\theta,$$
$$b(x,y)=\frac{\beta(\theta)}{r}\cos\theta+\alpha(r)\sin\theta,$$
$$c(x,y)=\frac{\gamma_1(r)+\gamma_2(\theta)}{r^2},$$

where α, β, γ_1 and γ_2 are assumed to be entire and real on the real axis. Transforming (1.1) to polar coordinates gives

$$\frac{1}{r}\frac{\partial}{\partial r}(ru_r)+\frac{1}{r^2}u_{\theta\theta}+\alpha(r)u_r+\frac{1}{r^2}\beta(\theta)u_\theta+\left(\frac{\gamma_1(r)+\gamma_2(\theta)}{r^2}\right)u=0. \tag{1.2}$$

If we put $u(r,\theta)=R(r)\Theta(\theta)$, we obtain

$$R''+\left(\frac{1}{r}+\alpha(r)\right)R'+\left(\frac{\gamma_1-\lambda}{r^2}\right)R=0, \tag{1.3}$$

$$\Theta''+\beta(\theta)\Theta'+(\gamma_2+\lambda)\Theta=0, \tag{1.4}$$

where λ is the separation constant. The substitution

$$\Theta(\theta)=w(\theta)\exp\left(-\frac{1}{2}\int_0^\theta\beta(\theta)\,d\theta\right)$$

reduces (1.4) to

$$\frac{d^2w}{d\theta^2}+(\lambda-q(\theta))w=0, \tag{1.5}$$

where

$$q(\theta)=\frac{-1}{4}\beta^2(\theta)+\frac{1}{2}\beta'(\theta)-\gamma_2(\theta), \qquad 0<\theta<2\pi. \tag{1.6}$$

As for the radial equation (1.3) we put $h(r)=r\alpha(r)$,

$$v(r)=r^{(1+h(0))/2}R(r)\exp\left(\frac{1}{2}\int_0^r\frac{h(r)-h(0)}{r}dr\right)$$

and reduce (1.3) to

$$v''+\left(\frac{g(r)-\lambda}{r^2}+\frac{k(r)}{r}\right)v=0,$$

where

$$g(r)=\frac{1}{4}(1-h^2(r))+\gamma_1(r), \qquad k(r)=-\frac{h'(r)}{2}.$$

A further substitution, $\rho=\log r$, $v=r^{1/2}z$ yields

$$\frac{d^2z}{d\rho^2}-(\lambda+p(\rho))z=0 \tag{1.7}$$

where $p(\rho)=\frac{1}{4}-g(e^{\rho})-e^{\rho}k(e^{\rho})$.

The case where $q(\theta)$ is periodic with period 2π and has an analytic continuation to the entire complex plane with $w'(0)-aw(0)=0$ and $w'(2\pi)+bw(2\pi)=0$, i.e. a regular Sturm–Liouville problem, was considered by Gilbert and Howard in a number of papers [12], [14] and [15]. They found the locations of the singularities of Sturm–Liouville (SL) expansions of the form

$$f(t)=\sum_{n=0}^{\infty}a_n\phi_n(t), \qquad \overline{\lim_{n\to\infty}}\,|a_n|^{1/n}<1,$$

where $\phi_n(t)$ are the normalized eigenfunctions of a regular SL problem, by relating them to the singularities of the associated power series

$$g(z)=\sum_0^{\infty}a_nz^n.$$

Their results have been extended by the authors to the case where $f(t)$ is a generalized function [23], [24], [25], [26], and [27]. Our main aim now is to extend their results to the case where $q(\theta)$ has a singularity at one of the end points of the interval $(0,2\pi)$. More precisely, we shall extend their results to a singular SL problem in which the spectrum is continuous or mixed. In this case the series must be replaced by an integral and the expansion be compared to an associated integral instead of an associated power series. The associated integral will be the Laplace transform of the coefficient function.

The singular points of an analytic function given by a Laplace transform have been extensively studied because of their importance in differential equations. The nature of the stability of the solution to a differential equation is determined by the location of those singularities [5]. In addition, their location is important in the study of entire functions since they determine the asymptotic behavior of entire functions of exponential type [17].

We consider the singular Sturm–Liouville problem with differential equation of the form

$$y''+(\lambda-q(x))y=0, \qquad x\in(0,\infty), \tag{1.8}$$

where $q(x)$ is analytic in the half-plane $\operatorname{Re}x>0$ and in $L^1(0,\infty)$. Each sufficiently nice function $f(x)$ on $(0,\infty)$ has an expansion of the form

$$f(x)=\int_{-\infty}^{\infty}F(\lambda)\phi(x,\lambda)\,d\rho(\lambda) \tag{1.9}$$

where $\phi(x,\lambda)$ are the eigenfunctions of (1.8), ρ is the spectral function and $F(\lambda)$ is the coefficient function. We shall describe these quantities more precisely in the next section. By using "Hadamard's argument" we relate the singularities of $f(x)$ to the singularities of

$$g(z)=\int_0^{\infty}F(\lambda)e^{isz}\,ds, \qquad \lambda=s^2. \tag{1.10}$$

This involves studying the integral operators which transform g into f and f into g and the singularities of their kernels. This will be done in §2 and in §3 we prove the main theorem and complete our investigation by giving some examples.

2. Preliminaries. Consider the singular Sturm–Liouville (SL) problem

$$y''(x,\lambda)+(\lambda-q(x))y(x,\lambda)=0 \tag{2.1}$$

with boundary conditions

$$y(0,\lambda)\cos\alpha+y'(0,\lambda)\sin\alpha=0 \tag{2.2}$$

and

$$|y(\infty,\lambda)|<\infty. \tag{2.3}$$

This problem is regarded as a limiting case of the regular Sturm–Liouville problem given by (2.1), (2.2) and

$$y(b,\lambda)\cos\beta+y'(b,\lambda)\sin\beta=0, \qquad 0<b<\infty \quad \text{as } b\to\infty. \tag{2.4}$$

Let us denote by $\lambda_{n,b}$ the eigenvalues of the regular SL problem and by $y_{n,b}(x)$ the corresponding eigenfunction. Put

$$\begin{aligned}\alpha_{n,b}^2&=\int_0^b y_{n,b}^2(x)\,dx,\\ \rho_b(\lambda)&=-\sum_{\lambda<\lambda_{n,b}\le 0}\frac{1}{\alpha_{n,b}^2}, \qquad \lambda\le 0,\end{aligned} \tag{2.5}$$

and

$$\rho_b(\lambda)=\sum_{0<\lambda_{n,b}\le\lambda}\frac{1}{\alpha_{n,b}^2}, \qquad 0<\lambda. \tag{2.6}$$

Then Parseval's equality takes the form

$$\int_0^b f^2(x)\,dx=\sum_n\frac{1}{\alpha_{n,b}^2}\left(\int_0^b f(x)y_{n,b}(x)\,dx\right)^2=\int_{-\infty}^{\infty}F^2(\lambda)\,d\rho_b(\lambda), \tag{2.7}$$

where

$$F(\lambda)=\int_0^b f(x)y(x,\lambda)\,dx. \tag{2.8}$$

It is known [19] that the sequence $\{\rho_b(\lambda)\}$ converges to a monotonic function $\rho(\lambda)$ as $b\to\infty$. Let $\phi(x)=\phi(x,\lambda)$, $\theta(x)=\theta(x,\lambda)$ be the solutions of (2.1) such that

$$\begin{aligned}\phi(0)&=\sin\alpha, & \phi'(0)&=-\cos\alpha,\\ \theta(0)&=\cos\alpha, & \theta'(0)&=\sin\alpha.\end{aligned} \tag{2.9}$$

Then for $f(x)\in L^2(0,\infty)$ we have

$$\int_0^\infty f^2(x)\,dx=\int_{-\infty}^{\infty}F^2(\lambda)\,d\rho(\lambda), \tag{2.10}$$

where

$$F(\lambda)=\underset{n\to\infty}{\mathrm{l\cdot i\cdot m}}\int_0^n f(x)\phi(x,\lambda)\,dx. \tag{2.11}$$

$F(\lambda)$ is called the generalized Fourier transform of $f(x)$. In addition, if we assume that $f(x)$ is continuous on $[0,\infty)$ and that the integral $\int_{-\infty}^{\infty}F(\lambda)\phi(x,\lambda)\,d\rho(\lambda)$ converges absolutely and uniformly with respect to x on every compact subset of $[0,\infty)$, then

$$f(x)=\int_{-\infty}^{\infty}F(\lambda)\phi(x,\lambda)\,d\rho(\lambda). \tag{2.12}$$

The general solution of (2.1) is of the form $\psi(x,\lambda)=\theta(x,\lambda)+m(\lambda)\phi(x,\lambda)$ where $m(\lambda)$ is analytic in the upper and lower half-planes $\mathrm{Im}\lambda\neq 0$. From now on we assume that $q(z)$ is:

i) analytic in the half-plane $\mathrm{Re}z>0$,

ii) real for real z,

iii) integrable over any line of the form $(c+i\infty,c-i\infty)$; $c>0$ and $(ic,\infty+ic)$.

We may extend $q(x)$ to $\mathrm{Re}z<0$ so that it satisfies similar conditions but otherwise is arbitrary. However, in most cases the even or odd extensions will be considered. Condition (i) implies that $\phi(z,\lambda)$ is analytic in $(\mathrm{Re}z>0)\times\mathbb{C}$ and condition (iii) implies that the spectrum of the SL problem is discrete and bounded from below for $\lambda\leq 0$.

We denote the space of all C^∞-functions in $\mathbb{R}$ with compact support by $\mathcal{D}$ and its dual space by $\mathcal{D}'$. The topology of $\mathcal{D}$ is the standard topology (see [3]). $\mathcal{E}$ will denote the space of all complex valued C^∞-functions on $\mathbb{R}$ and $\mathcal{E}'$ is its dual, i.e., the space of all generalized functions with compact support. For $f(x)\in\mathcal{E}'$ we define its analytic representation by

$$\hat{f}(z)=\frac{1}{2\pi i}\left\langle f(x),\,\frac{1}{x-z}\right\rangle,\qquad z\notin\mathrm{supp}f(x).$$

It is known that

$$\lim_{\varepsilon\to 0}\int_{-\infty}^{\infty}[\hat{f}(x+i\varepsilon)-\hat{f}(x-i\varepsilon)]\,\phi(x)\,dx=\langle f,\phi\rangle$$

for all $\phi\in\mathcal{E}$. For more details on generalized functions see [3].

We devote the remaining part of this section to proving some lemmas that will be needed in the following section. The first lemma gives a bound on the eigenfunctions $\phi(z,\lambda)$.

LEMMA 2.1. *Let* $\lambda=s^2$, $s=\sigma+i\tau$ *and* $z=x+iy$. *Then*

$$|\phi(z,\lambda)|\leq Ae^{|\mathrm{Im}\,sz|}=Ae^{|\tau|x+|y\sigma|}$$

in $(\mathrm{Re}z>0)\times(|s|\geq\delta>0)$.

Proof. The proof is similar to the one for real z [19, p. 206]. Since $\phi(z,\lambda)$ satisfies

$$\phi(z,\lambda)=\sin\alpha\cos(sz)-\cos\alpha\frac{\sin(sz)}{s}+\frac{1}{s}\int_0^z\sin\{s(z-t)\}\,q(t)\phi(t,\lambda)\,dt, \tag{2.13}$$

where the integral is a contour integral in the complex plane. But both $\cos(sz)$ and $\sin(sz)$ are bounded in absolute value by $\exp(|\tau|x+|y\sigma|)$ for $\operatorname{Re} z>0$. Set $\phi_1(z,\lambda)=\phi(z,\lambda)e^{-|\tau|x-|y\sigma|}$, whence we have

(2.14)

$$|\phi_1(z,\lambda)|<1+\frac{1}{|s|}+\frac{1}{|s|}\int_0^z|\sin\{s(z-t)\}|\,|q(t)|\,|\phi_1(t,\lambda)|e^{|\tau|(|t_1|-x)+|\sigma|(|t_2|-|y|)}\,|dt|$$

where $t=t_1+it_2$.

Since the integrand in (2.13) is analytic for $\operatorname{Re} z>0$, we can integrate along any path from 0 to z. We choose the line segments $(0,x)$ and $(x,x+iy)$. On this contour we have $0<t_1<x$ and $|t_2|<|y|$; therefore (2.14) yields

$$|\phi_1(z,\lambda)|\leq 1+\frac{1}{|s|}+\frac{1}{|s|}\int_0^z|q(t)|\,|\phi_1(t,\lambda)|\,|dt|.$$

By [19, Lemma 3.1, Ch. 4] we conclude that $|\phi_1(z,\lambda)|$ is bounded for $\operatorname{Re} z>0$ and $|s|\geq\delta>0$ and this proves the lemma. Q.E.D.

LEMMA 2.2. *Let $f(x)\in L^2[0,\infty)$. If $\int_0^\infty f(x)\phi(x,\lambda)\,dx$ converges uniformly for $\lambda\in(-\infty,\infty)$, then it is the generalized Fourier transform of $f(x)$.*

Proof. Let $F_N(\lambda)$ be given by

$$F_N(\lambda)=\int_0^N f(x)\phi(x,\lambda)\,dx.$$

Then by hypothesis

$$F_N(\lambda)\to G(\lambda)\quad\text{uniformly on bounded sets where } G(\lambda)=\int_0^\infty f(x)\phi(x,\lambda)\,dx.$$

Hence

$$F_N^{(-1)}(\lambda)=\int_0^\lambda F_N(\mu)\,d\rho(\mu)$$

also converges uniformly to $G^{(-1)}(\lambda)$ on bounded sets. But $F_N(\lambda)\to F(\lambda)$ in $L^2(d\rho)$ and therefore

$$|F_N^{(-1)}(\lambda)-F^{(-1)}(\lambda)|\leq\int_0^\lambda|F_N-F|d\rho$$

$$<\left\{\int_0^\lambda d\rho\right\}^{1/2}\left\{\int_0^\infty|F_N-F|^2d\rho\right\}^{1/2}\to 0\quad\text{on bounded sets.}$$

Hence $F^{(-1)}(\lambda)=G^{(-1)}(\lambda)$ and by differentiating with respect to λ, it follows that $F=G$ a.e. $(d\rho)$. Q.E.D.

LEMMA 2.3. *The even (or odd) extension of the integral*

$$\int_{-\infty}^\infty\phi(x,\lambda)\,d\rho(\lambda)$$

converges to a generalized function $\delta_1\in\mathscr{D}'(-\infty,\infty)$ concentrated at the origin, i.e. $\operatorname{supp}\delta_1=\{0\}$.

Proof. Let f be a C^∞-function on $\mathbb{R}^1$ with compact support. Then if $\sin\alpha\neq 0$, we have

$$f(x)=\int_{-\infty}^\infty F(\lambda)\phi(x,\lambda)\,d\rho(\lambda),\qquad s^2=\lambda,\tag{2.15}$$

where $F(\lambda)$ is rapidly decreasing. Hence

$$f(0)=\sin\alpha\int_{-\infty}^{\infty}F(\lambda)\,d\rho(\lambda)$$

which implies that

$$\frac{f(0)}{\sin\alpha}=\int_{-\infty}^{\infty}F(\lambda)\,d\rho=\lim_{M\to\infty}\int_{-\infty}^{M}F(\lambda)\,d\rho=\lim_{M\to\infty}\int_{0}^{\infty}f(x)\int_{-\infty}^{M}\phi(x,\lambda)\,d\rho(\lambda)\,dx.$$

If f has its support in $(0,\infty)$, then this last expression is just $\langle\delta_1,f\rangle$. The same is true if the support is in $(-\infty,0)$ provided the even extension of $\phi(x,\lambda)$ is used. In both cases $f(0)=0$. In general we have therefore

$$\delta_1=0\quad\text{on}\ \mathbb{R}-\{0\}.$$

If $\sin\alpha=0$, we differentiate both sides of (2.15) to get

$$f'(0)=-\cos\alpha\int_{-\infty}^{\infty}F(\lambda)\,d\rho(\lambda)$$

and then repeat the same argument.

COROLLARY 2.1. *The analytic representation of*

$$\delta_1(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\phi(x,\lambda)\,d\rho(\lambda)$$

is holomorphic everywhere in the finite complex plane except possibly at $z=0$.

Throughout the rest of this article we shall assume that the measure $d\rho(\lambda)$ is such that the generalized function

$$\delta_2(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-isx}\,d\rho(\lambda)$$

is also singular only at the origin, i.e., its analytic representation is holomorphic everywhere except possibly at the origin. This holds if $\rho'(s^2)$ is a rational function of s, since its Fourier transform in this case is a linear combination of the Dirac δ-function, its derivatives and functions which are holomorphic everywhere except possibly at the origin. Under this assumption it follows that the analytic representation of

$$\int_{-\infty}^{\infty}p(s)e^{-isx}\,d\rho(\lambda)$$

is holomorphic everywhere except possibly at the origin for every polynomial $p(s)$.

LEMMA 2.4. *Let*

$$K(t,z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\phi(t,s^2)e^{-isz}\,d\rho(\lambda)\tag{2.16}$$

and

$$L(t,z)=\frac{1}{\sqrt{2\pi}}\int_{0}^{\infty}\phi(t,s^2)e^{isz}\,ds.\tag{2.17}$$

Then

a) $K(t,z)$ *defines a holomorphic function in the region*

$$(\operatorname{Re} t>0)\times(\operatorname{Im} z<-|\operatorname{Im} t|).$$

b) *For t and z real* $K(t,z)$ *is a solution to the problem*

$$\frac{\partial^2 K}{\partial z^2}=\frac{\partial^2 K}{\partial t^2}-q(t)K, \qquad 0<t<\infty,$$
$$K(0,z)=\sin\alpha\,\delta_2(z), \qquad \left.\frac{\partial K}{\partial t}\right|_{t=0}=-\cos\alpha\,\delta_2(z),$$
$$K(t,0)=\delta_1(t).$$

c) $L(t,z)$ *defines a holomorphic function in the region*

$$(\operatorname{Re} t>0)\times(|\operatorname{Im} t|<\operatorname{Im} z).$$

d) *For t and z real* $L(t,z)$ *is a solution to the problem*

$$\frac{\partial^2 L}{\partial z^2}=\frac{\partial^2 L}{\partial t^2}-q(t)L, \qquad 0<t<\infty,$$
$$L(0,z)=\sin\alpha\,\delta^+(z), \qquad \left.\frac{\partial L}{\partial t}\right|_{t=0}=-\cos\alpha\,\delta^+(z).$$

Proof. a) Since [4, Lemma 2]

$$\rho'(\lambda)=\begin{cases}O(s) & \text{if } \sin\alpha=0\\ O\left(\dfrac{1}{s}\right) & \text{if } \sin\alpha\neq 0\end{cases} \qquad \text{as } \lambda\to\infty,$$

and by Lemma 2.1 $\phi(t,\lambda)=O(e^{|\operatorname{Im} t|s})$ as $s\to\infty$ it easily follows that the integral in (2.16) converges absolutely in the prescribed region and uniformly in any compact subset thereof.

b)

$$\frac{\partial^2 K}{\partial t^2}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\phi''(t,\lambda)e^{-isz}d\rho(\lambda)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(q(t)-s^2)\phi(t,\lambda)e^{-isz}d\rho(\lambda)$$
$$=q(t)K(t,z)+\frac{\partial^2 K}{\partial z^2}.$$

The boundary conditions can be easily verified. The proof of c) and d) is similar except for the fact that the Fourier transform of the Heaviside function is δ^+.

In the following lemma we continue both $K(t,z)$ and $L(t,z)$ beyond their original domains of definition as given by Lemma 2.4.

Lemma 2.5. *Let* $K(t,z)$ *and* $L(t,z)$ *be the even extensions of the expressions given in Lemma* 2.4. *Then as functions of two complex variables they are analytic everywhere except possibly on the manifolds* $t=\pm z$ *and* $\operatorname{Re} t=0$.

Proof. By Lemma 2.4, both K and L are solutions to the hyperbolic equation

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

for real x and t, $t>0$, with initial conditions:

$$u(0,x)=f_1(x), \qquad \frac{\partial u}{\partial t}(0,x)=f_2(x),$$

where f_1 and f_2 are generalized functions whose analytic representation is singular only at the origin. By using a well-known technique from partial differential equations, this problem may be converted to an integral equation, namely

$$u(t,x)=u_0(t,x)-\frac{1}{2}\int_0^t\int_{x-t+\tau}^{x+t-\tau}u(\tau,y)q(\tau)\,dy\,d\tau$$

where

$$u_0(t,x)=\frac{1}{2}\{f_1(x+t)+f_1(x-t)\}+\frac{1}{2}\int_{x-t}^{x+t}f_2(y)\,dy.$$

(See [19, p. 274].)

This is a Volterra equation of the second kind and hence may be solved by the Picard method of successive approximations. That is, starting with u_0, define u_1 as

$$\begin{aligned}(2.18)\qquad u_1(t,x)&=-\int_0^t\int_{x-t+\tau}^{x+t-\tau}\frac{q(\tau)}{2}u_0(\tau,\xi)\,d\xi\,d\tau\\&=\iint_{\Omega_{x,t}}G(t,x;\,\tau,\xi)u_0(\tau,\xi)\,d\xi\,d\tau\end{aligned}$$

where $G(t,x;\,\tau,\xi)=-q(\tau)/2$ on $\Omega_{x,t}=\{(\tau,\xi)|0\leqq\tau\leqq t,\ |x-\xi|<t-\tau\}$.

The function G, the kernel of the equation, defines a mapping from $L^2_{\text{loc}}(\mathbb{R}^2)$ into itself which we also denote by G. It can be extended to an operator on $\mathscr{D}'$ since it involves only multiplication by a C^∞-function and integration. Symbolically we have

$$u_1=Gu_0.$$

We then repeat the procedure by defining successively

$$u_2=Gu_1,\quad u_3=Gu_2,\quad \cdots,\quad u_n=Gu_{n-1},\quad \cdots$$

or

$$u_2=G^2u_0,\quad u_3=G^3u_0,\quad \cdots,\quad u_n=G^nu_0.$$

The operator G^n is also an integral operator with kernel $G_n(t,x;\,\tau,\xi)$. Moreover it is dominated by

$$(2.19)\qquad |G_n(t,x;\,\tau,\xi)|\leqq\frac{|q(\tau)|}{2^n}\frac{(x-\xi)^{n-1}}{(n-1)!}\frac{\left(\int_\tau^t|q|\right)^{n-1}}{(n-1)!},\qquad (\tau,\varepsilon)\in\Omega.$$

This ensures that the series defining the resolvent operator converges in the sense of L^2 on bounded domains and hence that

$$u=\left(\sum_{n=0}^{\infty}G^n\right)u_0$$

is a formal solution to the integral equation

$$u=u_0+Gu$$

for u_0 sufficiently well behaved.

However $f_1(x)$ and $f_2(x)$ are generalized functions which are not in L^2_{loc}, and hence we must use their analytic representations instead. Fortunately this presents no problem since the latter are singular at most at $x=0$. It follows that the analytic representation of $u_0(t,x)$, given by

$$\hat{u}_0(t,z)=\frac{1}{2\pi i}\frac{1}{2}\int_{-\infty}^{\infty}\left\{\frac{1}{x-t-z}+\frac{1}{x+t-z}\right\}\{f_1(x)+F_2(x)\}\,dx, \tag{2.20}$$

where F_2 is a generalized function such that $F_2'=f_2$, is singular at most at $z=\pm t$.

We now convert our integral equation (2.18) into an integral equation involving the analytic representations,

$$\hat{u}=\hat{u}_0+G\hat{u}.$$

But we have

$$\begin{aligned}\widehat{Gu}(t,z)&=\frac{-1}{2\pi i}\int_{-\infty}^{\infty}\frac{1}{x-z}\int_0^t\frac{q(\tau)}{2}\int_{x-t+\tau}^{x+t-\tau}u(\tau,\xi)\,d\xi\,d\tau\,dx\\&=\frac{-1}{2\pi i}\int_{-\infty}^{\infty}\int_0^t\frac{q(\tau)}{2}\int_{-t+\tau}^{t-\tau}\frac{u(\tau,x+s)}{x+s-s-z}\,ds\,d\tau\,dx\\&=-\int_0^t\frac{q(\tau)}{2}\int_{-t+\tau}^{t-\tau}\hat{u}(\tau,z+s)\,ds\,d\tau\\&=-\int_0^t\frac{q(\tau)}{2}\int_{z-t+\tau}^{z+t-\tau}\hat{u}(\tau,\xi)\,d\xi\,d\tau\\&=(G\hat{u})(t,z).\end{aligned} \tag{2.21}$$

We now apply the Picard method to $\hat{u}_0(t,z)$ and observe that $\hat{u}_1,\hat{u}_2,\cdots,\hat{u}_n,\cdots$ may have singularities at most at $z=\pm t$. For example $\hat{u}_0(\tau,\xi)$ has singularities at $\xi=\pm\tau$. Hence, its antiderivative $\hat{u}_0^{(-1)}$ does also, and $\hat{u}_0^{(-1)}(\tau,z+t-\tau)$ has singularities at most at $z+t-\tau=\pm\tau$. Multiplication by the holomorphic function $f(t)$ adds no new singularities nor does integration from 0 to t. Hence, the only possible singularities from this first term are at $z+t-t=\pm t$ and $z+t-0=0$, i.e. at $z=\pm t$. The same is true of the other term $\hat{u}_0^{(-1)}(\tau,z-t+\tau)$. Thus the series

$$\hat{u}(t,z)=\sum_{n=0}^{\infty}\hat{u}_n(t,z),$$

which gives us our solution, converges uniformly on bounded domains excluding these values to a function whose singular points at most coincide with those of $\hat{u}_0(t,z)$. The same is true of $u(t,z)$ since $\hat{u}$ is nonsingular everywhere u is.

3. The singularity theorem. The main result of this paper is the following theorem.

THEOREM. *Let the generalized Fourier transform* $F(\lambda)$ *of* $f(t)$ *be such that* $F(\lambda) = O(e^{-cs})$ *as* $s \to \infty$ *for some* $c > 0$. *In addition, let*

$$g(z) = \frac{1}{\sqrt{2\pi}} \int_0^\infty F(\lambda) e^{isz}\, ds. \tag{3.1}$$

Then, if $f(t)$ *has a singular point at* $t = \alpha (\operatorname{Re}\alpha > 0)$, $g(z)$ *has one at either* $z = \alpha$ *or* $z = -\alpha$ *according to* $\operatorname{Im}\alpha < 0$ *or* $\operatorname{Im}\alpha > 0$. *Conversely, if* $g(z)$ *has a singularity at* $z = \beta$, $f(t)$ *will have one at either* $t = \beta$ *or* $t = -\beta$ *depending on whether* $\operatorname{Im}\beta < 0$ *or* $\operatorname{Im}\beta > 0$.

Proof. First of all, the function $f(t)$ is holomorphic in the strip $(\operatorname{Re} t > 0) \times (|\operatorname{Im} t| < c)$. For we have

$$\left| \int_0^\infty F(\lambda)\phi(t,\lambda)\, d\rho \right| \leqq A \int_0^\infty e^{-cs} e^{|\operatorname{Im} t| s}\, ds < \infty$$

for $|\operatorname{Im} t| < c$. Moreover,

$$f(t) = \int_{-\infty}^0 F(\lambda)\phi(t,\lambda)\, d\rho(\lambda) + \int_0^\infty F(\lambda)\phi(t,\lambda)\, d\rho(\lambda). \tag{3.2}$$

Since the negative part of the spectrum is discrete and bounded below, the contribution from the first term, say $\eta(t)$, is holomorphic in $\operatorname{Re} t > 0$ because $\phi(t,\lambda)$ is holomorphic therein. The second integral has just been shown to be absolutely and uniformly convergent in $|\operatorname{Im} t| < c - \varepsilon$ for $\varepsilon > 0$, hence it defines a holomorphic function in the prescribed strip. It is readily seen that $g(z)$ is holomorphic in the half-plane $\operatorname{Im} z > -c$. Now we define an integral operator that maps g onto f. From the inversion formula for the Fourier transform we have

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty g(x+iy) e^{-is(x+iy)}\, dx = \begin{cases} F(\lambda), & \lambda > 0, \\ 0, & \lambda < 0, \end{cases} \tag{3.3}$$

or

$$F(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{ia-\infty}^{ia+\infty} g(z) e^{-isz}\, dz, \qquad \lambda > 0 \tag{3.4}$$

where $a > -c$. Upon substituting (3.4) into (3.2) we obtain

$$f(t) = \eta(t) + \int_{ia-\infty}^{ia+\infty} g(z) K(t,z)\, dz. \tag{3.5}$$

Interchanging the integrals is permissible by the uniform convergence. The line of integration can be replaced by any other contour $\gamma(z)$ going from $-\infty$ to ∞ provided that t is real and for all z, $\operatorname{Im} z > -c$. In fact, this representation of $f(t)$ holds for complex t as long as $-|\operatorname{Im} t| > \operatorname{Im} z > -c$. By using Hadamard's multiplication of singularities argument we can continue $f(t)$ beyond this initial domain of definition. As t moves in the complex t-plane ($\operatorname{Re} t > 0$) the singularities of the integrand move in the complex z-plane and the initial domain of definition of $f(t)$ is enlarged to contain all these points t for which the contour of integration γ can be deformed without a singularity of the integrand passing over γ. This process can be continued until we have a singularity of the integrand threatening to cross the contour and it is no longer possible to deform γ to avoid it. This happens whenever $g(z)$ and $K(t,z)$ have a common singular point. Therefore, if $g(z)$ has a singular point at $z = \alpha$, then $f(t)$ has a possible one at $t=\pm\alpha$ since by Lemma 2.5, $K(t,z)$ may have singularities only at $t=\pm z$.

Going the other direction we construct another integral transform that maps f onto g. We have

$$g(z)=\frac{1}{\sqrt{2\pi}}\int_0^\infty F(\lambda)e^{isz}\,ds=\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{isz}\,ds\int_0^\infty f(t)\phi(t,\lambda)\,dt \tag{3.6}$$
$$=\int_0^\infty f(t)L(t,z)\,dz.$$

It is easy to see that for t real and $\operatorname{Im} z>0$ the integrals converge absolutely and uniformly on compact subsets and hence interchanging the integrals is permissible. This representation of g may hold for other values of z. For example let z be real and positive; since $L(t,z)$ has a possible singularity at $t=z$, we deform the contour of integration to the following one $\gamma=(0,z-\delta)\cup\gamma'\cup(z+\delta,\infty)$ where γ' is the lower half of the circle that is centered at z and has radius δ such that $\delta<c$. Now we continue g to the lower half plane and by using the same reasoning as before we can show that the only possible singularities of $g(z)$ in $\operatorname{Re} z>0$ are the common singularities of $f(t)$ and $L(t,z)$. Therefore, if $f(t)$ has a singularity at $t=\alpha$ it follows from Lemma 2.5 that $g(z)$ may have one at $z=\pm\alpha$.

Furthermore we may deduce that g does in fact have a singularity at one of those points. Indeed if g were not singular at either, neither would f be singular at α by the first part of the proof. The same argument works in the opposite direction as well.

Q.E.D.

Briefly, the theorem states that $f(t)$ and $g(z)$ have the same singularities in the fourth quadrant and the singularities of $f(t)$ in the first quadrant are mapped into those of $g(z)$ in the third quadrant via the map $\alpha\to-\alpha$.

COROLLARY 3.1. *If $q(t)$ and $f(t)$ are even, then $f(t)$ and $\tilde{g}(z)=\int_{-\infty}^{0}F(\lambda)e^{ist}\,ds$ have the same singularities in the second quadrant and the singularities of $f(t)$ in the third quadrant give rise to singularities of $\tilde{g}(z)$ in the first quadrant.*

Proof. Similar to that of the main theorem.

We close this section by giving some examples.

Examples.

1) The simplest case is the Fourier-cosine transform i.e. $q(t)=0$ and $\alpha=\pi/2$. In this case

$$\rho'(\lambda)=\begin{cases}\dfrac{1}{s}, & \lambda>0,\\ 0, & \lambda<0,\end{cases}\qquad \phi(t,\lambda)=\cos(st),$$

$$F(\lambda)=\int_0^\infty f(t)\cos(st)\,dt,\qquad f(t)=\int_0^\infty F(\lambda)\cos(st)\,ds.$$

a) As an example consider

$$f(t)=\frac{2}{\pi}\frac{1}{t^2+\alpha^2},\qquad t>0,\ \operatorname{Re}\alpha>0.$$

Then

$$F(\lambda)=\frac{1}{\alpha}e^{-\alpha s},\qquad s>0,$$

$$g(z)=\frac{1}{\sqrt{2\pi}}\int_0^\infty\frac{1}{\alpha}e^{-\alpha s}e^{isz}\,ds=\frac{1}{\sqrt{2\pi}}\frac{1}{\alpha}\frac{\alpha+iz}{\alpha^2+z^2}=\frac{1}{\sqrt{2\pi}\alpha}\frac{1}{z+i\alpha}.$$

$f(t)$ has two singular points at $t = \pm i\alpha$ and since $\operatorname{Re}\alpha > 0$, $\operatorname{Im} i\alpha > 0$. If $\operatorname{Im}\alpha < 0$, then $f(t)$ has one singular point in the first quadrant at $i\alpha$ and one in the fourth quadrant at $-i\alpha$. In either case $g(z)$ has one singularity at $-i\alpha$.

b) More generally, for

$$f(t) = \frac{\beta}{\pi}\left[\frac{1}{\beta^2 + (\alpha - t)^2} + \frac{1}{\beta^2 + (\alpha + t)^2}\right]$$

where α is real and $\operatorname{Re}\beta > 0$, we have $F(\lambda) = e^{-\beta s}\cos\alpha s$ and

$$\begin{aligned} g(z) &= \frac{1}{2\sqrt{2\pi}}\left\{\frac{\beta + i(\alpha + z)}{\beta^2 + (\alpha + z)^2} + \frac{\beta + i(\alpha - z)}{\beta^2 + (\alpha - z)^2}\right\} \\ &= \frac{1}{2\sqrt{2\pi}}\left\{\frac{1}{\beta - i(\alpha + z)} + \frac{1}{\beta - i(\alpha - z)}\right\}. \end{aligned}$$

2) As for the case where $q(t) = 0$ and $0 \neq \alpha \neq \pi/2$, we have

$$\rho'(s) = \begin{cases} \dfrac{s}{\cos^2\alpha + s^2\sin^2\alpha} & (\lambda > 0), \\ 0 & (\lambda < 0), \end{cases}$$

$$\phi(t,\lambda) = \sin\alpha\cos(st) - \frac{\cos\alpha\sin(st)}{s}.$$

Thus,

$$F(\lambda) = \int_0^\infty f(t)\left(\sin\alpha\cos(st) - \frac{\cos\alpha\sin(st)}{s}\right)dt,$$

and for x real

$$g(x) = \sin\alpha\int_0^\infty f(t)\eta(x,t)\,dt - \cos\alpha\int_0^\infty f(t)\zeta(x,t)\,dt$$

where

$$\eta(x,t) = \frac{1}{\sqrt{2\pi}}\int_0^\infty e^{ixs}\cos(st)\,ds = \frac{ix}{(x + i0)^2 - t^2}$$

and

$$\zeta(x,t) = \frac{1}{\sqrt{2\pi}}\int_0^\infty e^{ixs}\frac{\sin(st)}{s}\,ds = \frac{1}{2i}\left[\ln(x - t + i0) - \ln(x + t + i0)\right].$$

Clearly ζ and η are interpreted as generalized functions. See [3]. Now by going to the analytic representations of ζ, η and using Hadamard's argument one can show that if $f(t)$ has a singularity at $t = \alpha$, then $g(z)$ has a possible one at $z = \pm\alpha$.

3) *The Hankel formula.* Let $q(t) = (\nu^2 - \frac{1}{4})/t^2$ with $\nu > 1$.

At the first glance it may appear that our technique does not apply to this case since it arises from a singular Sturm–Liouville problem on the whole real line or equivalently a singular SL problem on half-line with a singularity at the finite end point. However, a close analysis of the problem shows that our technique is still applicable with slight modifications, e.g. the role of $\phi(t,\lambda)$ is now played by $\psi(t,\lambda)$; see the sentence following (2.12) for the definition of $\psi(t,\lambda)$. Thus we have

$$F(\lambda)=\int_0^\infty f(t)\sqrt{t}\,J_\nu(ts)\,dt$$

where $J_\nu(x)$ is the Bessel function of the first kind and order ν.

a) Let

$$f(t)=\frac{\sqrt{t}}{(t^2+a^2)^{3/2}},\qquad \mathrm{Re}\,\alpha>0.$$

Then $F(\lambda)=(1/a)e^{-as}$ and

$$g(z)=\frac{i}{a}\frac{1}{z+ia}.$$

b) Let

$$f(t)=\frac{t^{\nu+1/2}}{(t^2+a^2)^{\nu+1/2}},\qquad \mathrm{Re}\,a>0,\ \mathrm{Re}\,\nu>0.$$

Then

$$F(\lambda)=\frac{\sqrt{\pi}\,s^{\nu-1}e^{-as}}{2^\nu\Gamma(\nu+1/2)},$$

$$g(z)=\frac{\Gamma(\nu)}{2^{\nu+1/2}\Gamma(\nu+1/2)}\,\frac{e^{i\nu\tan^{-1}(z/a)}}{(z^2+a^2)^{\nu/2}}.$$

c) Let

$$f(t)=t^{-\nu+1/2}\left\{\frac{\left[\sqrt{t^2+\beta^2}-\beta\right]^\nu}{\sqrt{t^2+\beta^2}}-\frac{\left[\sqrt{t^2+\alpha^2}-\alpha\right]^\nu}{\sqrt{t^2+\alpha^2}}\right\},$$

$$\mathrm{Re}\,\alpha,\ \mathrm{Re}\,\beta>0,\quad \mathrm{Re}\,\nu>-1,\quad \alpha\neq\beta.$$

Then $F(\lambda)=(e^{-\beta s}-e^{-\alpha s})/s$ and

$$g(z)=\frac{1}{\sqrt{2\pi}}\left\{\frac{1}{2}\ln\left(\frac{z^2+\alpha^2}{z^2+\beta^2}\right)+i\left(\tan^{-1}\frac{z}{\beta}-\tan^{-1}\frac{z}{\alpha}\right)\right\}.$$

4) Let

$$q(t)=\frac{\nu^2-1/4}{t^2}\quad \text{with } 0<\nu<1.$$

Then

$$\rho'(\lambda)=\frac{1}{c^2-2cs^{2\nu}\cos\nu\pi+s^{4\nu}},\qquad c<0,$$

$$\psi(t,\lambda)=\sqrt{t}\left[cJ_\nu(ts)-s^{2\nu}J_{-\nu}(ts)\right].$$

For

$$f(t)=\frac{t}{t^2+\alpha^2}, \qquad \operatorname{Re}\alpha>0, \qquad \nu=\frac{1}{2}$$

we have

$$F(\lambda)=c\sqrt{\alpha}\,K_{1/2}(as)-\frac{s}{\sqrt{a}}K_{-1/2}(as)$$

where $K_\nu(x)$ is the modified Bessel function of order ν. Hence

$$\begin{aligned} g(z)={}&\frac{c\sqrt{\pi}}{4\sqrt{z^2+a^2}}\left[\sqrt{z+u}\,(1+i)+\frac{a}{\sqrt{z+u}}-i\sqrt{u-z}\right] \\ &+\frac{1}{\sqrt{2\pi a}}\left[\frac{\Gamma(3/4)\Gamma(5/4)}{a^2}\,{}_2F_1\left(\frac{3}{4},\frac{5}{4};\frac{1}{2};\frac{-z^2}{a^2}\right)\right. \\ &\left.+\frac{2iz\Gamma(5/4)\Gamma(7/4)}{a^3}\,{}_2F_1\left(\frac{5}{4},\frac{7}{4};\frac{3}{2};-\frac{z^2}{a^2}\right)\right] \end{aligned}$$

where $u=\sqrt{z^2+a^2}$.

Acknowledgments. This research was done while the first author was visiting the Mathematics Department at the University of Wisconsin-Milwaukee and he would like to take the opportunity to thank the department for the hospitality and the stimulating atmosphere that were provided to him during his visit.

REFERENCES

[1] S. BERGMAN, *Integral Operators in the Theory of Linear Partial Differential Equations*, Ergeb. Math. N.S., 23, Springer, Berlin, 1961.

[2] ______, *Application of integral operators to singular differential equations and to computations of compressible fluid flows*, in Numerical Solutions of Partial Differential Equations, Academic Press, New York, 1966.

[3] H. BREMERMANN, *Distributions, Complex Variables and Fourier Transforms*, Addison Wesley, New York, 1965.

[4] H. DIAMOND, M. KON AND L. RAPHAEL, *Stable summation methods for a class of singular Sturm–Liouville expansions*, Proc. Amer. Math. Soc., 81 (1981), pp. 279–286.

[5] G. DOETSCH, *Handbuch der Laplace-transformationen-B*, Band II, Birkhäuser, Basel, 1971. (Russian original appeared in 1955.)

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

[7] R. GILBERT, *Singularities of three-dimensional harmonic functions*, Pacific J. Math., 10 (1960), pp. 1243–1255.

[8] ______, *A note on harmonic functions in $(p+2)$ variables*, Arch. Rational Mech. Anal., 8 (1961), pp. 223–227.

[9] ______, *On the location of singularities of a class of elliptic partial differential equations in four variables*, Canad. J. Math., 17 (1965), pp. 676–686.

[10] ______, *Function Theoretic Methods in Partial Differential Equations*, Academic Press, New York, 1969.

[11} R. GILBERT AND G. ROACH, *Constructive methods for metaparabolic and pseudoparabolic systems*, Bull. Math. Soc. Sci. Math. R. S. Roumaine (R.S), 20(68) 1976, no. 1-2, 97–109 (1977).

[12] R. GILBERT AND H. HOWARD, *A generalization of a theorem of Nehuri*, Bull. Amer. Math. Soc., 72 (1966), pp. 37–43.

[13] ______, *Role of the integral operator method in the theory of potential scattering*, J. Math. Phys., 8 (1967), pp. 141–148.

[14] R. Gilbert and H. Howard, *On the singularities of Sturm–Liouville expansions*, Proc. Symposium Analytic Methods in Mathematical Physics, Gordon and Breach, New York, 1970, pp. 443–452.

[15] ______, *The scope of the function theoretic approach for equations permitting a separation of variables*, J. Math. Anal. Appl., (1971), pp. 671–684.

[16] R. Gilbert and S. Shieh, *A new method in the theory of potential scattering*, J. Math. Phys., 7 (1966), pp. 431–433.

[17] J. Levin, *Distribution of zeros of entire functions*, AMS Trans., 5, American Mathematical Society, Providence, RI, 1972.

[18] B. Levitan and I. Sargsjan, *Introduction to Spectral Theory*, Math. Monos., 39, American Mathematical Society Providence, RI, 1975.

[19] Z. Nehari, *On the singularities of Legendre expansions*, J. Rational Mech. Anal., 5 (1956), pp. 987–991.

[20] R. Newton, *Scattering Theory of Waves of Particles*, McGraw-Hill, New York, 1966.

[21] R. Omnes and M. Froissart, *Mandelstam Theory and Regge Poles*, Benjamin, New York, 1963.

[22] F. Vekua, *Generalized Analytic Functions*, Addison-Wesley, Reading, MA, 1962.

[23] G. Walter, *On real singularities of Legendre expansions*, Proc. Amer. Math. Soc., 19 (1968), pp. 1407–1412.

[24] ______, *Singular points of Sturm-Liouville series*, this Journal, 2 (1971), pp. 393–401.

[25] A. Zayed, *On the singularities of Gegenbauer (ultraspherical) expansions*, Trans. Amer. Math. Soc., 262 (1980), pp. 487–503.

[26] ______, *Hyperfunctions as boundary values of generalized axially symmetric potentials*, Illinois J. Math., 25 (1981), pp. 306–317.

[27] A. Zayed and G. Walter, *Series of orthogonal polynomials as hyperfunctions*, this Journal, 13 (1982), pp. 664–675.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

FUNCTIONAL PERTURBATIONS OF SECOND ORDER DIFFERENTIAL EQUATIONS*

WILLIAM F. TRENCH[†]

Abstract. Conditions are given which imply that the functional differential equation

$$(r(t)x'(t))'+q(t)x(t)=f(t,x(g(t)))$$

has a solution $\bar{x}$ which behaves for large t in a precisely defined way like a given solution $\bar{y}$ of the ordinary differential equation

$$(r(t)y')'+q(t)y=0.$$

It is not assumed that $g(t)-t$ is sign-constant, and $f(t,u)$ need only be defined and continuous on a subset of the (t,u) plane which is near the curve $u=\bar{y}(g(t))$ in an appropriate sense for large t. The integral smallness conditions on $f(t,u)$ permit some of the improper integrals in question to converge conditionally. Separate treatments are given for the cases where the unperturbed equation is oscillatory or nonoscillatory. The results are new even in the case where $g(t)=t$.

1. Introduction. We present conditions implying that the functional differential equation

$$(1)\qquad (r(t)x'(t))'+q(t)x(t)=f(t,x(g(t)))$$

has a solution $\bar{x}$ which behaves for large t like a given solution $\bar{y}$ of the ordinary differential equation

$$(2)\qquad (r(t)y')'+q(t)y=0,\qquad t>a.$$

We give specific estimates of $\bar{x}-\bar{y}$ as $t\to\infty$. We do not require $g(t)-t$ to be sign constant, and the perturbing function $f=f(t,u)$ need be defined and continuous only on a subset of the (t,u) plane near the curve $u=\bar{y}(g(t))$ for large t, in a sense made precise below. We believe that our results are new even if $g(t)=t$. Our integral smallness conditions on the function $f(t,\bar{y}(g(t)))$ require only ordinary (i.e., perhaps conditional) convergence; however, we do impose conditions which imply absolute convergence of certain integrals involving differences

$$(3)\qquad f(t,x(g(t)))-f(t,\bar{y}(g(t))),$$

where x is a function near $\bar{y}$ in an appropriate sense. Since forcing functions (i.e., terms in $f(t,u)$ which are independent of u) obviously cancel out of (3), this means that our integral smallness conditions on them always allow conditional convergence; however, this is not the only way in which possibly conditional convergence enters into our hypotheses. Accordingly, all integrability assumptions below should be interpreted as allowing conditional convergence, except when the integrands in question are obviously nonnegative. Moreover, to avoid repetition, it is to be understood that whenever we write an improper integral in stating an assumption, we are assuming that it converges.

*Received by the editors October 4, 1983, and in revised form April 6, 1984.

[†] Department of Mathematical Sciences, Drexel University, Philadelphia, Pennsylvania 19104.

Since the asymptotic theory of (1) depends critically on whether (2) is oscillatory or nonoscillatory, we consider these two cases separately in §§2 and 3. Some of our results in §3 are related to results of Kusano and Naito [3] and Kusano and Onose [4]. Hallam [2] obtained related results, valid when (2) is either oscillatory of nonoscillatory, for the case where $r=1$ and $g(t)=t$.

To avoid repetition, we state here that three proofs below demonstrate the existence of a solution $\bar{x}$ of (1), with prescribed asymptotic properties, as a fixed point of a mapping $\mathscr{T}$defined on a closed convex subset D of the Frechet space $C[\tau_0,\infty)$ (for some $\tau_0\geqq a$), with the topology of uniform convergence on compact intervals. In this context we write

$$D\lim_{k\to\infty} x_k=x \tag{4}$$

to mean that $\{x_k\}$ is a sequence of functions in D which converges uniformly to x on compact subintervals of $[\tau_0,\infty)$.

The proof in each case consists of establishing the following:

(i) $\mathscr{T}(D)\subset D$.

(ii) $\mathscr{T}$is continuous; that is, (4) implies that

$$D\lim_{k\to\infty}\mathscr{T}x_k=\mathscr{T}x.$$

(iii) There is a continuous positive function ψ such that

$$|(\mathscr{T}x)'(t)|\leqq\psi(t),\qquad x\in D,\quad t\geqq\tau_0. \tag{5}$$

The last inequality implies that the function in $\mathscr{T}(D)$ are equicontinuous on compact intervals. Since it will be clear in all cases that the functions in D are uniformly bounded on compact intervals, this and (i) imply that $\mathscr{T}(D)$ has compact closure, by the Arzela–Ascoli theorem. The Schauder–Tykhonov fixed point theorem will then imply that $\mathscr{T}\bar{x}=\bar{x}$ for some $\bar{x}$ in D, and routine differentiation (which we omit) will show that $\bar{x}$ satisfies (1) on some interval (t_0,∞), with $t_0\geqq\tau_0$. We will call such a function a solution of (1).

All quantities are assumed to be real. The following assumption applies throughout.

Assumption A. The functions r,q, and g are continuous on $[a,\infty)$, $r>0$, and

$$g(t)\geqq a,\qquad \lim_{t\to\infty} g(t)=\infty. \tag{6}$$

The functions y_1 and y_2 are solutions of (2) such that

$$y_2'y_1-y_2y_1'=\frac{1}{r}, \tag{7}$$

and

$$\bar{y}=c_1y_1+c_2y_2\qquad (c_1,c_2=\text{constants}) \tag{8}$$

is a given solution of (2). The function ϕ is positive, continuous, and nonincreasing on $[a,\infty)$, and either

$$\lim_{t\to\infty}\phi(t)=0\quad\text{or }\phi=1. \tag{9}$$

2. Perturbations of an oscillatory equation. In this section,

$$z=\left(y_1^2+y_2^2\right)^{1/2}. \tag{10}$$

Our proofs here make no use of the assumption that (2) is oscillatory, so our results apply even if it is nonoscillatory; however, in the latter case, better results are obtained in §3.

THEOREM 1. *Suppose*

$$\overline{\lim_{t\to\infty}}\,(\phi(t))^{-1}\left|\int_t^\infty y_i(s)f(s,\bar{y}(g(s)))\,ds\right|=\alpha_i<\infty \tag{11}$$

and

$$\overline{\lim_{t\to\infty}}\,(\phi(t))^{-1}\int_t^\infty |y_i(s)|\sigma(s)\,ds=\beta_i<\infty \tag{12}$$

for $i=1,2$, where σ is positive and continuous on $[a,\infty)$. Suppose further that there are constants $T\geqq a$ and

$$M>K=\left[(\alpha_1+\beta_1)^2+(\alpha_2+\beta_2)^2\right]^{1/2} \tag{13}$$

such that f is continuous and

$$|f(t,u)-ft,\bar{y}(g(t))|\leqq\sigma(t) \tag{14}$$

on the set

$$\Omega=\left\{(t,u)\,\middle|\,t\geqq T,\,|u-\bar{y}(g(t))|\leqq M\phi(g(t))z(g(t))\right\}. \tag{15}$$

Then (1) *has a solution $\bar{x}$ such that*

$$\overline{\lim_{t\to\infty}}\,[\phi(t)z(t)]^{-1}|\bar{x}(t)-\bar{y}(t)|\leqq K. \tag{16}$$

Proof. Let

$$v_i(t)=\int_t^\infty |y_i(s)|\sigma(s)\,ds+\sup_{\tau\geqq t}\left|\int_\tau^\infty y_i(s)f(s,\bar{y}(g(s)))\,ds\right| \tag{17}$$

and

$$v=\left(v_1^2+v_2^2\right)^{1/2}; \tag{18}$$

then (11), (12), and (13) imply that

$$\overline{\lim_{t\to\infty}}\,(\phi(t))^{-1}v(t)\leqq K, \tag{19}$$

and that there is a $\tau_0\geqq T$ such that

$$v(t)\leqq M\phi(t),\qquad t\geqq\tau_0. \tag{20}$$

Let

$$D=\left\{x\in C[\tau_0,\infty)\,\middle|\,|\bar{x}(\tau)-\bar{y}(\tau)|\leqq M\phi(\tau)z(\tau),\tau\geqq\tau_0\right\} \tag{21}$$

(recall (10)), and choose $t_0 \geqq \tau_0$ so that

$$g(t) \geqq \tau_0, \qquad t \geqq t_0 \tag{22}$$

(recall (6)). Then (21), (22), and our assumptions on f imply that $f(t,x(g(t)))$ is continuous on $[t_0, \infty)$ if x is in D. Moreover, since

$$\int_t^\infty y_i(s)f(s,x(g(s)))\,ds = \int_t^\infty y_i(s)f(s,\bar{y}(g(s)))\,ds + \int_t^\infty y_i(s)[f(s,x(g(s))) - f(s,\bar{y}(g(s)))]\,ds,$$

(14) and (17) imply that

$$\left|\int_t^\infty y_i(s)f(s,x(g(s)))\,ds\right| \leqq \nu_i(t), \qquad t \geqq t_0, \quad x \in D. \tag{23}$$

Now define $\mathcal{T}$ on D by

(24)

$$(\mathcal{T}x)(t) = \begin{cases} \bar{y}(t) + \int_t^\infty [y_2(s)y_1(t) - y_1(s)y_2(t)]f(s,x(g(s)))\,ds, & t \geqq t_0, \\ \bar{y}(t) + \int_{t_0}^\infty [y_2(s)y_1(t) - y_1(s)y_2(t)]f(s,x(g(s)))\,ds, & \tau_0 \leqq t < t_0. \end{cases}$$

(The second line is vacuous if $\tau_0 > t_0$.) From (23) and Schwarz's inequality,

$$|(\mathcal{T}x)(t) - \bar{y}(t)| \leqq z(t)v(t), \qquad t \geqq \tau_0 \tag{25}$$

(to see this for $\tau_0 \leqq t \leqq t_0$, note that v is nonincreasing), which, with (20), implies that $\mathcal{T}(D) \subset D$.

If $D\lim_{k\to\infty} x_k = x$, then

$$\left|\int_t^\infty y_i(s)[f(s,x_k(g(s))) - f(s,x(g(s)))]\,ds\right| \leqq \int_{t_0}^\infty |y_i(s)||f(s,x_k(g(s))) - f(s,x(g(s)))|\,ds, \qquad t \geqq t_0,$$

where the integrand on the right converges pointwise to zero as $k \to \infty$, and is bounded for all k by $2|y_i(s)|\sigma(s)$ (recall (14)); hence, (12) and Lebesgue's dominated convergence theorem imply that the integral on the right approaches zero as $k \to \infty$. Therefore, if $\varepsilon > 0$ there is an N such that

$$\left|\int_t^\infty y_i(s)[f(s,x_k(g(s))) - f(s,x(g(s)))]\,ds\right| < \varepsilon, \qquad t \geqq t_0, \quad k \geqq N, \tag{26}$$

for $i = 1, 2$. From this, (24), and Schwarz's inequality,

$$|(\mathcal{T}x_k)t) - (\mathcal{T}t)(x)| \leqq \varepsilon\sqrt{2}\,z(t), \qquad t \geqq \tau_0, \quad k \geqq N.$$

This implies that $D\lim_{k\to\infty} \mathcal{T}x_k = \mathcal{T}x$.

By differentiating (24), we see from (18), (20), (23), Schwarz's inequality, and the monotonicity of ϕ that (5) holds, with

$$\psi=|\bar{y}'|+\left[(y_1')^2+(y_2')^2\right]^{1/2}M\phi,$$

and this completes the verification of (i), (ii), and (ii) of §1. Therefore, $\mathscr{T}$ has a fixed point (function) $\bar{x}$ in D. From (24),

$$\bar{x}(t)=\bar{y}(t)+\int_t^\infty [y_2(s)y_1(t)-y_1(s)y_2(t)]f(s,\bar{x}(g(s)))\,ds, \qquad t\geqq t_0, \tag{27}$$

so $\bar{x}$ satisfies (1) on (t_0,∞). Setting $x=\bar{x}$ in (25) and recalling that $\bar{x}=\mathscr{T}\bar{x}$, we see that

$$|\bar{x}(t)-\bar{y}(t)|\leqq z(t)v(t), \qquad t\geqq t_0,$$

so (19) implies (16). This completes the proof.

Taking $\phi=1$ in Theorem 1, so that obviously $\alpha_i=\beta_i=0$, yields the following corollary.

COROLLARY 1. *Suppose the integrals*

$$\int^\infty y_i(t)f(t,\bar{y}(g(t)))\,dt, \qquad i=1,2, \tag{28}$$

converge, and

$$\int^\infty z(t)\sigma(t)\,dt<\infty \tag{29}$$

with σ positive and continuous on $[a,\infty)$. Suppose also that there are constants $T\geqq a$ and $M>0$ such that f is continuous and satisfies (14) *on the set*

$$\bar{\Omega}=\{(t,u)|t\geqq T, |u-\bar{y}(g(t))|\leqq Mz(g(t))\}. \tag{30}$$

Then (1) *has a solution $\bar{x}$ such that*

$$\bar{x}(t)=\bar{y}(t)+o(z(t)). \tag{31}$$

Remark 1. Although (9) was not used in the proof of Theorem 1, it imposes no loss of generality, since Theorem 1 without (9) is easily shown to be equivalent to Corollary 1 if $\lim_{t\to\infty}\phi(t)>0$.

Remark 2. If, *in addition* to the assumptions of Corollary 1, the stronger integral conditions (11) and (12) hold (with $i=1,2$), then it is routine to verify that the solution $\bar{x}$ which satisfies (31) actually satisfies the stronger condition (16). However, this does not mean that Theorem 1 is only a trivial extension of Corollary 1. The hypotheses of Theorem 1 with $\lim_{t\to\infty}\phi(t)=0$ do not imply those of Corollary 1, since the set Ω in (15) on which (14) is required to hold in Theorem 1 is then smaller than the set $\bar{\Omega}$ in (30). Put another way, the hypotheses of Theorem 1 in this case imply the hypotheses of the Schauder–Tykhonov theorem for the subset D of $C[\tau_0,\infty)$ as defined by (21), but not for the larger subset $\bar{D}$ which would result if ϕ were replaced by one in (21). Example 2 below will illustrate this point.

Remarks similar to these apply to other results which follow.

THEOREM 2. *Suppose* (11) *holds with $i=1,2$. Let λ be nonnegative and continuous on $[a,\infty)$, and*

$$\overline{\lim_{t\to\infty}}\,(\phi(t))^{-1}\int_t^\infty |y_i(s)|\lambda(s)z(g(s))\phi(g(s))\,ds=b_i, \qquad i=1,2, \tag{32}$$

where

$$b_1^2+b_2^2<1. \tag{33}$$

Suppose further that there are constants $T\geqq a$ *and* $M>0$ *such that*

$$(\alpha_1+Mb_1)^2+(\alpha_2+Mb_2)^2<M^2 \tag{34}$$

and f is continuous and satisfies the inequality

$$|f(t,u)-f(t,\bar{y}(g(t)))|\leqq\lambda(t)|u-\bar{y}(g(t))| \tag{35}$$

on the set Ω *in* (15). *Then* (1) *has a solution* $\bar{x}$ *such that*

$$\overline{\lim_{t\to\infty}}\,[\phi(t)z(t)]^{-1}|\bar{x}(t)-\bar{y}(t)|\leqq\left[(\alpha_1+Mb_1)^2+(\alpha_2+Mb_2)^2\right]^{1/2}$$

Proof. (Note that (34) holds if M is sufficiently large, because of (33).) If f satisfies (35) on Ω, then it also satisfies (14) on Ω, with

$$\sigma(t)=M\lambda(t)z(g(t))\phi(g(t)).$$

This and (32) imply (12) with $\beta_i=Mb_i$, and then (34) implies (13). Hence, Theorem 1 implies the conclusion.

COROLLARY 2. *Suppose the integrals* (28) *converge, and*

$$\int^{\infty} z(t)\lambda(t)z(g(t))\,dt<\infty,$$

where λ *is continuous and positive on* $[a,\infty)$. *Suppose also that there are constants* $T\geqq a$ *and* $M>0$ *such that f is continuous and satisfies* (35) *on the set* $\overline{\Omega}$ *in* (30). *Then* (1) *has a solution* $\bar{x}$ *which satisfies* (31).

We now apply our results to the equation

$$(r(t)x'(t))'+q(t)x(t)=p(t)(x(g(t)))^{\gamma}+h(t), \tag{36}$$

which has the form of a generalized Emden–Fowler equation, but is unusual in that (2) may be oscillatory. (In §3 we consider (36) in the case where (2) is nonoscillatory.)

THEOREM 3. *Suppose* p, $h\in C[a,\infty)$, *and* γ *is positive and rational, with odd denominator. Suppose further that the integrals*

$$\int^{\infty} y_i(t)h(t)\,dt, \qquad i=1,2 \tag{37}$$

converge, and that

$$\int^{\infty} z(t)|p(t)|(z(g(t)))^{\gamma}dt<\infty. \tag{38}$$

Then (36) *has a solution* $\bar{x}$ *such that*

$$\bar{x}(t)=\bar{y}(t)+o(z(t)).$$

Proof. For (36), the function f in (1) is

$$f(t,u)=p(t)u^{\gamma}+h(t), \tag{39}$$

which is continuous on $\bar{\Omega}$ in (30) for any $T \geqq a$ and $M>0$. Moreover, if $(t,u)\in\bar{\Omega}$, then

$$(40) \qquad |u| \leqq |\bar{y}(g(t))| + Mz(g(t)) \leqq (C+M)z(g(t)),$$

where $C=(c_1^2+c_2^2)^{1/2}$. (See (8) and (10).) Since obviously

$$\begin{aligned} |f(t,u)-f(t,\bar{y}(g(t)))| &\leqq |p(t)|\left[|u|^{\gamma}+(\bar{y}(g(t)))^{\gamma}\right] \\ &\leqq |p(t)|\left[|u|^{\gamma}+C(z(g(t)))^{\gamma}\right], \end{aligned}$$

(40) implies (14) for $(t,u)\in\bar{\Omega}$, with

$$\sigma(t)=|p(t)|(z(g(t)))^{\gamma}\left[(C+M)^{\gamma}+C^{\gamma}\right],$$

Therefore (38) implies (29). Since (37) and (38) also imply that the integrals (28) converge, Corollary 1 implies the conclusion.

THEOREM 4. *Suppose* p, $h\in C[a,\infty)$ *and* $\gamma \geqq 1$ *is rational, with odd denominator. Suppose also that*

$$(41) \qquad \lim_{t\to\infty}\phi(t)=0,$$

$$\overline{\lim_{t\to\infty}}\,(\phi(t))^{-1}\left|\int_t^{\infty} y_i(s)\left[p(s)(\bar{y}(g(s)))^{\gamma}+h(s)\right]ds\right| = \alpha_i<\infty, \qquad i=1,2,$$

and

$$(42) \quad \overline{\lim_{t\to\infty}}\,(\phi(t))^{-1}\int_t^{\infty}|y_i(s)p(s)|(z(g(s)))^{\gamma}\phi(g(s))\,ds=B_i<\infty, \qquad i=1,2.$$

Finally, suppose that

$$\gamma C^{\gamma-1}\left(B_1^2+B_2^2\right)^{1/2}<1$$

and $M>0$ *satisfies the inequality*

$$\left(\alpha_1+M\gamma C^{\gamma-1}B_1\right)^2+\left(\alpha_2+M\gamma C^{\gamma-1}B_2\right)^2<M^2.$$

Then (36) *has a solution* $\bar{x}$ *such that*

$$\overline{\lim_{t\to\infty}}\,[\phi(t)z(t)]^{-1}|\bar{x}(t)-\bar{y}(t)| \leqq \left[\left(\alpha_1+M\gamma C^{\gamma-1}B_1\right)^2+\left(\alpha_2+M\gamma C^{\gamma-1}B_2\right)^2\right]^{1/2}$$

Proof. Again, f as in (39) is continuous on Ω in (15) for any $T\geqq a$ and $M>0$. As in (40),

$$|u| \leqq [C+M\phi(g(t))]\,z(g(t)), \qquad (t,u)\in\Omega.$$

Therefore, the mean value theorem implies (35) with $\lambda(t)=\gamma|p(t)|[C+M\phi(g(t))]^{\gamma-1}$. This, (9), (41), and (42) imply (32) with $b_i=\gamma C^{\gamma-1}B_i$, and this for any M. Now Theorem 2 implies the conclusion.

In the following examples, we take $y_1(t)=\cos t$ and $y_2(t)=\sin t$.

Example 1. Suppose p, $h\in[a,\infty)$ and γ is as in Theorem 3. Suppose further that

$$(43) \qquad \lim_{t\to\infty}h(t)=0, \qquad \int^{\infty}|h'(t)|\,dt<\infty,$$

and $\int^{\infty}|p(t)|\,dt<\infty$. Then Theorem 3 implies that the equation

$$x''(t)+x(t)=p(t)(x(g(t)))^{\gamma}+h(t)$$

has a solution $\bar{x}$ such that

$$\bar{x}(t)=c_1\cos t+c_2\sin t+o(1) \tag{44}$$

for any given constants c_1 and c_2. (Notice that (43) and Dirichlet's theorem imply the convergence of the integrals (37).)

Example 2. It is straightforward to verify that the equation

$$x''+x=\frac{x^2}{t} \tag{45}$$

satisfies the hypotheses of Theorem 4 with $g(t)=t$, $\phi(t)=1/t$, and B_1, $B_2\leqq 1$. Therefore, (45) has a solution $\bar{x}$ such that

$$\bar{x}(t)=c_1\cos t+c_2\sin t+O\left(\frac{1}{t}\right),$$

provided $c_1^2+c_2^2<\frac{1}{8}$. Notice that even though its conclusion would only be of the weaker form (44) anyway, Theorem 3 does not apply here, since (45) does not satisfy (38). This illustrates the point raised in Remark 2.

3. Perturbations of a nonoscillatory equation. If (2) is nonoscillatory, then it has a fundamental system which satisfies the following assumption on some semi-infinite interval, which we take—without loss of generality—to be $[a,\infty)$.

Assumption B. The functions y_1 and y_2 of Assumption A are also positive on $[a,\infty)$ and, if

$$\rho=\frac{y_2}{y_1} \tag{46}$$

then

$$\lim_{t\to\infty}\rho(t)=\infty. \tag{47}$$

Also, in all of the following, either (a) $i=2$ and $j=1$, or (b) $i=1$ and $j=2$. In Case (b) there is a number $\mu<1$ such that $\phi\rho^{\mu}$ is nondecreasing.

Assumptions A and B apply throughout the remainder of the paper.

Note that

$$\rho'=\frac{1}{ry_1^2}>0, \tag{48}$$

from (7) and (46).

The following lemma will be used to prove Theorem 5.

LEMMA 1. *Suppose $F\in C[t_0,\infty)$ for some $t_0\geqq a$ and $\int^{\infty}y_2(t)F(t)\,dt$ converges. Let*

$$\nu(t)=\sup_{\tau\geqq t}\left|\int_{\tau}^{\infty}y_2(s)F(s)\,ds\right|.$$

Then

$$\left|\int_{t_1}^{\infty}[y_2(s)y_1(t)-y_1(s)y_2(t)]F(s)\,ds\right|\leqq\nu(t)y_1(t),\qquad t_0\leqq t\leqq t_1,\tag{49}$$

and

$$\left|\int_t^{\infty}y_1(s)F(s)\,ds\right|\leqq 2\nu(t)/\rho(t),\qquad t\geqq t_0.\tag{50}$$

Proof. With $U(t)=\int_t^{\infty}y_2(s)F(s)\,ds$, integration by parts yields

$$\begin{aligned}&\int_{t_1}^{\infty}[y_2(s)y_1(t)-y_1(s)y_2(t)]F(s)\,ds\\ &\qquad=U(t_1)y_1(t)\left[1-\frac{\rho(t)}{\rho(t_1)}\right]+y_2(t)\int_{t_1}^{\infty}\frac{\rho'(s)}{\rho^2(s)}U(s)\,ds,\end{aligned}\tag{51}$$

where the integral on the right converges absolutely because of (47), (48), and the boundedness of U. Since $|U(s)|\leqq\nu(t_1)$ if $s\geqq t_1$, (51) and the monotonicity of ρ imply (49).

We obtain (50) by writing

$$\int_t^{\infty}y_1(s)F(s)\,ds=U(t)/\rho(t)-\int_t^{\infty}\frac{\rho'(s)}{\rho^2(s)}U(s)\,ds$$

and applying a similar argument. This completes the proof.

THEOREM 5. *Suppose* (11) *and* (12) *hold and there are constants* $T\geqq a$ *and* $M>0$ *such that* f *is continuous and satisfies* (14) *on the set*

$$\Omega_j=\left\{(t,u)\,|\,t\geqq T,\ |u-\bar{y}(g(t))|\leqq M\phi(g(t))y_j(g(t))\right\}.\tag{52}$$

Then: (a) *if* $i=2, j=1$, *and*

$$M>\alpha_2+\beta_2,\tag{53}$$

then (1) *has a solution* $\bar{x}$ *such that*

$$\overline{\lim_{t\to\infty}}\,[\phi(t)y_1(t)]^{-1}|\bar{x}(t)-\bar{y}(t)|\leqq\alpha_2+\beta_2.\tag{54}$$

(b) *If* $i=1, j=2$, *and*

$$M>(\alpha_1+\beta_1)/(1-\mu),\tag{55}$$

then (1) *has a solution* $\bar{x}$ *such that*

$$\overline{\lim_{t\to\infty}}\,[\phi(t)y_2(t)]^{-1}|\bar{x}(t)-\bar{y}(t)|\leqq\frac{\alpha_1+\beta_1}{1-\mu}.\tag{56}$$

Proof. From (11) and (12),

$$\overline{\lim_{t\to\infty}}\,(\phi(t))^{-1}\nu_i(t)\leqq\alpha_i+\beta_i.\tag{57}$$

(See (17).) If $\tau_0 \geqq T$, let

$$D_j = \left\{ x \in C[\tau_0, \infty) \,\middle|\, |\bar{x}(\tau) - \bar{y}(\tau)| \leqq M\phi(\tau) y_j(\tau), \tau \geqq \tau_0 \right\}.$$

We now consider Cases (a) and (b) separately.

(a) Choose $\tau_0 \geqq T$ so that

$$\nu_2(t) \leqq M\phi(t), \qquad t \geqq \tau_0, \tag{58}$$

which is possible because of (53) and (57) with $i=2$. Then choose $t_0 \geqq \tau_0$ to satisfy (22). As in the proof of Theorem 1, (23) holds with $i=2$, and, with $\mathcal{T}$as defined in (24), Lemma 1 implies that

$$|(\mathcal{T}x)(t) - \bar{y}(t)| \leqq \nu_2(t) y_1(t), \qquad t \geqq \tau_0, \quad x \in D_1. \tag{59}$$

This and (58) imply that $\mathcal{T}(D_1) \subset D_1$.

If $D_1 \lim_{k\to\infty} x_k = x$, then the argument given in the proof of Theorem 1 implies that for each $\varepsilon > 0$ there is an N such that (26) holds with $i=2$. This and Lemma 1 with

$$F(s) = f(s, x_k(g(s))) - f(s, x(g(s)))$$

imply that

$$|(\mathcal{T}x_k)(t) - (\mathcal{T}x)(t)| \leqq \varepsilon y_1(t), \qquad t \geqq \tau_0, \quad k \geqq N,$$

so $D_1 \lim_{k\to\infty} \mathcal{T}x_k = \mathcal{T}x$.

Since (23) holds with $i=2$, Lemma 1 (specifically, (50)) implies that

$$\left| \int_t^\infty y_1(s) f(s, x(g(s)))\, ds \right| \leqq \frac{2\nu_2(t)}{\rho(t)}, \qquad t \geqq t_0.$$

Therefore, differentiating (24) and applying routine estimates verifies (5), with

$$\psi = |\bar{y}'| + \nu_2 \left[|y_1'| + 2\frac{|y_2'|}{\rho} \right].$$

Now we conclude that $\mathcal{T}$has a fixed point (function) $\bar{x}$ which satisfies (27), and therefore (1), on (t_0, ∞). Setting $x = \bar{x}$ in (59) and recalling that $\mathcal{T}\bar{x} = \bar{x}$ yields the inequality

$$|\bar{x}(t) - \bar{y}(t)| \leqq \nu_2(t) y_1(t), \qquad t > t_0,$$

so (57) with $i=2$ implies (54). This completes the proof in Case (a).

(b) Choose $\tau_0 \geqq T$ so that

$$\nu_1(t) \leqq M(1-\mu)\phi(t), \qquad t \geqq \tau_0, \tag{60}$$

which is possible because of (55) and (57) with $i=1$. Then choose $t_0 \geqq \tau_0$ to satisfy (22). Now define $\mathcal{T}$on D_2 by

$$(\mathcal{T}x)(t) = \begin{cases} \bar{y}(t) - y_1(t) \displaystyle\int_{t_0}^t \rho'(\tau) \left(\int_\tau^\infty y_1(s) f(s, x(g(s)))\, ds \right) d\tau, & t \geqq t_0, \\ \bar{y}(t), \quad \tau_0 \leqq t \leqq t_0, \end{cases} \tag{61}$$

where the second line is vacuous if $\tau_0 > t_0$. (See (46) and (48).) Then

$$(\mathscr{T}x)(t)-\bar{y}(t)=0, \qquad \tau_0 \leqq t \leqq t_0, \tag{62}$$

while (23) with $i=1$ implies that

$$|(\mathscr{T}x)(t)-\bar{y}(t)| \leqq y_1(t)\int_{t_0}^{t} \rho'(\tau)\nu_1(\tau)\,d\tau, \qquad t \geqq t_0. \tag{63}$$

If $t \geqq a$, let

$$\hat{\nu}_1(t)=\sup_{\tau \geqq t} \nu_1(\tau)/\phi(\tau). \tag{64}$$

Then, if $t \geqq t_1 \geqq a$,

$$\begin{aligned}\int_{t_1}^{t} \rho'(\tau)\nu_1(\tau)\,d\tau &\leqq \hat{\nu}_1(t_1)\int_{t_1}^{t} \rho'(\tau)\phi(\tau)\,d\tau\\ &\leqq \hat{\nu}_1(t_1)(\rho(t))^{\mu}\phi(t)\int_{t_1}^{t} \rho'(\tau)(\rho(\tau))^{-\mu}\,d\tau,\end{aligned}$$

since $\phi\rho^{\mu}$ is nondecreasing. Since $\mu<1$ and $\rho'>0$, this implies that

$$\int_{t_1}^{t} \rho'(\tau)\nu_1(\tau)\,d\tau \leqq \hat{\nu}_1(t_1)\phi(t)\frac{\rho(t)}{1-\mu}. \tag{65}$$

Setting $t_1=t_0$ here and recalling (60), (63), and (64) shows that

$$|(\mathscr{T}x)(t)-\bar{y}(t)| \leqq M\phi(t)y_2(t), \qquad t \geqq t_0, \quad x \in D_2,$$

which, with (62), implies that $\mathscr{T}(D_2) \subset D_2$.

If $D_2 \lim_{k\to\infty} x_k = x$, the argument used in the proof of Theorem 1 again implies that for each $\varepsilon>0$ there is an N such that (26) holds with $i=1$. This and (61) imply that

$$|(\mathscr{T}x_k)(t)-(\mathscr{T}x)(t)| \leqq \varepsilon y_2(t), \qquad t \geqq \tau_0, \quad k \geqq N,$$

which implies that $D_2 \lim_{k\to\infty} \mathscr{T}x_k = \mathscr{T}x$.

Differentiating (61) and recalling (23) with $i=1$ shows that

$$|(\mathscr{T}x)'(t)| \leqq |\bar{y}'(t)| + |y_1'(t)|\int_{t_0}^{t} \rho'(\tau)\nu_1(\tau)\,d\tau + y_1(t)\rho'(t)\nu_1(t)$$

if $x \in D_2$ and $t \geqq t_0$. Since $\mathscr{T}x=\bar{y}$ on $[\tau_0, t_0]$ for every x in D_2, this is enough to imply the conclusion of (iii) in §1, so $\mathscr{T}$ has a fixed point (function) $\bar{x}$ in D_2. From (61),

$$\bar{x}(t)=\bar{y}(t)-y_1(t)\int_{t_0}^{t} \rho'(\tau)\left(\int_{\tau}^{\infty} y_1(s)f(s,\bar{x}(g(s)))\,ds\right)d\tau, \qquad t \geqq t_0.$$

This function satisfies (1) on (t_0,∞), and, from (63) with $x=\bar{x}(=\mathscr{T}\bar{x})$ and $i=1$,

$$|\bar{x}(t)-\bar{y}(t)| \leqq y_1(t)\int_{t_0}^{t} \rho'(\tau)\nu_1(\tau)\,d\tau.$$

This and (65) imply that

$$|\bar{x}(t)-\bar{y}(t)| \leqq y_1(t)\int_{t_0}^{t_1} \rho'(\tau)\nu_1(\tau)\,d\tau + \hat{\nu}_1(t_1)\phi(t)y_2(t)/(1-\mu) \tag{66}$$

if $t_0 \leqq t_1 \leqq t$. From (47), (48), and our assumption on $\phi\rho^{\mu}$, $\lim_{t\to\infty}\phi(t)\rho(t)=\infty$; hence, from (66),

$$\overline{\lim_{t\to\infty}}\,[\phi(t)y_2(t)]^{-1}|\bar{x}(t)-\bar{y}(t)| \leqq \hat{\nu}_1(t_1)/(1-\mu)$$

for every $t_1 \geqq t_0$. Letting $t_1\to\infty$ and recalling (57) (with $i=1$) and (64) therefore implies (56). This completes the proof.

Setting $\phi=1$ in Theorem 5, and noting again that this means that $\alpha_i=\beta_i=0$, yields the following corollary. (Here we reemphasize that Assumption B applies; specifically, that either $i=2$ and $j=1$ or $i=1$ and $j=2$.)

COROLLARY 3. *Suppose* $\int^\infty y_i(t)f(t,\bar{y}(g(t)))\,dt$ *converges and* $\int^\infty y_i(t)\sigma(t)\,dt<\infty$, *where* σ *is positive and continuous on* $[a,\infty)$. *Suppose also that there are constants* $T=a$ *and* $M>0$ *such that* f *is continuous and satisfies* (14) *on the set*

$$\bar{\Omega}_j=\left\{(t,u)\,\middle|\,t\geqq T,\,|u-\bar{y}(g(t))|\leqq My_j(g(t))\right\}.$$

Then (1) *has a solution* $\bar{x}$ *such that* $\bar{x}(t)=\bar{y}(t)+o(y_j(t))$.

COROLLARY 4. *Suppose* $h\in C[a,\infty)$, F *is continuous on* $[a,\infty)\times(0,\infty)$, *and* $|F(t,u)|$ *is either* (i) *nondecreasing in* u *for each* t, *or* (ii) *nonincreasing in* u *for each* t. *Suppose also that*

$$\int^\infty y_i(t)h(t)\,dt \tag{67}$$

converges, and

$$\int^\infty y_i(t)|F(t,\delta y_j(g(t)))|\,dt<\infty$$

for some $\delta>0$. *Then the equation*

$$(r(t)x'(t))'+q(t)x(t)=F(t,x(g(t)))+h(t)$$

has a solution $\bar{x}$ *such that*

$$\lim_{t\to\infty}\frac{\bar{x}(t)}{y_j(t)}=c_j, \tag{68}$$

provided $0<c_j<\delta$ *in Case* (i), *or* $c_j>\delta$ *in Case* (ii).

Proof. It is straightforward to verify that the present assumptions imply those of Corollary 3 with $\bar{y}=c_jy_j$, $f(t,u)=F(t,u)+h(t)$, and $\sigma(t)=2|F(t,\delta y_j(g(t)))|$. In Case (i), choose $M<\min\{c_j,\delta-c_j\}$; in Case II, $M<c_j-\delta$. In either case, let $T=a$.

Kusano and Naito [3] have given necessary and sufficient conditions for the equation

$$(r(t)x'(t))'=f(t,x(g(t))) \tag{69}$$

to have nonoscillatory solutions with specific asymptotic behavior, under the assumption that (69) is sublinear or superlinear (see [3] for definitions of these terms), where $g(t)\leqq t$ and $\lim_{t\to\infty}g(t)=\infty$. Kusano and Onose [4] have obtained analogous results for the case where $g(t)\geqq t$. Corollary 4 essentially contains the sufficiency halves of [3, Thms. 1,2] and [4, Thms. 1,2,5,6]. The reader who wishes to check this should let

$$y_1(t)=1,\qquad y_2(t)=\int_a^t(r(s))^{-1}\,ds\quad\text{if}\int^\infty(r(t))^{-1}\,dt=\infty,$$

or

$$y_1(t)=\int_t^\infty (r(s))^{-1}ds, \qquad y_2(t)=1 \quad \text{if} \int^\infty (r(t))^{-1}dt<\infty.$$

The next corollary follows easily from Corollary 4. It is perhaps noteworthy in that it deals with the generalized Emden–Fowler equation without the usual assumption that $\gamma>0$. (For other results concerning such equations with arbitrary γ, see [1], [5], [6], [7], and [8].) This observation applies also to Theorems 7 and 8 below.

COROLLARY 5. *Suppose* $p,\ h\in C[a,\infty)$, γ *is any real number, and* c_j *is any positive constant. Suppose further that* (67) *converges, and*

$$\int^\infty y_i(t)|p(t)|(y_j(g(t)))^\gamma dt<\infty.$$

Then (36) *has a solution* $\bar{x}$ *which satisfies* (68).

This corollary essentially contains the sufficiency halves of [4, Corollaries 1,4].

Theorem 5 implies the next result in much the same way that Theorem 1 implies Theorem 2. We omit the proof.

THEOREM 6. *Suppose* (11) *holds. Let* λ *be nonnegative and continuous on* $[a,\infty)$, *and*

$$\overline{\lim_{t\to\infty}}(\phi(t))^{-1}\int_t^\infty y_i(s)\lambda(s)y_j(g(s))\phi(g(s))\,ds=b_i. \tag{70}$$

Suppose further that there are constants $T\geqq a$ *and* M *such that* f *is continuous and satisfies* (35) *on the set* Ω_j *in* (52). *Then* (1) *has a solution* $\bar{x}$ *such that*

(a) $\overline{\lim}_{t\to\infty}[\phi(t)y_1(t)]^{-1}|\bar{x}(t)-\bar{y}(t)|\leqq\alpha_2+Mb_2$ *if* $i=1,\ j=1,\ b_2<1$, *and* $M>\alpha_2/(1-b_2)$; *or*

(b) $\overline{\lim}_{t\to\infty}[\phi(t)y_2(t)]^{-1}|\bar{x}(t)-\bar{y}(t)|\leqq(\alpha_1+Mb_1)/(1-\mu)$ *if* $i=1,\ j=2,\ b_1<1-\mu$, *and* $M>\alpha_1/(1-\mu-b_1)$.

We close by applying Theorem 6 to the generalized Emden–Fowler equation (36).

THEOREM 7. *Suppose* $p,\ h\in C[a,\infty)$, γ *is any real number, and*

$$\overline{\lim_{t\to\infty}}(\phi(t))^{-1}\int_t^\infty y_i(s)|p(s)|(y_j(g(s)))^\gamma\phi(g(s))\,ds=B_i<\infty. \tag{71}$$

Suppose also that

$$\int_t^\infty y_i(s)h(s)\,ds=O(\phi(t))$$

and

$$\int_t^\infty y_i(s)p(s)(y_j(g(s)))^\gamma ds=O(\phi(t)), \tag{72}$$

and let

$$\overline{\lim_{t\to\infty}}(\phi(t))^{-1}\left|\int_t^\infty y_i(s)\left[p(s)(c_jy_j(g(s)))^\gamma+h(s)\right]ds\right|=\alpha_i,$$

where c_j *is a given positive constant.*

(a) *If* $i=2$ *and* $j=1$, *suppose also that*

$$|\gamma|c_1^{\gamma-1}B_2<1$$

and

$$M > \alpha_2/\big(1-|\gamma|c_1^{\gamma-1}B_2\big). \tag{73}$$

Then (36) *has a solution* $\bar{x}$ *such that*

$$\varlimsup_{t\to\infty}[\phi(t)y_1(t)]^{-1}|\bar{x}(t)-c_1y_1(t)| \leqq \alpha_2+M|\gamma|c_1^{\gamma-1}B_2.$$

(b) *If* $i=1$ *and* $j=2$, *suppose also that*

$$|\gamma|c_2^{\gamma-1}B_1<1-\mu$$

and

$$M > \alpha_1\big(1-\mu-|\gamma|c_2^{\gamma-1}B_1\big). \tag{74}$$

Then (36) *has a solution* $\bar{x}$ *such that*

$$\varlimsup_{t\to\infty}[\phi(t)y_2(t)]^{-1}|\bar{x}(t)-c_2y_2(t)| \leqq \big(\alpha_1+M|\gamma|c_2^{\gamma-1}B_1\big)/(1-\mu).$$

Proof. Since Corollary 5 implies the conclusions if $\phi=1$, we assume that

$$\lim_{t\to\infty}\phi(t)=0. \tag{75}$$

Choose M to satisfy (73) or (74), whichever is appropriate, and then choose T so that $M\phi(g(t))<c_j$ if $t\geqq T$. (This is possible because of (6) and (75).) With this M and T and $\bar{y}=c_jy_j$, it is easy to show that if (t,u) is in Ω_j as defined in (52), then $0<[c_j-M\phi(g(t))]y_j(g(t))\leqq u\leqq[c_j+M\phi(g(t))]y_j(g(t))$. Therefore, the function f in (39) is continuous on Ω_j, and, by the mean value theorem, satisfies (35), with

$$\lambda(t)=|\gamma p(t)|\big[c_j\pm M\phi(g(t))\big]^{\gamma-1}\big(y_j(g(t))\big)^{\gamma-1}, \tag{76}$$

where the plus applies if $\gamma\geqq1$, the minus if $\gamma<1$. In either case, (6), (71), (75), and (76) imply (70) with $b_i=|\gamma|c_j^{\gamma-1}B_i$. Now Theorem 6 implies the stated conclusion.

THEOREM 8. *Suppose* p, $h\in C[a,\infty)$, γ *is an arbitrary real number, and*

$$\int^{\infty}\frac{\rho'(g(t))}{\rho^2(g(t))}|g'(t)|\,dt<\infty. \tag{77}$$

Suppose also that

$$\varlimsup_{t\to\infty}(\phi(t))^{-1}\int_t^{\infty}y_2(s)|p(s)|(y_2(g(s)))^{\gamma-1}y_1(g(s))\phi(g(s))\,ds=B_2<\infty, \tag{78}$$

$$\int_t^{\infty}y_2(s)h(s)\,ds=O(\phi(t)), \tag{79}$$

and

$$E(t)=\int_t^{\infty}y_2(s)p(s)(y_2(g(s)))^{\gamma}ds=O(\phi(t)). \tag{80}$$

Let c_2 *be a positive constant such that*

$$|\gamma|c_2^{\gamma-1}B_2<1,$$

and let $\bar{y}=c_1y_1+c_2y_2$, *where* c_1 *is arbitrary. Then the quantity*

$$\alpha_2=\overline{\lim_{t\to\infty}}(\phi(t))^{-1}\left|\int_t^{\infty}y_2(s)\left[p(s)(\bar{y}(g(s)))^{\gamma}+h(s)\right]ds\right| \tag{81}$$

exists and is finite; *moreover*, $\alpha_2=0$ *if* (79) *and* (80) *hold with "O" replaced by "o." Furthermore*, *if*

$$M>\alpha_2/\left(1-|\gamma|c_2^{\gamma-1}B_2\right), \tag{82}$$

then (36) *has a solution* $\bar{x}$ *such that*

$$\overline{\lim_{t\to\infty}}\left[\phi(t)y_1(t)\right]^{-1}|\bar{x}(t)-c_1y_1(t)-c_2y_2(t)|\leqq\alpha_2+M|\gamma|c_2^{\gamma-1}B_2.$$

Proof. From (80) and integration by parts,

(83)

$$\begin{aligned}&\int_t^{\infty}y_2(s)p(s)(\bar{y}(g(s)))^{\gamma}ds\\&\quad=-\int_t^{\infty}E'(s)\left(\frac{\bar{y}(g(s))}{y_2(g(s))}\right)^{\gamma}ds\\&\quad=E(t)\left(\frac{\bar{y}(g(t))}{y_2(g(t))}\right)^{\gamma}-\gamma c_1\int_t^{\infty}E(s)\left(\frac{\bar{y}(g(s))}{y_2(g(s))}\right)^{\gamma-1}\frac{\rho'(g(s))}{\rho^2(g(s))}g'(s)\,ds\end{aligned}$$

(see (46)), since $\lim_{t\to\infty}E(t)=0$ and

$$\lim_{t\to\infty}\frac{\bar{y}(g(t))}{y_2(g(t))}=c_2.$$

The integral on the right of (83) converges because of (77). From this and (79) it is easy to verify that α_2 in (81) has the stated properties.

Now choose M to satisfy (82), and then choose T so that

$$\bar{y}(g(t))\geqq M\phi(g(t))y_1(g(t)),\qquad t\geqq T.$$

(This is possible even if $\phi=1$, because of (6), (47), and the assumption that $c_2>0$.) With this M and T, it is easy to show that if (t,u) is in Ω_1 as defined in (52), then

$$0<\bar{y}(g(t))-M\phi(g(t))y_1(g(t))\leqq u\leqq\bar{y}(g(t))+M\phi(g(t))y_1(g(t)).$$

Therefore, the function f in (39) is continuous on Ω_1, and, again by the mean value theorem, satisfies (35) with

$$\lambda(t)=|\gamma p(t)|\left[\bar{y}(g(t))\pm M\phi(g(t))y_1(g(t))\right]^{\gamma-1},$$

where the plus applies if $\gamma\geqq1$, the minus if $\gamma<1$. In either case, since the quantity in brackets behaves asymptotically like

$$(c_2y_2(g(t)))^{\gamma-1},$$

(78) implies (70) with $i=2$ and $b_2=|\gamma|c_2^{\gamma-1}B_2$. Now part (a) of Theorem 6 implies the stated conclusion.

Remark 3. Since $\rho' > 0$ and $\lim_{t\to\infty}\rho(t)=\infty$, (77) is automatically satisfied if $g'(t)>0$ for t sufficiently large.

Remark 4. Theorems 7 and 8 show that integrability conditions involving other than forcing functions may permit conditionally convergent integrals, since (71) does not imply that the integral in (72) converges absolutely if $\lim_{t\to\infty}\phi(t)=0$, and (78) does not imply that the integral in (80) converges absolutely, even if $\phi=1$.

REFERENCES

[1] J. R. Graef, M. K. Grammatikopolous and P. W. Spikes, *On the positive solutions of a higher order differential equation with a discontinuity*, Internat. J. Math. Math. Sci., 5 (1982), pp. 263–273.

[2] T. G. Hallam, *Asymptotic relationships between the solutions of two second order differential equations*, Ann. Polon. Math., 24 (1971), pp. 295–300.

[3] T. Kusano and M. Naito, *Nonlinear oscillation of second order differential equations with retarded argument*, Ann. Mat. Pura Appl., 106 (1976), pp. 171–185.

[4] T. Kusano and H. Onose, *Nonlinear oscillation of second order functional differential equations with advanced argument*, J. Math. Soc. Japan, 29 (1977), pp. 541–559.

[5] T. Kusano and C. A. Swanson, *Asymptotic properties of semilinear elliptic equations*, Funkcial. Ekvac., 26 (1983), pp. 115–129.

[6] S. Taliaferro, *On the positive solutions of* $y''+\phi(t)y^{-\lambda}=0$, Nonlinear Anal., 2 (1978), pp. 437–446.

[7] W. F. Trench, *Asymptotic integration of* $y^{(n)}+P(t)y^{\gamma}=0$ *under mild integral conditions*, Funkcial. Ekvac., 26 (1983), pp. 197–209.

[8] ———, *Asymptotic behavior of solutions of an nth order differential equation*, Rocky Mount. J. Math., 14 (1984), pp. 441–450.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

THE CONDITION OF ORDINARY INTEGRAL CONVERGENCE IN THE ASYMPTOTIC THEORY OF LINEAR DIFFERENTIAL EQUATIONS WITH ALMOST CONSTANT COEFFICIENTS*

JAROMÍR ŠIMŠA†

Abstract. In this paper, we continue the work started in [2], wherethe asymptotic integration of nth order linear differential equations was considered under smallness conditions, expressed in terms of ordinary integral convergence. The method of integration in [2] was based on using the Banach contraction principle for a linear nonhomogeneous operator in a Banach space. In this paper, we use the same principle for a nonlinear operator which corresponds to the equation for the logarithmic derivative of the original dependent variable. The result proved concerns a class of equations considered in [2] under essential additional restrictions.

1. Introduction. This paper deals with asymptotic estimates for solutions of a scalar differential equation of the form

$$x^{(n)}+[a_1+p_1(t)]x^{(n-1)}+\cdots+[a_{n-1}+p_{n-1}(t)]x'+[a_n+p_n(t)]x=0, \tag{1.1}$$

where a_k are complex numbers and $p_k(t)$ are continuous complex-valued functions defined on the half-line $0\leqq t<\infty$. We assume that the functions $p_k(t)$ are "small" perturbations in the sense that the integrals

$$\int^{\infty} p_k(t)t^q dt \tag{1.2}$$

converge (possibly not absolutely) for some real $q\geqq 0$.

It is well known (see [1, Thm. 17.2]) that if the real parts of the roots λ_j of the characteristic equation

$$\lambda^n+a_1\lambda^{n-1}+\cdots+a_{n-1}\lambda+a_n=0 \tag{1.3}$$

are distinct and if the integrals (1.2) converge absolutely, then there exist n solutions $x_j(t)$ of (1.1) such that

$$x_j^{(k)}(t)=\left(\lambda_j^k+o(t^{-q})\right)\exp(\lambda_j t), \qquad 0\leqq k\leqq n-1, \quad \text{as } t\to\infty. \tag{1.4}$$

Our first attempt to extend the validity of this classical assertion to the class of equations (1.1) with conditionally convergent integrals (1.2) was made in [2], where the following theorem was established.

THEOREM A. *Suppose that the real parts of the roots $\lambda_1,\cdots,\lambda_n$ of* (1.3) *are distinct. Let the complex-valued functions $p_k(t)$ be continuous for $t\geqq 0$ and satisfy the following conditions*:

(i) $\int^{\infty}|p_1(t)|dt<\infty$.

(ii) *The integrals* (1.2) *converge* (*perhaps conditionally*) *for some nonnegative constant q.*

* Received by the editors April 5, 1983, and in revised form December 14, 1983.

† Department of Mathematical Analysis, J. E. Purkyně University, Janáčkovo nám. 2a, 662 95 Brno, Czechoslovakia.

(iii) *If* $0 \leqq q < 1$ *in* (ii), *then*

$$\int^{\infty} t^{-q}\left|\int_t^{\infty} p_k(s)s^q ds\right| dt < \infty, \qquad 2 \leqq k \leqq n.$$

Then (1.1) *has n solutions* $x_1(t), \cdots, x_n(t)$ *satisfying* (1.4).

As shown in [3], the condition (iii) cannot be omitted in Theorem A. For the reader's convenience, we state the counter-example given in [3] as the following theorem.

THEOREM B. *Let q and r be any real numbers satisfying* $2q-1<r<1$ *or* $0<r<1-2q$, *according as* $\frac{1}{2} \leqq q < 1$ *or* $0 \leqq q < \frac{1}{2}$. *Then there exists a real function* $p(t)$ *continuous for* $t \geqq 0$ *such that the integral* $\int^{\infty} p(t)t^q dt$ *converges and the equation* $x''-x=p(t)x$ *has a solution* $x(t)$ *satisfying*

$$x(t)=\begin{cases}(1+t^{-r}+o(t^{-r}))e^t & \text{if } \frac{1}{2} \leqq q < 1,\\ \exp(t-t^r+o(t^r)) & \text{if } 0 \leqq q < \frac{1}{2}.\end{cases}$$

We now state our main result.

THEOREM 1. *Assume that all conditions of Theorem* A *are satisfied except condition* (iii). *If* $0 \leqq q < 1$ *in* (ii), *then* (1.1) *has n solutions* $x_1(t), \cdots, x_n(t)$ *satisfying*

$$x_j(t)=\begin{cases}(1+o(t^{1-2q}))\exp(\lambda_j t) & \text{if } \frac{1}{2}<q<1,\\ \exp(\lambda_j t+o(\log t)) & \text{if } q=\frac{1}{2},\\ \exp(\lambda_j t+o(t^{1-2q})) & \text{if } 0 \leqq q < \frac{1}{2},\end{cases} \tag{1.5}$$

and

$$\frac{x_j^{(k)}(t)}{x_j(t)}=\lambda_j^k+o(t^{-q}) \qquad (1 \leqq k \leqq n-1). \tag{1.6}$$

Remark 1. Theorem A was proved in [2] under weaker conditions imposed on the roots λ_j of (1.3). Namely, we assumed that $\lambda_j \neq \lambda_m$ for any j and m instead of $\mathrm{Re}\lambda_j \neq \mathrm{Re}\lambda_m$. Then the assertion of Theorem A holds if the integrals (1.2) converge along with the integrals $\int^{\infty} p_k(t)\exp(i\beta t)t^q dt$ $(1 \leqq k \leqq n)$, where $\beta=\beta_{jm}=\mathrm{Im}(\lambda_j-\lambda_m)$ whenever $\mathrm{Re}\lambda_j=\mathrm{Re}\lambda_m$. Unfortunately, the condition $\mathrm{Re}\lambda_j \neq \mathrm{Re}\lambda_m$ seems to be "necessary" for our proof of Theorem 1. Thus the problem of an extension of Theorem 1 to the case of "multiple values" of $\mathrm{Re}\lambda_j$ is unsolved.

2. The main ideas of the proof. Our proof of Theorem 1 is based on using the Banach contraction principle [1, p. 404]. Although this principle is very simple, the proofs of needed estimates are rather complicated because of the weak condition of ordinary integral convergence. This is why we now restrict our attention to the basic arguments leading to the conclusion of Theorem 1. The propositions stated here as Lemmas 2.1–2.3 will be proved in detail in §3.

To avoid unnecessary subscripts, we let r be a fixed integer $(1 \leqq r \leqq n)$ throughout the proof. We will show that under the hypotheses of Theorem 1, there exists a solution $x=x_r$ of (1.1) satisfying (1.5) and (1.6) with $j=r$.

First we make a transformation of the form

$$(2.1) \qquad x=\exp(\lambda_r t)y, \qquad y(t)=C_0\exp\left(\int_{t_0}^{t}u(s)\,ds\right),$$

where C_0 is a nonzero constant, $u(t)$ is an unknown function and the number $t_0\geqq 1$ will be chosen later. In the following, assume that $t\geqq t_0$ and $t\to\infty$ in all asymptotics.

It follows from (2.1) that

$$(2.2) \qquad x^{(k)}=\exp(\lambda_r t)\sum_{j=0}^{k}\binom{k}{j}\lambda_r^{k-j}y^{(j)} \qquad (1\leqq k\leqq n)$$

and

$$(2.3) \qquad y^{(k)}=\left(u^{(k-1)}+g_k[u]\right)y \qquad (1\leqq k\leqq n),$$

where $g_1[u]=0$ and

$$(2.4) \qquad g_{k+1}[u]=uu^{(k-1)}+ug_k[u]+g_k'[u] \qquad (1\leqq k\leqq n-1).$$

So we have $g_2[u]=u^2$, $g_3[u]=u^3+3uu'$, $\cdots$. Note that the function $g_k[u]$ is a sum of expressions of the form

$$(2.5) \qquad D^{k}_{\alpha_1,\cdots,\alpha_m}u^{(\alpha_1)}u^{(\alpha_2)}\cdots u^{(\alpha_m)},$$

where $D^{k}_{\alpha_1,\cdots,\alpha_m}$ are constants, $0\leqq\alpha_j\leqq k-2$ $(1\leqq j\leqq m)$ and $m=2,3,\cdots,k$, for $k=2,3,\cdots,n$. This fact follows from (2.4) by induction.

Equation (1.1) is transformed by means of (2.1) to a nonlinear equation of $(n-1)$th order. By (2.2) and (2.3), this equation is of the form

$$(2.6) \qquad b_0u^{(n-1)}+b_1u^{(n-2)}+\cdots+b_{n-1}u=-R[u](t),$$

where

$$(2.7) \qquad b_k=\binom{n}{k}\lambda_r^k+\sum_{j=1}^{k}a_j\binom{n-j}{n-k}\lambda_r^{k-j} \qquad (0\leqq k\leqq n-1),$$

$$(2.8) \quad R[u](t)=\tilde{p}_n(t)+\sum_{k=1}^{n-1}\tilde{p}_k(t)u^{(n-k-1)}+\sum_{k=0}^{n-1}b_kg_{n-k}[u]+\sum_{k=1}^{n-1}\tilde{p}_k(t)g_{n-k}[u],$$

and

$$(2.9) \qquad \tilde{p}_k(t)=\sum_{j=1}^{k}\binom{n-j}{n-k}\lambda_r^{k-j}p_j(t) \qquad (1\leqq k\leqq n).$$

Note that the new functions $\tilde{p}_k(t)$ satisfy the same conditions stated in Theorem 1 as well as the functions $p_k(t)$. To simplify our notation, we shall write $p_k(t)$ instead of $\tilde{p}_k(t)$.

An easy computation based on (2.7) shows that the characteristic equation for linear differential operator on the left-hand side of (2.6) has $(n-1)$ distinct roots

$$(2.10) \qquad \mu_j=\lambda_j-\lambda_r, \qquad j\in J=\{1,2,\cdots,r-1,r+1,r+2,\cdots,n\}$$

with nonzero real parts.

In what follows we apply the method of the variation of constants [1, p. 64] to (2.6). Namely, we will consider an operator equation

$$u=T[u], \tag{2.11}$$

where

$$T[u](t)=-\sum_{j\in J} c_j \exp(\mu_j t) I_j[u](t) \tag{2.12}$$

and

$$I_j[u](t)=\int_{t_{0j}}^{t} R[u](s)\exp(-\mu_j s)\,ds \qquad (j\in J). \tag{2.13}$$

The numbers c_j in (2.12) satisfy the system

$$\sum_{j\in J} c_j\mu_j^k=\begin{cases}0 & \text{if } 0\leqq k\leqq n-3,\\ 1 & \text{if } k=n-2\end{cases} \tag{2.14}$$

and the limits t_{0j} in the integrals (2.13) are equal to t_0 or ∞, according to whether $\operatorname{Re}\mu_j<0$ or $\operatorname{Re}\mu_j>0$.

First we show that each solution of (2.11) is a solution of (2.6). In fact, if the integrals (2.13) exist for some function $u(t)$, then, by (2.12)–(2.14),

$$T^{(k)}[u](t)=-\sum_{j\in J} c_j\mu_j^k \exp(\mu_j t) I_j[u](t) \qquad (0\leqq k\leqq n-2), \tag{2.15}$$

$$T^{(n-1)}[u](t)=-\sum_{j\in J} c_j\mu_j^{n-1} \exp(\mu_j t) I_j[u](t)-R[u](t) \tag{2.16}$$

and, therefore,

$$T^{(n-1)}[u](t)+b_1T^{(n-2)}[u](t)+\cdots+b_{n-1}T[u](t)=-R[u](t).$$

Consequently, (2.11) implies (2.6).

We will consider the operator T in the space $U[t_0,\infty)$ of all functions $u(t)$ in $C^{n-2}[t_0,\infty)$ satisfying

$$u^{(k)}(t)=O(t^{-q}), \qquad 0\leqq k\leqq n-2,$$

which is a Banach space with respect to the norm

$$\|u\|=\sup_{t\geqq t_0}\sum_{k=0}^{n-2}|u^{(k)}(t)|t^q. \tag{2.17}$$

Denote by $S(t_0,\varepsilon)$ the closed sphere of the space $U[t_0,\infty)$

$$S(t_0,\varepsilon)=\{u\in U[t_0,\infty):\|u\|\leqq\varepsilon\},$$

where the number ε $(0<\varepsilon\leqq 1)$ will be chosen later.

We will show that the operator T is a contraction mapping of the set $S(t_0,\varepsilon)$ into itself, for suitable t_0 and ε. By (2.15), it will be sufficient to estimate the integrals

$I_j[u](t)$. According to (2.8) and (2.13), we may write

$$(2.18)\qquad I_j[u](t)=L_j(t)+\sum_{k=1}^{n-1}L_{jk}[u](t)+\sum_{k=0}^{n-1}b_kM_{jk}[u](t)+\sum_{k=1}^{n-1}N_{jk}[u](t),$$

and

$$(2.19)\qquad L_j(t)=\int_{t_{0j}}^{t}p_n(s)\exp(-\mu_js)\,ds,$$

$$(2.20)\qquad L_{jk}[u](t)=\int_{t_{0j}}^{t}p_k(s)u^{(n-k-1)}(s)\exp(-\mu_js)\,ds,$$

$$(2.21)\qquad M_{jk}[u](t)=\int_{t_{0j}}^{t}g_{n-k}[u](s)\exp(-\mu_js)\,ds$$

and

$$(2.22)\qquad N_{jk}[u](t)=\int_{0j}^{t}p_k(s)g_{n-k}[u](s)\exp(-\mu_js)\,ds.$$

In what follows, all constants $C_1, C_2, \cdots$ are independent of t_0 and ε, $0<\varepsilon\leqq 1\leqq t_0<\infty$.

LEMMA 2.1. *Under the hypotheses of Theorem* 1, *the functions* $L_j(t)$, $L_{jk}[u](t)$, $M_{jk}[\tilde{u}](t)$ *and* $N_{jk}[\tilde{u}](t)$ *are defined on the half-line* $t_0\leqq t<\infty$, *for any* u *in* $U[t_0,\infty)$ *and any* $\tilde{u}$ *in* $S(t_0,\varepsilon)$. *Moreover, there exists a function* $m(t_0,t)$ *of type* (*) (*see Definition* 1 *in* §3), *independent of* ε $(0<\varepsilon\leqq 1)$, *such that the estimates*

$$(2.23)\qquad |L_j(t)|\leqq m(t_0,t)\exp(\delta_jt)t^{-q},$$

$$(2.24)\qquad |L_{jk}[u](t)|\leqq \|u\|m(t_0,t)\exp(\delta_jt)t^{-q},$$

$$(2.25)\qquad |M_{jk}[\tilde{u}](t)-M_{jk}[\tilde{\tilde{u}}](t)|\leqq C_1\varepsilon\|\tilde{u}-\tilde{\tilde{u}}\|\exp(\delta_jt)t^{-2q}$$

and

$$(2.26)\qquad |N_{jk}[\tilde{u}](t)-N_{jk}[\tilde{\tilde{u}}](t)|\leqq \|\tilde{u}-\tilde{\tilde{u}}\|m(t_0,t)\exp(\delta_jt)t^{-3q}$$

hold for any u *in* $U[t_0,\infty)$ *and* $\tilde{u},\tilde{\tilde{u}}$ *in* $S(t_0,\varepsilon)$. *The numbers* δ_j *in* (2.23)–(2.26) *are defined by*

$$(2.27)\qquad \delta_j=-\operatorname{Re}\mu_j=\operatorname{Re}(\lambda_r-\lambda_j),\qquad j\in J.$$

The proof of Lemma 2.1 is given in §3.

From (2.15), (2.18) and from (2.23)–(2.26) with $\tilde{u}=u$ and $\tilde{\tilde{u}}=0$ we find that

$$(2.28)\qquad |T^{(k)}[u](t)|\leqq C_2(1+\|u\|)m(t_0,t)t^{-q}+C_3\varepsilon\|u\|t^{-2q},$$

for any u in $S(t_0,\varepsilon)$, $k=0,1,\cdots,n-2$. Since $m(t_0,t)$ is nonincreasing in t, it follows from (2.17) and (2.28) that the function $T[u](t)$ lies in $U[t_0,\infty)$ and its norm satisfies

$$(2.29)\qquad \|T[u]\|\leqq(n-1)\left[C_2(1+\|u\|)m(t_0,t_0)+C_3\varepsilon\|u\|\right].$$

Moreover, in the case $q>0$,

$$(2.30)\qquad T^{(k)}[u](t)=o(t^{-q})\qquad(0\leqq k\leqq n-2),$$

because $m(t_0,t)\to 0$ as $t\to\infty$.

The asymptotics (2.30) do not follow from (2.28) if $q=0$. Then we need a weaker property of the operator T stated as follows.

LEMMA 2.2. *Assume that* $q=0$. *Let*

$$\|u\|_t=\sup_{t_1\geqq t}\sum_{k=0}^{n-2}\left|u^{(k)}(t_1)\right| \qquad (t\geqq t_0), \tag{2.31}$$

and

$$\|u\|_\infty=\lim_{t\to\infty}\|u\|_t=\inf_{t\geqq t_0}\|u\|_t, \tag{2.32}$$

for any u *in* $U[t_0,\infty)$. *Then there exists a* C_4 *such that*

$$\|T[u]\|_\infty\leqq C_4(\|u\|_\infty)^2 \tag{2.33}$$

holds for any u *in* $S(t_0,\varepsilon)$.

The proof of Lemma 2.2 is given in §3.

Now we first choose a number ε and then a number t_0 so that

$$4(n-1)C_3\varepsilon<1, \qquad C_4\varepsilon<1 \quad (\text{if } q=0), \tag{2.34}$$

and

$$4(n-1)C_2m(t_0,t_0)\leqq\varepsilon, \tag{2.35}$$

which is possible because $m(t_0,t_0)\to 0$ as $t_0\to\infty$.

Since $\|u\|\leqq\varepsilon\leqq 1$ for any u in $S(t_0,\varepsilon)$, it follows from (2.29), (2.34) and (2.35) that $\|T[u]\||\leqq 3\varepsilon/4$, hence the operator T maps the set $S(t_0,\varepsilon)$ into itself. If we put $u=\tilde u-\tilde{\tilde u}$ in (2.24), then, by (2.15), (2.18) and (2.24)–(2.27), we find that

$$\left|T^{(k)}[\tilde u](t)-T^{(k)}[\tilde{\tilde u}](t)\right|\leqq\left(C_2m(t_0,t)+C_3\varepsilon t^{-q}\right)\|\tilde u-\tilde{\tilde u}\|t^{-q}, \qquad 0\leqq k\leqq n-2,$$

which along with (2.17), (2.34) and (2.35) leads to the estimate $\|T[\tilde u]-T[\tilde{\tilde u}]\||\leqq\|\tilde u-\tilde{\tilde u}\|/2$. Thus T is a contraction mapping of the closed set $S(t_0,\varepsilon)$ into itself. According to the Banach contraction principle, there exists u_r in $S(t_0,\varepsilon)$ such that $T[u_r]=u_r$. This means that $u_r(t)$ is a solution of (2.6) on the half-line $t\geqq t_0$.

It is easily seen that the solution $u_r(t)$ satisfies asymptotics

$$u_r^{(k)}(t)=o(t^{-q}) \qquad (0\leqq k\leqq n-2) \tag{2.36}$$

and

$$g_k[u_r](t)=o(t^{-2q}) \qquad (1\leqq k\leqq n). \tag{2.37}$$

Indeed, the functions $g_k[u]$ are sums of expressions of the form (2.5) and hence (2.37) follows a priori from (2.36). To prove (2.36), we distinguish two cases: $q>0$ and $q=0$. If $q>0$, then (2.36) follows from (2.30) with $u=T[u]=u_r$. If $q=0$, then we put $u=T[u]=u_r$ in (2.33). We obtain $\|u_r\|_\infty\leqq C_4(\|u_r\|_\infty)^2$. Thus we have either $\|u_r\|_\infty=0$ or $C_4\|u_4\|_\infty\geqq 1$. Since u_r is in $S(t_0,\varepsilon)$, we have $C_4\|u_r\|\leqq C_4\varepsilon<1$ (see (2.34)). Obviously $\|u_r\|\geqq\|u_r\|_\infty$ and, therefore, $\|u_r\|_\infty=0$, which proves (2.36) (see (2.31) and (2.32)).

We also need an integral estimate for the solution $u_r(t)$.

LEMMA 2.3. *If* $\frac{1}{2}<q<1$, *then the integral* $\int^\infty u_r(t)\,dt$ *converges and*

$$\int_t^\infty u_r(s)\,ds=o(t^{1-2q}). \tag{2.38}$$

If $0\leqq q\leqq\frac{1}{2}$, *then*

$$\int_{t_0}^t u_r(s)\,ds=\begin{cases} o(\log t) & \text{if } q=\frac{1}{2},\\ o(t^{1-2q}) & \text{if } 0\leqq q<\frac{1}{2}.\end{cases} \tag{2.39}$$

The proof of Lemma 2.3 is given in §3.

Now we are able to finish the proof of Theorem 1. According to (2.1), we put

$$x_r(t)=C_0\exp\left(\lambda_r t+\int_{t_0}^{t}u_r(s)\,ds\right),$$

where

$$C_0=\begin{cases}\exp\left(-\int_{t_0}^{\infty}u_r(s)\,ds\right) & \text{if } \frac{1}{2}<q<1,\\ 1 & \text{if } 0\leqq q\leqq\frac{1}{2}.\end{cases}$$

Then $x_r(t)$ is a solution of (1.1) on $[t_0,\infty)$. This solution can be extended to $[0,\infty)$. Using (2.2) and (2.3) with $x=x_r(t)$, $u=u_r(t)$ and $y=\exp(-\lambda_r t)x_r(t)$, we find from (2.36)–(2.39) that the function $x_r(t)$ satisfies (1.5) and (1.6) with $j=r$. This completes the proof of Theorem 1.

3. Integral estimates. In this section, we verify the integral estimates stated in the proof of Theorem 1 given in §2. To avoid unnecessary repetition, we first introduce a factor smallness condition in the following definition.

DEFINITION 1. Let $m(t_0,t)$ be a real function defined for all t_0 and t, $0<t_0\leqq t<\infty$, which is nonincreasing in t and $m(t_0,t)\to 0$ as $t\to\infty$, for any $t_0>0$. If in addition $m(t_0,t_0)\to 0$ as $t_0\to\infty$, then we say that the function $m(t_0,t)$ is of type $(*)$.

To prove Lemma 2.1, we express the corresponding integrals in the form

$$\int_{t_{0j}}^{t}K(s)h(s)\,ds, \tag{3.1}$$

where $h(t)$ is an integrable function and the function $K(t)$ satisfies

$$|K^{(m)}(t)|\leqq K_m\exp(\alpha t)t^{-\beta}\qquad (K_m=\text{const.},\ t_0\leqq t<\infty) \tag{3.2$_m$}$$

for $m=0$ and, if need be, for $m=1$. The general result concerning the integrals (3.1) is given in the two following lemmas. Their proofs are described in the end of this section.

LEMMA 3.1. *Let $h(t)$ be a real or complex-valued function in $C[0,\infty)$ such that $\int^{\infty}|h(t)|\,dt$ converges and let α and β be real numbers, $\beta\geqq 0$. Then there exists a function $m(t_0,t)$ of type $(*)$, having the following property: if a function $K(t)$ in $C[t_0,\infty)$, where $t_0>0$, satisfies* (3.2$_0$), *then*

$$\left|\int_{t_0(\alpha)}^{t}K(s)h(s)\,ds\right|\leqq K_0 m(t_0,t)\exp(\alpha t)t^{-\beta}\qquad (t_0\leqq t<\infty), \tag{3.3}$$

where $t_0(\alpha)=t_0$ or $t_0(\alpha)=\infty$, according as $\alpha>0$ or $\alpha\leqq 0$. If $\alpha\leqq 0$, then the convergence of the integral on the left-hand side of (3.3) *is a part of the assertion.*

LEMMA 3.2. *Let $h(t)$ be a real- or complex-valued function in $C[0,\infty)$ such that $\int^{\infty}h(t)\,dt$ converges (perhaps conditionally) and α and β be real numbers, $\alpha\neq 0$ and $\beta\geqq 0$. Then there exists a function $m(t_0,t)$ of type $(*)$, having the following property: if a function $K(t)$ in $C^1[t_0,\infty)$, where $t_0>0$, satisfies* (3.2$_m$) *with $m=0,1$, then* (3.3) *holds with K_0 replaced by K_0+K_1. Furthermore,*

$$\left|\int_{t}^{\infty}h(s)s^{-\beta}\,ds\right|\leqq m(t_0,t)t^{-\beta}\qquad (0<t_0\leqq t<\infty). \tag{3.4}$$

Remark 2. In the proof of Theorem 1, we consider a finite set of estimates of the form (3.3) or (3.4). To simplify our notation, we assume that all these estimates include a common smallness factor $m(t_0,t)$. In fact, if $m_k(t_0,t)$, $k=1,2,\cdots,N$ are all factors considered, we can replace each of them by

$$m(t_0,t)=\sum_{k=1}^{N} m_k(t_0,t),$$

because all estimates (3.3) and (3.4) remain valid and the function $m(t_0,t)$ is of type $(*)$ as well as each function $m_k(t_0,t)$.

Proof of Lemma 2.1. From (2.17),

$$|u^{(k)}(t)|\leqq\|u\|t^{-q} \qquad (0\leqq k\leqq n-2), \tag{3.5}$$

for any u in $U[t_0,\infty)$. We show that the estimates

$$|g_{n-k}[\tilde{u}](t)-g_{n-k}[\tilde{\tilde{u}}](t)|\leqq C_5\varepsilon\|\tilde{u}-\tilde{\tilde{u}}\|t^{-2q} \qquad (0\leqq k\leqq n-1) \tag{3.6}$$

and

$$|g'_{n-k}[\tilde{u}](t)-g'_{n-k}[\tilde{\tilde{u}}](t)|\leqq C_5\varepsilon\|\tilde{u}-\tilde{\tilde{u}}\|t^{-2q} \qquad (1\leqq k\leqq n-1) \tag{3.7}$$

hold for any functions $\tilde{u},\tilde{\tilde{u}}$ in $S(t_0,\varepsilon)$. Obviously, it is sufficient to verify

$$|g[\tilde{u}](t)-g[\tilde{\tilde{u}}](t)|\leqq C_6\varepsilon\|\tilde{u}-\tilde{\tilde{u}}\|t^{-2q} \tag{3.8}$$

for a finite set of expressions $g[u]=u^{(\alpha_1)}u^{(\alpha_2)}\cdots u^{(\alpha_m)}$, where $0\leqq\alpha_j\leqq n-2$ $(1\leqq j\leqq m)$ and $2\leqq m\leqq n$. Put $u=\tilde{u}-\tilde{\tilde{u}}$ for $\tilde{u}$, $\tilde{\tilde{u}}$ in $S(t_0,\varepsilon)$. Multiplying the identities $\tilde{u}^{(\alpha_j)}=\tilde{\tilde{u}}^{(\alpha_j)}+u^{(\alpha_j)}$ $(1\leqq j\leqq m)$, we obtain

$$\tilde{u}^{(\alpha_1)}u^{(\alpha_2)}\cdots u^{(\alpha_m)}=\tilde{\tilde{u}}^{(\alpha_1)}\tilde{\tilde{u}}^{(\alpha_2)}\cdots\tilde{\tilde{u}}^{(\alpha_m)}+\cdots, \tag{3.9}$$

where, on the right-hand side each of (2^m-1) nonwritten members is a product of m factors $\tilde{\tilde{u}}^{(\alpha_j)}$, $u^{(\alpha_j)}$ so that at least one factor is $u^{(\alpha_j)}$ in each member. Since $|\tilde{\tilde{u}}^{(\alpha_j)}(t)|\leqq\|\tilde{\tilde{u}}\|t^{-q}$, $|u^{(\alpha_j)}(t)|\leqq\|u\|t^{-q}$ and $\|u\|=\|\tilde{u}-\tilde{\tilde{u}}\|\leqq\|\tilde{u}\|+\|\tilde{\tilde{u}}\|\leqq 2\varepsilon\leqq 2$, each of the nonwritten members in (3.9) does not exceed $(2\varepsilon)^{m-1}\|u\|t^{-qm}$ in absolute value. Thus we have

$$|g[\tilde{u}](t)-g[\tilde{\tilde{u}}](t)|\leqq(2^m-1)(2\varepsilon)^{m-1}\|\tilde{u}-\tilde{\tilde{u}}\|t^{-qm},$$

which proves (3.8), because $2\leqq m\leqq n$ and $1\leqq t_0<t<\infty$. Consequently, the estimates (3.6) and (3.7) are valid for a sufficiently large C_5.

Now we are able to prove the estimates for Lemma 2.1. To prove (2.23), we express the integral $L_j(t)$ (see (2.19)) in the form (3.1) with $h(t)=p_n(t)t^q$ and $K(t)=\exp(-\mu_j t)t^q$. Since $h(t)$ is integrable and $K(t)$ satisfies (3.2) for $m=0,1$ with $\alpha=\delta_j\neq 0$, $\beta=q$, $K_0=1$ and $K_1=|\mu_j|+q$, (2.23) follows from Lemma 3.2. The integral $L_{jk}[u](t)$ (see (2.20)) is of the form (3.1) with $h(t)=p_k(t)t^q$ and $K(t)=u^{(n-k-1)}(t)\exp(-\mu_j t)t^{-q}$. By (3.5), $K(t)$ satisfies (3.2_m) for $m=0,1$ with $\alpha=\delta_j\neq 0$, $\beta=2q$, $K_0=\|u\|$ and $K_1=\|u\|(1+|\mu_j|+q)$ except the case $k=1$, because, in general, (3.5) does not hold for the $(n-1)$th derivative of $u(t)$. Consequently, in the case $k>1$, (2.24) follows from Lemma 3.2. If $k=1$, then we put $h(t)=p_1(t)$ and $K(t)=u^{(n-2)}(t)\exp(-\mu_j t)$. Since $h(t)$ is absolutely integrable and $K(t)$ satisfies (3.2_0) with $\alpha=\delta_j$, $\beta=q$ and $K_0=\|u\|$, (2.24) with $k=1$ follows from Lemma 3.1. (Here we need the assumption (i).) To prove (2.26), we note that $N_{jk}[\tilde{u}](t)-N_{jk}[\tilde{\tilde{u}}](t)$ is also an integral of the form (3.1) with $h(t)=p_k(t)t^q$ and $K(t)=(g_{n-k}[\tilde{u}](t)-g_{n-k}[\tilde{\tilde{u}}](t))\exp(-\mu_j t)t^{-q}$ (see (2.22)). Then, by

(3.6) and (3.7), $K(t)$ satisfies (3.2_m) for $m=0,1$ with $\alpha=\delta_j\neq 0$, $\beta=3q$, $K_0=C_1\varepsilon\|\tilde{u}-\tilde{\tilde{u}}\|$ and $K_1=C_1\varepsilon\|\tilde{u}-\tilde{\tilde{u}}\|(1+|\mu_j|+q)$. Consequently, (2.26) follows from Lemma 3.2. It remains to establish (2.25). From (2.21), (2.27) and (3.6),

$$|M_{jk}[\tilde{u}](t)-M_{jk}[\tilde{\tilde{u}}](t)|=\left|\int_{t_{0j}}^{t}(g_{n-k}[\tilde{u}](s)-g_{n-k}[\tilde{\tilde{u}}](s)\exp(-\mu_j s))\,ds\right|$$

$$\leqq C_5\varepsilon\|\tilde{u}-\tilde{\tilde{u}}\|\left|\int_{t_{0j}}^{t}\exp(\delta_j s)s^{-2q}ds\right|,$$

for any functions $\tilde{u},\tilde{\tilde{u}}$ in $S(t_0,\varepsilon)$. Thus (2.25) holds if there exists a C_7 such that

$$\exp(-\delta_j t)t^{2q}\left|\int_{t_{0j}}^{t}\exp(\delta_j s)s^{-2q}ds\right|\leqq C_7.$$

The last inequality is valid for a sufficiently large C_7, because, by L'Hospital's rule, its left-hand side converges to $|\delta_j|^{-1}$ as $t\to\infty$. Note that the convergence of the integrals $M_{jk}[\tilde{u}](t)$ and $N_{jk}[\tilde{u}](t)$ follows from the preceding with $\tilde{\tilde{u}}=0$, because $M_{jk}[0]=N_{jk}[0]=0$. This completes the proof of Lemma 2.1.

It is worth remarking that this proof is correct only in the case when $\delta_j=\operatorname{Re}(\lambda_r-\lambda_j)\neq 0$, for any $j\neq r$, because Lemma 3.2 is not applicable if a function $h(t)$ is conditionally integrable and $K(t)$ satisfies (3.2_m) with $\alpha=0$. See also Remark 1.

Proof of Lemma 2.2. From (2.31) we obtain $|u^{(k)}(t)|\leqq\|u\|_t$, $0\leqq k\leqq n-2$, which enables us to bound any expression (2.5) by $|D^k_{\alpha_1,\cdots,\alpha_m}|(\|u\|_t)^m$. Consequently, there exists a C_8 such that

$$(3.10)\qquad |g_k[u](t)|\leqq C_8(\|u\|_t)^2,\qquad 1\leqq k\leqq n,\quad t_0\leqq t<\infty,\quad u\in S(t_0,\varepsilon).$$

Using (3.10), we obtain a new estimate for the integral $M_{jk}[u](t)$. If $\delta_j=-\operatorname{Re}\mu_j<0$, then

$$(3.11)\qquad |M_{jk}[u](t)|=\left|\int_{t}^{\infty}g_{n-k}[\tilde{u}](s)\exp(-\mu_j s)\,ds\right|$$

$$\leqq C_8(\|u\|_t)^2\int_{t}^{\infty}\exp(\delta_j s)\,ds=C_8|\delta_j|^{-1}(\|u\|_t)^2\exp(\delta_j t).$$

If $\delta_j>0$, let $t_1=(t_0+t)/2$. Then

$$(3.12)\quad |M_{jk}[u](t)|=\left|\int_{t_0}^{t}g_{n-k}[u](s)\exp(-\mu_j s)\,ds\right|$$

$$\leqq\left|\int_{t_0}^{t_1}g_{n-k}[u](s)\exp(-\mu_j s)\,ds\right|+\left|\int_{t_1}^{t}g_{n-k}[u](s)\exp(-\mu_j s)\,ds\right|$$

$$\leqq C_8(\|u\|_{t_0})^2\int_{-\infty}^{t_1}\exp(\delta_j s)\,ds+C_8(\|u\|_{t_1})^2\int_{-\infty}^{t}\exp(\delta_j s)\,ds$$

$$=C_8\delta_j^{-1}\left(\|u\|^2_{t_0}\exp[\delta_j(t_1-t)]+\|u\|^2_{t_1}\right)\exp(\delta_j t).$$

Using (3.11) and (3.12) instead of (2.25), we obtain a new variant of (2.28)

$$\begin{aligned} |T^{(k)}[u](t)| \leqq {} & C_2(1+\|u\|)m(t_0,t)+C_9\|u\|_{t_1}^2 \\ & + C_{10}\|u\|_t^2 + C_{11}\|u\|_{t_0}^2 \sum_{\substack{j\in J \\ \delta_j>0}} \exp\left[\delta_j(t_1-t)\right] \end{aligned} \tag{3.13}$$

for any u in $S(t_0,\varepsilon)$, $0 \leqq k \leqq n-2$. Since R_{13}, the right-hand side of (3.13), is nonincreasing in t, it follows from (2.31) and (3.13) that $\|T[u]\|_t \leqq (n-1)R_{13}$. As $t \to \infty$ in the last inequality, we obtain (2.33) with $C_4=(n-1)(C_9+C_{10})$, because $m(t_0,t)\to 0$, $t_1 \to \infty$ and $t_1 - t \to \infty$ in R_{13} as $t \to \infty$. This completes the proof of Lemma 2.2.

Proof of Lemma 2.3. Let us integrate (2.6) with $u=u_r(t)$ between the limits t_0 and t. By (2.8), the result of integration may be written in the form

$$\begin{aligned} & \sum_{k=0}^{n-2} b_k\left(u_r^{(n-k-2)}(t)-u_r^{(n-k-2)}(t_0)\right)+b_{n-1}\int_{t_0}^t u_r(s)\,ds \\ & \quad = -\int_{t_0}^t p_n(s)\,ds - \sum_{k=1}^{n-1}\int_{t_0}^t p_k(s)u_r^{(n-k-1)}(s)\,ds - \sum_{k=0}^{n-1} b_k\int_{t_0}^t g_{n-k}[u_r](s)\,ds \\ & \qquad - \sum_{k=1}^{n-1}\int_{t_0}^t p_k(t)g_{n-k}u_r(s)\,ds. \end{aligned}$$

Since the characteristic equation for the left-hand side of (2.6) has no zero root, it holds that $b_{n-1}\neq 0$. Consequently, we can express the integral from the left-hand side of the last equality. Taking into account (2.36), we obtain

$$\begin{aligned} \int_{t_0}^t u_r(s)\,ds = {} & \text{const.} + o(t^{-q}) \\ & + b_{n-1}^{-1}\Bigg\{ P_n(t) - \sum_{k=0}^{n-1} b_k\int_{t_0}^t tg_{n-k}[u_r](s)\,ds \\ & + \sum_{k=1}^{n-1}\int_{t_0}^t P_k'(s)u_r^{(n-k-1)}(s)\,ds + \sum_{k=1}^{n-1}\int_{t_0}^t P_k'(s)g_{n-k}[u_r](s)\,ds \Bigg\}, \end{aligned}$$

where, in view of assumption (ii) and (3.4),

$$P_k(t)=\int_t^\infty p_k(s)\,ds=o(t^{-q}) \qquad (1\leqq k\leqq n). \tag{3.14}$$

Hence Lemma 2.3 will be proved if we show that the integrals of the following three types

$$\int_{t_0}^t P_k'(s)u_r^{(n-k-1)}(s)\,ds, \quad \int_{t_0}^t g_{n-k}[u_r](s)\,ds, \quad \int_{t_0}^t P_k'(s)g_{n-k}[u_r](s)\,ds$$

have either the property (2.38) or (2.39). Since

$$\int_{t_0}^t o(s^\beta)\,ds = \begin{cases} \text{const.} + o(t^{1+\beta}) & \text{if } \beta \neq -1, \\ o(\log t) & \text{if } \beta = -1, \end{cases} \tag{3.15}$$

the above-mentioned property of the middle type of integral follows from (2.37). Integrating by parts yields

$$(3.16)\qquad \int_{t_0}^{t} P_k'(s)u^{(n-k-1)}(s)\,ds = P_k(s)u_r^{(n-k-1)}(s)\Big|_{t_0}^{t} - \int_{t_0}^{t} P_k(s)u_r^{(n-k)}(s)\,ds$$

and

$$(3.17)\qquad \int_{t_0}^{t} P_{t_0}' P_k'(s)g_{n-k}[u_r](s)\,ds = P_k(s)g_{n-k}[u_r](s)\Big|_{t_0}^{t} - \int_{t_0}^{t} P_k(s)g_{n-k}'[u_r](s)\,ds.$$

Besides (2.37) we now need the asymptotics

$$(3.18)\qquad g_k'[u_r](t) = O(t^{-2q}) \qquad 1 \leqq k \leqq n-1,$$

which follow from (2.36) as well as (2.37). In view of (2.36), (3.14) and (3.18), the integrals on the right of (3.16) and (3.17) may be estimated by means of rule (3.15). Consequently, the needed property is established except for the integral (3.16) with $k=1$, because, in general, $u_r^{(n-1)}(t) \neq o(t^{-q})$ and hence integration by parts is useless. However, we can estimate the integral directly by using (2.36) with $k=n-2$ and assumption (i):

$$\int_t^\infty |P_1'(s)u_r^{(n-2)}(s)|\,ds = \int_t^\infty |p_1(s)o(s^{-q})|\,ds = o(t^{-q}).$$

This completes the proof of Lemma 2.3.

A sketch of the proofs of Lemmas 3.1 *and* 3.2. Denote

$$H(t) = \int_t^\infty h(s)\,ds, \quad H_1(t) = \int_t^\infty |h(s)|\,ds \quad \text{and} \quad H_2(t) = \sup_{t_1 \geqq t} |H(t_1)|.$$

If $\alpha \leqq 0$ (<0), then Lemmas 3.1 and 3.2 follow from the estimates

$$\left|\int_t^\infty K(s)h(s)\,ds\right| \leqq \int_t^\infty |K(s)h(s)|\,ds \leqq K_0 \exp(\alpha t)t^{-\beta}H_1(t)$$

and

$$\begin{aligned}\left|\int_t^\infty K(s)h(s)\,ds\right| &\leqq |K(t)H(t)| + \int_t^\infty |K'(s)H(s)|\,ds \\ &\leqq K_0\exp(\alpha t)t^{-\beta}H_2(t) + K_1 t^{-\beta}H_2(t)\int_t^\infty \exp(\alpha s)\,ds \\ &= \left(K_0 + |\alpha|^{-1}K_1\right)H_2(t)\exp(\alpha t)t^{-\beta}.\end{aligned}$$

If $\alpha > 0$, denote $t_1 = (t_0 + t)/2$. Under the hypotheses of Lemma 3.1, we can write

$$\begin{aligned}\left|\int_{t_0}^t K(s)h(s)\,ds\right| &\leqq \int_{t_0}^{t_1} |K(s)h(s)|\,ds + \int_{t_1}^{t} |K(s)h(s)|\,ds \\ &\leqq K_0\left[H_1(t_0)\exp(\alpha t_1)t_0^{-\beta} + H_1(t_1)\exp(\alpha t)t_1^{-\beta}\right].\end{aligned}$$

Therefore, the estimate (3.3) holds if

$$m(t_0,t) \geqq H_1(t_0)\exp[\alpha(t_1-t)]\left(\frac{t}{t_0}\right)^{\beta} + H_1(t_1)\left(\frac{t}{t_1}\right)^{\beta}.$$

Since $t/t_1 = 2t/(t_0+t) \leqq 2$ and the function $2^{\beta}H_1((t_0+t)/2)$ is of type $(*)$, the function $m(t_0,t)$ is of the form $m(t_0,t) = m_1(t_0,t) + 2^{\beta}H_1(t_1)$, where $m_1(t_0,t)$ is a function of type $(*)$ such that

$$m_1(t_0,t) \geqq H_1(t_0)\exp[\alpha(t_1-t)]\left(\frac{t}{t_0}\right)^{\beta}. \tag{3.19}$$

This inequality is satisfied by a (finite) function

$$m_1(t_0,t) = H_1(t_0)\sup_{t_2 \geqq t}\left\{\exp[\alpha(t_0-t_2)/2](t_2/t_0)^{\beta}\right\},$$

which is nonincreasing in t and converges to 0 as $t \to \infty$, for any $t_0 > 0$. For $m_1(t_0,t)$ to be of type $(*)$, it remains to show that $m_1(t_0,t_0) \to 0$ as $t_0 \to \infty$. Since $H_1(t_0) \to 0$ as $t_0 \to \infty$, the last fact will follow from the inequality

$$\sup_{t_2 \geqq t_0}\left\{\exp[\alpha(t_0-t_2)/2](t_2/t_0)^{\beta}\right\} \leqq 2^{\beta}(1+MT_0^{-\beta}), \tag{3.20}$$

where M is an upper bound of the function $\exp(-\alpha t/2)t^{\beta}$ on the half-line $0 \leqq t < \infty$. To prove (3.20), we note that for $t_2 \in [t_0, 2t_0]$ we have $\exp[\alpha(t_0-t_2)/2](t_2/t_0)^{\beta} \leqq 2^{\beta}$. If $t_2 \geqq 2t_0$, then, by the definition of the bound M, $\exp[\alpha(t_0-t_2)/2] \leqq M(t_2-t_0)^{-\beta}$ and thus $\exp[\alpha(t_0-t_2)/2](t_2/t_0)^{\beta} \leqq Mt_0^{-\beta}[t_2/(t_2-t_0)]^{\beta} \leqq 2^{\beta}Mt_0^{-\beta}$, because $t_2/(t_2-t_0) \leqq 2$. This completes the proof of (3.20). To prove Lemma 3.2 in the case $\alpha > 0$, we find that

$$\left|\int_{t_0}^{t} K(s)h(s)\,ds\right| \leqq |K(t_0)H(t_0)| + |K(t)H(t)| + \int_{t_0}^{t} |K'(s)H(s)|\,ds,$$

$$|K(t)H(t)| \leqq K_0H_2(t)\exp(\alpha t)t^{-\beta}$$

and

$$\begin{aligned}\int_{t_0}^{t}|K'(s)H(s)|\,ds &= \int_{t_0}^{t_1}|K'(s)H(s)|\,ds + \int_{t_1}^{t}|K'(s)H(s)|\,ds\\ &\leqq K_1H_2(t_0)t_0^{-\beta}\int_{-\infty}^{t_1}\exp(\alpha s)\,ds + K_1H_2(t_1)t_1^{-\beta}\int_{-\infty}^{t}\exp(\alpha s)\,ds\\ &= K_1\alpha^{-1}\left[H_2(t_0)\exp(\alpha t_1)t_0^{-\beta} + H_2(t_1)\exp(\alpha t)t_1^{-\beta}\right].\end{aligned}$$

Therefore, the estimate (3.3) with K_0 replaced by $K_0 + K_1$ holds if

$$\begin{aligned} m(t_0,t) \geqq (1+\alpha^{-1})\Bigg\{ & H_2(t_0)\exp[\alpha(t_0-t)]\left(\frac{t}{t_0}\right)^{\beta} + H_2(t)\\ & + H_2(t_0)\exp[\alpha(t_1-t)]\left(\frac{t}{t_0}\right)^{\beta} + H_2(t_1)\left(\frac{t}{t_1}\right)^{\beta}\Bigg\}.\end{aligned}$$

Since we have $\alpha(t_0-t)\leqq\alpha(t_0-t)/2=\alpha(t_1-t)$ and $t/t_1=2t/(t_0+t)<2$, it is sufficient to require

$$m(t_0,t)\geqq(1+\alpha^{-1})\left\{2H_2(t_0)\exp[\alpha(t_1-t)]\left(\frac{t}{t_0}\right)^{\beta}+2^{\beta}H_2(t_1)+H_2(t)\right\}.$$

Now we can put

$$m(t_0,t)=(1+\alpha^{-1})\{2m_1(t_0,t)+2^{\beta}H_2(t_1)+H_2(t)\},$$

where $m_1(t_0,t)$ is a function of type (*) satisfying (3.19) with $H_1(t_0)$ replaced by $H_2(t_0)$. The existence of the function $m_1(t_0,t)$ in (3.19) was proved before. This completes the proof of both lemmas.

REFERENCES

[1] P. Hartman, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[2] J. Šimša, *Asymptotic integration of perturbed linear differential equations under conditions involving ordinary integral convergence*, this Journal, 15 (1984), pp. 116–123.

[3] ———, *The second order differential equation with an oscillatory coefficient*, Arch. Math. (Brno), 18 (1982), pp. 95–100.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

ASYMPTOTIC BEHAVIOR OF SOLUTIONS OF A CLASS OF HIGHER ORDER ORDINARY DIFFERENTIAL EQUATIONS*

TAKAŜI KUSANO[†] AND BHAGAT SINGH[‡]

Abstract. The asymptotic behavior of solutions of perturbed disconjugate differential equations of the type

$$\text{(A)}\qquad \begin{aligned} &L_nx+F(t,L_0x,L_1x,\cdots,L_{n-1}x)=f(t),\\ &L_0x=x/p_0,\qquad L_ix=(L_{i-1}x)'/p_i,\qquad 1\leqq i\leqq n,\end{aligned}$$

is studied. Bounds on the growth of solutions are given, criteria for all solutions to be oscillatory or nonoscillatory are established, and conditions are presented which guarantee that every solution tends to a limit as $t\to\infty$. Our results for (A) can be applied to the qualitative study of a class of elliptic partial differential equations.

1. Introduction. In this paper we are concerned with nonlinear ordinary differential equations of the type

$$\text{(A)}\qquad L_nx+F(t,L_0x,L_1x,\cdots,L_{n-1}x)=f(t)$$

in an infinite interval $[a,\infty)$, where L_i, $0\leqq i\leqq n$, denote the differential operators defined by

$$\text{(1)}\qquad L_0x(t)=\frac{x(t)}{p_0(t)},\qquad L_ix(t)=\frac{1}{p_i(t)}\frac{d}{dt}L_{i-1}x(t),\qquad 1\leqq i\leqq n,$$

in terms of the positive continuous functions $p_i(t)$, $0\leqq i\leqq n$, on $[a,\infty)$. Here $f(t)$ and $F(t,u_0,u_1,\cdots,u_{n-1})$ are continuous functions defined on $[a,\infty)$ and $[a,\infty)\times R^n$, respectively. Notice that in the particular case where $p_i(t)\equiv 1$, $0\leqq i\leqq n$, equation (A) reduces to

$$\text{(B)}\qquad x^{(n)}+F(t,x,x',\cdots,x^{(n-1)})=f(t).$$

By a solution of equation (A) we mean a function $x(t)$ which has the continuous "quasi-derivatives" $L_ix(t)$, $0\leqq i\leqq n$, and satisfies (A) on some half-line $[T_x,\infty)$. Such a solution is called oscillatory if it has arbitrarily large zeros; otherwise it is called nonoscillatory.

Our main purpose is to study the asymptotic behavior of solutions of equation (A). More specifically, bounds on the growth of solutions will be given in §2; conditions for all solutions to be oscillatory or nonoscillatory are presented in §3, and §4 is devoted to establishing criteria which force every solution to approach a limit as $t\to\infty$. Roughly speaking, our consideration is focused on the situation in which the "perturbation" $F(t,L_0x,L_1x,\cdots,L_{n-1}x)$ is so small that the solutions of (A) essentially behave as those of $L_nx=f(t)$ asymptotically. Equations of the form (A) with large perturbations F will be the object of a forthcoming paper.

* Received by the editors April 26, 1983, and in revised form December 17, 1983.

[†] Department of Mathematics, Faculty of Science, Hiroshima University, Hiroshima 730, Japan.

[‡] Department of Mathematics, University of Wisconsin Center, Manitowoc, Wisconsin 54220.

This work combines the existence theory with the asymptotic theory of oscillatory solutions. Existence of nonoscillatory solutions has been established using the fixed point methods in a number of papers. Since the fixed point theory is not applicable to the existence of oscillatory solutions, we believe that this work is a step toward filling this vacuum.

In recent years there has been an increasing interest in studying the qualitative behavior of solutions of equations of the type (A); see, for example, the papers [2]–[7], [11]–[28], [31]–[33]. Most of the literature, however, has been concerned with equations of the form $L_nx+F(t,x)=f(t)$, and it seems that very little is known about equations which really involve the lower quasi-derivatives $L_1x,\cdots,L_{n-1}x$. The present paper is an attempt to make a systematic study of such general differential equations. It should be emphasized that most of our results seem to be new even when specialized to equation (B). For typical results on the asymptotic theory for (B) the reader is referred to the papers [8]–[10], [29], [30].

2. Bounds on the growth of solutions. Unless otherwise stated explicitly the following conditions are assumed to hold throughout the paper:

$$\int_a^\infty p_i(t)\,dt=\infty \quad \text{for } 1\leqq i\leqq n-1, \tag{2}$$

$$|F(t,u_0,u_1,\cdots,u_{n-1})|\leqq \sum_{i=0}^{n-1} q_i(t)|u_i|^{r_i} \tag{3}$$

for all $(t,u_0,u_1,\cdots,u_{n-1})\in[a,\infty)\times R^n$, where r_i are nonnegative constants and $q_i(t)$ are nonnegative continuous functions on $[a,\infty)$.

Let $h_1(t)$, $h_2(t),\cdots$ be continuous functions on $[a,\infty)$. We define for $t,s\in[a,\infty)$

$$\begin{aligned} &I_0=I,\\ &I_i(t,s;h_i,\cdots,h_1)=\int_s^t h_i(r)I_{i-1}(r,s;h_{i-1},\cdots,h_1)\,dr, \qquad i=1,2,\cdots. \end{aligned} \tag{4}$$

The following identities are easily verified:

$$I_i(t,s;h_i,\cdots,h_1)=(-1)^iI_i(s,t;h_1,\cdots,h_i), \tag{5}$$

$$I_i(t,s;h_i,\cdots,h_1)=\int_s^t h_1(r)I_{i-1}(t,r;h_i,\cdots,h_2)\,dr, \tag{6}$$

$$I_i(t,s;h_i,\cdots,h_1)=\sum_{j=0}^{i} I_{i-j}(t,r;h_i,\cdots,h_{j+1})I_j(r,s;h_j,\cdots,h_1). \tag{7}$$

LEMMA 1 (Hallam [9]). *If $a_i\geqq 0$, $b_i\geqq 0$, $r_i\geqq 0$, $1\leqq i\leqq n$, and if $b_i>1$ for some i, then*

$$\sum_{i=1}^{n} a_ib_i^{r_i}\leqq\left[\sum_{i=1}^{n} a_i\right]\left[\sum_{i=1}^{n} b_i\right]^r \quad \textit{where } r=\max_i r_i. \tag{8}$$

LEMMA 2 (Bihari [1]). *Let $x(t)$ and $k(t)$ be nonnegative continuous functions on $[t_0,T)$, $T\leqq\infty$, and let $g(x)>0$ be a continuous nondecreasing function on $(0,\infty)$. If*

$$x(t)\leqq M+\int_{t_0}^t k(s)g(x(s))\,ds, \qquad t\in[t_0,T), \tag{9}$$

for some nonnegative constant M, then

$$(10)\qquad x(t) \leqq G^{-1}\left(G(M)+\int_{t_0}^{t} k(s)\,ds\right), \qquad t\in[t_0,T),$$

provided that both sides of (10) *are defined, where $G(x)$ is an antiderivative of* $1/g(x)$ *and $G^{-1}(u)$ is the inverse function of $G(x)$.*

Let $t_0 > a$ (an initial time) be fixed and let $x(t)$ be a solution of equation (A) satisfying the initial condition

$$(11)\qquad L_i x(t_0)=\xi_i, \qquad 0\leqq i\leqq n-1,$$

where ξ_i are given constants. Explicit bounds on the growth of $x(t)$ will be given in the following theorem which is one of the main results of this paper.

THEOREM 1. *Let the functions $\varphi_i(t)$, $0\leqq i\leqq n-1$, be defined by*

$$(12)\qquad \varphi_i(t)=\max\left\{I_{n-i-1}(t,a;p_{i+1},\cdots,p_{n-1}),\, I_{n-i}\left(t,a;p_{i+1},\cdots,p_{n-1},p_n|f|\right)\right\}.$$

Suppose that

$$(13)\qquad \int_a^{\infty} p_n(t)q_i(t)[\varphi_i(t)]^{r_i}dt<\infty, \qquad 0\leqq i\leqq n-1.$$

Then, if $r=\max_i r_i\leqq 1$, all solutions $x(t)$ of (A) *satisfy*

$$(14)\qquad L_i x(t)=O(\varphi_i(t)) \quad as\ t\to\infty, \quad 0\leqq i\leqq n-1,$$

and if $r>1$, solutions $x(t)$ of (A) *satisfy* (14) *provided t_0 is sufficiently large and ξ_i are sufficiently small.*

Proof. The method of Hallam [9, Thm. 1] will be adapted. Integrating (A), we have

$$(15)\qquad L_i x(t)=\sum_{j=i}^{n-1} L_j x(t_0) I_{j-i}(t,t_0;p_{i+1},\cdots,p_j)$$
$$+I_{n-i}(t,t_0;p_{i+1},\cdots,p_{n-1},p_n f)-I_{n-i}(t,t_0;p_{i+1},\cdots,p_{n-1},p_n F[x])$$

for $0\leqq i\leqq n-1$, where

$$(16)\qquad F[x]=F(t,L_0x,L_1x,\cdots,L_{n-1}x).$$

Dividing (15) by $\varphi_i(t)$ and noting that $I_{j-i}(t,t_0;p_{i+1},\cdots,p_j)/\varphi_i(t)$, $i\leqq j\leqq n-1$, are bounded and

$$I_{n-1}\left(t,t_0,p_{i+1},\cdots,p_{n-1},p_n|F[x]|\right)/\varphi_i(t)$$
$$\leqq \frac{I_{n-i}\left(t,t_0;p_{i+1},\cdots,p_{n-1},p_n|F[x]|\right)}{I_{n-i-1}\left(t,a;p_{i+1},\cdots,p_{n-1}\right)} \leqq \int_{t_0}^{t} p_n(s)|F[x](s)|\,ds$$
$$\leqq \int_{t_0}^{t}\sum_{k=0}^{n-1} p_n(s)q_k(s)|L_k x(s)|^{r_k}ds, \qquad t\geqq t_0,$$

we obtain

$$(17)\qquad \frac{|L_i x(t)|}{\varphi_i(t)} \leqq A_i+\int_{t_0}^{t}\sum_{k=0}^{n-1} p_n(s)q_k(s)[\varphi_k(s)]^{r_k}\left(\frac{|L_k x(s)|}{\varphi_k(s)}\right)^{r_k} ds$$

for $t \geqq t_0$ and $0 \leqq i \leqq n-1$, where

$$A_i = 1 + \sup_{t \geqq t_0} \sum_{j=i}^{n-1} |\xi_j| I_{j-i}(t, t_0; p_{i+1}, \cdots, p_j)/\varphi_i(t).$$

From (17) it can be shown that there exists a constant $A > 0$ such that

$$(18) \qquad \sum_{i=0}^{n-1} \frac{|L_i x(t)|}{\varphi_i(t)} \leqq A + n \int_{t_0}^{t} \left[\sum_{k=0}^{n-1} p_n(s) q_k(s) [\varphi_k(s)]^{r_k} \right] \left(\sum_{k=0}^{n-1} \frac{|L_k x(s)|}{\varphi_k(s)} \right)^r ds$$

for $t \geqq t_0$, where $r = \max_i r_i$. To see this, let $P(s)$ and $Q(s)$ denote the integrands in (17) and (18), respectively, and put

$$R(s) = \sum_{k=0}^{n-1} p_n(s) q_k(s) [\varphi_k(s)]^{r_k}.$$

Then, obviously $P(s) \leqq R(s)$ if

$$(19) \qquad |L_i x(s)|/\varphi_i(s) \leqq 1, \qquad 0 \leqq i \leqq n-1,$$

while $P(s) \leqq Q(s)$ if (19) does not hold, by Lemma 1. Therefore

$$(20) \qquad P(s) \leqq Q(s) + R(s)$$

in either case. Since $\int_{t_0}^{\infty} R(s) ds < \infty$ because of (13), combining (17) with (20) implies (18) as claimed, with A defined as

$$A = \sum_{i=0}^{n-1} A_i + n \int_{t_0}^{\infty} R(s).$$

Applying now Lemma 2 to (18), we conclude that if $r \leqq 1$, then $|L_i x(t)|/\varphi_i(t)$, $0 \leqq i \leqq n-1$, are bounded for any values of ξ_i, and that if $r > 1$, then the same is true of $x(t)$ provided the initial time t_0 is large enough and the initial values ξ_i are small enough so that

$$(21) \qquad (r-1) n A^{r-1} \int_{t_0}^{\infty} \sum_{k=0}^{n-1} p_n(s) q_k(s) [\varphi_k(s)]^{r_k} ds < 1.$$

This completes the proof.

Remark 1. Using L'Hospital's rule, we have

$$(22) \qquad \lim_{t \to \infty} \frac{I_{n-i}(t, a; p_{i+1}, \cdots, p_{n-1}, p_n |f|)}{I_{n-i-1}(t, a; p_{i+1}, \cdots, p_{n-1})} = \int_a^{\infty} p_n(t) |f(t)| \, dt.$$

Consequently, if

$$(23) \qquad \int_a^{\infty} p_n(t) |f(t)| \, dt < \infty,$$

then the functions $\varphi_i(t)$ in (12) can be taken to be

$$(24) \qquad \varphi_i(t) = I_{n-i-1}(t, a; p_{i+1}, \cdots, p_{n-1}), \qquad 0 \leqq i \leqq n-1,$$

and if

$$\int_a^\infty p_n(t)|f(t)|\,dt = \infty, \tag{25}$$

then the functions $\varphi_i(t)$ can be taken to be

$$\varphi_i(t) = I_{n-i}(t,a;p_{i+1},\cdots,p_{n-1},p_n|f|), \qquad 0 \leqq i \leqq n-1. \tag{26}$$

Remark 2. In the particular case where $p_i(t) \equiv 1$ for $0 \leqq i \leqq n$ we obtain

$$I_{n-i-1}(t,a;p_{i+1},\cdots,p_{n-1}) = \frac{(t-a)^{n-i-1}}{(n-i-1)!},$$

$$I_{n-i}(t,a;p_{i+1},\cdots,p_{n-1},p_n f) = \int_a^t \frac{(t-s)^{n-i-1}}{(n-i-1)!} f(s)\,ds.$$

It is easy to see that:

(i) if $\lim_{t\to\infty} f(t)/t^m = \text{constant} \neq 0$ for some $m > -1$, then

$$\lim_{t\to\infty} \int_a^t \frac{(t-s)^{n-i-1}}{(n-i-1)!} f(s)\,ds/t^{m+n-i} = \text{constant} \neq 0;$$

(ii) if $\lim_{t\to\infty} f(t)/t^{-1} = \text{constant} \neq 0$, then

$$\lim_{t\to\infty} \int_a^t \frac{(t-s)^{n-i-1}}{(n-i-1)!} f(s)\,ds/t^{n-i-1}\log t = \text{constant} \neq 0;$$

(iii) if $\lim_{t\to\infty} f(t)/t^m e^{bt} = \text{constant} \neq 0$ for some m and $b > 0$, then

$$\lim_{t\to\infty} \int_a^t \frac{(t-s)^{n-i-1}}{(n-i-1)!} f(s)\,ds/t^m e^{bt} = \text{constant} \neq 0.$$

From Theorem 1 and Remarks 1 and 2 we have the following corollary.

COROLLARY 1. (i) *Suppose that*

$$\int_a^\infty |f(t)|\,dt < \infty \tag{27}$$

and

$$\int_a^\infty t^{(n-i-1)r_i} q_i(t)\,dt < \infty, \qquad 0 \leqq i \leqq n-1. \tag{28}$$

Then, if $r = \max_i r_i \leqq 1$, *all solutions* $x(t)$ *of* (B) *satisfy*

$$x^{(i)}(t) = O(t^{n-i-1}) \quad \text{as } t \to \infty, \quad 0 \leqq i \leqq n-1, \tag{29}$$

and if $r > 1$, *solutions* $x(t)$ *of* (B) *with large initial time and small initial values satisfy* (29).

(ii) *Suppose that* $|f(t)| \leqq f^*(t)$, *where* $f^*(t)$ *is continuous,*

$$\lim_{t\to\infty} f^*(t)/t^m = \text{constant} > 0 \quad \textit{for some } m > -1 \tag{30}$$

and

$$(31)\qquad \int_a^\infty t^{(m+n-i)r_i} q_i(t)\,dt < \infty, \qquad 0 \leqq i \leqq n-1.$$

Then, if $r \leqq 1$ *all solutions* $x(t)$ *of* (B) *satisfy*

$$(32)\qquad x^{(i)}(t) = O(t^{m+n-i}) \quad \text{as } t \to \infty, \quad 0 \leqq i \leqq n-1,$$

and if $r > 1$, *all solutions* $x(t)$ *of* (B) *with large initial time and small initial values satisfy* (32).

(iii) *Suppose that* $|f(t)| \leqq f^*(t)$, *where* $f^*(t)$ *is continuous,*

$$(33)\qquad \lim_{t\to\infty} f^*(t)/t^m e^{bt} = \text{constant} > 0 \quad \text{for some } m \text{ and } b > 0$$

and

$$(34)\qquad \int_a^\infty (t^m e^{bt})^{r_i} q_i(t) dt < \infty, \qquad 0 \leqq i \leqq n-1.$$

Then, if $r \leqq 1$, *all solutions* $x(t)$ *of* (B) *satisfy*

$$(35)\qquad x^{(i)}(t) = O(t^m e^{bt}) \quad \text{as } t \to \infty, \quad 0 \leqq i \leqq n-1,$$

and if $r > 1$, *solutions* $x(t)$ *of* (B) *with large initial time and small initial values satisfy* (35).

3. Oscillation and nonoscillation of solutions. On the basis of Theorem 1 we can establish conditions which guarantee that all solutions of (A) are oscillatory or nonoscillatory.

THEOREM 2. *Suppose that* $f(t)$ *is eventually positive* (*or negative*) *and*

$$(36)\qquad \int_a^\infty p_n(t) f(t)\,dt = \infty \qquad (or\ -\infty).$$

Suppose moreover that (13) *holds with* $\varphi_i(t)$ *defined by* (26). *Then, if* $r = \max_i r_i \leqq 1$, *all solutions of* (A) *are nonoscillatory, and if* $r > 1$, *solutions of* (A) *with sufficiently large initial time and sufficiently small initial values are nonoscillatory.*

Proof. We may assume that $f(t)$ is eventually positive. Let $x(t)$ be a solution of (A) (with large initial time and small initial values if $r > 1$). We have (15) (with $i = 0$):

$$(37)\quad L_0 x(t) = \sum_{j=0}^{n-1} L_j x(t_0) I_j(t, t_0; p_1, \cdots, p_j) + I_n(t, t_0; p_1, \cdots, p_{n-1}, p_n f) - I_n(t, t_0; p_1, \cdots, p_{n-1}, p_n F[x]).$$

Dividing (37) by $\varphi_0(t)$ and letting $t \to \infty$, we conclude that $L_0 x(t)/\varphi_0(t) \to 1$ as $t \to \infty$, which shows that $x(t)$ is nonoscillatory. To see this it suffices to verify

$$(38)\qquad I_j(t, t_0; p_1, \cdots, p_j)/\varphi_0(t) \to 0, \qquad 0 \leqq j \leqq n-1,$$

$$(39)\qquad I_n(t, t_0; p_1, \cdots, p_{n-1}, p_n f)/\varphi_0(t) \to 1,$$

and

$$(40)\qquad I_n(t, t_0; p_1, \cdots, p_{n-1}, p_n F[x])/\varphi_0(t) \to 0,$$

as $t \to \infty$. Relations (38) and (39) follow with the use of L'Hospital's rule. By Theorem 1 there are positive constants c_i such that $|L_i x(t)| \leqq c_i \varphi_i(t)$, $t \geqq t_0$, for $0 \leqq i \leqq n-1$.

Using this inequality and (38), we can prove (40) as follows:

$$
\begin{aligned}
&I_n\big(t,t_0;p_1,\cdots,p_{n-1},p_n|F[x]|\big)/\varphi_0(t)\\
&\quad=\frac{I_{n-1}(t,t_0;p_1,\cdots,p_{n-1})}{\varphi_0(t)}\,\frac{I_n\big(t,t_0;p_1,\cdots,p_{n-1},p_n|F[x]|\big)}{I_{n-1}(t,t_0;p_1,\cdots,p_{n-1})}\\
&\quad\leqq\frac{I_{n-1}(t,t_0;p_1,\cdots,p_{n-1})}{\varphi_0(t)}\int_{t_0}^{t}p_n(s)|F[x](s)|\,ds\\
&\quad\leqq\frac{I_{n-1}(t,t_0;p_1,\cdots,p_{n-1})}{\varphi_0(t)}\int_{t_0}^{t}\sum_{k=0}^{n-1}p_n(s)q_k(s)|L_kx(s)|^{r_k}ds\\
&\quad\leqq\frac{I_{n-1}(t,t_0;p_1,\cdots,p_{n-1})}{\varphi_0(t)}\int_{t_0}^{\infty}\sum_{k=0}^{n-1}p_n(s)q_k(s)[c_k\varphi_k(s)]^{r_k}ds\to 0
\end{aligned}
$$

as $t\to\infty$. This completes the proof.

Remark 3. Starting from (15) with any i, $0\leqq i\leqq n-1$, and arguing as in the above proof, we can show that the solution $x(t)$ under consideration satisfies

$$\lim_{t\to\infty} L_ix(t)/\varphi_i(t)=1 \quad\text{or } -1,\qquad 0\leqq i\leqq n-1,\tag{41}$$

according as $f(t)$ is eventually positive or negative.

COROLLARY 2. (i) *Suppose that*

$$\lim_{t\to\infty} f(t)/t^m=\text{constant}\neq 0\quad \textit{for some } m>-1\tag{42}$$

and (31) *holds. Then, if* $r=\max_i r_i\leqq 1$, *all solutions* $x(t)$ *of* (B) *satisfy*

$$\lim_{t\to\infty} x^{(i)}(t)/t^{m+n-i}=\text{constant}\neq 0,\qquad 0\leqq i\leqq n-1,\tag{43}$$

and if $r>1$, *solutions* $x(t)$ *of* (B) *with large initial time and small initial data satisfy* (43).

(ii) *Suppose that*

$$\lim_{t\to\infty} f(t)/t^me^{bt}=\text{constant}\neq 0\quad \textit{for some } m \textit{ and } b>0\tag{44}$$

and (34) *holds. Then, if* $r\leqq 1$, *all solutions* $x(t)$ *of* (B) *satisfy*

$$\lim_{t\to\infty} x^{(i)}(t)/t^me^{bt}=\text{constant}\neq 0,\qquad 0\leqq i\leqq n-1,\tag{45}$$

and if $r>1$, *solutions of* (B) *with large initial time and small initial data satisfy* (45).

Remark 4. Corollaries 1 and 2 unify some of the main results of Hallam [9].

THEOREM 3. *Suppose that for any* $t_0\geqq a$

$$\limsup_{t\to\infty}\frac{I_n(t,t_0;p_1,\cdots,p_{n-1},p_nf)}{I_{n-1}(t,t_0;p_1,\cdots,p_{n-1})}=\infty,\tag{46}$$

$$\liminf_{t\to\infty}\frac{I_n(t,t_0;p_1,\cdots,p_{n-1},p_nf)}{I_{n-1}(t,t_0;p_1,\cdots,p_{n-1})}=-\infty.\tag{47}$$

Suppose moreover that (13) *holds with* $\varphi_i(t)$ *defined by* (26). *Then, if* $r=\max r_i\leqq 1$, *all solutions of* (A) *are oscillatory, and if* $r>1$, *solutions of* (A) *with sufficiently large initial time and sufficiently small initial values are oscillatory.*

Proof. First observe that (46) and (47) imply that $\int_a^\infty p_n(t)|f(t)|dt=\infty$. By Theorem 1 and Remark 1 there are positive constants c_i such that $|L_ix(t)|\leqq c_i\varphi_i(t)$, $t\geqq t_0$, for $0\leqq i\leqq n-1$. We divide (37) by $I_{n-1}(t,t_0;p_1,\cdots,p_{n-1})$ and let $t\to\infty$. Since $I_j(t,t_0;p_1,\cdots,p_j)/I_{n-1}(t,t_0;p_1,\cdots,p_{n-1})$, $0\leqq j\leqq n-1$, are bounded as $t\to\infty$ and

$$\frac{I_n\big(t,t_0;p_1,\cdots,p_{n-1},p_n|F[x]|\big)}{I_{n-1}\big(t,t_0;p_1,\cdots,p_{n-1}\big)}\leqq\int_{t_0}^t p_n(s)|F[x](s)|\,ds$$

$$\leqq\int_{t_0}^\infty\sum_{k=0}^{n-1}p_n(s)q_k(s)[c_k\varphi_k(s)]^{r_k}ds<\infty,$$

we then conclude that

$$\limsup_{t\to\infty}L_0x(t)/I_{n-1}(t,t_0;p_1,\cdots,p_{n-1})=\infty,$$
$$\liminf_{t\to\infty}L_0x(t)/I_{n-1}(t,t_0;p_1,\cdots,p_{n-1})=-\infty.$$

This shows that $x(t)$ is oscillatory and the proof is complete.

Remark 5. We now replace condition (3) by

$$(48)\qquad u_0F(t,u_0,u_1,\cdots,u_{n-1})\geqq 0\quad\text{for }(t,u_0,u_1,\cdots,u_{n-1})\in[a,\infty)\times R^n.$$

Then only the conditions (46) and (47) (without requiring (13)) guarantee the oscillation of all solutions of (A) that are continuable to ∞. In fact, suppose to the contrary that (A) has a nonoscillatory solution $x(t)$ which is eventually positive. Then, since $F[x](t)\geqq 0$, $t\geqq t_0$, t_0 being sufficiently large, it follows from (37) that

$$L_0x(t)\leqq\sum_{j=0}^{n-1}L_jx(t_0)I_j(t,t_0;p_1,\cdots,p_j)+I_n(t,t_0;p_1,\cdots,p_{n-1},p_nf)$$

for $t\geqq t_0$. Dividing the above inequality by $I_{n-1}(t,t_0;p_1,\cdots,p_{n-1})$ and letting $t\to\infty$, we see with the aid of (47) that

$$\liminf_{t\to\infty}L_0x(t)/I_{n-1}(t,t_0;p_1,\cdots,p_{n-1})=-\infty,$$

which contradicts the eventual positivity of $x(t)$. Likewise we are led to a contradiction if we assume the existence of an eventually negative solution of (A).

Remark 6. Oscillation results similar to but weaker than Theorem 3 have been obtained in the papers by Kusano [16] and Singh and Kusano [26].

Example 1. Consider the equation

$$(49)\qquad x^{(n)}+c(t)|x|^\gamma\operatorname{sgn}x=t^m\sin t,\qquad t\geqq 1,$$

where γ and m are positive constants and $c(t)$ is a continuous function on $[a,\infty)$. Assume that $m>n-1$ and

$$(50)\qquad\int_1^\infty t^{(m+n)\gamma}|c(t)|\,dt<\infty.$$

From Theorem 3 it follows that if $\gamma\leqq 1$, then all solutions of (49) are oscillatory and if $\gamma>1$, then solutions of (49) with large initial time and small initial values are oscillatory. In view of Remark 5, if $c(t)\geqq 0$ on $[a,\infty)$, then the same conclusion holds without condition (50).

Remark 7. Theorems 2 and 3 are concerned with the case where the integral of $p_n(t)|f(t)|$ diverges. The situation is different if $p_n(t)|f(t)|$ is integrable on $[a,\infty)$ or satisfies a stronger integrability condition. In fact, if $\int_a^\infty p_n(t)|f(t)|\,dt<\infty$ and if (13) holds with $\varphi_i(t)$ defined by (24), then there exists a nonoscillatory solution $x(t)$ of (A) such that $\lim_{t\to\infty} L_i x(t)/\varphi_i(t)=1$ for $0\leqq i\leqq n-1$. This result is contained in the following theorem due to Fink and Kusano [5].

THEOREM 4. *Let* k, $0\leqq k\leqq n-1$, *be fixed. Suppose that*

$$\int^\infty I_{n-k-1}(t,a;p_{n-1},\cdots,p_{k+1})p_n(t)|f(t)|\,dt<\infty, \tag{51}$$

$$\int^\infty I_{n-k-1}(t,a;p_{n-1},\cdots,p_{k+1})[I_{k-i}(t,a;p_{i+1},\cdots,p_k)]^{r_i}p_n(t)q_i(t)\,dt<\infty \tag{52}$$

for $0\leqq i\leqq k$,

$$\int^\infty I_{n-k-1}(t,a;p_{n-1},\cdots,p_{k+1})[I_{i-k}(t,a;p_i,\cdots,p_{k+1})]^{-r_i}p_n(t)q_i(t)\,dt<\infty \tag{53}$$

for $k+1\leqq i\leqq n-1$.

Then (A) *has a nonoscillatory solution* $x(t)$ *with the property*

$$\lim_{t\to\infty} L_i x(t)/I_{k-i}(t,a;p_{i+1},\cdots,p_k)=\text{constant}\neq 0,\qquad 0\leqq i\leqq k, \tag{54}$$

$$\lim_{t\to\infty} L_i x(t)\cdot I_{i-k}(t,a;p_i,\cdots,p_{k+1})=0,\qquad k+1\leqq i\leqq n-1. \tag{55}$$

The desired solution $x(t)$ is obtained via the Schauder–Tykhonov fixed-point theorem as a solution of the integral equation

$$\begin{aligned}L_0x(t)={}&cI_k(t,T;p_1,\cdots,p_k)\\&+(-1)^{n-k}\int_T^t I_{k-1}(t,s;p_1,\cdots,p_{k-1})p_k(s)\\&\qquad\int_s^\infty I_{n-k-1}(r,s;p_{n-1},\cdots,p_{k+1})\\&\qquad\qquad\cdot p_n(r)[f(r)-F[x](r)]\,dr\,ds\quad\text{if } 0<k\leqq n-1,\end{aligned}$$

$$L_0x(t)=c+(-1)^n\int_t^\infty I_{n-1}(s,t;p_{n-1},\cdots,p_1)p_n(s)[f(s)-F[x](s)]\,ds\quad\text{if } k=0,$$

where $c>0$ is any given constant, $T>a$ is a suitably chosen large constant and $F[x]$ is defined by (16). That $x(t)$ satisfies (54) and (55) follows by differentiation of this integral equation; identity (7) is needed in order to verify (55).

4. Asymptotic behavior of solutions. The purpose of this section is to establish conditions which guarantee that every solution of (A) tends to a limit as $t\to\infty$.

THEOREM 5. *Suppose that*

$$\int^\infty I_{n-1}(t,a;p_{n-1},\cdots,p_1)p_n(t)|f(t)|\,dt<\infty, \tag{56}$$

$$\int^\infty I_{n-1}(t,a;p_{n-1},\cdots,p_1)[I_{n-i-1}(t,a;p_{i+1},\cdots,p_{n-1})]^{r_i}p_n(t)q_i(t)\,dt<\infty \tag{57}$$

for $0\leqq i\leqq n-1$.

Then, if $r=\max_i r_i \leqq 1$, *for every solution* $x(t)$ *of* (A), $L_0x(t)$ *tends to a limit (finite or infinite), and if* $r>1$, *for every solution* $x(t)$ *with sufficiently large initial time and sufficiently small initial values,* $L_0x(t)$ *tends to a limit as* $t\to\infty$. *In particular, for every oscillatory solution* $x(t)$ *of* (A), $L_0x(t)$ *tends to zero as* $t\to\infty$.

Proof. Our proof patterns after that of Singh and Kusano [24]. Let $x(t)$ be a solution of (A) (with large initial time and small initial values if $r>1$). Since (56) and (57) imply (13) with $\varphi_i(t)=I_{n-i-1}(t,a;p_{i+1},\cdots,p_{n-1})$ (see Remark 1), by Theorem 1, there exist positive constants c_i such that

$$(58)\qquad |L_ix(t)|/I_{n-i-1}(t,a;p_{i+1},\cdots,p_{n-1})\leqq c_i,\qquad 0\leqq i\leqq n-1.$$

Suppose to the contrary that $L_0x(t)$ does not approach a limit as $t\to\infty$. Then, there are two constants ξ and η such that

$$(59)\qquad \liminf_{t\to\infty} L_0x(t)<\xi<\eta<\limsup_{t\to\infty} L_0x(t).$$

Let $T>a$ be large enough so that

$$(60)\qquad \int_T^\infty I_{n-1}(t,a;p_{n-1},\cdots,p_1)p_n(t)$$
$$\left[|f(t)|+\sum_{i=0}^{n-1}\left[c_iI_{n-i-1}(t,a;p_{i+1},\cdots,p_{n-1})\right]^{r_i}q_i(t)\right]dt<\frac{\eta-\xi}{2}.$$

Choose $T<S_0<T_0<S_1<T_1$ so that

$$L_0x(S_0)<\xi<\eta<L_0x(T_0)$$

and

$$L_0x(S_1)<\xi<\eta<L_0x(T_1).$$

Let $[s_1,s_2]$ be the smallest interval containing T_0 such that

$$(61)\qquad \begin{aligned}&L_0x(s_1)=L_0x(s_2)=\xi,\quad\text{and}\\ &\max\{L_0x(t):t\in[s_1,s_2]\}=L_0x(s')>\eta.\end{aligned}$$

Clearly, $T<s_1<s'<s_2$. Let $s_2\leqq t_1\leqq t_2\leqq\cdots\leqq t_{n-1}$ be such that

$$(62)\qquad L_ix(t_i)=0,\qquad 1\leqq i\leqq n-1.$$

Such t_i exist, since $L_ix(t)$, $1\leqq i\leqq n-1$, are oscillatory because of (59).

Integrating (A) $n-1$ times and using (62), we obtain

$$(63)$$
$$L_1x(t)=(-1)^{n-1}\int_t^{t_1}p_2(r_2)\int_{r_2}^{t_2}\cdots$$
$$\int_{r_{n-2}}^{t_{n-2}}p_{n-1}(r_{n-1})\int_{r_{n-1}}^{t_{n-1}}p_n(r_n)[f(r_n)-F[x](r_n)]\,dr_n\,dr_{n-1}\cdots dr_2,$$

where $F[x]$ is defined by (16). Multiplying (63) by $p_1(t)$ and integrating from s_1 to s', we have in view of (61)

$$\eta-\xi<\int_{s_1}^{s'}p_1(r_1)\int_{r_1}^{t_1}p_2(r_2)\int_{r_2}^{t_2}\cdots$$

$$\int_{r_{n-2}}^{t_{n-2}}p_{n-1}(r_{n-1})\int_{r_{n-1}}^{t_{n-1}}p_n(r_n)\big[|f(r_n)|+|F[x](r_n)|\big]\,dr_n\,dr_{n-1}\cdots dr_2\,dr_1$$

$$\leqq\int_{s_1}^{t_{n-1}}p_1(r_1)\int_{r_1}^{t_{n-1}}p_2(r_2)\int_{r_2}^{t_{n-1}}\cdots$$

$$\int_{r_{n-2}}^{t_{n-1}}p_{n-1}(r_{n-1})\int_{r_{n-1}}^{t_{n-1}}p_n(r_n)\big[|f(r_n)|+|F[x](r_n)|\big]\,dr_n\,dr_{n-1}\cdots dr_2\,dr_1$$

$$=\int_{s_1}^{t_{n-1}}I_{n-1}(r,s_1;p_{n-1},\cdots,p_1)p_n(r)\big[|f(r)|+|F[x](r)|\big]\,dr$$

$$<\int_{s_1}^{t_{n-1}}I_{n-1}(r,s_1;p_{n-1},\cdots,p_1)p_n(r)$$

$$\cdot\left[|f(r)|+\sum_{i=0}^{n-1}\left[c_iI_{n-i-1}(r,a;p_{i+1},\cdots,p_{n-1})\right]^{r_i}q_i(r)\right]dr<\frac{\eta-\xi}{2},$$

where we have used (60). This contradiction completes the proof.

As a byproduct we have the following nonoscillation theorem for the homogeneous version of (A), i.e.

$$(\mathrm{A}_0)\qquad L_nx+F(t,L_0x,L_1x,\cdots,L_{n-1}x)=0.$$

THEOREM 6. *Suppose that conditions* (3) *and* (57) *are satisfied with* $r_i=1$, $0\leqq i\leqq n-1$. *Then, all nontrivial solutions of* (A_0) *are nonoscillatory.*

Proof. Suppose to the contrary that (A_0) has an oscillatory solution $x(t)$. By Theorem 5, $L_0x(t)\to 0$ as $t\to\infty$. It can be shown that $L_ix(t)\to 0$ as $t\to\infty$ for $1\leqq i\leqq n-1$. In fact, let t be fixed arbitrarily and take t_i, $0\leqq i\leqq n-1$, so that $t\leqq t_0\leqq t_1\leqq\cdots\leqq t_{n-1}$ and (62) holds and $L_0x(t_0)=0$. Then, integration of (A), for any i, $1\leqq i\leqq n-1$, yields

$$L_ix(t)=(-1)^{n-i}\int_t^{t_i}p_{i+1}(r_{i+1})\int_{r_{i+1}}^{t_{i+1}}\cdots$$

$$\int_{r_{n-2}}^{t_{n-2}}p_{n-1}(r_{n-1})\int_{r_{n-1}}^{t_{n-1}}p_n(r_n)F[x](r_n)\,dr_n\,dr_{n-1}\cdots dr_{i+1}.$$

Therefore,

$$(64)\qquad |L_ix(t)|\leqq\int_t^{t_{n-1}}I_{n-i-1}(r,t;p_{n-1},\cdots,p_{i+1})p_n(r)|F[x](r)|\,dr$$

$$\leqq\int_t^{t_{n-1}}I_{n-i-1}(r,t;p_{n-1},\cdots,p_{i+1})p_n(r)\sum_{k=0}^{n-1}q_k(r)|L_kx(r)|\,dr$$

$$\leqq\int_t^{\infty}I_{n-i-1}(r,t;p_{n-1},\cdots,p_{i+1})p_n(r)\sum_{k=0}^{n-1}q_k(r)|L_kx(r)|\,dr,$$

which implies that $L_ix(t)\to 0$ as $t\to\infty$, $1\leqq i\leqq n-1$. Inequality (64) also holds for $i=0$. From the above observation it follows that

$$(65)\qquad |L_ix(t)|\leqq\int_t^\infty I_{n-i-1}(r,t;p_{n-1},\cdots,p_{i+1})p_n(r)\sum_{k=0}^{n-1}q_k(r)|L_kx(r)|\,dr$$

for $0\leqq i\leqq n-1$. If we put $X(t)=\Sigma_{i=0}^{n-1}|L_ix(t)|$, then we conclude from (65) that

$$X(t)\leqq\sup_{s\geqq t}X(s)\cdot\sum_{i,k=0}^{n-1}\int_t^\infty I_{n-i-1}(r,t;p_{n-1},\cdots,p_{i+1})p_n(r)q_k(r)\,dr,$$

and consequently

$$\sup_{s\geqq t}X(s)\leqq\sup_{s\geqq t}X(s)\cdot\sum_{i,k=0}^{n-1}\int_t^\infty I_{n-i-1}(r,t;p_{n-1},\cdots,p_{i+1})p_n(r)q_k(r)\,dr.$$

Since, in view of (58) and $r_i=1$, the sum of the last integrals tends to zero as $t\to\infty$, we have a contradiction. This finishes the proof.

Example 2. Consider the equation

$$(66)\quad (tx'')''+\frac{1}{t^4}x''+\frac{1}{t^6}x=\frac{2\sin(\log t)-3\cos(\log t)}{t^7}-\frac{10\sin(\log t)+20\cos(\log t)}{t^4}.$$

The conditions of Theorem 5 are satisfied, and so all solutions of (66) approach limits (finite or infinite) as $t\to\infty$. In particular every oscillatory solution of (66) tends to zero as $t\to\infty$. One such solution is $x(t)=\sin(\log t)/t$. Equation (66) also has nonoscillatory solutions. In fact, Theorem 4 is applicable and one sees that there exist solutions $x_i(t)$, $1\leqq i\leqq 4$, with the following properties:

$$\lim_{t\to\infty}x_1(t)=\text{constant}\neq 0,\qquad \lim_{t\to\infty}x_2(t)/t=\text{constant}\neq 0,$$
$$\lim_{t\to\infty}x_3(t)/t\log t=\text{constant}\neq 0,\qquad \lim_{t\to\infty}x_4(t)/t^2=\text{constant}\neq 0.$$

COROLLARY 3. *Suppose that*

$$(67)\qquad \int^\infty t^{n-1}|f(t)|\,dt<\infty,$$

$$(68)\qquad \int^\infty t^{n-1+(n-i-1)r_i}q_i(t)\,dt<\infty,\qquad 0\leqq i\leqq n-1.$$

Then, if $r\leqq 1$, *every solution of* (B) *tends to a limit* (*finite or infinite*) *as* $t\to\infty$, *and if* $r>1$, *every solution with large initial time and small initial values tends to a limit as* $t\to\infty$. *In particular, every oscillatory solution of* (B) *tends to zero as* $t\to\infty$.

COROLLARY 4. *Consider the equation*

$$(\mathrm{B}_0)\qquad x^{(n)}+F(t,x,x',\cdots,x^{(n-1)})=0,$$

where $F(t,u_0,u_1,\cdots,u_{n-1})$ *satisfies* (3) *with* $r_i=1$, $0\leqq i\leqq n-1$. *If* (68) *holds with* $r_i=1$, $0\leqq i\leqq n-1$, *then all solutions of* (B_0) *are nonoscillatory.*

We conclude with an example showing that the above results concerning the ordinary differential equation (A) can be applied to the qualitative study of a particular class of partial differential equations.

Example 3. Consider the fourth order elliptic equation

$$\Delta^2 u+\alpha(|y|)\Delta u+\beta(|y|)u=\gamma(|y|) \tag{69}$$

in an exterior domain Ω in R^3, where $y=(y_1,y_2,y_3)\in R^3$, $|y|$ is the Euclidean length of y, Δ is the Laplace operator, and $\alpha(t)$, $\beta(t)$ and $\gamma(t)$ are continuous functions on $[a,\infty)$ for some $a>0$. Noting that for a spherically symmetric function $v(|y|)$

$$\Delta v(|y|)=t^{-1}\frac{d}{dt}t^2\frac{dv}{dt}=t^{-1}\frac{d^2}{dt^2}(tv), \qquad t=|y|,$$

we see that a spherically symmetric function $u(|y|)$ is a solution of (69) in some exterior domain $\Omega_T=\{y\in R^3:|y|\geqq T\}$, $T\geqq a$, if and only if $u(t)$ is a solution of the ordinary differential equation

$$t^{-1}\frac{d^4}{dt^4}(tu)+\alpha(t)t^{-1}\frac{d^2}{dt^2}(tu)+\beta(t)u=\gamma(t), \qquad t\geqq T, \tag{70}$$

which is a special case of (A) satisfying (2) and (3) with $p_0(t)=t^{-1}$, $p_1(t)=p_2(t)=p_3(t)=1$, $p_4(t)=t$, $q_0(t)=t^{-1}|\beta(t)|$, $q_1(t)=0$, $q_2(t)=t^{-1}|\alpha(t)|$, $q_3(t)=0$, $f(t)=\gamma(t)$.

Applying Theorems 1–3 and 5 to (70) we have the following results concerning spherically symmetric solutions of (69) defined in some exterior domain in R^3:

(i) If we suppose that

$$\int_1^\infty t|\alpha(t)|\,dt<\infty, \quad \int_1^\infty t^3|\beta(t)|dt<\infty, \quad \int_1^\infty t|\gamma(t)|\,dt<\infty, \tag{71}$$

then all spherically symmetric solutions $u(|y|)$ of (69) satisfy $u(|y|)=O(|y|^2)$ as $|y|\to\infty$.

(ii) If we replace (71) by

$$\int_1^\infty t^4|\alpha(t)|\,dt<\infty, \quad \int_1^\infty t^6|\beta(t)|\,dt<\infty, \quad \int_1^\infty t^4|\gamma(t)|\,dt<\infty, \tag{72}$$

then, for every spherically symmetric solution $u(|y|)$ of (69), $|y|u(|y|)$ tends to a limit (finite or infinite) as $|y|\to\infty$; in particular, $|y|u(|y|)\to 0$ as $|y|\to\infty$ for every oscillatory solution $u(|y|)$ of (69).

(iii) Suppose that $\gamma(t)$ is eventually of constant sign and

$$\int_1^\infty t\gamma(t)\,dt=\infty \quad \text{or} \ -\infty. \tag{73}$$

If in addition

$$\int_1^\infty |\alpha(t)|\left(\int_1^t (t-s)s|\gamma(s)|\,ds\right)dt<\infty, \tag{74}$$

$$\int_1^\infty |\beta(t)|\left(\int_1^t (t-s)^3 s|\gamma(s)|\,ds\right)dt<\infty, \tag{75}$$

then all spherically symmetric solutions of (69) are nonoscillatory.

(iv) Suppose that for any $T\geqq a$

$$\limsup_{t\to\infty}\int_T^t (t-s)^3 s\gamma(s)\,ds/t^3=\infty, \tag{76}$$

$$\liminf_{t\to\infty}\int_T^t (t-s)^3 s\gamma(s)\,ds/t^3=-\infty. \tag{77}$$

If in addition (74) and (75) hold, then all spherically symmetric solutions of (69) are oscillatory.

Acknowledgment. The authors would like to express their sincere thanks to the referee for very helpful comments and suggestions.

REFERENCES

[1] I. BIHARI, *A generalization of a lemma of Bellman and its application to uniqueness problems of differential equations*, Acta Math. Acad. Sci. Hungar., 7 (1956), pp. 81–94.

[2] T. A. ČANTURIJA, *On monotone and oscillatory solutions of higher order ordinary differential equations*, Ann. Polon. Math., 37 (1980), pp. 93–111. (In Russian.)

[3] U. ELIAS, *Oscillatory solutions and extremal points for linear differential equations*, Arch. Rational Mech. Anal., 71 (1979), pp. 177–198.

[4] A. M. FINK AND T. KUSANO, *Nonoscillation theorems for differential equations with general deviating arguments*, Lecture Notes in Mathematics 1032, Springer-Verlag, New York, 1983, pp. 224–239.

[5] ———, *Nonoscillation theorems for a class of perturbed disconjugate differential equations*, Japan. J. Math. (New Series), 9 (1983), pp. 277–291.

[6] A. GRANATA, *Singular Cauchy problems and asymptotic behavior for a class of nth order differential equations*, Funkcial. Ekvac., 20 (1977), pp. 193–212.

[7] ———, *Canonical factorizations of disconjugate differential operators*, this Journal, 11 (1980), pp. 160–172.

[8] J. K. HALE AND N. ONUCHIC, *On the asymptotic behavior of solutions of a class of differential equations*, Contributions to Differential Equations, 2 (1963), pp. 61–75.

[9] T. G. HALLAM, *Asymptotic behavior of the solutions of an nth order nonhomogeneous ordinary differential equation*, Trans. Amer. Math. Soc., 122 (1966), pp. 177–194.

[10] ———, *Asymptotic expansions of the subdominant solutions of a class of nonhomogeneous differential equations*, J. Differential Equations, 6 (1969), pp. 125–141.

[11] G. W. JOHNSON, *A bounded nonoscillatory solution of an even order linear differential equation*, J. Differential Equations, 15 (1974), pp. 172–177.

[12] W. J. KIM, *Properties of disconjugate linear differential operators*, J. Differential Equations, 43 (1982), pp. 369–398.

[13] Y. KITAMURA AND T. KUSANO, *Oscillation criteria for semilinear metaharmonic equations in exterior domains*, Arch. Rational Mech. Anal., 75 (1980), pp. 79–90.

[14] ———, *Nonlinear oscillation of higher-order functional differential equations with a general deviating argument*, J. Math. Anal. Appl., 77 (1980), pp. 100–119.

[15] K. KREITH AND T. KUSANO, *Extremal solutions of general nonlinear differential equations*, Hiroshima Math. J., 10 (1980), pp. 141–152.

[16] T. KUSANO, *Oscillation theory of higher-order ordinary and functional differential equations with forcing terms*, Proc. Fifth Czechoslovak Conference on Differential Equations and Their Applications, Bratislava, August 24–28, 1981, M. Greguš, ed., Teubner, Leipzig, 1982, pp. 218–221.

[17] T. KUSANO AND M. NAITO, *Comparison theorems for functional differential equations with deviating arguments*, J. Math. Soc. Japan, 33 (1981), pp. 509–532.

[18] ———, *Boundedness of solutions of a class of higher order ordinary differential equations*, J. Differential Equations, 46 (1982), pp. 32–45.

[19] D. L. LOVELADY, *On the oscillatory behavior of bounded solutions of higher order differential equations*, J. Differential Equations, 19 (1975), pp. 167–175.

[20] Z. NEHARI, *Disconjugate linear differential operators*, Trans. Amer. Math. Soc., 129 (1967), pp. 500–516.

[21] R. OLAH, *Note on the oscillatory behavior of bounded solutions of higher order differential equations with retarded argument*, Arch. Math. (Brno), 14 (1978), pp. 171–174.

[22] H. ONOSE, *Asymptotic behavior of nonoscillatory solutions of a higher order functional differential equation*, Bull. Austral. Math. Soc., 24 (1981), pp. 85–92.

[23] CH. G. PHILOS AND V. A. STAIKOS, *Boundedness and oscillation of solutions of differential equations with deviating argument*, An. Sti. Univ. "Al. I. Cuza" Iaşi Sect. I a Mat., 26 (1980), pp. 307–317.

[24] B. SINGH AND T. KUSANO, *On asymptotic limits of nonoscillations in functional equations with retarded arguments*, Hiroshima Math. J., 10 (1980), pp. 557–565.

[25] ———, *Asymptotic behavior of oscillatory solutions of a differential equation with deviating arguments*, J. Math. Anal. Appl., 83 (1981), pp. 395–407.

[26] B. SINGH AND T. KUSANO, *Forced oscillations in functional differential equations with deviating arguments*, Arch. Math. (Brno), 19 (1983), pp. 9–17.

[27] V. A. STAIKOS AND CH. G. PHILOS, *Nonoscillatory phenomena and damped oscillations*, Nonlinear Anal., 2 (1978), pp. 197–210.

[28] M. ŠVEC, *Behavior of nonoscillatory solutions of some nonlinear differential equations*, Acta Math. Univ. Comenian., 39 (1980), pp. 115–130.

[29] W. F. TRENCH, *Asymptotic behavior of solutions of* $Lu=g(t,u,\cdots,u^{(k-1)})$, J. Differential Equations, 11 (1972), pp. 38–48.

[30] ______, *Asymptotic behavior of solutions of perturbed disconjugate equations*, J. Differential Equations, 11 (1972), pp. 661–671.

[31] ______, *Canonical forms and principal systems for general disconjugate equations*, Trans. Amer. Math. Soc., 189 (1974), pp. 319–327.

[32] ______, *Oscillation properties of perturbed disconjugate equations*, Proc. Amer. Math. Soc., 52 (1975), pp. 147–155.

[33] ______, *Asymptotic theory of perturbed general disconjugate equations*, Hiroshima Math. J., 12 (1982), pp. 43–58.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

010

ORDER STARS, APPROXIMATIONS AND FINITE DIFFERENCES II. THEOREMS IN APPROXIMATION THEORY*

A. ISERLES†

Abstract. The theory of order stars is applied to two problems in approximation theory: determination of upper bounds on the size of blocks in the Padé tableau of an analytic function and the task of highest-order approximation of a stable analytic function by a contraction. It is shown that in both cases global properties of the underlying function—the loci and the nature of essential singularities, zeros and poles—lead to realistic upper bounds. The given theory is applied to a range of examples that illustrate the potential of this approach.

1. Introduction. The subject of the present paper is the application of order stars to some problems in approximation theory. The theory of order stars started with a paper by Wanner, Hairer and Nørsett [11], who addressed themselves to stability properties of rational approximations to the exponential function. It was subsequently applied to rational functions that arise in the discretization of hyperbolic equations by finite differences [7]. In paper Part I of this series [8] the present author generalizes the theory, giving a framework for the analysis of arbitrary function, analytic except for isolated poles and essential singularities, by other functions of similar type.

Our first problem is an approximation of analytic functions by rational functions. Given a function f, analytic in a neighbourhood of the origin, and a rational function $R \in \pi_{m/n}$. where

$$\pi_{m/n} := \left\{ \frac{P}{Q} : P, Q \text{ polynomials}, \deg P = m, \deg Q = n, Q(0) = 1 \right\},$$

we say that R is an approximation of order p if

$$R(z) = f(z) + cz^{p+1} + O\left(|z|^{p+2}\right), \qquad c \neq 0.$$

Let us consider an approximation $R_{m/n}$ that attains the maximal possible order in $\pi_{m/n}$. If its order is at least $m+n$, then it is called the $[m/n]$ *Padé approximation* [2]. Such an approximation is unique. An arrangement of the $R_{m/n}$'s for m, $n \geqq 0$ in an infinite matrix gives the *Padé tableau* of f. As is well known, the Padé tableau is composed out of square blocks of identical approximations. Furthermore, if $R_{m/n}$ appears at the northwestern corner of a q-by-q block, then it is of order $m+n+q-1$. The question we pose is how to deduce the size of the maximal block in a Padé tableau from analytic properties of the underlying function f. More formally, let $p(m/n)$ denote the order of $R_{m/n}$. Then

$$\beta(f) := \max\{ p(m/n) - m - n + 1 : m, n \geqq 0 \} \tag{1}$$

gives the maximal size of a block in the Padé tableau of f. We call it the *block number* of f. It is known [2] that, unless f itself is a rational function, $p(m/n)$ is bounded for every m, $n \geqq 0$. However, it may well happen that $\beta(f) = \infty$.

*Received by the editors June 14, 1983, and in revised form October 24, 1983.

† King's College, University of Cambridge, Cambridge CB2 1ST, England

In §2 we use order stars to obtain realistic bounds on the size of $\beta(f)$. It turns out that the nature of singularities of f determinees a bound on the block number and, accompanied by simple technical manipulation, frequently gives explicitly the value of $\beta(f)$.

Section 3 is devoted to numerous examples of functions whose block number can be determined by the theorems of §2—Bessel functions, Mittag–Leffler functions, trigonometric functions, a theta-like function, etc.

In §4 we address ourselves to a different problem in approximation theory, namely the contractive approximation of stable functions. Given a complex domain V and a function f, analytic in V and such that $|f(z)|=1$ for every $z\in\partial V$, with the possible exception of essential singularities, and $|f(z)<1$ within V, we ask what is the highest degree of interpolation of f by an arbitrary analytic function R in cl V which preserves the property that $|R(z)|\leqq 1$ for every $z\in \mathrm{cl}\, V$. Once again, the answer comes from the order star theory, bounding the maximal degree of interpolation in terms of the number of zeros of f in V and the nature of its essential singularities on ∂V.

Examples of contractive approximations are given in §5. Among other results we derive there the celebrated result of Wanner, Hairer and Nørsett [11], namely the proof of the first Ehle conjecture, as a special case of one of our theorems. We also establish the connection between our results and the classical Pick theorem [10].

Our analysis consists mainly of an application of the theory that has been developed in [8]. In particular, we extensively use Propositions 1–6 therein. These propositions, as well as definitions and concepts of the order star theory, will be mentioned in the text with no further reference to [8].

2. Bounds on the block number. Let f be a function, analytic in $\mathrm{cl}\,\mathbb{C}=\mathbb{C}\cup\{\infty\}$ with the possible exception of a finite number of essential singularities and at most a countable number of poles. We further assume that the origin is an analytic point of f and, to avoid spurious blocks in the first few columns of the Padé tableau, that $f(0)\neq 0$.

Given that $z_1,\cdots,z_L$ are all the essential singularities of f, we set

$$I(f):=\sum_{j=1}^{L}\mathrm{ind}(z_j).$$

All the theorems of this section depend upon counting the number of sectors of A (and possibly of D) that may approach the origin and invoking Proposition 1 to obtain upper bounds on order.

THEOREM 1. *If f is analytic and nonzero in $\mathbb{C}/\{z_1,\cdots,z_L\}$ and has essential singularities at $z_1,\cdots,z_L$, where, without loss of generality, $z_1=\infty$, then*

$$\beta(f)\leqq I(f)+L-1. \tag{2}$$

Proof. Let $R\in\pi_{m/n}$ be a given approximation of order p to f. By Proposition 1 there are exactly $p+1$ sectors of A and $p+1$ sectors of D approach the origin, since $f(0)\neq 0$.

Since the poles of $\sigma(z):=R(z)/f(z)$ and of $R(z)$ coincide, there are exactly n poles of σ. Hence, it follows from Proposition 2 that at most n sectors of A that reach the origin may belong to analytic A-regions.

The remaining sectors belong to regions that have essential singularities on their boundary. Each z_j is approached by at most $\mathrm{ind}(z_j)$ sectors of A. If $2\leqq j\leqq L$, then this adds at most $\mathrm{ind}(z_j)+1$ to our count of sectors at the origin, since one of the regions may encircle z_j (cf. Fig. 1). No region encircles ∞ and $z_1=\infty$ adds at most $\mathrm{ind}(z_1)$ such sectors. Finally, there is the possibility for sectors of D to bisect sectors of A inside a nonanalytic A-region. Such sectors of D must necessarily belong to analytic D-regions. Since σ has exactly m zeros, Proposition 2 implies that this may result in at most m sectors of A.

This exhausts all the possibilities of sectors of A reaching the origin, giving the upper bound

$$I(f)+L+m+n-1$$

on the number of such sectors. Proposition 1 now gives

$$p \leqq I(f)+L+m+n-2.$$

This is true for every $R \in \pi_{m/n}$, in particular,

$$p(m/n) \leqq I(f)+L+m+n-2 \tag{3}$$

and the upper bound (2) follows at once from the definition (1) of the block number. □

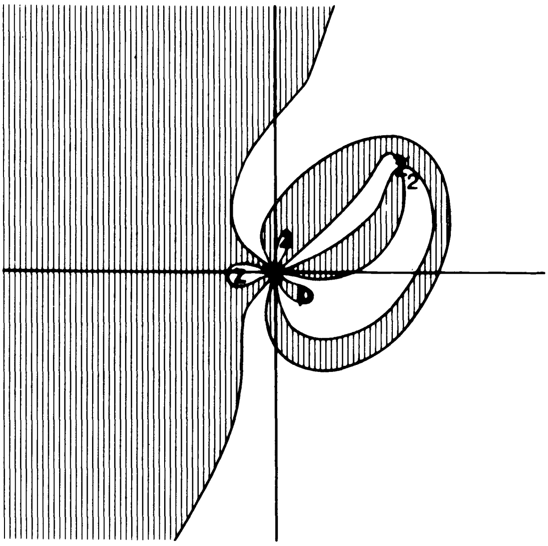


FIG. 1. *A schematic order star with* $L=2$, $\tilde{z}_1=\infty$, $\text{ind}(\tilde{z}_1)=1$, $\text{ind}(\tilde{z}_2)=2$, $m=2$, $n=1$, *that satisfies* (3) *with an equality. The dark-shaded area denotes* A *and* "p", "z" *denote poles and zeros, respectively, of* σ.

Note that the essential singularity at ∞ has a distinct role in the proof of the last theorem. Indeed, if $z_1, \cdots, z_L$ are finite, then the bound on $\beta(f)$ is slightly more generous:

THEOREM 2. *If* f *is analytic and nonzero in* $\mathbb{C}/\{z_1, \cdots, z_L\}$, *has essential singularities at* $z_1, \cdots, z_L$ *and is analytic at* ∞ *then*

$$\beta(f) \leqq (f)+L. \tag{4}$$

Proof. Once again, we count sectors that may approach the origin, given $R \in \pi_{m/n}$ of order p.

If $n \geqq m$, we count sectors of A. At most n of them may be "contributed" by poles of R, $I(f)+L$ by essential singularities and m by sectors of D that bisect sectors of A and belong to analytic bounded D-regions—all this like in the proof of Theorem 1. Note that, although σ has a zero of multiplicity $n-m$ at ∞, this zero contributes nothing to our count. By substituting the upper bound on the number of sectors into Proposition 1 we obtain

$$p \leqq I(f)+m+n+L-1.$$

The theorem follows at once from the definition (1).

The second case, $m \geqq n+1$, is dealt with by counting sectors of D, instead of sectors of A, in an identical manner. This leads, once again, to the upper bound (4).
□

Theorems 1 and 2 can be readily extended to functions f that are analytic and nonzero in $\mathrm{cl}\,\mathbb{C}/\{z_1,\cdots,z_L;\xi_1,\cdots,\xi_M;\eta_1,\cdots,\eta_N\}$, possess essential singularities at $z_1,\cdots,z_L$, zeros at $\xi_1,\cdots,\xi_M$ and poles at $\eta_1,\cdots,\eta_N$. Let the sums of the multiplicities of the zeros and of the poles be denoted by M^* and N^* respectively. It follows at once from the method of proof of the last theorem that an approximation of f at the origin by a function from $\pi_{m/n}$ leads to a similar bound on $\beta(f)$ as an approximation of a function f^*, analytic and nonzero in $\mathrm{cl}\,\mathbb{C}/\{z_1,\cdots,z_L\}$ by a function from $\pi_{m+N^*/n+M^*}$. In other words, the poles (zeros) of an approximation R play the same role as the zeros (poles) of f in our proof. This leads at once to a more general upper bound:

THEOREM 3. *If f is analytic and nonzero in $\mathrm{cl}\,\mathbb{C}/\{z_1,\cdots,z_L;\xi_1,\cdots,\xi_M;\eta_1,\cdots,\eta_N\}$, has essential singularities at z_j, $1 \leqq j \leqq L$, zeros of multiplicity α_j at ξ_j, $1 \leqq j \leqq M$ and poles of multiplicity β_j at η_j, $1 \leqq j \leqq N$, then*

$$\beta(f) \leqq I(f)+L+\sum_{j=1}^{M}\alpha_j+\sum_{j=1}^{N}\beta_j-K, \tag{5}$$

where $K=0$ if all the z_j's are finite and $K=1$ otherwise.

The bound (5) is not useful if f has an infinite number of zeros, say. However, if all but a finite number of zeros of f are real and f has only a finite number of poles, then a much better upper bound is available.

Let a function f of that type be given. We set $J=0$. If a sequence $\{\zeta_j\}_{j=1}^{\infty}$ of real numbers exists such that

$$\lim_{j\to\infty}\zeta_j=+\infty, \qquad \lim_{j\to\infty}|f(\zeta_j)|<1,$$

we increase J by one and we do likewise subject to the existence of a sequence $\{\mu_j\}_{j=1}^{\infty}$ such that

$$\lim_{j\to\infty}\mu_j=-\infty, \qquad \lim_{j\to\infty}|f(\mu_j)|<1.$$

Hence J is an integer, $0 \leqq J \leqq 2$.

THEOREM 4. *Let the following conditions be satisfied*:

(i) *f is analytic and nonzero in $\mathbb{C}/\{z_1,\cdots,z_L;\xi_1,\cdots,\xi_M;\eta_1,\cdots,\eta_N;\kappa_1,\kappa_2,\cdots\}$, with essential singularities at z_j, $1 \leqq j \leqq L$, where $z_1=\infty$.*

(ii) *f has a pole of multiplicty β_j at each $\eta_j \in \mathbb{C}$, $1 \leqq j \leqq N$.*

(iii) *f has a zero at each $\kappa_j \in \mathbb{R}$, $1 \leqq j$.*

(iv) *$f(\bar{z})=\overline{f(z)}$ for every $z \in \mathbb{C}$.*

Then

$$\beta(f) \leqq I(f) - J + L + \sum_{j=1}^{N} \beta_j + 1. \tag{6}$$

Proof. We count sectors of A that reach the origin, given $R \in \pi_{m/n}$. It is central to our analysis that, because of (iv), the order star is symmetric with respect to the real axis.

Let us suppose that q_+ and q_- sectors of A that reach the origin contain positive or negative poles of σ respectively. Since the order star is symmetric with respect to the real axis, it means that these sectors enclose $q_+ + q_- - 2$ sectors of D (we assume here that q_-, $q_+ \geqq 1$, otherwise our bound can be further reduced). Each A-region that reaches the origin and contains points on the real axis must be either unbounded or contain some of the $q_+ + q_-$ sectors. Hence, the "contribution" of $\kappa_1, \kappa_2, \cdots$ is expressed only in these sectors.

In addition, at most $I(f) + L - 1$ sectors are "justified" by essential singularities (the proof is identical to that of Theorem 1), and at most $n + \sum_{j=1}^{M} \alpha_j$ are accounted for by poles of σ. Furthermore, we have $m + \sum_{j=1}^{N} \beta_j - (q_+ + q_- - 2)$ zeros left and they can contribute to sectors of D that bisect sectors of A at the origin. Finally, J sectors of A that reach the origin from ∞ via the real axis may be bisected by sectors of D, but that "costs" additional $2J$ zeros.

Proposition 1 gives

$$\begin{aligned} p + 1 &\leqq (q_+ + q_-) + (I(f) + L - 1) + \left(n + \sum_{j=1}^{M} \alpha_j \right) \\ &\quad + \left(m + \sum_{j=1}^{N} \beta_j - q_+ - q_- + 2 \right) + J - 2J \\ &= I(f) + L + m + n + \sum_{j=1}^{M} \alpha_j + \sum_{j=1}^{N} \beta_j - J + 1, \end{aligned}$$

and (6) now follows from (1). □

The role of zeros in increasing $\beta(f)$ can be even further suppressed in some instances:

THEOREM 5. *Let f be an entire function such that* $\operatorname{ind}(\infty) = 1$, $f(\bar{z}) = \overline{f(z)}$ *for every* $z \in \mathbf{C}$, *all zeros of f are negative and a positive number r exists such that the ray* (r, ∞) *belongs to D in the order star of $\sigma = R/f$, R rational. Then* $\beta(f) = 1$ *and the Padé tableau is normal.*

Proof. Let R be a function in $\pi_{m/n}$. It is obvious that a single unbounded A-region exists in the order star. We denote it by A_∞ and distinguish between two cases:

(a) No sector of A at the origin belongs to A_∞: n sectors of A can be accounted for by poles of R. All other such sectors must belong to A-regions that contain portions of the real negative ray. By the symmetry of the order star about the real axis, if there are q such sectors, say, they must enclose $q - 1$ sectors of D that belong to bounded (*ergo* analytic) D-regions. Therefore $q - 1 \leqq m$ and, the order being p, Proposition 1 gives $p \leqq m + n$.

(b) Some sectors of A at the origin belong to A_∞: Given that q such sectors belong to A_∞, they enclose $q - 1$ sectors of D that belong to bounded D-regions. Therefore, by

Proposition 2, $q \leqq m+1$. Since at most n additional sectors of A may belong to bounded A-regions, the theorem follows by Proposition 1. □

It is possible to extend further the present theory in three directions: firstly, to obtain theorems that cater for more general functions. Secondly. to obtain stricter inequalities. Thirdly, to examine different patterns of interpolation—two-point Padé approximations, N-point interpolations, approximation with loci of zeros or poles being restricted etc. See [9] and [11] for examples of theorems that restrict the degree of interpolation of $f(z)=e^z$ by rational functions of various forms.

3. Block numbers of certain functions. In the present section we apply the theorems that bound the block number to a range of functions. In all the cases we are, in fact, in a position to find from analytic considerations the exact value of $\beta(f)$.

(a) A nonzero entire function of bounded perfect order of growth ρ is necessarily of the form

$$f(z)=e^{g(z)},$$

where g is a polynomial, $\deg g=\rho$ [1]. It is easy to extend this result and to show that the conditions of Theorem 1 amount to f being of the form

$$f(z)=\exp\Big(g_1(z)+g_2\big((z-z_2)^{-1}\big)+g_3\big((z-z_3)^{-1}\big)+\cdots+g_L(z-z_L)^{-1}\Big).$$

Hence, subject to $I(f)<\infty$, Proposition 5 implies that each g must be a polynomial of degree $\operatorname{ind}(z_j)$, $1\leqq j<L$. The bound (2) is actually attained by the function

$$f(z)=e^{z^K},$$

with $L=1$, $z_1=\infty, \operatorname{ind}(z_1)=K$.

(b)
$$f(z)=e^z\sum_{k=0}^{M}\frac{(-1)^k}{k!}z^k.$$

We are within the conditions of Theorem 3. f has a single essential singularity at ∞ and, by Proposition 5, $\operatorname{ind}(\infty)=1$. It is entire and has M complex zeros. Hence

$$L=1,\quad I(f)=1,\quad \sum_{j=1}^{M}\alpha_j=M,\quad \sum_{j=1}^{N}\beta_j=0,\quad K=1$$

and (5) gives

$$\beta(f)\leqq M+1.$$

However,

$$f(z)=1+\frac{(-1)^M}{(M+1)!}z^{M+1}+O\big(|z|^{M+2}\big)$$

and $p(0/0)=M$. Thus, (1) gives $M+1\leqq\beta(f)$ and, consequently,

$$\beta(f)=M+1.$$

A similar result is attained for

$$f(z)=\frac{e^z}{\sum_{k=0}^{N}(1/k!)z^k}.$$

We have now $\sum_{j=1}^{M}\alpha_j=0$, $\sum_{j=1}^{N}\beta_j=N$,

$$f(z)=1+\frac{1}{(N+1)!}z^{N+1}+O\left(|z|^{N+2}\right)$$

and (5), together with (1), give

$$\beta(f)=N+1.$$

(c) *Trigonometric functions*. Let $f(z)=(\sin z)/z$. It is an entire function, $f(\bar{z})=\overline{f(z)}$, all the zeros are real and $I(f)=2$ (cf. [8]). Furthermore, $J=2$. Hence, by Theorem 4, $\beta(f)\leqq 2$. It now follows at once that

$$\beta(f)=2,$$

since f is even (cf. [8, Fig. 2]).

A similar result can be readily obtained for $f(z)=\cos z$.

(d) $f(z)=(e^2\sin z)/z$.

Let $z=re^{i\theta}$, $r\gg 0$. Then

$$|f(re^{i\theta})|=\frac{1}{2r}|e^{r\cos\theta}||e^{r\sin\theta}-e^{-r\sin\theta}|(1+\sigma(1)).$$

Therefore $p(f)=1$ and, by Proposition 4, ind$(\infty)\leqq 2$. The rays $\theta=\frac{3}{4}\pi$and $\theta=\frac{5}{4}\pi$ are asymptotics of ∂ when $r\to\infty$ and there is an unbounded D-region in the order star of $\sigma=R/f$, R rational, to the right of these rays. Furthermore, since there are poles of σ (zeros of f) along the positive axis with an accumulation point at infinity. $\mathbb{P}(\infty)=\{1,2\}$ and ind$(\infty)=1$. Substitution in (6) gives $\beta(f)\leqq 2$. It follows from the Taylor expansion

$$f(z)=1+z+\frac{1}{3}z^2+\frac{1}{30}z^4+O\left(|z|^5\right)$$

that $p(2/0)=3$. Therefore (1) gives $\beta(f)=2$.

(e) $f(z)=J_\nu(z)/z^{|\nu|}$, *ν integer*. J_ν is the Bessel function and the factor $z^{-|\nu|}$ caters for the $|\nu|$-fold zero at the origin. As is well known, all the zeros of f are real and $f(\bar{z})=\overline{f(z)}$. Moreover, by the asymptotic formula for Bessel functions [4]

$$f(z)=\left(\frac{\pi}{2}\right)^{1/2}z^{-|\nu|+1/2}\cos\left(z-\left(\frac{1}{2}\nu+\frac{1}{4}\right)\pi\right)(1+o(1))$$

for $|z|\gg 0$, $|\arg z|<\pi$. This, together with

$$J_\nu(-z)=(-1)^\nu J_\nu(z),$$

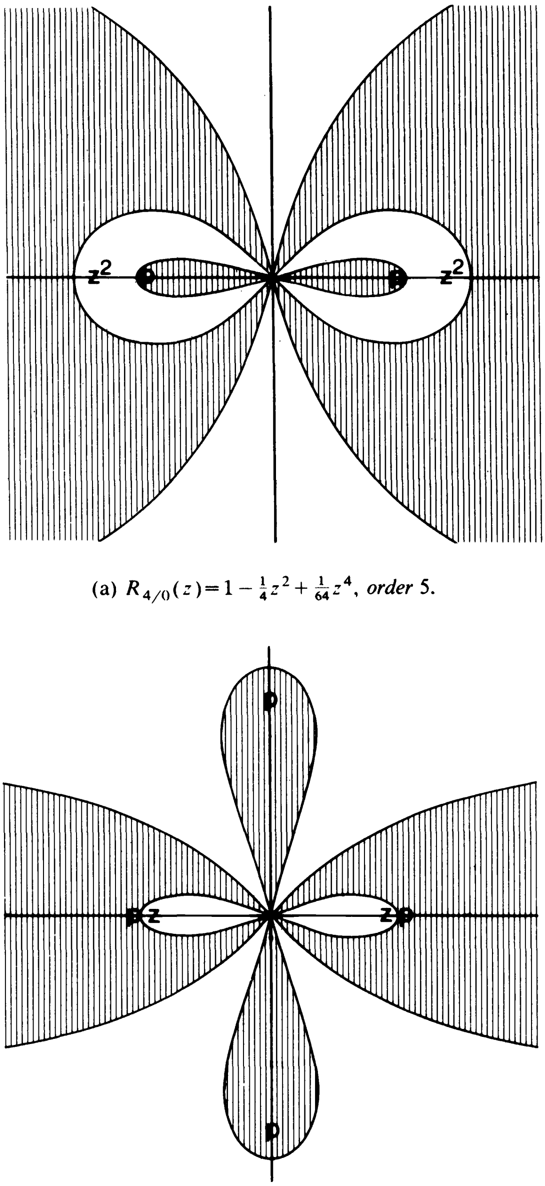
implies that ind$(\infty)=2$. Theorem 4 holds with $I(f)=2$, $L=1$, $J=2$, yielding $\beta(f)\leqq 2$. It follows at once that $\beta(f)=2$, since f is even (cf. Fig. 2).

(f) $$f(z)=E_\alpha(z)=\sum_{k=0}^{\infty}\frac{z^k}{\Gamma(\alpha k+1)},\qquad \alpha>2.$$

E_α is the Mittag–Leffler function. It is entire, $\rho(f)=1/\alpha<1/2$, all the zeros are in $(-\infty,0)$ and $f(\bar{z})=\overline{f(z)}$ [4]. Proposition 4 gives $I(f)=\operatorname{ind}(\infty)=0$ and Theorem 4 is satisfied with $J=1$. Therefore $\beta(f)=1$ and the Padé tableau of E_α is normal for $\alpha>2$.

(g) $f(z)=\prod_{k=1}^{\infty}(1-q^kz)$, $0<q<1$. This is a theta-like function and it is entire [9]. It is easy to show that

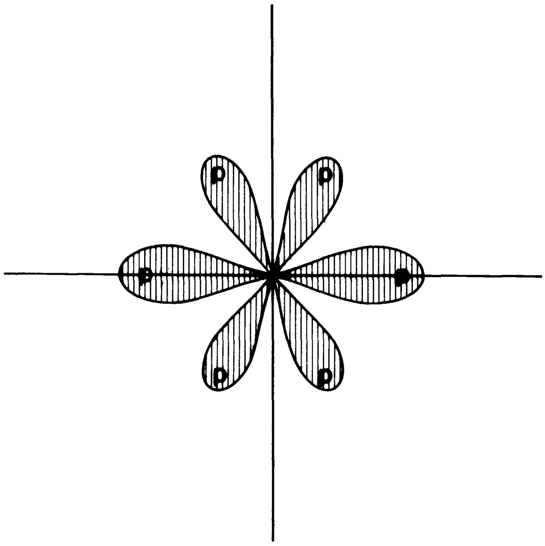
$$f(z)=\sum_{k=0}^{\infty}(-1)^k\frac{k(k+1)/q^2}{[q]_k}z^k, \tag{7}$$



(a) $R_{4/0}(z)=1-\frac{1}{4}z^2+\frac{1}{64}z^4$, *order* 5.

(b) $R_{2/2}(z)=(16+3z^2)/(16+z^2)$, *order* 5.

FIG. 2. *Padé approximations to* $f(z)=J_0(z)$. "z^2" *denotes a zero of* σ *of multiplicity* 2.



(c) $R_{0/4}(z)=1/(1+z^2/4+3z^4/64)$, *order* 5.

FIG. 2. (continued)

where $[q]_0:=1$, $[q]_k=(1-q)(1-q^2)\cdots(1-q^k)$ for $k\geqq 1$. The explicit form of Padé approximations to f is known [5].

$$R_{m/n}(z)=\frac{\sum_{k=0}^{m}(-1)^k\begin{bmatrix} m \\ k \end{bmatrix}\frac{[q]_{m+n-k}}{[q]_{m+n}}q^{k(k+1)/2}z^k}{\sum_{k=0}^{n}\begin{bmatrix} n \\ k \end{bmatrix}\frac{[q]_{m+n-k}}{[q]_{m+n}}z^k},$$

where

$$\begin{bmatrix} j \\ k \end{bmatrix}:=\frac{[q]_j}{[q]_k[q]_{j-k}},\qquad 0\leqq k\leqq j.$$

Therefore, since every two Padé approximations are different, $\beta(f)=1$ and the Padé tableau is normal.

Here we give a proof of that fact based on Theorem 4: since

$$\rho(f)=\limsup_{n\to\infty}\frac{n\log n}{\log|c_n|^{-1}},$$

where c_n is the nth coefficient in the Taylor expansion of f, it follows from (7) and Proposition 4 that

$$I(f)=\operatorname{ind}(\infty)=\rho(f)=0.$$

We are within tbe conditions of Theorem 4 with $J=1$ and (6) yields at once $\beta(f)=1$.

(h) Let f be an entire function, $\rho(f)=1$, with the representation

$$f(z)=\prod_{n=1}^{\infty}(1+c_nz)$$

where $c_n>0$, $N\geqq 1$, and $\lim_{n\to\infty}c_n=0$.

LEMMA 6. *If f is as above then*$\operatorname{ind}(\infty)=1$, ∞ *is a regular point of the order star of* $\sigma=R/f$, *where R is an arbitrary rational function, and there exist exactly one unbounded A-region and one unbounded D-region whose joint boundary asymptotically approaches* $\mathbb{R}$, *the A-region being to the left.*

Proof. Let $z=re^{i\theta}$. Then

$$|f(z)|^2=\prod_{n=1}^{\infty}\left\{1+2c_nr\cos\theta+c_n^2r^2\right\}.$$

Since $\lim_{n\to\infty}c_n=0$, $c_n>0$, for every $\varepsilon>0$ there exists N such that, given $n\geqq N$, $0<c_nr\leqq\varepsilon$. Therefore

$$\begin{aligned}&\cos\theta>\varepsilon,\,n\geqq N\quad\text{implies}\quad 1+2c_nr\cos\theta+c_n^2r^2>1+3\varepsilon^2>1;\\&\cos\theta<-\varepsilon,\,n\geqq N\quad\text{implies}\quad 1+2c_nr\cos\theta+c_n^2r^2<1-\varepsilon^2<1.\end{aligned}\tag{8}$$

Thus $\cos\theta>\varepsilon$ implies that $z\in D$, whereas if $\cos\theta<-\varepsilon$, then $z\in A$. There is still a possibility for "thin" sectors of A and D to fit into the order star near $\mathbb{R}$. However, since the order star is symmetric with respect to $\mathbb{R}$, there must be an equal number of such sectors of A in the upper and lower half-plane. In other words, if such sectors exist, then $\operatorname{ind}(\infty)\geqq 3$. However, Proposition 4 implies that $\operatorname{ind}(\infty)\leqq 2\rho(f)=2$. Therefore $\operatorname{ind}(\infty)=1$ and the lemma follows from (8). □

As a consequence of the last lemma, the function f satisfies the conditions of Theorem 5 and has a normal Padé tableau (i.e. $\beta(f)=1$). An example of such a function is

$$f(z)=\prod_{n=2}^{\infty}\left(1+\frac{z}{n(\log n)^2}\right).$$

By [6] it is indeed of perfect order of growth 1. Given a function f with the factorization

$$f(z)=\prod_{n=1}^{\infty}(1+c_nz),$$

$c_n>0$ for every $n\geqq 1$, it is easy to verify whether indeed $\rho(f)=1$ and we are within the conditions of Lemma 6—this will happen if and only if

$$\sum_{n=1}^{\infty}c_n<\infty$$

and

$$\sum_{n=1}^{\infty}c_n^{\alpha}=\infty$$

for every $\alpha<1$ [6].

4. Contractive approximation of stable functions. In the present section we analyse the following problem: given an open subset V of the extended complex plane with a Jordan boundary, consider a function f, analytic inside V and such that $|f(z)|\equiv 1$ identically along ∂V, with the possible exception of essential singularities. If essential singularities on ∂V occur, we demand that $|f(z)<1$ for every z along the restriction of a circle of radius ε around each essential singularity to V, for every $0<\varepsilon\ll 1$. In that case it follows at once that $|f(z)\leqq 1$ for every $z\in\operatorname{cl}V$. It is often vital to preserve this important property, frequently called *stability*, in approximating f—numerical analysis of differential equations abounds with such examples.

Following [8] we say that a function R is a *V-contraction* if R is analytic in cl V and $|R(z)|\leqq 1$ there. Let us assume that a point $\tilde{z}\in \text{cl}\, V$ has been specified and we wish to interpolate f by R at $\tilde{z}$,

$$R(z)=f(z)+c(z-\tilde{z})^p+O\left(|z-\tilde{z}|^{p+1}\right), \qquad c\neq 0. \tag{9}$$

We consider the problem of finding realistic upper bounds on p, subject to V-contractivity of R. Note that R need not be rational.

It is assumed throughout this section that f is not identically a constant. We denote by L the number of zeros of f in V. If f is analytic along ∂V, then $1\leqq L<\infty$: If $L=0$, then also $1/f$ would have been analytic in cl V and $|f(z)|\equiv 1$ along ∂V, together with the maximum principle, would have led to $1/|f(z)\leqq 1$ in cl V. But this is impossible, since f is not a constant, implying that $1\leqq L$. Moreover, $L<\infty$, otherwise there will be an accumulation point of zeros of f in cl V.

Given $\tilde{z}$ in cl V, we set

$$\alpha(\tilde{z}):=2\pi$$

if $\tilde{z}\in V$. If $\tilde{z}\in\partial V$, we denote by $\alpha(\tilde{z})$ the angle (from within V) that is spanned by the left and right tangents to ∂V at $\tilde{z}$. This definition is meaningful, since the boundary is rectifiable. Note that if $\tilde{z}$ is a smooth boundary point, then $\alpha(\tilde{z})=\pi$. In general $0\leqq\alpha(\tilde{z})\leqq 2\pi$, the extremal values being attained when $\tilde{z}$ is at an apex of a cusp.

THEOREM 7. *If f is analytic along ∂V and $\alpha(\tilde{z})>0$, then*

$$p\leqq\frac{(2L+1)\pi}{\alpha(\tilde{z})} \tag{10}$$

for every V-contractive approximation R.

Proof. By Proposition 3 V-contractivity is equivalent to R being analytic in cl V and $\partial V\cap A=0$. Therefore, given V-contractive R, ∂V separates A-regions. Moreover, since R is analytic in cl V, σ has L poles there. Consequently, the sum of mulitplicities of the A-regions within V is, by Proposition 2, exactly L.

Given (9), it follows from Proposition 1 that $\text{ind}(\tilde{z})=p$ and that $\tilde{z}$ is a regular point of ∂. Therefore, if $\tilde{z}\in V$, then p sectors of A adjoin $\tilde{z}$ inside V. To count sectors of A if $\tilde{z}\in\partial V$, we exploit regularity. If q sectors of A adjoin $\tilde{z}\in\partial V$ from within V, then

$$\frac{2q+1}{p}\pi\geqq\alpha(\tilde{z})\geqq\frac{2q-1}{p}\pi$$

(cf. Fig. 3). Hence, regardless of whether $\tilde{z}\in V$ or $\tilde{z}\in\partial V$.

$$p\leqq\frac{2q+1}{\alpha(\tilde{z})}\pi, \tag{11}$$

q being the number of sectors of A adjoining $\tilde{z}$ from within V. It follows at once from the definition that $q\leqq p$. This, together with (11), completes the proof of the theorem. □

It is interesting to generalize the upper bound (10) to a more elaborate pattern of interpolation. Let $\{\tilde{z}_j\}_{j\in M}$ be given in V (we prohibit $\tilde{z}_j$'s on the boundary to avoid too complicated and messy formulas) and suppose that

$$R(z)=f(z)+c_j(z-\tilde{z}_j)^{p_j}+O\left(|z-\tilde{z}_j|^{p_j+1}\right), \qquad c_j\neq 0, \quad p_j\geqq 0, \quad j\in M.$$

Note that M may be either finite or infinite. If $p_j\geqq 1$, then $\tilde{z}_j$ is an interpolation point. We define

$$P:=\sum_{j\in M} p_j.$$

THEOREM 8. *Given the V-contractive approximation R to a function f that is analytic in* cl V, *it is true that*

$$P\leqq L. \tag{12}$$

Proof. Follows in an identical manner to the proof of Theorem 7, by counting multiplicities of A-regions within V and the number of sectors of A that adjoin the points $\tilde{z}_j$ when $p_j\geqq 1$ (cf. Fig. 4). □

We now turn our attention to a function f that possesses essential singularities on ∂V. It is an immediate consequence of $|f(z)|<1$, $z\in V$, that each such essential singularity is approached by exactly one A-region from within V in the order star of $\sigma=R/f$, R analytic in cl V.

THEOREM 9. *Let f have K different essential singularities along* ∂V. *Then, if R is a V-contractive function that satisfies* (9) *for* $\tilde{z}\in$ cl V, *it holds that*

$$p\leqq\frac{2(L+K+m)+1}{\alpha(\tilde{z})}\pi, \tag{13}$$

where m is the number of zeros of R in V.

Proof. We evaluate an upper bound on the sum of multiplicities of the A-regions that are, by Proposition 3, enclosed within V and contribute sectors that reach $\tilde{z}$. Zeros of f may lie, by Proposition 2, in analytic A-regions whose combined multiplicities may not exceed L. In addition, each essential singularity on ∂V is approached by one A-region inside V. If that A-region contributes $r\geqq 1$ sectors, say, that reach $\tilde{z}$ in V then it encloses analytic D-regions, sum of whose multiplicities is at least $r-1$. Since, by Proposition 2, the sum of all the multplicities of analytic D-regions that lie wholly in V may not exceed m, the number of zeros of R (and of σ) in V, at most $K+m$ sectors of A that reach $\tilde{z}$ within V may belong to nonanalytic A-regions. We now proceed like in the proof of Theorem 7 to derive the inequality (13) (cf. Fig. 5). □

THEOREM 10. *Given V-contractive approximation R to a function f that has K essential singularities on* ∂V, *it is true that*

$$P\leqq L+K+m,$$

where m is the number of zeros of R in V.

Proof. Similar to the proof of Theorem 8, but with L replaced by $L+K+m$. □

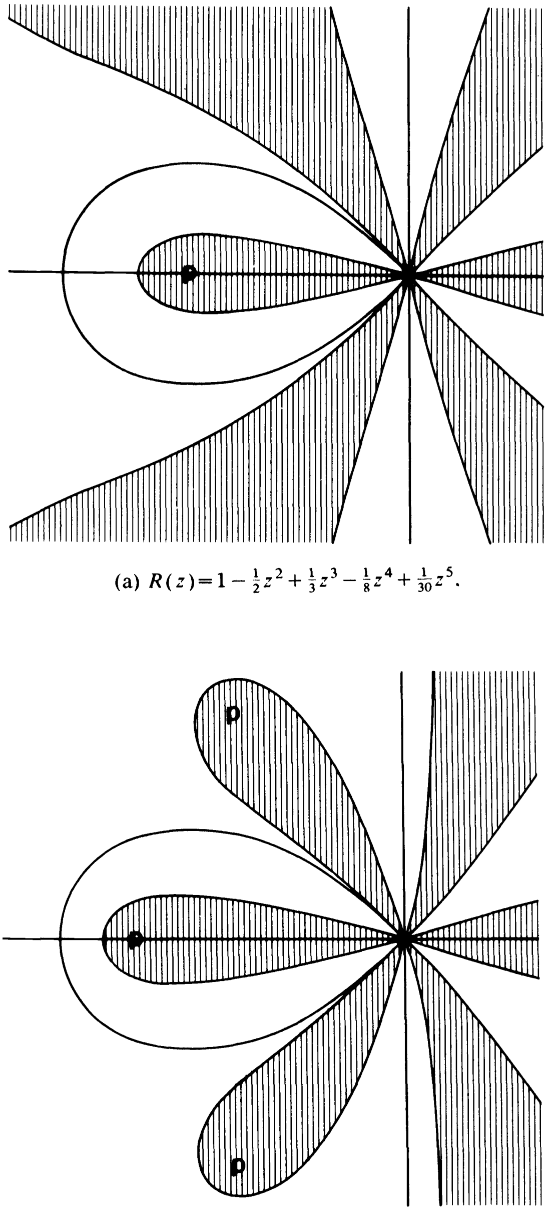
5. Examples of contractive approximations.

(i) $f(z)=(1+z)e^{-z}$, $\tilde{z}=0$. The set V is given in Fig. 3. The origin is a corner of ∂V and $\alpha(0)=\pi/2$ (this can be seen at once from Proposition 1, since $1+z=e^z-\frac{1}{2}z^2+O(z^3)$). Hence, since f is analytic along ∂V and has a single zero in V, (10) gives $p\leqq 6$.

Let

$$R(z)=1-\frac{1}{2}z^2+\frac{1}{3}z^3-\frac{1}{8}z^4+\frac{1}{30}z^5,$$

the [5/0] Padé approximation to f. It can be shown that it is V-contractive. Figure 3(a) gives the order star.



(a) $R(z)=1-\frac{1}{2}z^2+\frac{1}{3}z^3-\frac{1}{8}z^4+\frac{1}{30}z^5$.

(b) $R(z)=\dfrac{1}{1+\frac{1}{2}z^2-\frac{1}{3}z^3+\frac{3}{8}z^4-\frac{11}{30}z^5}$.

FIG. 3. *Approximations to* $f(z)=(1+z)e^{-z}$ *at the origin. The boundary of* V *is denoted by "&".*

Another example of a V-contractive function that attains $p=6$ is the [0/5] Padé approximation,

$$R(z)=\frac{1}{1+\frac{1}{2}z^2-\frac{1}{3}z^3+\frac{3}{8}z^4-\frac{11}{30}z^5}.$$

It is given in Fig. 3(b).

(j) Let V be a simply-connected domain with smooth boundary. Then it is possible to show that subject to a single exception, there exists an approximation R that attains the bound of Theorem 8. This is done with the help of the *Pick theorem* [10]. Let W denote the open complex unit disk. Given $z_0,\cdots,z_p\in W$ and complex numbers $\omega_0,\cdots,\omega_p$, there exists a unique q-tuple $(\kappa_1,\cdots,\kappa_q)$, $q\leqq p$, of numbers in W and a $\lambda\in\mathbb{C}$ such that the function

$$g(z)=\lambda B_q(z;\kappa_1,\cdots,\kappa_q)$$

satisfies

$$(14)\qquad g(z_j)=\omega_j,\qquad 0\leqq j\leqq p,$$

and has the minimal L_∞ norm in cl W amongst all the analytic functions that satisfy (14). B_q denots the Blaschke product,

$$B_q(z;\kappa_1,\cdots,\kappa_q):=\prod_{j=1}^{q}\frac{z-\kappa_j}{1-\bar{\kappa}_j z}.$$

The theorem also gives an explicit formula to evaluate $|\lambda|=\|g\|_{\infty,W}$, but this is not important to our argument.

Let f be a function, analytic in cl W and such that $|f(z)|\equiv 1$ along ∂W. We wish to interpolate it by a W-contraction at $z_0,\cdots,z_p$, say. We set $\omega_j=f(z_j)$, $0\leqq j\leqq p$. There are just two possibilities:

(i) f is the function g from the Pick theorem. In that case every other analytic function R that satisfies (14) must have an L_∞ norm exceeding 1 and no W-contractive approximation is possible.

(ii) f differs from g. we set $R\equiv g$. Since

$$\|g\|_{\infty,W}<\|f\|_{\infty,W}=1,$$

g is a W-contraction. Note that, by Theorem 8, $L\geqq p\geqq q$.

In other words, unless f is the minimal interpolant from the Pick theorem, the upper bound (12) is always attainable in W by a scaled Blaschke product. By the way, f itself is necessarily a Blaschke product, scaled by a constant of unit modulus.

Our analysis can be readily extended to Hermitian interpolation by a limiting process and to an arbitrary simply connected domain V with smooth boundary by conformally mapping V onto W.

As an example we give in Fig. 4 the order star for $V=W$,

$$f(z)=\left(\frac{z-\sqrt{2}/2}{1-(\sqrt{2}/2)z}\right)^2,$$

$$R(z)=-\frac{\sqrt{6}+\sqrt{2}}{4}\,\frac{z-(\sqrt{6}-\sqrt{2})/2}{1-((\sqrt{6}-\sqrt{2})/2)z},$$

$$R(z)=f(z)+O(z^2).$$

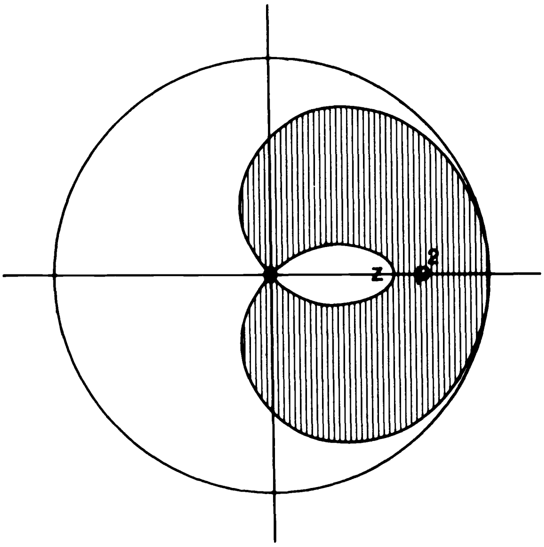


FIG. 4. *An approximation to* $f(z)=(z-\sqrt{2}/2)(1-(\sqrt{2}/2)z)^{-2}$ *at the origin.*

R is the Pick approximant, with the L_∞ norm

$$\|R\|_{\infty,V}=\frac{\sqrt{6}+\sqrt{2}}{5}=0.965925826\cdots.$$

(k) $f(z)=\exp(z/(z^2-1))$. The set V is the half-disk $\{z\in\mathbb{C}\colon |z|<1, \operatorname{Re} z>0\}$ and there is an essential singularity at $1\in\partial V$. f is being approximated at $\tilde{z}=0$. Therefore, by Theorem 9,

$$p\leqq 2m+3, \tag{15}$$

since $L=0$, $K=1$ and $\alpha(\tilde{z})=\pi$.

Figure 5 gives order stars for three different approximations;

$$R(z)=R_{1/1}(z)=\frac{1-z/2}{1+z/2};$$

$$R(z)=R_{0/2}(z)=\frac{1}{1+z+z^2/2};$$

$$R(z)=R_{2/0}(z)=1-z+z^2/2.$$

In all three cases $m=0$ and $p=3$, that is to say (15) holds as an equality. Note that only the first two approximations are V-contractions—indeed, (15) is a necessary condition but it is far from being sufficient for V-contractivity!

(l) $f(z)=\exp(2z/(z+1))$. V is the unit disk and f has an essential singularity at $-1\in\partial V$. Approximating at $\tilde{z}=0$, Theorem 9 gives the upper bound

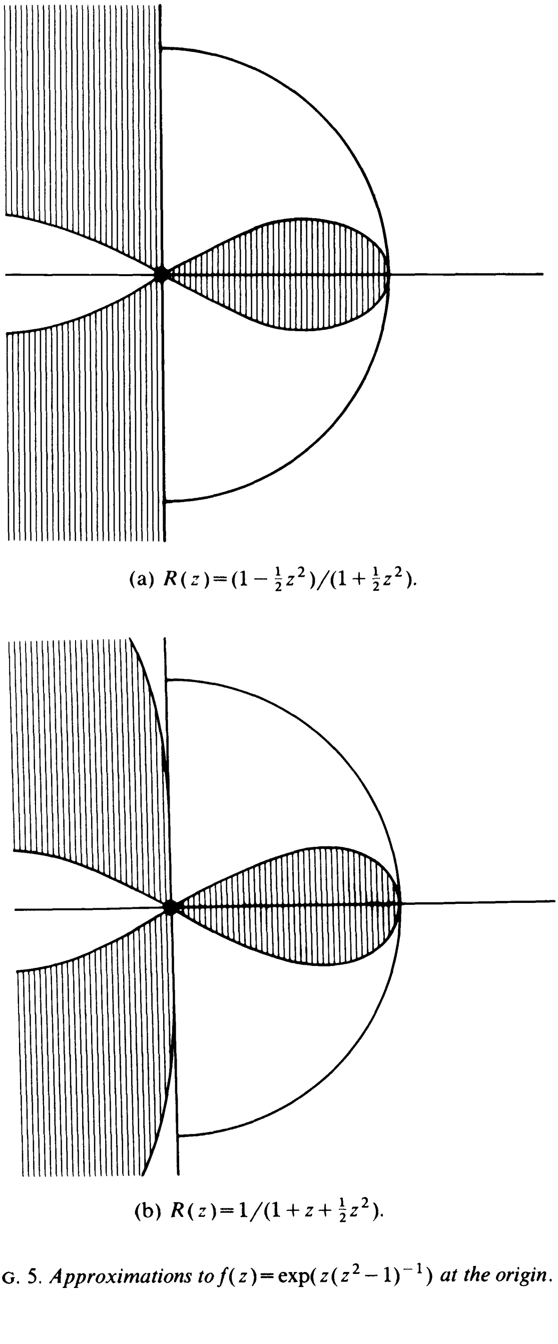
$$p\leqq m+1. \tag{16}$$

Consider the following approximations:

$$R(z)=R_{0/2}(z)=\frac{1}{1-2z+4z^2}, \qquad m=0, \quad p=3,$$

$$R(z)=R_{3/0}(z)=1+2z-\frac{2}{3}z^3, \qquad m=1, \quad p=4,$$

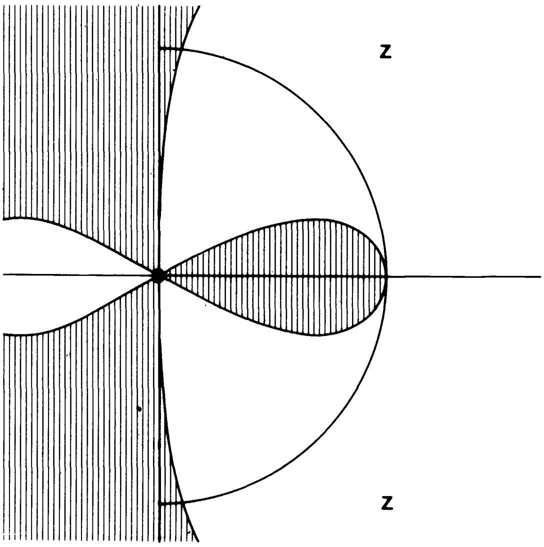
$$R(z)=\frac{1+5z/2}{(1+z/4)^2}, \qquad m=1, \quad p=2.$$



(a) $R(z)=(1-\frac{1}{2}z^2)/(1+\frac{1}{2}z^2)$.

(b) $R(z)=1/(1+z+\frac{1}{2}z^2)$.

FIG. 5. *Approximations to* $f(z)=\exp(z(z^2-1)^{-1})$ *at the origin.*



(c) $R(z)=1-z+\frac{1}{2}z^2$.

FIG. 5. (continued)

The first two approximations fail to satisfy (16) and so cannot be V-contractive. It is easy to verify that the infringement of V-contractivity occurs for different reasons—while the first function has poles in V and is not analytic there, the second exceeds 1 in modulus along ∂V. The third function is in agreement with (16) (as an equality) and is contractive.

(m) The first Ehle conjecture: it has been conjectured by Ehle (3) that the $[m/n]$ Padé approximation to e^z can be A-acceptable only if $m \leqq n \leqq m+2$. The proof of this conjecture by Wanner, Hairer and Nørsett [11] was the first success of the order star theory. Here we derive it from Theorem 9. We approximate $f(z)=e^z$ by $R_{m/n} \in \pi_{m/n}$ at $\tilde{z}=0$.

The set V coincides with $\mathbb{C}^- := \{z \in \mathbb{C}\colon \operatorname{Re} z \leqq 0\}$ and $p=m+n+1$ (the degree of interpolation always exceeds the order by 1, cf. (9)). There is a single essential singularity of f along V, at ∞. Furthermore, $L=0$ and $\alpha(\tilde{z})=\pi$. It now follows from (13) that A-acceptability (identical to V-contractivity in our framework) implies

$$n \leqq m+2.$$

As the inequality $m \leqq n$ is obvious, the proof of Ehle's conjecture follows as a straightforward corollary of our results.

REFERENCES

[1] L. V. AHLFORS, *Complex Analysis*, McGraw-Hill, New York, 1966.

[2] G. A. BAKER, *Essentials of Padé Approximants*, Academic Press, New York, 1975.

[3] B. L. EHLE, *A-stable methods and Padé approximants to the exponential*, SIAM J. Math. Analysis, 4 (1973), pp. 671–680.

[4] A. ERDELYI et al., *Higher Transcendental Functions*, Vol. III, McGraw-Hill, New York, 1955.

[5] W. B. GRAGG, *private communication*.

[6] E. Hille, *Analytic Function Theory*, Vol. II, Blaisdell, Waltham, MA, 1962.

[7] A. Iserles, *Order stars and a saturation theorem for first-order hyperbolics*, IMA J. Numer. Anal., 2 (1982), pp. 49–61.

[8] ______, *Order stars, approximations and finite differences* I. *The general theory of order stars*, this Journal, 16 (1985), pp. 559–576.

[9] A. Iserles and M. J. D. Powell, *On the A-acceptability of rational approximations that interpolate the exponential function*, IMA J. Numer. Anal., 1 (1981), pp. 241–251.

[10] G. Pick, *Über die bescharankungen analytischer funktionen, welche durch vorgegebene funcktionswerte bewirkt werden*, Math. Ann., 77 (1916), pp. 7–237.

[11] G. Wanner, E. Hairer and S. P. Nørsett, *Order stars and stability theorems*, BIT, 18 (1978), pp. 475–489.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

011

n-WIDTHS AND OPTIMAL INTERPOLATION OF TIME- AND BAND-LIMITED FUNCTIONS II*

AVRAHAM A. MELKMAN[†]

Abstract. Denote by $B(\sigma, T)$ the class of entire functions of exponential type σ which are bounded by 1 on the real axis outside $(-T, T)$. It is shown that this class, considered as a subset of $C[-T, T]$, has approximate dimension $2\sigma T/\pi$ in analogy to the Landau–Pollak–Slepian dimension theorem. More generally, the optimal subspaces and corresponding worst functions for the n-widths of $B(\sigma, T)$ are characterized. Prominently featured is the fact that it is possible to achieve the n-widths via interpolation, provided the sampling points are adroitly chosen. However, the interpolating functions differ from the standard ones.

1. Introduction. Denote by $B(\sigma, T)$ the class of entire functions of exponential type σ which are bounded by 1 on $(-\infty, -T)\cup(T, \infty)$. This paper mainly concerns the following problems.

(a) Given $\{t_i\}_1^n$ with $-T \leqq t_i \leqq T$, find an algorithm A^*: $C^n \to C[-T, T]$ which estimates $f \in B(\sigma, T)$ from the data $y = \{f(t_i)\}_1^n$ optimally in the sense of Micchelli and Rivlin [10], i.e. it achieves

$$E(t_1, \cdots, t_n) = \inf_{A} \max_{f \in B(\sigma, T)} \|f - Ay\|_T,$$

where $\|\cdot\|_T$ is the max norm on $[-T, T]$, and A is any map $C^n \to C[-T, T]$.

(b) Find the n-widths of $B(\sigma, T)$ with respect to $C[-T, T]$,

$$d_n(\sigma, T) = d_n(B(\sigma, T), C[-T, T])$$
$$= \min_{X_n \subset C[-T, T]} \max_{f \in B(\sigma, T)} \min_{y \in X_n} \|f - y\|_T,$$

where X_n is an n-dimensional subspace.

Clearly then $E(t_1, \cdots, t_n) \geqq d_n(\sigma, T)$; we will show that equality is achieved with the optimal choice of sampling points in (a). Of particular interest is $N(\sigma, T)$, the least n for which the n-width is 1 or less. This $N(\sigma, T)$ may be regarded as the "approximate dimension" of $B(\sigma, T)$, though its dimension is of course infinite. This point of view is best explained within the context of the original problem in which it arose. Regard $\varepsilon B(\sigma, T)$, $\varepsilon > 0$, as the set of those functions which are band-limited to (i.e. with Fourier transform supported on) $(-\sigma, \sigma)$ and simultaneously time-limited to $[-T, T]$ to within measurement accuracy ε; after all, outside $(-T, T)$ any $f \in \varepsilon B(\sigma, T)$ is pointwise indistinguishable from 0 within accuracy ε. Thus it is reasonable to define the approximate dimension of $\varepsilon B(\sigma, T)$ to be the dimension of the smallest subspace which contains for each $f \in \varepsilon B(\sigma, T)$ an element pointwise indistinguishable from f in $[-T, T]$ within accuracy ε.

The dimensionality problem is therefore the L_∞ version of the L_2 problem raised and dealt with by Landau and Pollak [6] and Slepian [12], of which we gave an account in a previous paper [9]. Analogously to the results found there we prove that the

* Received by the editors February 3, 1983, and in revised form February 27, 1984. This research was supported in part by the U.S. Army through its European office under contract DAJA 37-81-C-0234.

[†] Department of Mathematics, Ben-Gurion University, Beer-sheva, Israel.

approximate dimension of $B(\sigma,T)$ is $2\sigma T/\pi$ and that the approximation may proceed by interpolation.

Logan [7] analyzed the $n=0$ case, which is to find the function maximally concentrated in $[-T,T]$; his methods stimulated ours. Another source of inspiration has been the work of Boas and Schaeffer [3] (see also Ahiezer [1]). In fact, §2 is mostly devoted to showing how their arguments may be modified to solve the more general problem of characterizing the extremal f_0 for $\max(Lf{:}f\in B(\sigma,T))$, with L a linear functional. This characterization is of use for problem (a) when taking L of the form

$$Lf=f(t)-\sum_{i=1}^{n}\alpha_i f(t_i),$$

thereby obtaining the error in the pointwise estimation of $f(t)$ from the data $y=\{f(t_i)\}_1^n$ by means of the linear algorithm $Ay=\sum_{i=1}^{n}\alpha_i f(t_i)$. Moreover, Micchelli and Rivlin [10] have shown that for pointwise optimal estimation there always is such a linear algorithm which is optimal and

$$\min_{\{\alpha_i\}_1^n}\max_{f\in B(\sigma,T)}\left|f(t)-\sum_{i=1}\alpha_i f(t_i)\right|=\max_{\substack{f\in B(\sigma,T)\\ f(t_i)=0,\, i=1,\cdots,n}}|f(t)|.$$

Section 3 uses this observation to analyze the algorithm for pointwise optimal estimation and shows that it is also a globally optimal one.

In §4 it is shown that this linear optimal estimation based on the best set of interpolation points actually yields the n-widths, i.e.

$$d_n(\sigma,T)=\min_{\{t_i\}_1^n}\min_{A}\max_{f\in B(\sigma,T)}\|f-Ay\|_T,$$

a result similar to Micchelli, Rivlin and Winograd [11]. Finally in §5 the dimensionality of $B(\sigma,T)$ is calculated.

2. The maximum of a linear functional. This section summarizes the main results of Boas and Schaeffer [3] and the slight variations pertinent to the present setting.

Like them we are interested in the linear functional

$$Lf=\sum_{i=1}^{m}\sum_{j=0}^{n_i}a_i^{(j)}f^{(j)}(x_i) \tag{2.1}$$

with $a_i^{(j)}$ given real numbers, and x_i real points, and denote $l=\sum_{i=1}^{m}(n_i+1)$. The only differences are:

(i) we want to maximize Lf over the class $B_R(\sigma,T)$ of entire functions of exponential type σ which are real on the real axis and bounded by 1 on $(-\infty,-T)$ and (T,∞) instead of their $B_R(\sigma,0)$;

(ii) we require $x_i\in[-T,T]$, $i=1,\cdots,m$.

In the following theorem we gather all the information needed in later sections.

THEOREM 2.1. *The element f of $B_R(\sigma,T)$ for which*

$$Lf=\sup Lg \tag{2.2}$$

is unique and either constant or of type σ exactly. If not constant it satisfies a differential equation

$$\frac{\{f'(z)\}^2}{1-\{f(z)\}^2}=\frac{\sigma^2\{p(z)\}^2}{q(z)}$$

and is therefore of the form

$$f(z)=\sin\psi(z),\qquad \psi(z)=\sigma\int_0^z\{q(w)\}^{-1/2}p(w)\,dw+\sin^{-1}f(0).$$

Here $p(z)$ and $q(z)$ are monic polynomials with real coefficients.

Denote by $\cdots<\lambda_{-2}<\lambda_{-1}\leqq -T$, $T\leqq\lambda_1<\lambda_2<\cdots$ the points in $(-\infty,-T]$ and $[T,\infty)$ at which $|f(\lambda_i)|=1$. If $\lambda_{\pm1}=\pm T$ and $f'(\pm T)\neq 0$ set $s_\pm(z)=z\mp T$, otherwise $s_\pm(z)=1$. Let

$$s(z)=s_+(z)s_-(z),\qquad \nu=\text{degree } s. \tag{2.3}$$

(1) *The degree of p is at most $l-2+\nu$ and the zeros of $p(z)$ are precisely those zeros of f' different from the λ_i. Thus f nearly equioscillates outside $(-T,T)$ in the sense that $f(\lambda_i)f(\lambda_{i+1})=-1$ for all $i\leqq -2$, $i\geq 1$ with the exception of at most $l-2+\nu$ values.*

(2) *f vanishes simply between successive λ_i such that $f(\lambda_i)f(\lambda_{i+1})=-1$ and has at most $l-1$ zeros in addition.*

Proof. We merely sketch the main points of the proof, referring to Boas and Schaeffer for details, whenever possible.

a. [3, Lemma 2.2]. sup $|Lg|$ is finite, positive and attained. Note: the proof of this in [3] needs only be supplemented by the fact, Logan [5], that if f is bounded by 1 on $(-\infty,-T)$ and (T,∞) then it is bounded by $\cosh\sigma T$ on $(-T,T)$.

b. [3, Lemma 3.3]. Let f be an extremal for (2.2). Then $|f(x)|=1$ for at least one x in $(-\infty,T]\cup[T,\infty)$; and if $g\in B_R(\sigma,T)$ satisfies $g(\lambda_i)=0$ $i=\pm1,\pm2,\cdots$ then $Lg=0$.

c. [3, Lemma 3.7]. If $g\in B_R(\sigma,T)$ satisfies $g(\lambda_i)=0$, $i=\pm1,\pm2,\cdots$ and $g(x)=O(|x|^{-(l-1)})$ as $|x|\to\infty$ then $g\equiv 0$. Note. Our requirement $x_i\in[-T,T]$, $i=1,\cdots,m$ removes the obstacle noted in [3], to completing their proof in case $n_i=0$, $i=1,\cdots,m$.

d. [3, p. 863]. In particular consider $f'(z)s(z)$ which vanishes at all λ_i. If it has r_1 additional zeros then let $p(z)$ be the monic polynomial with precisely these zeros. The function $f'(z)s(z)/p(z)$ is in $B_R(\sigma,T)$ and behaves as $O(|x|^{-(r_1-m)})$. Thus $r_1\leqq l-2+m$.

e. [3, Lemma 3.2]. Similarly, if $(1-\{f(z)\}^2)s(z)$ has r_0 zeros in addition to the double ones at the λ_i then there is a function $g\in B_R(\sigma,T)$ vanishing at all λ_i and behaving like $O(|x|^{(r_0-m)/2})$. Thus $(r_0-m)/2<l-1$.

f. [3, p. 863]. Combining (d) and (e) shows that

$$\left\{\frac{f'(z)}{p(z)}\right\}^2\Big/\left\{\frac{1-f(z)^2}{q(z)}\right\}$$

is a zero-free entire function of exponential type bounded on the real axis and therefore constant.

g. In order to prove (2) let the additional zeros of f be comprised of α zeros in $[-T,T]$; 2β complex zeros which must come in pairs since $f(z)$ is real for real z; 2γ real zeros between successive λ_i, which must be even in number in order to preserve the sign change. Let $r(z)$ be the monic polynomial of degree $k=\alpha+2\beta+2\gamma$ with these zeros. Now if $k \geqq l$ let h be a monic polynomial of degree k with all its zeros in $(-T,T)$ and such that $h^{(j)}(t_i)=0\, j=0,\cdots,n_i$, $i=1,\cdots,m$. Set $d(z)=f(z)h(z)/r(z)$. Then sign $d(t)=\operatorname{sign} f(t)$ for $|t|\geqq T$ while $d(t)\to f(t)$ as $|t|\to\infty$. Thus $|f(t)-\varepsilon d(t)|<1$, $|t|\geqq T$, for small enough ε. However $L(f-\varepsilon d)=Lf$ contradicting the maximality of f.

3. Optimal estimation of a function from given data. Consider the optimal estimation of $f(t)$, $f\in B(\sigma,T)$, from its values $f(t_i)$, $i=1,\cdots,n$ with $-T\leqq t_1\leqq t_2\leqq\cdots\leqq t_n\leqq T$ where, in case of coincident points, appropriate derivative evaluations should be taken. Since attention can be restricted to linear recoveries, e.g., Micchelli and Rivlin [10], the search is for optimal coefficients a_i^* and the extremal function f_0 achieving

$$\left|f_0(t)-\sum_{i=1}^{n} a_i^* f_0(t_i)\right| = \min_{a_i\in C}\ \max_{g\in B(\sigma,T)}\left|g(t)-\sum_{i=1}^{n} a_i g(t_i)\right|. \tag{3.1}$$

As noted before, it is a result of Micchelli and Rivlin [10] that f_0 is at the same time the extremal for

$$\max\big(|g(t)| : g\in B(\sigma,T),\, g(t_i)=0, i=1,\cdots,n\big). \tag{3.2}$$

Since f_0 may be assumed real on the real axis it is sufficient to consider only $g\in B_R(\sigma,T)$ in (3.1) and (3.2), and hence also only real a_i. Thus the knowledge that $f_0(t_i)=0, i=1,\cdots,n$ can be combined with the properties f_0 is endowed with as the extremal of a problem of type (2.2).

PROPOSITION 3.1. *Let f_0 be an extremal of problem* (3.1). *Then*:

(1) *f_0 vanishes at the t_i while all its other zeros are real, simple and outside* $[-T,T]$.

(2) *f_0 equioscillates outside $(-T,T)$, i.e. denoting by*

$$\cdots<\lambda_{-2}<\lambda_{-1}\leqq -T,\quad T\leqq\lambda_1<\lambda_2<\cdots$$

the points at which $|f_0(\lambda_i)|=1$ then $f_0(\lambda_i)=(-1)^{\rho+i+1}$ with $\rho=0$ for $i>0$ and $\rho=n$ for $i<0$ (assuming $f_0(\lambda_1)=1$).

(3) *f_0' vanishes in $(-T,T)$ at precisely $n-1+\nu$ points μ_i, $i=1,\cdots,n-1+\nu$ separating the t_i, with $-T<\mu_1<t_1$ if $s_-(z)\not\equiv 1$, $t_n<\mu_{n-1+\nu}<T$ if $s_+(z)\not\equiv 1$.*

Proof. Invoking Theorem 2.1 (2) with $l=n+1$ it follows that, in addition to the real, simple zeros between successive λ_i, f_0 vanishes only at the t_i. Therefore by Laguerre's theorem, Boas [2, 2.8], f_0' too has only real simple zeros separating those of f_0. Additional information on f_0' can be gleaned from Theorem 2.1 (1), to the effect that f_0' has in $(-T,T)$ at most ν zeros in addition to the $n-1$ zeros between t_i. If, for example, $s_-(z)\not\equiv 1$ meaning $f_0(-T)=1$, $f_0'(-T)>0$ (since $|f_0(t)|\leqq 1$ for $t\leqq -T$) then f_0' must vanish between $-T$ and t_1. A similar phenomenon occurs at T causing f_0' to possess precisely $n-1+\nu$ zeros in $(-T,T)$.

With notation as in Proposition 2.1 set

$$h(z)=f_0'(z)s(z)\Big/\prod_{i=1}^{n-1+\nu}(z-\mu_i). \tag{3.3}$$

Thus $h(z)$ vanishes only at λ_i, $i=\pm1,\pm2,\cdots$

THEOREM 3.1. *The extremal f_0 of problem* (3.2) *is unique, up to a sign, and independent of t. The optimal estimate of $g(t)$, $|t|\leqq T$, is effected by interpolating the data $g(t_i)$, $i=1,\cdots,n$ with a function in*

$$\operatorname{span}\{h(t)t^i\}_0^{n-1} \tag{3.4}$$

and the estimate errs at most by $|f_0(t)|$, with equality only for f_0.

Proof. First we show that, with f_0 an extremal for (3.2), any function $g \in B(\sigma, T)$ has the representation

$$(3.5) \qquad g(z)=\sum_{i=1}^{n} g(t_i)\frac{h(z)}{h(t_i)}L_i(z)+\sum_{|k|=1}^{\infty} g(\lambda_k)\frac{h(z)\omega(z)}{h'(\lambda_k)\omega(\lambda_k)(t-\lambda_k)}$$

where $\omega(z)=\Pi_{i=1}^{n}(z-t_i)$ and $L_i(z)$ are the Lagrange polynomials, $L_i(t_j)=\delta_{ij}$.

Indeed, let C_k be the rectangular contour consisting of $|y|=k$, $x=\eta_k \pm k$ where η_k is the point at which the maximum of f_0' in $(\lambda_k, \lambda_{k+1})$ is achieved. Then

$$I_k=\int_{C_k}\frac{f_0(z)}{h(z)\omega(z)}\frac{1}{z-t}dz$$

satisfies $\lim_{k\to\infty} I_k=0$ because, by Theorem 1, $h(z)$ behaves for large z like $\sin \sigma z/z^{n-1}$. Thus Cauchy's theorem yields the desired representation.

Now for $k \geqq 1$ $\operatorname{sign} h'(\lambda_k)=\operatorname{sign} f_0''(\lambda_k)=-\operatorname{sign} f_0(\lambda_k)$ since λ_k is an extremum of f_0 (except for λ_1 if $f_0(T)=1$, but then $f_0'(T)<0$ leads to the same conclusion). Observe in addition that $\omega(\lambda_k)>0$ while $t-\lambda_k<0$ for $t \leqq T$. Combined with similar results for $k \leqq -1$ this yields

$$f_0(t)=h(t)\omega(t)\sum_{|k|\geqq 1}|h'(\lambda_k)\omega(\lambda_k)(t-\lambda_k)|^{-1}, \qquad t \leqq T.$$

Thus for any $g \in B_R(\sigma, T)$ so that $|g(\lambda_k)| \leqq 1$ and $t \in [-T, T]$

$$\left|g(t)-\sum_{i=1}^{n} g(t_i)\frac{h(t)}{h(t_i)}L_i(t)\right| \leqq |f_0(t)|$$

as claimed.

To prove uniqueness use the representation (3.5) for any $g \in B_R(\sigma, T)$ vanishing at the t_i. We have

$$|g(t)|=\left|h(t)\omega(t)\sum_{|k|=1}^{\infty}\frac{g(\lambda_k)}{h'(\lambda_k)\omega(\lambda_k)(t-\lambda_k)}\right| \leqq |f_0(t)|,$$

since $|g(\lambda_k)| \leqq 1$. Moreover equality occurs for $|t|<T$ if and only if $g(\lambda_k)=\delta f_0(\lambda_k)$ with $|\delta|=1$ in which case formula (3.5) yields $g(t)=\delta f_0(t)$.

Up to now we have dealt with the estimate of a function at a point. Another approach is to attempt estimation of the function as a whole on $[-T, T]$. One looks then at algorithms A: $C^n \to C[-T, T]$, which map the data $y=(g(t_1), \cdots, g(t_n))$ to a complex valued continuous function, and the optimal one achieves

$$E(t_1, \cdots, t_n)=\min_{A}\max_{g \in B(\sigma, T)}\|g-Ay\|_T$$

where $\|g\|_T=\max(|g(t)|: -T \leqq t < T)$.

However, the pointwise optimal estimate consisted of an interpolant which was in fact of exponential type σ and bounded on the real axis. Thus the following corollary is immediate.

COROLLARY. *The pointwise optimal estimate of $g(t)$ given in Theorem* 3.1 *is also the global optimal estimate of g on* $[-T, T]$ *and*

$$E(t_1, \cdots, t_n)=\|f_0\|_T.$$

4. Optimal sampling and *n*-widths. The previous section closed with a description of an optimal procedure for recovering a function from its values sampled at a given set of points. In this setting it is natural to ask for the optimal set of points at which to sample, i.e., those points which minimize $E(t_1,\cdots,t_n)$,

$$\begin{aligned} E\left(t_1^*,\cdots,t_n^*\right) &= \min_{t_i} E(t_1,\cdots,t_n) \\ &= \min_{t_i}\left(\max\|g\|_T: g\in B(\sigma,T),\, g(t_i)=0, i=1,\cdots,n\right). \end{aligned}$$

The following theorem answers this question and at the same time provides a characterization of the n-widths of $B(\sigma,T)$. Before stating the theorem, let us briefly describe the notion of n-widths; for a fuller description consult Lorentz [8] or Pinkus [12].

The procedure of interpolating a function $g\in B(\sigma,T)$, at a fixed set of n points by the n functions (3.4) is one particular kind of linear approximation. It is conceivable, and indeed sometimes the case that a different approximation process from some n-dimensional subspace X_n of functions will yields a smaller worst case error. Thus one is led to the notion of the n-width (in the sense of Kolmogorov) defined as

$$d_n(\sigma,T)=d_n(B(\sigma,T);\, C[-T,T])=\min_{X_n\subset C[-T,T]}\ \max_{g\in B(\sigma,T)}\ \min_{h\in X_n}\|g-h\|_T.$$

The n-width is therefore the minimum possible worst case error incurred in approximating the set $B(\sigma,T)$ with a set of n functions. Of particular interest, of course, is the set of functions that achieves the n-width.

THEOREM 4.1. *There exists a unique function $F_n\in B_R(\sigma,T)$ with the following properties*:

(1) *F_n equioscillates in $[-T,T]$ between the values $\pm\|F_n\|_T$ exactly $n+1$ times at the points $\rho_1<\cdots<\rho_{r+1}$, i.e. $F_n(\rho_i)=(-1)^{n+1-i}\|F_n\|$ $i=1,\cdots,n+1$.*

(2) *F_n equioscillates outside $(-T,T)$ between ± 1.*

(3) *If $\|F_n\|_T<1$ then $|F_n(\pm T)|=\|F_n\|_T$ and otherwise $|F_n(\pm T)|=1$.*

(4) *F_n has only the real simple zeros implied by* (1) *and* (2). *This function is the unique solution to the problem*

$$\min_{t_i}\ \max_{\substack{g\in B(\sigma,T)\\ g(t_i)=0}}\|g\|_T. \tag{4.1}$$

The zeros t_i of F_n in $[-T,T]$ are an optimal sampling point set and

$$d_n(B(\sigma,T);\, C[-T,T])=E\left(t_1^*,\cdots,t_n^*\right)=\|F\|_T.$$

Furthermore the n-width is achieved through the approximation process of interpolation at the points t_i^* by the set of functions $\{h(t)t^j\}_0^{n-1}$, where

$$h(t)=\begin{cases} F_n'(t)(t^2-T^2)/\prod\limits_{i=1}^{n+1}(t-\rho_i) & \text{if } \|F_n\|_T>1, \\ F_n'(T)/\prod\limits_{i=2}^{n}(t-\rho_i) & \text{if } \|F_n\|_T\leqq 1. \end{cases}$$

Proof. For ease of reading the proof is divided up into several lemmas. First we prove the existence of F_n as claimed and then its uniqueness. This already solves the optimal sampling problem. Since the concomittant interpolation procedure is a form of linear approximation it also follows that $d_n \leqq \|F_n\|_T$. The proof is then completed by showing $d_n \geqq \|F_n\|_T$.

LEMMA 4.1. *There exists a function $F \in B_R(\sigma, T)$ with the properties claimed for it in Theorem* 4.1 (1)–(4).

Proof. We employ, almost verbatim, the method of Karlin and Studden [5, Thm. 10.1]. Given any $\xi = (\xi_0, \xi_1, \cdots, \xi_n)$ within the simplex

$$\xi_i \geqq 0, \quad i = 0, \cdots, n, \qquad \sum_{i=0}^{n} \xi_i = 2T \tag{4.2}$$

construct interpolation points $t_i = -T + \sum_{k=0}^{i-1} \xi_k$, $i = 1, \cdots, n$.

By Theorem 3.1 there exists a unique function f_ξ such that

$$\|f_\xi\|_T = \max\left(\|g\|_T \colon g \in B(\sigma, T), g(t_i) = 0, i = 1, \cdots, n\right) \tag{4.3}$$

with the normalization $f(T+\varepsilon) > 0$ for small ε. From Theorem 2.1 f_ξ already has properties (2), (4). It remains to show that ξ can be chosen so that properties (1) and (3) hold.

With $t_0 = -T$, $t_{n+1} = T$ let

$$\delta_i(\xi) = \max_{t_i \leqq t \leqq t_{i+1}} |f_\xi(t)|, \qquad i = 0, \cdots, n$$

and, at the suggestion of A. Pinkus,

$$\delta(\xi) = \max_i \delta_i(\xi),$$
$$e_i(\xi) = \delta(\xi) - \delta_i(\xi), \quad i = 0, \cdots, n, \qquad e_{n+1}(\xi) = e_0(\xi).$$

Note that $\delta_i(\xi) = 0$ if and only if $\xi_i = 0$, i.e. $t_i = t_{i+1}$, and that $e_i(\xi) \geqq 0$ with equality for at least one i. Now, any f_ξ can equioscillate at most $n+1$ times in $[-T, T]$, by Proposition 3.1 (3). Thus the existence of F is equivalent to the existence of a ξ such that $\sum_{k=0}^{n} e_k(\xi) = 0$.

Suppose to the contrary that

$$e(\xi) = \sum_{k=0}^{n} e_k(\xi) > 0 \quad \text{for all } \xi.$$

Then the mapping $\xi \to \xi'$ given by

$$\xi_i' = \frac{2Te_i(\xi)}{e(\xi)}, \qquad i = 0, \cdots, n \tag{4.4}$$

is well defined on the whole simplex (4.2). Moreover the mapping is continuous because, roughly speaking, a sequence of extremals converges to an extremal. More precisely, suppose ξ^k converge to ξ, and let $\{t_i^k\}$, $\{t_i\}$ be the corresponding points. Denote the extremal of (4.3) for ξ^k by f_k. Then f_k converges to f because $B(\sigma, T)$ is a normal family. Our claim is that $f = f_\xi$. Because of the uniqueness of f_ξ it suffices to prove that if $g \in B(\sigma, T)$, and $g(t_i) = 0$ $i - 1, \cdots, n$, then $|g(t)| \leqq |f(t)|$. Indeed, given any $\delta > 0$ choose k large enough so that

$$|t_i - t_i^k| \leqq \delta \min_j |t - t_j|, \quad i = 1, \cdots, n, \qquad |f_k(t)| \leqq (1+\delta)|f(t)|.$$

Consider then

$$g_k(t)=g(t)\prod_{i=1}^{n}\left(\frac{t-t_i^k}{t-t_i}\right).$$

Since $g_k(t_i^k)=0$, $i=1,\cdots,k$. It follows that $|g_k(t)|\leqq|f_k(t)|$. But

$$|g_k(t)|\geqq|g(t)|\prod_{i=1}^{n}\left(1-\left|\frac{t_i-t_i^k}{t-t_i}\right|\right)\geqq|g(t)|(1-\delta)^n$$

implying $|g(t)|\leqq|f(t)|(1+\delta)(1-\delta)^{-n}$ for any δ and therefore $|g(t)|\leqq|f(t)|$.

Thus the mapping (4.4) has a fixed point ξ^*

$$\xi_i^*=\frac{2Te_i(\xi^*)}{e(\xi^*)}.$$

However $e_i(\xi^*)=0$ for some i, hence for that i, $\xi_i^*=0$ and therefore $\delta_i(\xi^*)=0$. But also $\delta(\xi^*)-\delta_i(\xi^*)=0$ and so $\delta(\xi^*)=0$, a contradiction.

LEMMA 4.2. *Let f_n be the extremal of problem* (3.2),

$$\max\bigl(|g(t)|: g\in B(\sigma,T),\ g(t_i)=0, i=1,\cdots,n\bigr).$$

If $f\in B(\sigma,T)$ oscillates $n+1$ times in $[-T,T]$, i.e. there exist extrema $-T\leqq e_1<e_2<\cdots<e_{n+1}\leqq T$ such that $(-1)^{n+1-i}f(e_i)>0$, then either $\min_i|f(e_i)|<\|f_n\|_T$ *or f is a constant multiple of f_n.*

Proof. Assume to the contrary that $\min_i|f(e_i)|\geqq\|f_n\|_T$. Using the representation formula (3.5), based on f_n,

$$|f(t)-h(t)p(t)|\leqq|f_n(t)|,\qquad -T\leqq t\leqq T,$$

where $h(t)p(t)$ is the interpolant to f based on the points t_i, from the set $\{h(t)t^k\}_0^{n-1}$. In particular

$$|f(e_i)-h(e_i)p(e_i)|\leqq\|f_n\|_T$$

and therefore

$$\operatorname{sign}h(e_i)p(e_i)=\operatorname{sign}f(e_i)=(-1)^{n+1-i}.$$

If now $h(e_i)>0$, $i=1,\cdots,n+1$, as has to be the case if $\min|f(e_i)|>\|f_n\|_T$, then $(-1)^{n+1-i}p(e_i)\geqq0$ which implies $p\equiv0$ because p is a polynomial of degree $n-1$. But $h(t)p(t)$ coincides with f at the t_i, i.e. $f(t_i)=0$, $i=1,\cdots,n$, and therefore by Theorem 3.1 $|f(t)|<|f_n(t)|$, $-T\leqq t\leqq T$, a contradiction.

Note that in any case $h(t)>0$ for $-T<t<T$. Thus a modification of the previous argument may be required only for $e_{n+1}=T$ with $f(T)=f_n(T)=\|f_n\|_T$ (or the analogous case $e_1=-T$). If $f_n(T)<1$ then from (3.3) $h(T)>0$ and no modification is called for. It remains therefore to investigate $f_n(T)=\|f_n\|_T=1$. In this case $h(t)=f_n'(t)/\prod_{i=1}^{n-1}(t-\mu_i)$ so that $h'(T)\neq0$. Differentiating (3.5) and evaluating at T yields

$$f'(T)=0=h'(T)p(T)+R_0+h'(T)R_1,$$

$$f_n'(T)=0=R_0+h'(T)S_1,$$

where the $k=1$ term has been split off as R_0. It is easily shown, by the same means employed in Theorem 3.1, that $|R_1|\leqq S_1$. Hence

$$h'(T)p(T)=h'(T)(S_1-R_1)$$

proves that again $p(T)\geqq 0$.

LEMMA 4.3. *The function F_n obtained in Lemma* 4.1 *is the unique solution to problem* (4.1).

Proof. Let $s_i, i=1,\cdots,n$ be the zeros of F_n in $(-T,T)$. Then the method of Theorem 3.1 shows that there is a representation formula of the form (3.5) based on F_n and s_i. In particular it follows that F_n is the extremal of problem (3.2), i.e. if $g\in B(\sigma,T)$ satisfies $g(s_i)=0$, $i=1,\cdots,n$ then $|g(t)|\leqq|F_n(t)|$, $-T\leqq t\leqq T$. Thus F_n is a candidate for the solution of (4.1).

On the other hand, if f_n is any other candidate, then the previous lemma shows

$$\min_i |F_n(\rho_i)|=\|F_n\|_T<\|f_n\|_T.$$

LEMMA 4.4. $d_n(\sigma,T)\geqq\|F_n\|_T$.

Proof. We rely on Lorentz [8, Lemma 9.1] to the effect that if for each choice of complex signs $\sigma_i, i=1,\cdots,n+1$ there is a function $g\in B(\sigma,T)$ such that $g(\rho_i)=\sigma_i\|F\|_T$ then $d_n\geqq\|F\|_T$.

Let $t_i, i=1,\cdots,n$ be the n zeros of F in $[-T,T]$ and denote $\omega(t)=\prod_{i=1}^n(t-t_i)$. Let $p(t)$ be the polynomial of degree n such that

$$p(\rho_i)=\sigma_i(-1)^{n+1-i}\omega(\rho_i), \qquad i=1,\cdots,n+1.$$

Write p in Lagrange form (thanks to T. Rivlin for pointing out this tack)

$$p(t)=\sum_{i=1}^{n+1}\sigma_i(-1)^{n+1-i}\omega(\rho_i)L_i(t).$$

Note that for $t>T$, $\operatorname{sign}(\omega(t)L_i(t))=\operatorname{sign}\omega(\rho_i)=(-1)^{n+1-i}$. Thus

$$\begin{aligned}|p(t)|&\leqq\sum_{i=1}^{n+1}|\omega(\rho_i)L_i(t)|\\&=(\operatorname{sign}\omega(t))\sum_{i=1}^{n+1}\omega(\rho_i)L_i(t)=|\omega(t)|, \qquad t\geqq T.\end{aligned}$$

Therefore the function

$$g(t)=\frac{F_n(t)p(t)}{\omega(t)}$$

satisfies

$$\begin{aligned}&g(\rho_i)=\sigma_i(-1)^{n+1-i}F_n(\rho_i)=\sigma_i\|F_n\|_T,\\&|g(t)|\leqq|F_n(t)|\leqq 1 \quad \text{for } |t|\geqq T,\end{aligned}$$

and hence $g\in B(\sigma,T)$.

5. The dimensionality of time- and band-limited functions. For some small ε consider the set $\varepsilon B(\sigma,T)$. Any function in this set is of exponential type σ or less, and may therefore be regarded as band-limited to $(-\sigma,\sigma)$. Moreover such a function is almost time-limited to $(-T,T)$ because outside this interval it is indistinguishable from 0 to within accuracy ε.

Define the dimension, $N(\sigma,T)$, of this set to be the least n such that $d_n(\varepsilon B(\sigma,T)) = \varepsilon d_n(\sigma,T) \leq \varepsilon$. $N(\sigma,T)$ is therefore the dimension of the smallest subspace of $C[-T,T]$ which contains for each $f \in B(\sigma,T)$ an element approximating f in the max norm on $[-T,T]$ to within ε. Such an approximant is indistinguishable from f at the same level of accuracy as f is known to be time-limited.

This notion of the dimension of the class of almost time- and band-limited functions was introduced and investigated in the $\mathscr{L}_2$ setting by Landau and Pollak [6] and Slepian [13]. A slightly improved version of their results is contained in [9]. The main result, that $N(\sigma,T) = 2\sigma T/\pi$, is replicated here. We will see however that the corresponding natural basis is somewhat different. But first a preparatory proposition.

Proposition 5.1. a. $d_{n+1}(\sigma,T) < d_n(\sigma,T)$.

b. *If* $T_1 < T_2$ *then* $d_n(\sigma,T_1) < d_n(\sigma,T_2)$.

Proof.

a. Let F_n, F_{n+1} be the extremals corresponding to d_n, d_{n+1}, e.g. $d_n = \|F_n\|_T$. Recall that F_n (F_{n+1}) equioscillates precisely $n+1$ $(n+2)$ times. By Lemma 4.2 $d_{n+1}(\sigma,T) = \|F_{n+1}\|_T < \|F_n\|_T = d_n(\sigma,T)$.

b. Let F_n, G_n be the extremals corresponding to $d_n(\sigma,T_1)$, $d_n(\sigma,T_2)$. Note that $T_1 < T_2$ implies $B(\sigma,T_1) \subseteq B(\sigma,T_2)$ and therefore $F_n \in B(\sigma,T_2)$. Applying Lemma 4.2 yields $d_n(\sigma,T_1) = \|F_n\|_T < \|G_n\|_T = d_n(\sigma,T_2)$ unless $F_n = G_n$. The latter however is impossible: if $\|G_n\|_T > 1$ then $G_n(t) > 1 \geq F_n(t)$ for $T_1 < t < T_2$; if $\|G_n\|_T \leqq 1$ then $\pm T_2$ must be two of its equioscillation points while F_n has all of its equioscillation points in $[-T_1,T_1]$.

Theorem 5.1. *Let* $N(\sigma,T)$ *be the least* n *such that* $d_n(\sigma,T) \leqq 1$. *Then* $N(\sigma,T) = \lceil 2\sigma T/\pi \rceil =$ *the least integer not less than* $2\sigma T/\pi$. *In case* $2\sigma T/\pi$ *is an integer an optimal* N*-dimensional approximating subspace is spanned by the functions*

$$(5.1)\qquad \sin\sigma t \text{ and } \frac{\sin(\sigma t - k\pi)}{\sigma t - k\pi}, \quad k = 0, \pm 1, \cdots, \pm(m-1) \qquad \text{if } \frac{2\sigma T}{\pi} = 2m,$$

$$(5.2)\qquad \cos\sigma t \text{ and } \frac{\cos(\sigma t - r\pi)}{\sigma t - r\pi}, \quad r = \pm\frac{1}{2}, \frac{3}{2}, \cdots, \pm\left(m - \frac{3}{2}\right) \qquad \text{if } \frac{2\sigma T}{\pi} = 2m - 1.$$

Proof. Observe that if $2\sigma T/\pi = 2m$ then $F_{2m}(t) = \cos\sigma t$ has properties (1)–(4) of Theorem 4.1 with $\|F_{2m}\|_T = 1$ and therefore $d_{2m}(\sigma, m\pi/\sigma) = 1$. From Proposition 5.1 $d_{2m-1}(\sigma, m\pi/\sigma) > 1$ and hence $N(\sigma, m\pi/\sigma) = 2m$. Theorem 4.1 also yields the result that an optimal $2m$-dimensional subspace is spanned by

$$(t^j \sin\sigma t) \Big/ \prod_{i=-(k-1)}^{k-1} (\sigma t - i\pi) \qquad j = 0, \cdots, 2k-1$$

which is equivalent to (5.1). Note that the approximation may proceed by interpolation at the points $r\pi/\sigma$, $r = \pm\frac{1}{2}, \cdots, \pm(m - \frac{1}{2})$. The case $2\sigma T/\pi = 2m-1$ is proved analogously. If now $2\sigma T/\pi \leqq 2m+1$ then from Proposition 5.1

$$1 = d_{2m}(\sigma, m\pi/\sigma) < d_{2m}(\sigma,T),$$
$$d_{2m+1}(\sigma,T) < d_{2m+1}\left(\sigma, \left(m + \frac{1}{2}\right)\pi/\sigma\right) = 1,$$

proving $N(\sigma,T) = 2m+1 = \lceil 2\sigma T/\pi \rceil$. The proof of the case $2m-1 < 2\sigma T/\pi \leqq 2m$ proceeds similarly.

Finally, let us mention those explicit n-widths that are easily derived:

1. $d_0(\sigma, T) = \cosh \sigma T$, extremal: $F_0(t) = \cos \sigma\sqrt{t^2 - T^2}$;

2. for $0 < \sigma T \leqq \pi/2$, $d_1(\sigma, T) = \sin \sigma T$, extremal: $F_1(t) = \sin \sigma t$;

3. for $0 < \sigma T \leqq \pi$ $d_2(\sigma, T) = \sin((\sigma T)^2/2\pi)$, extremal: $F_2(t) = \cos \sigma\sqrt{t^2 + \alpha^2}$ with $\sigma\alpha = \pi/2 - (\sigma T)^2/2\pi$; an optimal subspace is spanned by

$$\frac{\sin \sigma\sqrt{t^2 + \alpha^2}}{\sigma\sqrt{t^2 + \alpha^2}} t^j, \qquad j = 0, 1$$

with interpolation at the points $\pm(T/2)\sqrt{2 - (\sigma T/\pi)^2}$.

Furthermore, as noted by Jagerman [4], an easy asymptotic estimate is obtained via polynomial interpolation, namely

$$d_n(\sigma, T) \leqq \text{const} \frac{1}{n!} \left(\frac{\sigma T}{2} \right)^n.$$

A better estimate for small $n > 2\sigma T/\pi$ is $c_1 \exp[c_2(2\sigma T/\pi - n)]$, c_1 and c_2 constants. This estimate can be deduced from H. J. Landau's results as described in his recent manuscript *Extrapolating a band-limited function from its samples taken in a finite interval.*

Acknowledgments. This work would not have come to fruition without the stimulating conversations with Professors R. Boas and S. Fisher and the congenial hospitality of Northwestern University.

REFERENCES

[1] N. I. AHIEZER, *Extremal properties of entire functions of exponential type*, Amer. Math. Soc. Transl., 86 (1970), pp. 1–30.

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

[3] R. P. BOAS AND A. C. SCHAEFFER, *Variational methods in entire functions*, Amer. J. Math., 79 (1957), pp. 857–884.

[4] D. JAGERMAN, *Information theory and approximation of band-limited functions*, Bell System Tech. J., 49 (1970), pp. 1911–1941.

[5] S. KARLIN AND W. J. STUDDEN, *Tchebycheff Systems: With Applications in Analysis and Statistics*, John Wiley, New York, 1966.

[6] H. J. LANDAU AND H. O. POLLAK, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*, III, Bell System Tech. J., 41 (1962), pp. 1295–1336.

[7] B. LOGAN, JR., *Properties of high-pass signals*, Ph.D. dissertation, Columbia Univ., New York, 1965.

[8] G. G. LORENTZ, *Approximation of Functions*, Holt, Rinehart and Winston, New York, 1966.

[9] A. A. MELKMAN, *n-widths and optimal interpolation of time- and band-limited functions*, in Optimal Estimation in Approximation Theory, C. A. Micchelli and T. J. Rivlin, eds., Plenum Press, New York, 1977, pp. 55–68.

[10] C. A. MICCHELLI AND T. J. RIVLIN, *A survey of optimal recovery* in Optimal Estimation in Approximation Theory, C. A. Micchelli and T. J. Rivlin, eds., Plenum Press, New York, 1977, pp. 1–53.

[11] C. A. MICCHELLI, T. J. RIVLIN AND S. WINOGRAD, *The optimal recovery of smooth functions*, Numer. Math., 26 (1976), pp. 191–200.

[12] A. PINKUS, *n-Widths in Approximation Theory*, to be published by Springer-Verlag, New York.

[13] D. SLEPIAN, *On band width*, Proc. IEEE, 64 (1976), pp. 292–300.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

012

RATIONAL APPROXIMATIONS TO THE EXPONENTIAL FUNCTION WITH TWO COMPLEX CONJUGATE INTERPOLATION POINTS*

ERNST HAIRER[†], ARIEH ISERLES[‡] AND SYVERT P. NØRSETT[§]

Abstract. We examine rational approximations to the exponential function which satisfy $R(z)=\exp(z)+O(z^{2m-1})$ and $R(z_0)=\exp(z_0)$, $R(\bar{z}_0)=\exp(\bar{z}_0)$, where m denotes the maximal degree of numerator and denominator. It is proved that a curve bisects the complex plane so that if z_0 lies to its left the approximation is A-acceptable, while if it is to the right of this curve then $R(z)$ is non-A-acceptable.

1. Introduction. The investigation of rational approximations to the exponential function attracted much attention in recent years. Such approximations are of central importance in the design and analysis of numerical methods for stiff ordinary differential equations.

Let R be a function in $\pi_{m,n}:=\{p/q : p,q$ real polynomials, $\deg p\le m$, $\deg q\le n$, $q\not\equiv 0\}$. According to the maximal interpolation theorem (Iserles [1979]) the number of real zeros of the equation

$$R(x)=e^x,$$

counted with their multiplicity, may not exceed $n+m+1$. Using the technique of order stars, developed by Wanner, Hairer and Nørsett [1978], all A-acceptable approximations which attain the bound of the maximal interpolation theorem are characterized by Iserles and Powell [1981].

The analysis of A-acceptability of rational approximations with pre-assigned complex conjugate interpolation points is considerably more complicated. It follows at once from the order star theory that the number of complex zeros of the equation

$$R(z)=e^z$$

is infinite for every choice of $R\in\pi_{m,n}$ and that $\pm i\infty$ are their accumulation points.

The present paper is a successor of a report by Iserles and Nørsett [1982b] and contains similar results. The proofs of this article, however, are based on the characterization of A-acceptable approximations as given by Hairer [1982] and are more elegant. We examine approximations from $\pi_{m/m}$ which possess order $2m-2$ (i.e. have an interpolation point of multiplicity $2m-1$ at the origin) and, in addition, interpolate the exponential at a specified pair of complex conjugate points. There are two sound reasons to restrict our attention to this particular case. First, it is likely in numerical applications to use two degrees of freedom to exponentially fit a damped oscillation, reserving the remaining degrees of freedom to attain the highest possible order. Second, this type of approximation nicely fits into the framework of the Iserles–Powell result. In other words, if the conjugate pair approaches the real axis within the complex left half plane the resulting approximation is A-acceptable.

In a previous work Iserles and Nørsett [1982a] studied the case of conjugate pair being pure imaginary. In this case the approximation is A-acceptable if and only if the

* Received by the editors October 28, 1983.

[†] Universität Heidelberg, Institut für Angewandte Mathematik, Im Neuenheimer Feld 294, 6900 Heidelberg 1, Germany.

[‡] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, England CB3 9EW.

[§] Norges Teknikske Høgskole, Institutt for Numerisk Matematikk, 7034 Trondheim, Norway.

interpolation points belong to certain intervals whose end-points are related to zeros of spherical Bessel functions.

Our present model was considered by Liniger and Willoughby [1970] for the case of m equaling 2. They experimentally show that a certain curve bisects the complex plane near the imaginary axis. Conjugate pairs which lie to the left of this curve give raise to A-acceptable approximations, while pairs to the right cause non-A-acceptability. In the present paper we prove that this is typical to all approximations of this form, regardless of the value of $m \geq 2$.

2. Exponential fitting. Let $N_m(z)/N_m(-z)$ be the diagonal Padé approximation to the exponential function. Then

$$N_m(z)=\sum_{k=0}^{m}\frac{(2m-k)!}{(2m)!}\binom{m}{k}z^k. \tag{2.1}$$

It follows from Hairer [1982, Thm. 1] that any approximation from $\pi_{m/m}$ which has order at least $2m-2$ must necessarily be of the form

$$R(z)=\frac{N_{m-1}(z)(1-f_1z)+z^2\xi_{m-1}^2N_{m-2}(z)g_0}{N_{m-1}(-z)(1-f_1z)+z^2\xi_{m-1}^2N_{m-2}(-z)g_0},\qquad \xi_j^2:=\frac{1}{4(4j^2-1)}, \tag{2.2}$$

where f_1 and g_0 are arbitrary real constants.

Our goal is to examine the satisfaction of $R(z_0)=e^{z_0}$ (and hence also $R(\bar{z}_0)=e^{\bar{z}_0}$) by rational functions (2.2), subject to a choice of real numbers g_0 and f_1. For this purpose we introduce (compare Iserles and Nørsett [1982a]) the function

$$\psi_m(z)=N_m(z)-e^zN_m(-z). \tag{2.3}$$

The following relations will be used frequently:

$$\psi_m(z)-\psi_{m-1}(z)=z^2\xi_{m-1}^2\psi_{m-2}(z), \tag{2.4}$$

$$\psi_m'(z)=\frac{\psi_m(z)}{2}-\frac{z}{4(2m-1)}\psi_{m-1}(z). \tag{2.5}$$

They are an immediate consequence of the fact that the same formulas are valid for the polynomials $N_m(z)$ and are easily verified. Furthermore, from Iserles and Nørsett [1982a] we have

$$\psi_m(iy)=-i\sqrt{2\pi}\,\frac{m!}{(2m)!}e^{iy/2}\left(\frac{y}{2}\right)^{m+1/2}J_{m+1/2}\left(\frac{y}{2}\right) \tag{2.6}$$

where $J_{m+1/2}$ denotes the spherical Bessel function of the first kind. This implies

LEMMA 1. *All roots of $\psi_m(z)$ lie on the imaginary axis. Besides $z=0$ they are given by iy where y is a root of $J_{m+1/2}(y/2)$.*

Proof. The roots of $\psi_m(z)$ are just the interpolation points of the diagonal Padé approximation $N_m(z)/N_m(-z)$. Its order star (see Wanner, Hairer and Nørsett [1978]) immediately implies that all roots of $\psi_m(z)$ must lie on the imaginary axis. The rest follows from (2.6). □

We are now able to prove the next theorem.

THEOREM 2. *Suppose that* $\operatorname{Im} z_0 \neq 0$, $\psi_{m-1}(z_0)\neq 0$ *and let*

$$r_m(z)=|z|^2\xi_{m-1}^2\operatorname{Im}\left\{z\frac{\psi_{m-2}(z)}{\psi_{m-1}(z)}\right\}. \tag{2.7}$$

For $r_m(z_0)\neq 0$ the approximation (2.2) *satisfies $R(z_0)=e^{z_0}$ if and only if*

$$\begin{aligned} g_0&=\operatorname{Im} z_0/r_m(z_0),\\ f_1&=\operatorname{Im}(\psi_m(z_0)/\psi_{m-1}(z_0))/r_m(z_0). \end{aligned} \tag{2.8}$$

Proof. The condition $R(z_0)=e^{z_0}$ is equivalent to

$$\psi_{m-1}(z_0)(1-f_1z_0)+z_0^2\xi_{m-1}^2\psi_{m-2}(z_0)g_0=0. \tag{2.9}$$

Dividing by $\psi_{m-1}(z_0)$ and multiplying by $\bar z_0$ we get

$$\bar z_0-f_1|z_0|^2+|z_0|^2\xi_{m-1}^2z_0\frac{\psi_{m-2}(z_0)}{\psi_{m-1}(z_0)}\cdot g_0=0.$$

Since g_0 is real, the imaginary part of this equation yields the formula for g_0. Using (2.4), (2.9) can be written as

$$1-f_1z_0+\left(\frac{\psi_m(z_0)}{\psi_{m-1}(z_0)}-1\right)g_0=0.$$

Again the imaginary part yields the formula for f_1. □

3. *A*-acceptability. The aim of this section is to investigate for which z_0 the rational approximation (2.2) with g_0,f_1 given by (2.8) is A-acceptable. We shall use the following result of Hairer [1982].

THEOREM 3. *Let $R(z)$ be given by* (2.2) *with $g_0\neq 0$. Then*

a) *$R(z)$ is I-acceptable* (*i.e.* $|R(iy)|\le 1$ *for every* $y\in\mathbb{R}$) *iff* $g_0\cdot f_1\ge 0$.

b) *$R(z)$ is A-acceptable* (*i.e.* $|R(z)|\le 1$ *for* $\operatorname{Re} z\le 0$) *iff* $g_0>0$ *and* $f_1\ge 0$.

In order to apply this theorem to the parameters (2.8) the following lemma is useful.

LEMMA 4. *Excluding the points with $\psi_{m-1}(z)=0$ we have*

$$\operatorname{Im}\left(\frac{\psi_m(z)}{\psi_{m-1}(z)}\right)\begin{cases}=0 & \textit{if } \operatorname{Re}z\operatorname{Im}z=0,\\ >0 & \textit{if } \operatorname{Re}z\operatorname{Im}z<0,\\ <0 & \textit{if } \operatorname{Re}z\operatorname{Im}z>0.\end{cases}$$

Proof. The statement is trivial for $\operatorname{Im}z=0$. For $\operatorname{Re}z=0$ (i.e. $z=iy$) the statement follows from

$$\psi_m(iy)\psi_{m-1}(-iy)=2\operatorname{Re}\{N_m(iy)N_{m-1}(-iy)-e^{iy}N_m(-iy)N_{m-1}(-iy)\},$$

which is an immediate consequence of the definition of $\psi_m(z)$. For $z=re^{it}$ ($r\to\infty$) a simple analysis shows that

$$\psi_m(z)/\psi_{m-1}(z)\sim\begin{cases}\dfrac{1}{2(2m-1)}\cdot re^{it} & \text{if } \cos t<0,\\[2ex] -\dfrac{1}{2(2m-1)}\cdot re^{it} & \text{if } \cos t>0.\end{cases}$$

Finally we have to investigate the behaviour of $\operatorname{Im}(\psi_m(z)\psi_{m-1}(z))$ near the singularities which we denote by iy_k, $k=1,2,\cdots$ (cf. Lemma 1; the origin is a nonsingular point since $\psi_m(z)/\psi_{m-1}(z)=\text{Const}\cdot z^2$ for $z\to 0$). With the help of (2.4) and (2.5) we get for $z=iy_k+\varepsilon e^{it}$, $\varepsilon\to 0$

$$\operatorname{Im}\left(\frac{\psi_m(z)}{\psi_{m-1}(z)}\right)\sim -\frac{y_k}{\varepsilon}\cdot\frac{\cos t}{(2m-1)}.$$

Now we are ready to apply the maximum principle for harmonic functions in the four domains

$$\left\{z\in\mathbf{C}\,|\operatorname{Re}z\gtrless 0, \operatorname{Im}z\gtrless 0|, |z|<R, |z-iy_k|>\varepsilon, k=1,2,\cdots\right\}. \qquad \square$$

THEOREM 5. *Let* $\operatorname{Im}z_0\neq 0$, $\psi_{m-1}(z_0)\neq 0$, $r_m(z_0)\neq 0$ *and* g_0, f_1 *be given by* (2.8). *Then* $R(z)$ *is I-acceptable* $\Leftrightarrow$ $\operatorname{Re}z_0\leq 0$.
$R(z)$ *is A-acceptable* $\Leftrightarrow$ $\operatorname{Re}z_0\leq 0$ *and* $\operatorname{Im}z_0\cdot r_m(z_0)>0$.
Proof. This is an immediate consequence of Theorem 3 and Lemma 4. □

4. Domain of A-acceptability. In order to get the right feel for the domain of A-acceptability we have to study the condition $\operatorname{Im}z_0\cdot r_m(z_0)>0$ (compare Theorem 5) in more detail. The following lemma presents some properties of the function $r_m(z)$.

LEMMA 6. *It holds that*

(4.1) $\quad r_m(x+iy_0)=r_m(-x+iy_0)$,

(4.2) $\quad \lim\limits_{|x|\to\infty} y_0\cdot r_m(x+iy_0)>0$, *for* $y_0\neq 0$,

(4.3) $\quad r_m(iy)=y^3\xi^2_{m-1}(\psi_{m-2}(iy)/\psi_{m-1}(iy))$,

(4.4) $\quad$ *for fixed* $y_0\neq 0$, $r(x+iy_0)$ *has at most one root in* $(-\infty,0)$.

Proof. The symmetry relation (4.1) follows from

$$\psi_m(-x+iy)=-e^{-x+iy}\cdot\overline{\psi_m(x+iy)}$$

which can easily be verified. Similar to the proof of Lemma 4 we obtain for $z=x+iy_0$, $|x|\to\infty$

$$z\frac{\psi_{m-2}(z)}{\psi_{m-1}(z)}=\operatorname{sign}(x)\cdot 2(2m-3)-4(m-1)(2m-3)\frac{1}{z}+\operatorname{sign}(x)\frac{c}{z^2}+O\left(\frac{1}{z^3}\right)$$

where c is some real constant. Taking the imaginary part of this formula leads to

$$y_0r_m(x+iy_0)=\frac{(m-1)}{(2m-1)}y_0^2+O\left(\frac{1}{x}\right).$$

Statement (4.3) is trivial, since it follows from Lemma 4 that $\psi_{m-2}(iy)/\psi_{m-1}(iy)$ is real. For (4.4) it is sufficient to prove that

(4.5) $\quad r_m(x^*+iy_0)=0$ and $x^*<0$ imply$y_0\dfrac{d}{dx}r_m(x^*+iy_0)<0$.

With $g(z)$ given by

$$g(z)=\operatorname{Im}(z\psi_{m-2}(z)/\psi_{m-1}(z))$$

it holds that

$$g(x^*+iy_0)=0, \qquad \frac{d}{dx}r_m(x^*+iy_0)=\left(x^{*2}+y_0^2\right)\xi^2_{m-1}g'(x^*+iy_0).$$

A simple analysis using (2.4) and (2.5) yields

$$g'(z)=(2m-2)\mathrm{Im}\left\{\frac{\psi_{m-2}(z)}{\psi_{m-1}(z)}\right\}+\frac{1}{4(2m-3)}\mathrm{Im}\left\{\left(z\,\frac{\psi_{m-2}(z)}{\psi_{m-1}(z)}\right)^{2}\right\}.$$

The second term of this expression vanishes at $z=x^*+iy_0$ since $g(x^*+iy_0)=0$ implies that $(x^*+iy_0)\cdot\psi_{m-2}(x^*+iy_0)/\psi_{m-1}(x^*+iy_0)$ is real. The right sign of the first term at $z=x^*+iy_0$ follows from Lemma 4. □

Denote the nonzero roots of $\psi_{m-1}(y)$ by $\pm iy_1^{(m-1)}$, $\pm iy_2^{(m-1)},\cdots$; then

$$0<y_1^{(m-2)}<y_1^{(m-1)}<y_2^{(m-2)}<y_2^{(m-1)}<y_3^{(m-2)}<\cdots.$$

This follows from Lemma 1 since the zeros of spherical Bessel functions interlace (Abramowitz and Stegun [1965, p. 370]).

Observe further that for $z\to 0$

$$\psi_m(z)=(-1)^{m+1}c_m z^{2m+1}+O(z^{2m+2})\quad\text{with } c_m>0$$

so that

$$r_m(iy)=D_m y+O(y^2)\quad\text{for } y\to 0,\quad\text{with } D_m>0.$$

Hence, by (4.3), $r_m(iy_0)$ is negative if and only if

$$y_0\in\left\{\,y\,|\,y_k^{(m-2)}<y<y_k^{(m-1)}\text{ for some } k\geq 1\right\}.$$

Only for such values of y_0 $r(x+iy_0)$ has a negative and a positive root. The curves consisting of the roots of $r_m(z)$ are plotted in Fig. 1 (full lines). The shaded region corresponds to the z_0 with A-acceptable $R(z)$. (See also Fig. 2.)

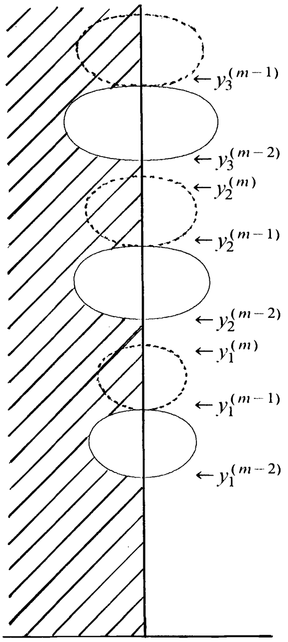


FIG 1.

Finally let us study the limiting cases $\psi_{m-1}(z_0)=0$ and $r_m(z_0)=0$ which are not covered by Theorem 5 (assume $\operatorname{Im} z_0 \neq 0$).

A) $z_0 = \pm i y_k^{(m-1)}$ (i.e. $\psi_{m-1}(z_0)=0$): In this case we have $g_0=0$. The rational function (2.2) is reducible and equivalent to $N_{m-1}(z)/N_{m-1}(-z)$. This is known to be A-stable and of order $2m-2$.

B) $z = \pm i y_k^{(m-2)}$ (i.e. $\psi_{m-2}(z_0)=0$): Here we have $|g_0| \to \infty$ and $f_1/g_0 \to 0$, so that (2.2) becomes $N_{m-2}(z)/N_{m-2}(-z)$. Again this is A-stable, but only of order $2m-4$.

C) $r_m(z_0)=0$, $\operatorname{Re} z_0 \neq 0$: The parameters g_0, f_1 satisfy

$$|g_0| \to \infty, \qquad f_1/g_0 \to \operatorname{Im}\left(\frac{\psi_m(z_0)}{\psi_{m-1}(z_0)}\right)/\operatorname{Im} z_0 \neq 0.$$

With the help of the analogue of (2.4) for $N_m(z)$, $R(z)$ can be written in this case as

$$R(z) = \frac{N_{m-2}(z)\left(1-(g_0/f_1)\xi_{m-1}^2 z\right) + z^2 \xi_{m-2}^2 N_{m-3}(z)}{N_{m-2}(-z)\left(1-(g_0/f_1)\xi_{m-1}^2 z\right) + z^2 \xi_{m-2}^2 N_{m-3}(-z)}.$$

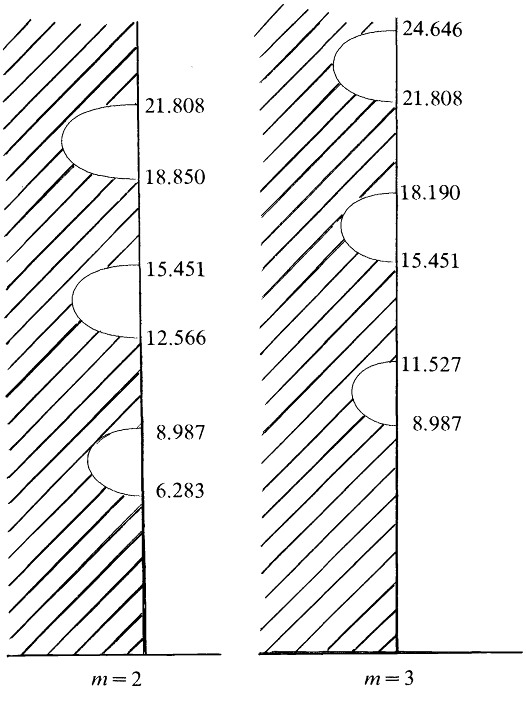
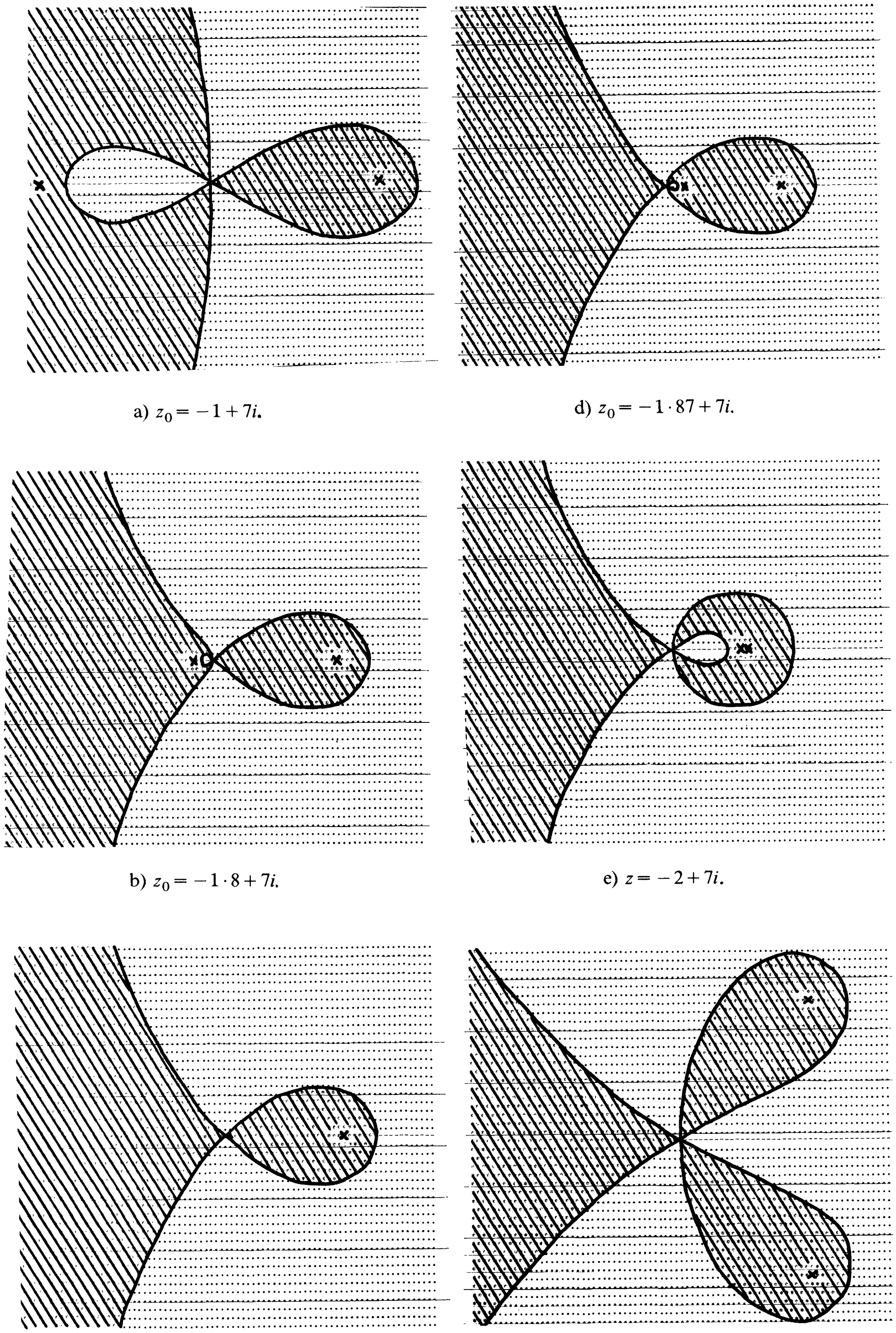


FIG 2.

The results of Hairer [1982] show that this approximation is of exact order $2m-3$ and in addition A-acceptable for $\operatorname{Re} z_0 < 0$. For the sake of completeness let us mention for which values of z_0 the order of $R(z)$ is larger than $2m-2$. By Hairer [1982, Cor. 3] this is the case if $g_0 = 1$ or equivalently if $\operatorname{Im} z_0 = r_m(z_0)$. Using (2.4), $r_m(z_0)$ can be rewritten as

$$r_m(z_0) = |z_0|^2 \cdot \operatorname{Im}\left\{\frac{\psi_m(z_0)}{z_0 \psi_{m-1}(z_0)}\right\} - \operatorname{Im} z_0$$



a) $z_0 = -1 + 7i$.

b) $z_0 = -1 \cdot 8 + 7i$.

c) $z_0 = -1 \cdot 834651861 + 7i$.

d) $z_0 = -1 \cdot 87 + 7i$.

e) $z = -2 + 7i$.

f) $z = -4 + 7i$.

FIG 3.

so that the condition $g_0 = 1$ is equivalent to

$$\text{(4.6)} \qquad \operatorname{Im}\left\{ \frac{\psi_m(z)}{z_0 \psi_{m-1}(z_0)} \right\} = 0.$$

D) $z_0 = \pm i y_k^{(m)}$ (i.e. $\psi_m(z_0) = 0$): We have $g_0 = 1$ and $f_1 = 0$. $R(z)$ becomes $N_m(z)/N_m(-z)$ of order $2m$. As it is A-acceptable, we must necessarly have $y_k^{(m-1)} < y_k^{(m)} < y_{k+1}^{(m-2)}$.

E) $r_{m+1}(z_0) = 0$, $\operatorname{Re} z_0 \neq 0$: By definition of $r_{m+1}(z)$ (4.6) is satisfied in this case. We therefore have $g_0 = 1$ but $f_1 \neq 0$. Thus, the approximation $R(z)$ has exact order $2m - 1$ (this is denoted by dotted lines in Fig. 1).

Appendix. The order star of a rational approximation to $\exp(z)$ shows all its essential properties (cf. Wanner, Hairer and Nørsett [1978]). We depict in Fig. 3 the evolution of the order star of $R_2(z)$ with g_0, f_1 given by (2.8) and $z_0 = x + 7i$ for a range of values of $x < 0$. Note that with $x = -1.834651861$ we have $r_2(z_0) = 0$. At this point a drop in the order can be perceived, as predicted by the above considerations.

REFERENCES

M. Abramowitz and J. A. Stegun [1965], *Handbook of Mathematical Functions*, Dover, New York.

E. Hairer [1982], *Constructive characterization of A-stable approximations to* exp(z) *and...*, Numer. Math., 39, pp. 247–258.

A. Iserles [1979], *On the generalized Padé approximations to the exponential function*, SIAM J. Numer. Anal., 16, pp. 631–636.

A. Iserles and S. P. Nørsett [1982a], *Frequency fitting of rational approximations to the exponential function*, Math. Comp., 40 (1983), pp. 547–559.

______ [1982b], *Rational approximations to the exponential function with two complex conjugate interpolation points*, Report 7/82, Univ. Trondheim, Trondheim, Norway.

A. Iserles and M. J. D. Powell [1981], *On the A-acceptability of rational approximations that interpolate the exponential function*, IMA J. Numer. Anal., 1, pp. 241–251.

W. Liniger and R. A. Willoughby [1970], *Efficient integration methods for stiff systems of ordinary differential equations*, SIAM J. Numer. Anal. 7, pp. 47–66.

G. Wanner, E. Hairer and S. P. Nørsett [1978], *Order stars and stability theorems*, BIT, 18, pp. 475–489.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

013

DIRECTIONALLY DEPENDENT ASYMPTOTIC BEHAVIOR OF BIHARMONIC FUNCTIONS WITH APPLICATIONS TO ELASTICITY*

KENNETH B. HOWELL†

Abstract. The behavior of biharmonic functions defined on infinite domains is investigated with interest focused on obtaining local bounds on the gradients of these functions based on assumed local bounds on the original biharmonic functions. The assumed bounds involve two independent distances, each raised to some arbitrarily chosen exponent. One of the two distances is the distance to some fairly arbitrary subset of the closure of the domain (e.g., the boundary) while the other is the distance to some arbitrarily chosen plane. The derived bounds on the gradients reflect this "directional dependency". In addition, as the first distance increases, the derived bounds on the gradients tend to either increase more slowly or decrease more rapidly than the assumed bounds on the original biharmonic functions.

Several classes of problems from classical elasticity are then discussed. These problems involve unbounded domains and either periodic or "slightly periodic" boundary data. Using the results from the first part of the paper "physically reasonable" assumptions are shown to insure appropriate periodicity in the solutions to periodic boundary value problems and appropriate uniqueness in the solutions to "slightly periodic" boundary value problems.

1. Introduction. When dealing with problems on infinite domains, one must have some concern about the asymptotic behavior of the functions and their derivatives "near infinity". Often, for example, it is desired—and assumed—that one or more of the functions and their derivatives rapidly and uniformly approach well-defined (and computable) limits "at infinity". Muskhelishvili [8] and Gurtin and Sternberg [2] have shown the extent to which this very type of asymptotic behavior can be expected in classical elastostatic problems on domains exterior to some compact set (see, also, Knops and Payne [7, Chap. 6] for a discussion of similar problems on the whole- and half-plane). Unfortunately, it is not always clear that the asymptotic behavior one would desire or expect can be guaranteed on more complex domains. Indeed, in many cases, just determining what behavior should be desired or expected is a significant problem in itself. The problem is often complicated by the fact that the asymptotic behavior may be strongly dependent on the direction along which "infinity" is approached. These are the sort of problems we shall examine in this and a subsequent paper. In the first part of this paper, local bounds will be derived describing the asymptotic behavior of the gradient of a biharmonic function based on the asymptotic behavior assumed for the original function. This differs from previous work (cf. Knops and Payne [7, Chap. 6]) in that the assumed asymptotic behavior will be bounded by linear combinations of terms of the form $x^\gamma\rho^\beta$ where ρ denotes the distance from some fairly arbitrary set of points (in practice, the boundary of the domain), x denotes the distance from some arbitrarily chosen plane, and γ and β are any fixed pair of real constants with $\gamma > -1$. The derived bounds on the gradient will reflect this dependency on direction. The exact form of the derived bounds will depend on whether γ is positive or not. In either case β may be positive, negative, or zero. Also, we shall allow the domain to be a bit more general then the half-space.

* Received by the editors December 7, 1982. This research was supported in part by an Incentive Fund Research Grant from the University of Alabama in Huntsville.

† Department of Mathematics, University of Alabama in Huntsville, Huntsville, Alabama 35899.

The investigation of the first part culminates in Theorem 4.3. This theorem should be considered the main result of this paper, and will be the theorem most commonly referred to in the sequels.

The second part of this paper will be devoted to a discussion of the uniqueness and some general properties of the solutions to various classes of problems in classical elasticity. By making use of the results developed in the first part of the paper, it will be possible to base this discussion on more satisfying and "physically reasonable" assumptions than previously possible (cf. Howell [5]).

The work discussed in the second part of this paper makes fairly straightforward use of the results from the first part. A somewhat deeper use of the results from the first part, however, leads to a very strong Saint-Venant's principle for certain classes of problems in elasticity. For these problems, it can be shown that if the stress grows no faster than some arbitrary polynomial, then, in fact, the "stress at infinity" (actually, an affine tensor field) is completely determined by four "computable" constants. Moreover, the rate at which an elastic state approaches its asymptotic form is on the order of $O(\rho^{-\beta})$ where ρ is the distance to the boundary of the domain and β is any arbitrarily fixed real number. These results go beyond the scope of the paper at hand and will be found in the "subsequent paper" alluded to previously (Howell [6]).

2. Preliminaries. The reader is reminded that a function, ϕ, is biharmonic on some open region in Euclidean space, Ω, if it is in $C^4(\Omega)$ and satisfies $\Delta\Delta\phi=0$ throughout Ω. (In this paper, Δ will denote the Laplacian, while ∇^n will denote n successive applications of the gradient operator.) Attention shall be restricted to biharmonic functions on two- and three-dimensional domains. Thus, the term "space" will henceforth be taken as synonymous with the expression "two- or three-dimensional Euclidean space". When convenient, points in k-dimensional space will be identified with vectors in $\mathbb{R}^k$ in the standard manner—i.e., through the agency of a suitably chosen Cartesian coordinate system. The induced orthonormal frame of vectors will be denoted by $\{\mathbf{e}^1,\mathbf{e}^2,\cdots,\mathbf{e}^k\}$. As usual, if $\mathbf{v}$ and T are, respectively, a vector and a (second rank) tensor, then v_i will denote $\mathbf{v}\cdot\mathbf{e}^i$ and T_{ij} will denote $\mathbf{e}^i\cdot T\mathbf{e}^j$.

All calculations will be done assuming that the domain in question is three-dimensional. Corresponding results for functions on two-dimensional domains will follow immediately—and without much comment—by the fact that any biharmonic function $\phi(x_1,x_2)$, on a two-dimensional domain, Ω, can be viewed as the biharmonic function $\psi(x_1,x_2,x_3)$ on the three-dimensional domain $\{(x_1,x_2,x_3)\colon (x_1,x_2)\in\Omega\}$ where ψ is defined by $\psi(x_1,x_2,x_3)=\phi(x_1,x_2)$.

Given any domain Ω in space, $\operatorname{cl}\Omega$ will denote the closure of that domain, and $\mathbf{n}$ will denote the outward unit normal vector field on the domain's boundary. Two types of domains will be of particular interest: open balls, which will be denoted by $\mathscr{D}$, and "cones" (defined at the beginning of §4), which will be denoted by $\mathscr{K}$.

3. Some local bounds. The investigation begins with the following well-known mean-value theorem for the biharmonic functions (Nicolesco [9]) and a slightly less well-known corollary. For completeness, we include the proof of the corollary.

THEOREM 3.1. *Let $\mathscr{D}_\rho$ be the ball of radius ρ centered at the point $\mathbf{x}$ in three-dimensional space. Let ϕ be biharmonic in a domain containing the closure of $\mathscr{D}_\rho$. Then*

$$\phi(\mathbf{x})=\frac{3}{8\pi}\left[\frac{5}{\rho^3}\int_{\mathscr{D}_\rho}\phi\,dv-\frac{1}{\rho^2}\int_{\partial\mathscr{D}_\rho}\phi\,da\right].$$

COROLLARY 3.2. *Let ϕ be biharmonic on a two- or three-dimensional domain containing the closed ball of radius ρ centered at the point* $\mathbf{x}$. *Then there is a universal constant, K, independent of the choice of* $\mathbf{x}$ *and ρ, such that*

$$|\nabla\phi(\mathbf{x})| \leqq K\rho^{-1} \sup_{|\mathbf{x}-\mathbf{y}|\leqq\rho} |\phi(\mathbf{y})|. \tag{3.1}$$

In particular, if the domain is three-dimensional

$$\nabla\phi(\mathbf{x}) = \frac{3}{2\pi}\rho^{-4}\left[6\int_{\mathscr{D}_\rho} \phi\mathbf{r}\,dv - \rho\int_{\partial\mathscr{D}_\rho} \phi\mathbf{n}\,da\right] \tag{3.2}$$

where

$$\mathbf{r} = \mathbf{r}(\mathbf{y}) = \frac{\mathbf{y}-\mathbf{x}}{|\mathbf{y}-\mathbf{x}|}.$$

Proof. In that (3.1) follows easily from (3.2), only (3.2) shall be derived.

Let $\mathscr{D}_\lambda$ denote the ball of radius λ about $\mathbf{x}$. Since ϕ is biharmonic, so is $\nabla\phi$. Applying Theorem 3.1 along with the divergence theorem to $\nabla\phi$ leads to

$$\frac{8\pi}{3}\lambda^3\nabla\phi(\mathbf{x}) = 5\int_{\partial\mathscr{D}_\lambda} \phi\mathbf{n}\,da - \lambda\int_{\partial\mathscr{D}_\lambda} \nabla\phi\,da$$

for each $\lambda \leqq \rho$. Integrating this with respect to λ from 0 to ρ and employing the divergence theorem, again, yields

$$\frac{2\pi}{3}\rho^4\nabla\phi(\mathbf{x}) = 5\int_{\mathscr{D}_\rho} \phi\mathbf{r}\,da - \left[\rho\int_{\mathscr{D}_\rho} \nabla\phi\,dv - \int_0^\rho \int_{\mathscr{D}_\lambda} \nabla\phi\,dv\,d\lambda\right]$$

$$= 6\int_{\mathscr{D}_\rho} \phi\mathbf{r}\,dv - \rho\int_{\partial\mathscr{D}_\rho} \phi\mathbf{n}\,da,$$

which is equivalent to (3.2). □

In deriving the formulas describing the asymptotic behavior of $\nabla\phi$ based on the asymptotic behavior of ϕ, estimates will have to be computed for the integrals appearing in formula (3.2). Some of the more tedious computations required are established in the following lemma.

LEMMA 3.3. *Let γ be a fixed real number greater than -1, and let* $\mathbf{m}$ *be some fixed unit vector.*

If $\gamma \leqq 0$ then there is a constant, C_γ, such that for any given point $\bar{\mathbf{x}}$, and any positive number, σ:

$$\int_{\partial\mathscr{D}(\bar{\mathbf{x}},\sigma)} (1+|\mathbf{x}\cdot\mathbf{m}|)^\gamma da \leqq C_\gamma\sigma^2(1+|\bar{\mathbf{x}}\cdot\mathbf{m}|)^\gamma \tag{3.3}$$

where $\mathscr{D}(\bar{\mathbf{x}},\sigma)$ is the ball of radius σ centered at $\bar{x}$.

On the other hand, if $\gamma > 0$, then there is a constant, C_γ, such that for any given point, $\bar{\mathbf{x}}$, *and any positive number σ*:

$$\int_{\partial\mathscr{D}(\bar{\mathbf{x}},\sigma)} (1+|\mathbf{x}\cdot\mathbf{m}|)^\gamma da \leqq C_\gamma\left[\sigma^2(1+|\bar{\mathbf{x}}\cdot\mathbf{m}|)^\gamma + \sigma^{\gamma+2}\right] \tag{3.4}$$

where $\mathscr{D}(\bar{\mathbf{x}},\sigma)$ is the ball of radius σ centered at $\bar{\mathbf{x}}$.

Proof. Clearly it suffices to derive the above assuming that $\bar{\mathbf{x}}$ is a point in three-space with $\bar{\mathbf{x}}\cdot\mathbf{m}\geqq 0$. For convenience $\bar{\mathbf{x}}\cdot\mathbf{m}$ and $\mathbf{x}\cdot\mathbf{m}$ will often be denoted by $\bar{x}$ and x, respectively and $\mathscr{D}$ will denote the unit ball about $\mathbf{0}$. In addition, J will occasionally be used to denote the integral appearing in the statement of this lemma.

Three cases must be considered.

Case 1. $\gamma>-1$ and $0<\sigma\leqq\bar{\mathbf{x}}\cdot\mathbf{m}=\bar{x}$.

Since $\sigma\leqq\bar{x}$, the following manipulations are valid:

$$\begin{aligned} J&=\int_{\partial\mathscr{D}(\bar{\mathbf{x}},\sigma)}(1+|\mathbf{x}\cdot\mathbf{m}|)^{\gamma}da\\ &=\sigma^2\int_{\partial\mathscr{D}}(1+\bar{x}+\sigma\mathbf{x}\cdot\mathbf{m})^{\gamma}da\\ &=(1+\bar{x})^{\gamma}\sigma^2\int_{\partial\mathscr{D}}\left[1+\frac{\sigma\mathbf{x}\cdot\mathbf{m}}{1+\bar{x}}\right]^{\gamma}da. \end{aligned}$$

Letting $\mathbf{m}$ define the polar axis of a spherical coordinate system leads to an easy evaluation of the last integral. The result is

$$J=\frac{2\pi}{\gamma+1}(1+\bar{x})^{\gamma+1}\sigma\left\{\left[1+\frac{\sigma}{1+\bar{x}}\right]^{\gamma+1}-\left[1-\frac{\sigma}{1+\bar{x}}\right]^{\gamma+1}\right\}$$

which can be rewritten as

$$J=\frac{2\pi}{\gamma+1}(1+\bar{x})^{\gamma}\sigma^2F(\omega), \tag{3.5}$$

where $\omega=\sigma/(1+\bar{x})$ and

$$F(\omega)=\frac{(1+\omega)^{\gamma+1}-(1-\omega)^{\gamma+1}}{\omega}.$$

By the assumptions on σ and $\bar{x}$, ω lies in the interval $(0,1)$. Now, $F(\omega)$ is not only clearly continuous on $(0,1)$ but satisfies

$$\lim_{\omega\to 0}F(\omega)=2(\gamma+1),\qquad \lim_{\omega\to 1}F(\omega)=2^{\gamma+1}.$$

Thus, $F(\omega)$ is bounded by some finite constant on $(0,1)$. Replacing $F(\omega)$ with this constant in (3.5) completes the proof of the lemma for Case 1.

Case 2. $-1<\gamma\leqq 0$ and $0\leqq\bar{\mathbf{x}}\cdot\mathbf{m}=\bar{x}\leqq\sigma$. Straightforward integration quickly verifies that

$$\int_{\partial\mathscr{D}^-}(1+|\mathbf{x}\cdot\mathbf{m}|)^{\gamma}da\leqq\int_{\partial\mathscr{D}^+}(1+|\mathbf{x}\cdot\mathbf{m}|)^{\gamma}da$$

where

$$\begin{aligned}\mathscr{D}^-&=\{\mathbf{x}\in\mathscr{D}(\bar{\mathbf{x}},\sigma)\colon\ \mathbf{x}\cdot\mathbf{m}<0\},\\ \mathscr{D}^+&=\{\mathbf{x}\in\mathscr{D}(\bar{\mathbf{x}},\sigma)\colon\ \mathbf{x}\cdot\mathbf{m}>0\}.\end{aligned}$$

Thus, using a spherical coordinate system centered at $\bar{\mathbf{x}}$ with polar axis defined by $\mathbf{m}$,

$$J=\int_{\partial\mathscr{D}(\bar{\mathbf{x}},\sigma)}(1+|\mathbf{x}\cdot\mathbf{m}|)^{\gamma}da \leqq 2\int_{\partial\mathscr{D}^{+}}(1+|\mathbf{x}\cdot\mathbf{m}|)^{\gamma}da \tag{3.6}$$

$$=4\pi\sigma^2\int_0^{\psi_0}(1+\bar{x}+\sigma\cos\psi)^{\gamma}\sin\psi\,d\psi=J_1+J_2.$$

Here ψ_0 denotes the angle between $\pi/2$ and π satisfying $\bar{x}+\sigma\cos\psi_0=0$ and J_1 and J_2 are given by

$$J_1=4\pi\sigma^2\int_0^{\pi/2}(1+\bar{x}+\sigma\cos\psi)^{\gamma}\sin\psi\,d\psi,$$

$$J_2=4\pi\sigma^2\int_{\pi/2}^{\psi_0}(1+\bar{x}+\sigma\cos\psi)^{\gamma}\sin\psi\,d\psi.$$

Since $\gamma\leqq 0$ it is obvious that

$$J_1\leqq 4\pi\sigma^2\int_0^{\pi/2}(1+\bar{x})^{\gamma}\sin\psi\,d\psi=4\pi\sigma^2(1+\bar{x})^{\gamma}. \tag{3.7}$$

Evaluation of the integral defining J_2 and the fact that $\bar{x}^{-1}\geqq\sigma^{-1}$ yields

$$J_2=\frac{4\pi\sigma}{\gamma+1}\left[(1+\bar{x})^{\gamma+1}+1\right] \tag{3.8}$$

$$=\frac{4\pi}{\gamma+1}\sigma^2(1+\bar{x})^{\gamma}\left[\frac{1+\bar{x}-(1+\bar{x})^{-\gamma}}{\sigma}\right]$$

$$\leqq\frac{4\pi}{\gamma+1}\sigma^2(1+\bar{x})^{\gamma}G(\bar{x}),$$

where

$$G(\bar{x})=\frac{1+\bar{x}-(1+\bar{x})^{-\gamma}}{\bar{x}}.$$

Clearly, G is continuous on $(0,\infty)$. Further, by L'Hôpital and the fact that $-1<\gamma\leqq 0$

$$\lim_{\bar{x}\to 0}G(\bar{x})=1+\gamma,\qquad \lim_{\bar{x}\to\infty}G(\bar{x})=1,$$

we see that $G(\bar{x})$ is bounded, say by G_{γ}. Substituting this into (3.8) and combining with (3.7) and (3.6) yields

$$J\leqq 4\pi\left[1+\frac{G_{\gamma}}{(1+\gamma)}\right](1+\bar{x})^{\gamma}\sigma^2$$

as claimed in the lemma.

Case 3. $0<\gamma$ and $0\leqq \bar{\mathbf{x}}\cdot\mathbf{m}=\bar{x}<\sigma$.

In this case, it is clear that

$$J=\int_{\partial\mathcal{D}(\bar{\mathbf{x}},\sigma)}(1+|\mathbf{x}\cdot\mathbf{m}|)^{\gamma}da$$

$$\leqq 2\int_{\substack{\partial\mathcal{D}(\bar{\mathbf{x}},\sigma)\\ x>\bar{x}}}(1+|\mathbf{x}\cdot\mathbf{m}|)^{\gamma}da$$

$$=4\pi\sigma^{2}\int_{0}^{\pi/2}(1+\bar{x}+\sigma\cos\psi)^{\gamma}\sin\psi\, d\psi$$

$$=\frac{4\pi\sigma}{\gamma+1}\left[(1+\bar{x}+\sigma)^{\gamma+1}-(1+\bar{x})^{\gamma+1}\right]$$

$$=\frac{4\pi\sigma}{\gamma+1}(1+\bar{x})^{\gamma+1}\left[\left(1+\frac{\sigma}{1+\bar{x}}\right)^{\gamma+1}-1\right].$$

Now, let N be the greatest integer which is less than or equal to γ, and let $\alpha=\gamma-N$. Applying simple algebra and the binomial theorem to the above gives the following:

$$(3.9)\qquad J\leqq\frac{4\pi}{\gamma+1}(1+\bar{x})^{\gamma+1}\sigma\left\{\left[1+\frac{\sigma}{1+\bar{x}}\right]^{\alpha-1}\left[1+\frac{\sigma}{1+\bar{x}}\right]^{N+2}-1\right\}$$

$$=\frac{4\pi}{\gamma+1}(1+\bar{x})^{\gamma+1}\sigma\left\{\left[1+\frac{\sigma}{1+\bar{x}}\right]^{\alpha-1}\sum_{k=0}^{N+2}C_{N+2,k}\left(\frac{\sigma}{1+\bar{x}}\right)^{k}-1\right\}.$$

However, since $\alpha<1$,

$$\left[1+\frac{\sigma}{1+\bar{x}}\right]^{\alpha-1}\leqq\min\left\{1,\left(\frac{\sigma}{1+\bar{x}}\right)^{\alpha-1}\right\}$$

and

$$C_{N+2,0}\left(\frac{\sigma}{1+\bar{x}}\right)^{0}\left[1+\frac{\sigma}{1+\bar{x}}\right]^{\alpha-1}-1=\left[1+\frac{\sigma}{1+\bar{x}}\right]^{\alpha-1}-1\leqq 0.$$

Combining these inequalities with (3.9) leads to the following string of inequalities:

$$J\leqq\frac{4\pi}{\gamma+1}(1+\bar{x})^{\gamma+1}\sigma\left\{\left[1+\frac{\sigma}{1+\bar{x}}\right]^{\alpha-1}\sum_{k=1}^{N+2}C_{N+2,k}\left(\frac{\sigma}{1+\bar{x}}\right)^{k}\right\}$$

$$\leqq\frac{4\pi}{\gamma+1}(1+\bar{x})^{\gamma+1}\sigma\left\{\sum_{k=1}^{N+1}C_{N+2,k}\left(\frac{\sigma}{1+\bar{x}}\right)^{k}+\left(\frac{\sigma}{1+\bar{x}}\right)^{\alpha-1}C_{N+2,N+2}\left(\frac{\sigma}{1+\bar{x}}\right)^{N+2}\right\}$$

$$=\frac{4\pi}{\gamma+1}\left\{\sigma^{\gamma+2}+\sum_{k=1}^{N+1}C_{N+2,k}(1+\bar{x})^{\gamma+1-k}\sigma^{k+1}\right\}$$

$$=\frac{4\pi}{\gamma+1}\left\{\sigma^{\gamma+2}+\sum_{k=0}^{N}C_{N+2,k+1}(1+\bar{x})^{\gamma-k}\sigma^{k+2}\right\}.$$

Finally, it should be noted that, for $0 \leqq k \leqq N$, either

$$(1+\bar{x})^{\gamma-k}\sigma^{k+2} \leqq \sigma^{\gamma+2}$$

or

$$(1+\bar{x})^{\gamma-k}\sigma^{k+2} \leqq (1+\bar{x})^{\gamma}\sigma^{2},$$

depending on whether $1+\bar{x}$ or σ is the greater. This, and the choosing of

$$C_N = \frac{4\pi N}{\gamma+1}\max\{C_{N+2,k}: k=0,1,\cdots,N+1\}$$

allows the last bound on J to be simplified to

$$J \leqq C_N\{\sigma^{\gamma+2}+(1+\bar{x})^{\gamma}\sigma^{2}\}.$$

This completes the proof of the lemma. □

4. Asymptotic behavior. At this point, it is convenient to explain the term "cone" as used in this paper. Let $\mathbf{x}$ be a fixed point in space, $\mathbf{v}$ a unit vector, and θ a scalar satisfying $0<\theta \leqq \pi/2$. The corresponding cone, $\mathcal{K}=\mathcal{K}(\mathbf{x},\mathbf{v},\theta)$, is the region in space given by

$$\mathcal{K}=\{\mathbf{y}: (\mathbf{y}-\mathbf{x})\cdot\mathbf{v} < |\mathbf{y}-\mathbf{x}|\cos\theta\}.$$

This definition corresponds, of course, to the standard notion of a solid cone with vertex at $\mathbf{x}$, axis in the direction of $\mathbf{v}$, and whose sides make an angle of θ with the axis.

If, instead, one has a set of points in space, Σ, and a unit vector field on Σ, $\mathbf{v}(\mathbf{x})$, and a single fixed θ with $0<\theta \leqq \pi/2$, then the corresponding cone, $\mathcal{K}=\mathcal{K}(\Sigma,\mathbf{v},\theta)$, is given by

$$\mathcal{K}(\Sigma,\mathbf{v},\theta)=\bigcup_{\mathbf{x}\in\Sigma}\mathcal{K}(\mathbf{x},\mathbf{v}(\mathbf{x}),\theta).$$

Finally, if n is a nonnegative integer, the nth subcone, $\mathcal{K}_n$, of $\mathcal{K}(\Sigma,\mathbf{v},\theta)$ is defined by

$$\mathcal{K}_n=\mathcal{K}(\Sigma,\mathbf{v},2^{-n}\theta).$$

It may be noted that $\mathcal{K}_0=\mathcal{K}$.

Obviously, many domains of interest contain cones. In fact, when given an unbounded domain in space, Ω, it is often possible to choose $\Sigma, \mathbf{v}$, and θ so that $\Omega=\mathcal{K}=\mathcal{K}_1=\mathcal{K}_2=\cdots$ or so that for "most" $\mathbf{x}$ in Ω the distance from $\mathbf{x}$ to $\partial\Omega$ is closely approximated by the distance from $\mathbf{x}$ to Σ. Some of the more simple examples of such domains would be half-spaces and exterior domains.

The next lemma deals with biharmonic functions on the simplest type of cone.

LEMMA 4.1. *Let* $\mathbf{x}^0$ *be a point in space,* $\mathbf{v}$ *a unit vector,* θ *a constant with* $0<\theta\leqq\frac{1}{2}\pi$ *and* $\mathcal{K}=\mathcal{K}(\mathbf{x}^0,\mathbf{v},\theta)$. *Let* f *be a positive-valued, locally integrable, nonincreasing function on* $[0,\infty)$ *and let* β,γ, *and* c *be real constants with* c *positive and* γ *greater than* -1. *Let* $\mathbf{m}$ *denote some fixed unit vector.*

Suppose ϕ *is a biharmonic function on* $\mathcal{K}$ *which satisfies the following bound for each* $\mathbf{x}$ *in* $\mathcal{K}$

$$|\phi(\mathbf{x})| \leqq c(1+|\mathbf{m}\cdot\mathbf{x}|)^{\gamma}|\mathbf{x}-\mathbf{x}^0|^{\beta}f(|\mathbf{x}-\mathbf{x}^0|). \tag{4.1}$$

If $-1<\gamma\leqq 0$, *then, for each* $\mathbf{x}$ *in* $\mathscr{K}_1=\mathscr{K}(\mathbf{x}^0,\mathbf{v},\frac{1}{2}\theta)$, $\nabla\phi(\mathbf{x})$ *satisfies*

$$|\nabla\phi(\mathbf{x})|\leqq Bc(1+|\mathbf{m}\cdot\mathbf{v}|)^{\gamma}|\mathbf{x}-\mathbf{x}^0|^{\beta-1}f\left(\frac{1}{2}|\mathbf{x}-\mathbf{x}^0|\right).$$

If $0<\gamma$, *then, for each* $\mathbf{x}$ *in* $\mathscr{K}_1=\mathscr{K}(\mathbf{x}^0,\mathbf{v},\frac{1}{2}\theta)$, $\nabla\phi(\mathbf{x})$ *satisfies*

$$(4.2)\qquad |\nabla\phi(\mathbf{x})|\leqq Bc\left\{|\mathbf{x}-\mathbf{x}^0|^{\beta+\gamma-1}+(1+|\mathbf{x}\cdot\mathbf{m}|)^{\gamma}|\mathbf{x}-\mathbf{x}^0|^{\beta-1}\right\}f\left(\frac{1}{2}|\mathbf{x}-\mathbf{x}^0|\right).$$

In either case B is the constant, dependent on β,γ, *and* θ *given by*

$$B=\frac{2^{|\beta|}9}{\sin(\theta/2)}C_{\gamma},$$

where C_γ *is the constant, dependent on* γ, *from Lemma* 3.3.

Proof. As already discussed, it suffices to only consider the case where $\mathscr{K}$ is three-dimensional.

Let $\mathbf{x}$ be fixed in $\mathscr{K}$ and let

$$\sigma_0=\frac{1}{2}|\mathbf{x}-\mathbf{x}^0|\sin\left(\frac{1}{2}\theta\right).$$

By elementary geometry it is easily seen that if

$$|\mathbf{x}-\mathbf{y}|\leqq\sigma_0,$$

then $\mathbf{y}$ is in $\mathscr{K}$ and the following inequalities hold:

$$\frac{1}{2}|\mathbf{x}-\mathbf{x}^0|\leqq|\mathbf{y}-\mathbf{x}^0|\leqq\frac{3}{2}|\mathbf{x}-\mathbf{x}^0|$$

and so

$$|\mathbf{y}-\mathbf{x}^0|^{\beta}\leqq 2^{|\beta|}|\mathbf{x}-\mathbf{x}^0|^{\beta}.$$

Let $0\leqq\sigma\leqq\sigma_0$ and let $\mathscr{D}_\sigma$ be the ball of radius σ about $\mathbf{x}$. By the bounds assumed on ϕ and the above

$$(4.3)\qquad \begin{aligned}\int_{\partial\mathscr{D}_\sigma}|\phi|\,da&\leqq c\int_{\partial\mathscr{D}_\sigma}(1+|\mathbf{m}\cdot\mathbf{y}|)^{\gamma}|\mathbf{y}-\mathbf{x}^0|^{\beta}f(|\mathbf{y}-\mathbf{x}^0|)\,da_{\mathbf{y}}\\ &\leqq c2^{|\beta|}|\mathbf{x}-\mathbf{x}^0|^{\beta}f\left(\frac{1}{2}|\mathbf{x}-\mathbf{x}^0|\right)\int_{\partial\mathscr{D}_\sigma}(1+|\mathbf{m}\cdot\mathbf{y}|)^{\gamma}da_{\mathbf{y}}.\end{aligned}$$

Assume first that $-1<\gamma\leqq 0$. Formula (3.3) from Lemma 3.3 and (4.3) together imply that

$$(4.4)\qquad \int_{\partial\mathscr{D}_\sigma}|\phi|\,da\leqq cC_\gamma 2^{|\beta|}|\mathbf{x}-\mathbf{x}^0|^{\beta}\sigma^2(1+|\mathbf{m}\cdot\mathbf{x}|)^{\gamma}f\left(\frac{1}{2}|\mathbf{x}-\mathbf{x}^0|\right).$$

In particular, for $\sigma=\sigma_0=\frac{1}{2}|\mathbf{x}-\mathbf{x}^0|\sin(\frac{1}{2}\theta)$

$$(4.5)\qquad \frac{1}{\sigma_0^3}\int_{\partial\mathscr{D}_{\sigma_0}}|\phi|\,da\leqq\left[\frac{c2^{|\beta|+1}}{\sin(\theta/2)}C_\gamma\right]|\mathbf{x}-\mathbf{x}^0|^{\beta-1}(1+|\mathbf{m}\cdot\mathbf{x}|)^{\gamma}f\left(\frac{1}{2}|\mathbf{x}-\mathbf{x}^0|\right),$$

and by integrating expression (4.5)

$$(4.6)\qquad \frac{1}{\sigma_0^4}\int_{\mathscr{D}_{\sigma_0}} |\phi|\,dv = \frac{1}{\sigma_0^4}\int_{\sigma=0}^{\sigma_0}\int_{\partial\mathscr{D}_\sigma} |\phi|\,da\,d\sigma$$

$$\leqq \left[\frac{c2^{|\beta|+1}}{3\sin(\theta/2)}C_\gamma\right]|\mathbf{x}-\mathbf{x}^0|^{\beta-1}(1+|\mathbf{m}\cdot\mathbf{x}|)^\gamma f\left(\frac{1}{2}|\mathbf{x}-\mathbf{x}^0|\right).$$

On the other hand, it follows from Corollary 3.2 that

$$(4.7)\qquad |\nabla\phi(\mathbf{x})| \leqq \frac{3}{2\pi}\sigma_0^{-4}\left[6\int_{\mathscr{D}_{\sigma_0}}|\phi|\,dv + \sigma_0\int_{\partial\mathscr{D}_{\sigma_0}}|\phi|\,da\right].$$

Inequality (4.1) of this lemma now follows immediately after substituting (4.5) and (4.6) into (4.7).

The remaining inequality of this lemma, (4.2), follows in the same manner using (3.4) from Lemma 3.3 instead of (3.3). □

On occasion it will be desirable to take several derivatives of the biharmonic function. The next lemma is designed for such occasions. The proof will not be given, since the lemma is actually an immediate consequence of Lemma 4.1.

LEMMA 4.2. *Let* $\mathbf{x}^0$, $\mathbf{v}, \theta, \mathscr{K}=\mathscr{K}(\mathbf{x}^0,\mathbf{v},\theta)$, f, β, γ, c, *and* $\mathbf{m}$ *be as in Lemma* 4.1 *with* $\gamma > 0$. *Let* ϕ *be a biharmonic function on* $\mathscr{K}$ *which satisfies*

$$|\phi(\mathbf{x})| \leqq c\left\{|\mathbf{x}-\mathbf{x}^0|^{\beta+\gamma} + (1+|\mathbf{x}\cdot\mathbf{m}|)^\gamma|\mathbf{x}-\mathbf{x}^0|^\beta\right\}f(|\mathbf{x}-\mathbf{x}^0|)$$

for each $\mathbf{x}$ *in* $\mathscr{K}$. *Then, for each* $\mathbf{x}$ *in* $\mathscr{K}_1$,

$$|\nabla\phi(\mathbf{x})| \leqq Mc\left\{|\mathbf{x}-\mathbf{x}^0|^{\beta+\gamma-1} + (1+|\mathbf{x}\cdot\mathbf{m}|)^\gamma|\mathbf{x}-\mathbf{x}^0|^{\beta-1}\right\}f\left(\frac{1}{2}|\mathbf{x}-\mathbf{x}^0|\right),$$

where

$$M = \frac{9}{\sin(\theta/2)}\left[2^{|\beta+\gamma|}C_0 + 2^{|\beta|}C_\gamma\right]$$

and C_0 *and* C_γ *are the constants from Lemma* 3.3.

The main results of this paper are contained in the following theorem.

THEOREM 4.3. *Let* Σ *be a set of points in space,* $\mathbf{v}$ *a unit vector field on* Σ, θ *a constant with* $0<\theta\leqq\pi/2$, *and* $\mathscr{K}=\mathscr{K}(\Sigma,\mathbf{v},\theta)$. *Assume that*

$$\Sigma \subseteq \mathrm{cl}\mathscr{K}.$$

Let f *be a positive-valued, locally integrable, nonincreasing function on* $[0,\infty)$; *let* $\mathbf{m}$ *be some fixed unit vector, and let* β, γ, *and* c *be real constants with* c *positive and* $-1<\gamma$. *Finally, let* $\rho(\mathbf{x})$ *denote the minimum distance from* $\mathbf{x}$ *to the closure of* Σ.

i) *If* $-1<\gamma\leqq 0$ *and* ϕ *is a biharmonic function on* $\mathscr{K}$ *satisfying*

$$|\phi(\mathbf{x})| \leqq c(1+|\mathbf{m}\cdot\mathbf{x}|)^\gamma[\rho(\mathbf{x})]^\beta f(\rho(\mathbf{x}))$$

for all $\mathbf{x}$ *in* $\mathscr{K}$, *then*

$$|\nabla\phi(\mathbf{x})| \leqq Bc(1+|\mathbf{m}\cdot\mathbf{x}|)^\gamma[\rho(\mathbf{x})]^{\beta-1}f\left(\frac{1}{2}\alpha\rho(\mathbf{x})\right)$$

for all $\mathbf{x}$ *in* $\mathscr{K}_1$, *where*

$$\alpha=\sin\left(\frac{1}{2}\theta\right),\qquad B=\frac{9}{2}\left[\frac{2}{\alpha}\right]^{|\beta|+1}C_\gamma,$$

and C_γ *is the constant, depending on* γ *only, from Lemma* 3.3.

ii) *If* $0<\gamma$ *and* ϕ *and* ψ *are biharmonic functions on* $\mathscr{K}$ *satisfying*

$$|\phi(\mathbf{x})|\leqq c(1+|\mathbf{m}\cdot\mathbf{x}|)^\gamma[\rho(\mathbf{x})]^\beta f(\rho(\mathbf{x})),$$

$$|\psi(\mathbf{x})|\leqq c\{[\rho(\mathbf{x})]^{\beta+\gamma}+(1+|\mathbf{m}\cdot\mathbf{x}|)^\gamma[\rho(\mathbf{x})]^\beta\}f(\rho(\mathbf{x})),$$

for all $\mathbf{x}$ *in* $\mathscr{K}$, *then*

$$|\nabla\phi(\mathbf{x})|\leqq Bc\{[\rho(\mathbf{x})]^{\beta+\gamma-1}+(1+|\mathbf{m}\cdot\mathbf{x}|)^\gamma[\rho(\mathbf{x})]^{\beta-1}\}f\left(\frac{1}{2}\alpha\rho(\mathbf{x})\right),$$

$$|\nabla\psi(\mathbf{x})|\leqq Mc\{[\rho(\mathbf{x})]^{\beta+\gamma-1}+(1+|\mathbf{m}\cdot\mathbf{x}|)^\gamma[\rho(\mathbf{x})]^{\beta-1}\}f\left(\frac{1}{2}\alpha\rho(\mathbf{x})\right),$$

for all $\mathbf{x}$ *in* $\mathscr{K}_1$, *where*

$$\alpha=\sin\left(\frac{1}{2}\theta\right),$$

$$B=\frac{9}{2^{\gamma+1}}\left[\frac{2}{\alpha}\right]^{|\beta|+\gamma+1}C_\gamma,$$

$$M=\frac{9\left[2^{|\beta|+\gamma}C_0+2^{|\beta|}C_\gamma\right]}{2^{|\beta|+\gamma+1}},$$

and C_0 *and* C_γ *are the constants, depending only on* γ, *from Lemma* 3.3.

Proof. Choose any $\mathbf{x}^0$ in Σ. The bounds claimed by the theorem are easily verified on $\mathscr{K}(\mathbf{x}^0,\mathbf{v}(\mathbf{x}^0),\frac{1}{2}\theta)$ by invoking the previous two lemmas and making the obvious use of the easily derived fact that, if $\mathbf{x}$ is in $\mathscr{K}(\mathbf{x}^0,\mathbf{v}(\mathbf{x}^0),\frac{1}{2}\theta)$,

$$\sin\left(\frac{1}{2}\theta\right)|\mathbf{x}-\mathbf{x}^0|\leqq\rho(\mathbf{x})\leqq|\mathbf{x}-\mathbf{x}^0|.$$

But $\mathscr{K}_1$ is the union of all cones of the form $\mathscr{K}(\mathbf{x}^0,\mathbf{v}(\mathbf{x}^0),\frac{1}{2}\theta)$ where $\mathbf{x}^0$ is in Σ. Thus, the bounds claimed in the theorem must hold throughout $\mathscr{K}_1$. □

As a corollary, it is quite easy to show that any function which is biharmonic on all of space and is bounded by some polynomial is, itself, a polynomial. Although we will be much more interested in biharmonic functions which are not defined everywhere, this corollary does have some application in whole- and half-space problems in elasticity and is of some interest in its own right. Also, it gives a nice simple application of Theorem 4.3. For these reasons we shall go ahead and present the corollary and its proof below.

COROLLARY 4.4. *Let* ϕ *be biharmonic on all of space and suppose there is a positive integer, m, such that*

$$\phi(\mathbf{x})=o(|\mathbf{x}|^m)\quad as\ |\mathbf{x}|\to\infty.$$

Then ϕ *is given by a polynomial on* $\mathbb{R}^k$ *of degree strictly less than m.*

Proof. Let Σ be the surface of any small ball about the origin; let $\mathbf{v}$ be the outward normal vector field on Σ, and let θ be any suitably fixed angle. It is obvious that

$\mathscr{K}=\mathscr{K}_1=\mathscr{K}_2=\cdots$, and that ϕ satisfies the conditions of Theorem 4.3 with $\gamma=0$, $\beta=m$, and with f chosen so that $f(a)$ approaches zero as a approaches infinity. Successive applications of Theorem 4.3 yield

$$\begin{aligned} \nabla\phi(\mathbf{x}) &= o\left(|\mathbf{x}|^{m-1}\right) && \text{as } |\mathbf{x}|\to\infty, \\ \nabla^2\phi(\mathbf{x}) &= o\left(|\mathbf{x}|^{m-2}\right) && \text{as } |\mathbf{x}|\to\infty, \\ &\vdots \\ \nabla^m\phi(\mathbf{x}) &= o(1) && \text{as } |\mathbf{x}|\to\infty, \\ \Delta\nabla^m\phi(\mathbf{x}) &= o\left(|\mathbf{x}|^{-2}\right) && \text{as } |\mathbf{x}|\to\infty. \end{aligned}$$

Now, since ϕ is biharmonic, so is $\nabla^m\phi$, and, hence, $\Delta\nabla^m\phi$ is a harmonic function on all of space which, by the above, vanishes uniformly "at infinity". The venerable maximum principle for harmonic functions tells us that $\Delta\nabla^m\phi$ vanishes everywhere. But this means that $\nabla^m\phi$ is, in fact, also harmonic, and, like $\Delta\nabla^m\phi$, the vanishing of $\nabla^m\phi$ "at infinity" forces the vanishing of $\nabla^m\phi$ everywhere.

Solving $\nabla^m\phi=0$ by elementary methods completes the proof of the corollary. □

5. Elastostatics: preliminaries. The (elastic) body will be denoted by $\mathscr{B}$ and its closure by $\mathrm{cl}\,\mathscr{B}$. It occupies an open connected subset of k-dimensional space such that, if $\mathscr{D}$ is any ball of finite radius, then $\partial(\mathscr{B}\cap\mathscr{D})$ consists of a finite union of smooth $(k-1)$-dimensional manifolds with boundary. For the remainder of this paper, it shall be assumed that $\mathscr{B}$ is 3-dimensional (i.e., $k=3$). This will simplify the discussion. Also, some special features of the corresponding two-dimensional problems would lead to a discussion considerably beyond the scope of this paper (see Howell [6]).

An elastic state on $\mathscr{B}$ consists of the displacement (vector) field, the strain field, and the stress field, denoted, respectively, by $\mathbf{u}$, E, and S. They are related throughout $\mathscr{B}$ by the following system:

$$E=\operatorname{sym}\nabla\mathbf{u}, \tag{5.1}$$

$$S=C[E]=2\mu E+\lambda(\operatorname{tr}E)\mathrm{Id}, \tag{5.2}$$

$$\operatorname{div}S+\mathbf{b}=\mathbf{0}, \tag{5.3}$$

where μ and λ are real constants (the Lamé moduli) and $\mathbf{b}$ denotes the body forces. If $\mathbf{n}$ is defined, then $\mathbf{s}$ will denote the corresponding surface traction, $S\mathbf{n}$.

Those familiar with the subject have already realized that it is being assumed that $\mathscr{B}$ is composed of homogeneous isotropic material. These assumptions lead to the formula given in (5.2) for the "elasticity field", C. At times additional assumptions will be made on C. The most common assumption will be that C is positive definite, that is, there is a positive constant, c, such that, given any strain field, E, on $\mathscr{B}$

$$E\cdot C[E]\geq c|E|^2$$

at every point in $\mathscr{B}$. C being positive definite is equivalent to the Lamé moduli satisfying both $\mu>0$ and $3\lambda+2\mu>0$. Occasionally, the slightly weaker assumption that C is strongly elliptic will be made. In this case the Lamé moduli satisfy $\mu>0$ and $2\mu+\lambda>0$.

The standard continuity and differentiability conditions assumed in classical treatments will be assumed here.

The reader is reminded of the following three well-known facts:

1. The strain field, E, vanishes throughout a region if and only if $\mathbf{u}$ is a rigid displacement in that region (i.e., $\nabla \mathbf{u}$ is given by a single fixed skew tensor throughout the region).

2. If $\mathbf{b}$ is both curl free and divergence free on $\mathscr{B}$ and $\mu(\lambda+2\mu)\neq 0$, then $\mathbf{u}, E$, and S are biharmonic and infinitely differentiable on $\mathscr{B}$. Thus, the results from the previous sections are applicable.

3. The function C, as defined by formula (5.2), is an invertible mapping of symmetric tensors to symmetric tensors provided both $\mu\neq 0$ and $2\mu+3\lambda\neq 0$.

For a given (mixed) boundary value problem, the Lamé moduli, μ and λ and the body force field, $\mathbf{b}$ are specified and the boundary of $\mathscr{B}$ is partitioned into two subsurfaces, $\mathscr{S}_1$ and $\mathscr{S}_2$, each of which is the domain of a predetermined vector field, $\hat{\mathbf{u}}$ and $\hat{\mathbf{s}}$, respectively. A solution to the boundary value problem consists of an elastic state, $(\mathbf{u}, E, S)$ satisfying

$$\mathbf{u}=\hat{\mathbf{u}} \quad \text{on} \, \mathscr{S}_1, \qquad \mathbf{s}=\hat{\mathbf{s}} \quad \text{on} \, \mathscr{S}_2.$$

If $\mathscr{S}_1$ consists of the entire boundary of $\mathscr{B}$, the problem is referred to as a (surface) displacement problem, while if $\mathscr{S}_2$ equals the boundary of $\mathscr{B}$ (up to a set of surface measure zero) then the problem is called a (surface) traction problem. If the body is unbounded, one also usually desires that the elastic state at the point $\mathbf{x}$ behaves "reasonably" as $|\mathbf{x}|$ approaches infinity along one or more paths in $\mathscr{B}$. Precisely what is meant by "reasonable" depends on the particular problem at hand and will be the object of much of the subsequent investigation presented in this paper.

Given any mixed boundary value problem, there exists the corresponding null boundary value problem in which $\hat{\mathbf{u}}, \hat{\mathbf{s}}$, and $\mathbf{b}$ all vanish on their respective domains ($C, \mathscr{B}$, $\mathscr{S}_1$, and $\mathscr{S}_2$ are are in the original problem). An obvious, but important, fact is that the difference between two solutions to the same mixed boundary value problem is a solution to the corresponding null boundary value problem.

Other types of boundary value problems can be described. In this paper the term "boundary value problem" implies that $\mathbf{b}, \mu$, and λ are prescribed and that if $(\mathbf{u}, E, S)$ is the difference between any two solutions to the problem then $\mathbf{u}\cdot\mathbf{s}$ vanishes almost everywhere on the boundary of $\mathscr{B}$. It is trivial to verify that this includes the mixed boundary value problems described above. For this more general class of boundary value problems, the corresponding null boundary value problem is defined in the obvious manner.

The following theorem, proved under somewhat weaker assumptions than assumed here, may be found in Howell [3, Thm. 2.1]. It deals with the uniqueness of the solutions to boundary value problems on unbounded bodies.

THEOREM 5.1. *Let $\mathscr{B}$ be an unbounded body with a positive definite elasticity field. Suppose that $(\mathbf{u}, E, S)$ is the difference between two solutions to the same general boundary value problem $\mathscr{B}$ and that*

$$\int_{\mathscr{B}\cap\partial\mathscr{D}_R} |\mathbf{u}|^2 da = O(R) \quad \text{as } R\to\infty, \tag{5.4}$$

where $\mathscr{D}_R$ is the ball of radius R about the origin. Then E and S vanish on $\mathscr{B}$ and $\mathbf{u}$ is a rigid displacement.

6. Periodicity: preliminaries. Let $\mathbf{p}$ be a fixed nonzero vector. A scalar-, vector-, or tensor-valued function, ϕ, with domain Ω is said to be periodic (with period $\mathbf{p}$) if both of the following hold:

1. $\mathbf{x}$ is in Ω if and only if $\mathbf{x}+\mathbf{p}$ is in Ω.
2. $\phi(\mathbf{x}+\mathbf{p})=\phi(\mathbf{x})$ for every $\mathbf{x}$ in Ω.

For convenience, the following conventions will be implicit for the remainder of this paper:

1. The cartesian coordinate system mentioned in the first paragraph of §2 is chosen so that

$$\mathbf{p}=p\mathbf{e}^1$$

where $p=|\mathbf{p}|$.

2. Unless otherwise stated, all periodic functions have the same period, $\mathbf{p}$.

A periodic boundary value problem is a boundary value problem in which the body force, $\mathbf{b}$, and the boundary data, are given by periodic functions. An elastic state, $(\mathbf{u},E,S)$, on $\mathscr{B}$ may have periodic displacement, periodic strain, or periodic stress. Clearly, if the displacement is periodic so is the strain, and if the strain is periodic so is the stress. Likewise, if $\mu(2\mu+3\lambda)\neq 0$, then periodic stress implies periodic strain. Periodic strain, however, does not necessarily imply periodic displacement. A trivial example would be an elastic state in which $\mathbf{u}$ is a rigid displacement and E vanishes everywhere. Less trivial examples can be found in Howell [4] and [5]. Perhaps even more disconcerting is an example of a solution to a periodic boundary value problem in which the strain is *not* periodic. This, also, can be found in Howell [5].

A somewhat more general class of problems is the class of *slightly* periodic boundary value problems. A problem is said to be "slightly periodic" if its corresponding null boundary value problem—but not necessarily the original problem—is a periodic boundary value problem. An obvious example of a slightly periodic problem would be any traction problem on a half space.

For many slightly periodic problems it will be convenient to define corresponding "period sections". Let $\mathbf{x}$ be any point in space and $\mathscr{L}_{\mathbf{x}}$ be any plane through $\mathbf{x}$ which is not parallel to $\mathbf{p}$. The corresponding "period section", denoted either by $\mathscr{P}_{\mathbf{x}}$ or $\mathscr{P}$, is the following subset of $\mathscr{B}$:

$$\mathscr{P}=\{\mathbf{y}\in\mathscr{B}\colon \text{for some } 0<\alpha<1, \mathbf{y}-\alpha\mathbf{p}\in\mathscr{L}_{\mathbf{x}}\}.$$

Because of periodicity, the actual choice of $\mathbf{x}$, and $\mathscr{L}_{\mathbf{x}}$, will usually be irrelevant.

Proofs of the following theorems may be found in Howell [5]. The first is concerned with the uniqueness of solutions to periodic boundary value problems. The second theorem discusses the extent to which the displacement corresponding to a periodic strain state may, itself, fail to be periodic.

THEOREM 6.1. *Assume that the elasticity field, C, is positive definite. Let $(\mathbf{u},E,S)$ be the difference between two solutions to the same boundary value problem on $\mathscr{B}$ and assume that $\mathbf{u}$ is periodic on $\mathscr{B}$. If, letting $\mathscr{C}_R$ be the circular cylinder of radius R about the X_1-axis,*

$$\int_{\mathscr{P}\cap\partial\mathscr{C}_R}|\mathbf{u}|^2da=O(R)\quad \text{as } R\to\infty, \tag{6.1}$$

then E and S vanish on $\mathscr{B}$ and $\mathbf{u}$ is a rigid displacement.

THEOREM 6.2. *Let* $(\mathbf{u}, E, S)$ *be an elastic state on* $\mathscr{B}$ *with periodic strain*, E. *There is then a fixed skew tensor*, $\overline{W}$, *a constant*, κ, *and a fixed vector*, $\bar{\mathbf{u}}$, *such that*

$$\mathbf{p}\cdot\bar{\mathbf{u}}=0$$

and, for each $\mathbf{x}$ *in* $\mathscr{B}$,

$$\mathbf{u}(\mathbf{x}+\mathbf{p})-\mathbf{u}(\mathbf{x})=\overline{W}\mathbf{x}+\kappa\mathbf{p}+\bar{\mathbf{u}}.$$

Furthermore, if $\mathbf{w}$ *is the rigid displacement given by*

$$\mathbf{w}(\mathbf{x})=p^{-2}[(\mathbf{p}\otimes\bar{\mathbf{u}})\mathbf{x}-(\bar{\mathbf{u}}\otimes\mathbf{p})\mathbf{x}]$$

and $\tilde{\mathbf{u}}$ *is given by*

$$\tilde{\mathbf{u}}(x)=\mathbf{u}(\mathbf{x})+\mathbf{w}(\mathbf{x}),$$

then, for every $\mathbf{x}$ *in* $\mathscr{B}$,

$$\tilde{\mathbf{u}}(\mathbf{x}+\mathbf{p})-\tilde{\mathbf{u}}(\mathbf{x})=\overline{W}\mathbf{x}+\kappa\mathbf{p}.$$

Henceforth, whenever $(\mathbf{u}, E, S)$ is a periodic strain state, $\overline{W}, \kappa$, and $\mathbf{u}$ will denote, respectively, the skew tensor, constant, and vector whose existence is guaranteed by the preceding theorem.

7. Applications. We shall now employ the results from §4 to investigate two issues which arise in the study of elasticity. The first is the determination of the conditions which insure the periodicity of solutions to periodic boundary value problems. The second is the determination of the conditions which insure the uniqueness of the solutions to slightly periodic boundary value problems. The most striking aspect of the results presented here, perhaps, is the lack of strong assumptions concerning the behavior of solutions near infinity. For example, in several theorems it will merely be assumed that

$$E(\mathbf{x}+\mathbf{p})-E(\mathbf{x})=O(\rho^n) \quad \text{as } \rho\to\infty,$$

where ρ is the distance from $\mathbf{x}$ to the boundary of the body and n is an arbitrary positive constant.

Notation will be simplified by use of the "difference operator", δ. If ϕ is a suitable function, then $\delta\phi$, $\delta^2\phi$, $\delta^3\phi$, etc. are defined by

$$\begin{aligned}&\delta\phi(\mathbf{x})=\phi(\mathbf{x}+\mathbf{p})-\phi(\mathbf{x}),\\&\delta^2\phi(\mathbf{x})=\delta[\delta\phi](\mathbf{x})=\delta\phi(\mathbf{x}+\mathbf{p})-\delta\phi(\mathbf{x}),\\&\delta^3\phi(\mathbf{x})=\delta[\delta^2\phi](\mathbf{x})=\delta^2\phi(\mathbf{x}+\mathbf{p})-\delta^2\phi(\mathbf{x}).\end{aligned}$$

Clearly, ϕ is periodic on some domain if and only if $\delta\phi$ vanishes throughout that domain. Also, there is a close relationship between δ and ∇. This relationship comes from

$$\delta\phi(\mathbf{x})=\phi(\mathbf{x}+\mathbf{p})-\phi(\mathbf{x})=\int_{\mathbf{x}}^{\mathbf{x}+\mathbf{p}}\nabla\phi\cdot d\sigma.$$

Thus, if M is the maximum of $|\nabla\phi|$ along some path from $\mathbf{x}^0$ to $\mathbf{x}^0+\mathbf{p}$ and L is the length of that path, then

$$|\delta\phi(\mathbf{x}^0)|\leqq LM. \tag{7.1}$$

For several reasons, attention will be focused on problems involving two broad classes of bodies: periodic exterior bodies and periodic fractional spaces. A body, $\mathscr{B}$, is said to be a periodic exterior body if it satisfies all of the following conditions:

i) $\mathscr{B}$ is three-dimensional.

ii) The characteristic function of $\mathscr{B}$ is periodic.

iii) There is a cylinder of finite radius about the X_1-axis which contains the complement of $\mathscr{B}$.

Given any point, $\mathbf{x}$, of a periodic exterior body, $r=r(\mathbf{x})$ will denote distance from $\mathbf{x}$ to the X_1-axis and x will be used instead of x_1 to denote $\mathbf{x}\cdot\mathbf{e}^1$.

The concept of a periodic fractional space may be viewed as a generalization of the concept of a half space. $\mathscr{B}$ will be said to be a periodic fractional space if all of the following are satisfied:

i) $\mathscr{B}$ is three-dimensional.

ii) The characteristic function of $\mathscr{B}$ is periodic.

iii) There are a straight line, $\mathscr{I}$, which is parallel to the X_1-axis, an angle $0<\theta\leqq\pi/2$, and a constant, $d_0<\infty$, such that whenever $\mathbf{x}$ is a point in $\partial\mathscr{B}$, then there is a $\mathbf{y}$ in $\partial\mathscr{K}(\mathscr{I},\mathbf{e}^2,\theta)$ for which $|\mathbf{x}-\mathbf{y}|\leqq d_0$.

In general, the main distance of interest associated with any point, $\mathbf{x}$, in a periodic fractional space will be the distance from $\mathbf{x}$ to the boundary of $\mathscr{B}$. This distance will be denoted by either ρ or $\rho(\mathbf{x})$.

The periodicity of solution proofs all follow the same pattern. One observes that if $(\mathbf{u},E,S)$ is a solution to a given periodic boundary value problem, then so is $(\mathbf{u}^*,E^*,S^*)$ where

$$\mathbf{u}^*(\mathbf{x})=\mathbf{u}(\mathbf{x}+\mathbf{p}),$$
$$E^*(\mathbf{x})=E(\mathbf{x}+\mathbf{p}),$$
$$S^*(\mathbf{x})=S(\mathbf{x}+\mathbf{p}).$$

Thus, $(\delta\mathbf{u},\delta E,\delta S)$, which can be written $(\mathbf{u}^*-\mathbf{u},E^*-E,S^*-S)$, is the difference between two solutions to the same boundary value problem. Likewise, for $m=1,2,3,\cdots$, $(\delta^m\mathbf{u},\delta^m E,\delta^m S)$ is the difference between two solutions to the same boundary value problem. Successive applications of the results from §4 and the observed relationship between δ and ∇ will lead, eventually, to the existence of a positive integer, M, such that

$$\int_{\mathscr{B}\cap\partial\mathscr{D}_R}|\delta^M\mathbf{u}|\,da=O(R)\quad\text{as } R\to\infty.$$

At this point the general uniqueness Theorem 5.1 can be invoked showing that $\delta^M E$ vanishes on $\mathscr{B}$. Hence, $\delta^{M-1}E$ is periodic. If M is greater than one, other assumptions, along with, possibly, Theorem 6.1, are used to show that $\delta^{M-1}E$ also vanishes on $\mathscr{B}$ and, thus, $\delta^{M-2}E$ is periodic. Repeating these arguments as much as necessary finally leads to the fact that E must be periodic.

By definition, if $(\mathbf{u},E,S)$ is the difference between two solutions to the same slightly periodic boundary value problem, then $(\mathbf{u},E,S)$ is a solution to the corresponding periodic null boundary value problem. It is, therefore, natural that each "periodicity" theorem be followed by a corresponding uniqueness theorem for slightly periodic boundary value problems. The proofs of the first such pair of theorems will be given. Details of most of the remaining theorems will be left to the reader.

8. Applications involving periodic exterior bodies.

THEOREM 8.1. *Let $\mathscr{B}$ be a periodic exterior body with positive definite elasticity. Suppose* $(\mathbf{u}, E, S)$ *is a solution to a periodic boundary value problem on $\mathscr{B}$ which satisfies all of the following*:

i) *For some n*

$$\delta E(\mathbf{x}) = O(r^n) \quad \text{as } r \to \infty. \tag{8.1}$$

ii) *For each* $\mathbf{x}$ *in* $\mathscr{B}$

$$\delta E(\mathbf{x} + k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \quad \text{on } \mathbb{N}. \tag{8.2}$$

iii) *Given any $R > 0$ there is a constant, a_R such that*

$$|\delta^3 \mathbf{u}(\mathbf{x})| \leqq a_R \tag{8.3}$$

whenever $r = r(\mathbf{x}) \leq R$. Then E is periodic.

Proof. We may, without any loss of generality, assume that n is a positive integer and that $\partial\mathscr{B}$ is bounded (on each period section) by a circular cylinder about the X_1-axis of radius 1.

Let $\mathscr{T}_1, \mathscr{T}_2, \mathscr{E}_1$, and $\mathscr{E}_2$ denote the folowing subsets of $\mathscr{B}$:

$$\mathscr{T}_1 = \{\mathbf{x} \in \mathscr{B}: r(\mathbf{x}) < 1\},$$
$$\mathscr{T}_2 = \{\mathbf{x} \in \mathscr{B}: r(\mathbf{x}) < 2\},$$
$$\mathscr{E}_1 = \{\mathbf{x} \in \mathscr{B}: r(\mathbf{x}) > 1\},$$
$$\mathscr{E}_2 = \{\mathbf{x} \in \mathscr{B}: r(\mathbf{x}) > 2\}.$$

By assumption iii), $|\delta^3 \mathbf{u}|$ is bounded on $\mathscr{T}_2$. It should be clear then that if M is any integer greater than 3, then there is a constant a_M such that

$$|\delta^M \mathbf{u}(\mathbf{x})| \leqq a_M$$

whenever $\mathbf{x}$ is in $\mathscr{T}_2$.

We now make the trivial observation that $\mathscr{T}_1 = \mathscr{K} = \mathscr{K}_n$ for $n = 0, 1, 2, \cdots$ where

$$\mathscr{K} = \mathscr{K}\left(\partial\mathscr{T}_1, \mathbf{x} - x_1\mathbf{e}^1, \frac{1}{2}\pi\right).$$

For each $\mathbf{x}$, let $\rho = \rho(\mathbf{x})$ denote the minimum distance from $\mathbf{x}$ to $\partial\mathscr{T}_1$. It should be clear that assumption i) (and the fact that if $\mathbf{x} \in \mathscr{E}_1$ then ρ actually equals $r - 1$) implies the existence of a locally integrable, nonincreasing function, f_n, such that

$$|\delta E(\mathbf{x})| \leqq \rho^n f_n(\rho) \tag{8.4}$$

for every $\mathbf{x}$ in $\mathscr{T}_1$

Observe that for any elastic state $(\hat{\mathbf{u}}, \hat{E}, \hat{S})$

$$\hat{u}_{i,jk} = \hat{E}_{ij,k} + \hat{E}_{ik,j} + \hat{E}_{kj,i}.$$

Thus,

$$|\nabla^2 \hat{\mathbf{u}}(\mathbf{x})| \leqq 3|\nabla \hat{E}(\mathbf{x})|$$

and, using assumption i) and Theorem 4.3

$$(8.5)\qquad |\nabla^2\boldsymbol{\delta}\mathbf{u}(\mathbf{x})|\leqq 3|\nabla\delta E(\mathbf{x})|\leqq 3B\rho^{n-1}f_n\left(\frac{1}{2}\alpha\rho\right)$$

for every $\mathbf{x}$ in $\mathscr{E}_1$ (B and α are fixed constants from Theorem 4.3). Moreover, using (8.5) and the relationship between δ and ∇ discussed in the previous section, it is seen that:

$$\begin{aligned}|\delta^2\nabla\mathbf{u}(\mathbf{x})|&\leqq p\sup\{|\delta\nabla^2\mathbf{u}(\mathbf{y})|:\rho(\mathbf{y})=\rho(\mathbf{x})\}\\&\leqq 3pB^{n-1}f_n\left(\frac{1}{2}\alpha\rho\right)\end{aligned}$$

and

$$\begin{aligned}|\delta^3\mathbf{u}(\mathbf{x})|&\leqq p\sup\{|\delta^2\nabla\mathbf{u}(\mathbf{y})|:\rho(\mathbf{y})=\rho(\mathbf{x})\}\\&\leqq 3p^2B^{n-1}f_n\left(\frac{1}{2}\alpha\rho\right)\end{aligned}$$

for each $\mathbf{x}$ in $\mathscr{E}_1$.

But, inequality (7.1) and Theorem 4.3 can be combined again yielding

$$\begin{aligned}|\delta^4\mathbf{u}(\mathbf{x})|&\leqq p\sup\{|\nabla\delta^3\mathbf{u}(\mathbf{y})|:\rho(\mathbf{y})=\rho(\mathbf{x})\}\\&\leqq p^3(3B)^2\rho^{n-2}f_n\left(\left(\frac{1}{2}\alpha\right)^2\rho\right).\end{aligned}$$

Repeating this last sequence of computations $n-1$ more times leads to the conclusion that

$$(8.6)\qquad |\delta^{3+n}\mathbf{u}(\mathbf{x})|\leqq C\rho^{-1}f_n(\beta\rho)$$

where

$$C=p^{n+3}(3B)^{n+1}\quad\text{and}\quad\beta=\left(\frac{1}{2}\alpha\right)^{n+1}.$$

Now, for any $\mathbf{x}$ in $\mathscr{E}_2$, $r>2$ and $2\rho=2(r-1)>r$. This, inequality (8.6), and the fact that f_n is nonincreasing yields

$$(8.7)\qquad |\delta^{3+n}\mathbf{u}(\mathbf{x})|\leqq 2Cr^{-1}f_n(2\beta)$$

for each $\mathbf{x}$ in $\mathscr{E}_2$.

Next, consider the following integral:

$$\int_{\mathscr{B}\cap\partial\mathscr{D}_R}|\delta^{3+n}\mathbf{u}|^2da=\int_{\mathscr{T}_2\cap\partial\mathscr{D}_R}|\delta^{3+n}\mathbf{u}|^2da+\int_{\mathscr{E}_2\cap\partial\mathscr{D}_R}|\delta^{3+n}\mathbf{u}|^2da.$$

It is easily verified that, using (8.7) and (8.3),

$$(8.8)\qquad \int_{\mathscr{B}\cap\partial\mathscr{D}_R}|\delta^{3+n}\mathbf{u}|^2da=O(R)\quad\text{as } R\to\infty.$$

Recall now that $(\mathbf{u}, E, S)$ is the solution to a periodic boundary value problem. By periodicity, $(\mathbf{u}^*, E^*, S^*)$ is also a solution to that problem provided

$$\begin{aligned} \mathbf{u}^*(\mathbf{x}) &= \mathbf{u}(\mathbf{x}+\mathbf{p}), \\ E^*(\mathbf{x}) &= E(\mathbf{x}+\mathbf{p}), \\ S^*(\mathbf{x}) &= S(\mathbf{x}+\mathbf{p}). \end{aligned}$$

Thus, $(\delta\mathbf{u}, \delta E, \delta S) = (\mathbf{u}^* - \mathbf{u}, E^* - E, S^* - S)$ is the difference between two solutions to the same boundary value problem on $\mathscr{B}$. Continuing in this vein $(\delta^M\mathbf{u}, \delta^M E, \delta^M S)$, for $M = 1, 2, 3, \cdots$, all clearly represent the difference between two solutions to the same boundary value problem on $\mathscr{B}$. For $M = 3+n$, however, (8.8) holds, and, uniqueness Theorem 5.1 informs us that $\delta^{3+n}E$ vanishes on $\mathscr{B}$. But this is equivalent to $\delta^{2+n}E$ being periodic on $\mathscr{B}$.

Assumption ii) now becomes important. It clearly implies that for each $\mathbf{x}$ in $\mathscr{B}$

$$\delta^{2+n}E(\mathbf{x}+k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \text{ on } \mathbb{N},$$

which is compatible with $\delta^{2+n}E$ being periodic on $\mathscr{B}$ only if $\delta^{2+n}E$ vanishes on $\mathscr{B}$. Thus, $\delta^{1+n}E$ must be periodic on $\mathscr{B}$ and by assumption ii) satisfies

$$\delta^{1+n}E(\mathbf{x}+k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \text{ on } \mathbb{N}$$

and, hence, $\delta^{1+n}E$ also vanishes on $\mathscr{B}$. Successively repeating this argument n more times finally leads to the conclusion that δE must vanish on $\mathscr{B}$. Thus E is periodic on $\mathscr{B}$. □

THEOREM 8.2. *Let $\mathscr{B}$ be a periodic exterior body with positive definite elasticity. Suppose $(\mathbf{u}, E, S)$ is the difference between two solutions to the same slightly periodic boundary value problem on $\mathscr{B}$ which satisfies both of the following:*

i) *For some n*

$$\delta E(\mathbf{x}) = O(r^n) \quad \text{as } r \to \infty.$$

ii) *Given any $R > 0$ there is a constant, a_R, such that if $r(\mathbf{x}) < R$ then*

$$|\delta^3\mathbf{u}(\mathbf{x})| \leqq a_R.$$

If, in addition, $(\mathbf{u}, E, S)$ satisfies any of the following conditions, then E and S vanish on $\mathscr{B}$ and **u** *is a rigid displacement:*

1) *For each* **x** *in $\mathscr{B}$*

$$S(\mathbf{x}+k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \quad \text{on } \mathbb{N}.$$

2) *For each* **x** *in $\mathscr{B}$*

$$\nabla S(\mathbf{x}+k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \text{ on } \mathbb{N}$$

and

$$S(\mathbf{x}) = o(1) \quad \text{as } r \to \infty.$$

3) *For each* **x** *in $\mathscr{B}$*

$$\nabla S(\mathbf{x}+k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \text{ on } \mathbb{N}$$

and there exist three points on $\mathscr{S}_2$ (the surface on which **s** *is prescribed) such that the normals to $\mathscr{S}_2$ at those points span $\mathbb{R}^3$.*

4) *For some nonnegative integer, n, and each* $\mathbf{x}$ *in* $\mathscr{B}$.

$$\nabla^n \mathbf{u}(\mathbf{x}+k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \text{ on } \mathbb{N}$$

and $\mathscr{S}_1$ *(the surface on which* $\mathbf{u}$ *is prescribed) is nondegenerate, that is, the projection of* $\mathscr{S}_1$ *onto the* X_2*-*X_3 *plane has nontrivial interior.*

Proof. Reviewing the proof of the previous theorem, it is clear that (8.2) can be replaced with the stronger assumption of condition 1 from this theorem. Thus condition 1 not only implies that $S(\mathbf{x}+k\mathbf{p})$ vanishes for each $\mathbf{x}$ in $\mathscr{B}$ as $k \to \infty$, but that S is periodic. This is impossible unless E and S vanish.

By the relationship between δ and ∇ it should be clear that

$$\nabla S(\mathbf{x}+k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \text{ on } \mathbb{N} \tag{8.9}$$

also implies (8.2). By Theorem 8.1, S and, thus, ∇S are periodic. But, of course, ∇S cannot be periodic and satisfy (8.9) without vanishing on $\mathscr{B}$. This means that S is then given by some fixed symmetric tensor S^0. If S satisfies condition 2 of Theorem 8.2, then, since S must vanish "at $r=\infty$", S^0 must be the zero tensor. If S satisfies condition 3 of Theorem 8.2, then, since $(\mathbf{u}, E, S)$ satisfies the null boundary value problem, $S^0\mathbf{n} = \mathbf{0}$ for a set of $\mathbf{n}$'s which span all of $\mathbb{R}^3$. This also forces S^0 to be the zero tensor.

By very similar arguments it is easily seen that if

$$\nabla^n \mathbf{u}(\mathbf{x}+k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \quad \text{on } \mathbb{N}$$

for each $\mathbf{x}$ in $\mathscr{B}$, then $\nabla^n \mathbf{u}$ vanishes on $\mathscr{B}$. Hence $\mathbf{u}$ is given by a polynomial of degree n or less, say

$$\mathbf{u}(x,y,z) = \sum_{j=0}^{n} x^j \mathbf{P}_j(y,z)$$

where each $\mathbf{P}_j$ denotes a (vector valued) polynomial in two variables. But $(\mathbf{u}, E, S)$ is the difference between two solutions to the same slightly periodic boundary value problem, and, so, for each $\mathbf{x}^0 = (x^0, y^0, z^0)$ in $\mathscr{S}_1$ and each integer, k,

$$\mathbf{0} = \mathbf{u}(\mathbf{x}^0) = \mathbf{u}(\mathbf{x}^0 + k\mathbf{p}) = \sum_{j=0}^{n} (x^0 + kp)^j \mathbf{P}_j(y^0, z^0).$$

Thus, $\mathbf{u}(\mathbf{x}^0 + \alpha\mathbf{p})$ is a polynomial in α with infinitely many zeros, $\alpha = 0, \pm 1, \pm 2, \cdots$, which is possible only if each $\mathbf{P}_j(y^0, z^0)$ vanishes whenever (y^0, z^0) is in the projection of $\mathscr{S}_1$ on the $X_2 - X_3$ axis. The additional assumption that this projection be nondegenerate then forces each polynomial $\mathbf{P}_j$, to vanish on an open subset of $\mathbb{R}^2$, which as is well known, forces each $\mathbf{P}_j$ to vanish throughout $\mathbb{R}^2$. Consequently, $\mathbf{u}$ (and, hence, E and S) must vanish throughout $\mathscr{B}$. □

THEOREM 8.3. *Let* $\mathscr{B}$ *be a periodic exterior body with positive definite elasticity. Suppose* $(\mathbf{u}, E, S)$ *is a solution to a periodic boundary value problem such that, on any given* $\mathscr{P}$

$$\mathbf{u}(\mathbf{x}+\mathbf{p}) - \mathbf{u}(\mathbf{x}) = O(1) \quad \textit{as } |\mathbf{x}| \to \infty \textit{ on } \mathscr{P} \tag{8.10}$$

and, for some nonnegative integer, n, either

$$\delta^n \mathbf{u}(\mathbf{x}) = O\left(|\mathbf{x}|^{1/2}\right) \quad \textit{as } |\mathbf{x}| \to \infty \textit{ on } \mathscr{B} \tag{8.11}$$

or

$$(8.12)\qquad \delta^n \nabla E = O\left(|\mathbf{x}|^{1/2}\right) \quad \textit{as } |\mathbf{x}| \to \infty \textit{ on } \mathscr{B}.$$

If, in addition, either of the following holds, then E is periodic with $\overline{W}=0$.

1) $\mathscr{S}_1$, *the surface on which* **u** *is prescribed, is nontrivial.*

2) *The problem is a traction problem and* $(\mathbf{u}, E, S)$ *satisfies any one of the following*:

i) $\delta u_1(\mathbf{x}+\mathbf{p}) - \delta \mathrm{u}_1(\mathbf{x}) = o(1)$ *as* $|\mathbf{x}| \to \infty$ *on* $\mathscr{P}$.

ii) $E_{11}(\mathbf{x}+\mathbf{p}) - E_{11}(\mathbf{x}) = o(1)$ *as* $|\mathbf{x}| \to \infty$ *on* $\mathscr{P}$.

iii) *There are an* $R_0 > 0$ *and a nonincreasing function*, $f(r)$, *which approaches zero as r approaches infinity, such that*

$$|E_{11}(\mathbf{x})| \leqq [r + 1 + |x|] f(r)$$

for all **x** *in* $\mathscr{B}$ *with* $r(\mathbf{x}) = r \geqq R_0$ *and* $x = \mathbf{x}\cdot\mathbf{e}^1$.

iv) *There are an* $R_0 > 0$ *and a nonincreasing function*, $f(r)$, *which approaches zero as r approaches infinity, such that*

$$|u_1(\mathbf{x})| \leqq [r^2 + (1 + |x|) r] f(r)$$

for all **x** *in* $\mathscr{B}$ *with* $r(\mathbf{x}) = r \geqq R_0$ *and* $x = \mathbf{x}\cdot\mathbf{e}^1$.

Sketch of proof. First it should be observed that several of the above assumptions are implied by their alternatives. For example, if 2iv) holds then the second part of Theorem 4.3 (with $\gamma = 1$, $\beta = 1$) clearly implies that for some other nonincreasing function, $g(r)$, which also approaches zero as r approaches infinity,

$$(8.13)\qquad |E_{11}(\mathbf{x})| = |u_{1,1}(\mathbf{x})| \leqq [r + (1 + |x|)]\, g(r)$$

for all **x** in $\mathscr{B}$ with $r = r(\mathbf{x}) > R_0$. A second application of Theorem 4.3 (with $\gamma = 1$ and $\beta = 0$) and the relationship between δ and ∇ shows the existence of another nonincreasing function, $h(r)$, which also approaches zero as r approaches infinity, such that

$$|\delta E_{11}(\mathbf{x})| \leqq [r^0 + (1 + |x|) r^{-1}]\, h(r)$$

for all **x** in $\mathscr{B}$ with $r = r(\mathbf{x}) > R_0$. Recalling the definition of $\mathscr{P}$, we see that this implies that

$$\delta E_{11}(\mathbf{x}) = o(1) \quad \text{as } |\mathbf{x}| \to \infty \text{ on } \mathscr{P}.$$

This, in turn (using the relationship between δ and ∇ and the fact that $E_{11} = u_{1,1}$) implies that

$$\delta^2 u_1(\mathbf{x}) = o(1) \quad \text{as } |\mathbf{x}| \to \infty \text{ on } \mathscr{P}.$$

Thus, in the alternatives listed under condition 2, we see that

$$\text{iv)} \Rightarrow \text{iii)} \Rightarrow \text{ii)} \Rightarrow \text{i)}.$$

In addition, using arguments demonstrated in the proof of Theorem 8.1, it can be shown that (8.12) implies (8.11). Hence for the remainder of this proof, it may be assumed that (8.11) and either assumption 1) or 2i) holds.

Now, by arguments similar to those above and those given in the proof of Theorem 8.1, it is apparent that by (8.11) there exist positive constants M, A, and R_0

such that for all $\mathbf{x}$ in $\mathscr{B}$

$$|\delta^{n+1}\mathbf{u}(\mathbf{x})| \leqq M\left[r^{-1/2} + (1+|x|)^{1/2} r^{-1}\right],$$

and, if $r(\mathbf{x}) < R_0$,

$$|\delta^{n+1}\mathbf{u}(\mathbf{x})| \leqq A(1+|x|)^{1/2}.$$

It is not hard to verify that on a periodic exterior body the above bounds imply that

$$\int_{\mathscr{B}\cap\partial\mathscr{D}_R} |\delta^{n+1}\mathbf{u}|^2 da = O(R) \quad \text{as } R \to \infty$$

and so by the general uniqueness theorem $\delta^{n+1}E$ vanishes on $\mathscr{B}$. Thus $\delta^n E$ is periodic.

Let $(\mathbf{u}^*, E^*, S^*)$ denote $(\delta^n\mathbf{u}, \delta^n E, \delta^n S)$ and let $\overline{W}^*$, κ^*, and $\bar{\mathbf{u}}^*$ be the skew tensor, constant, and vector such that

$$\delta\mathbf{u}^*(\mathbf{x}) = \mathbf{u}^*(\mathbf{x}+\mathbf{p}) - \mathbf{u}^*(\mathbf{x}) = \overline{W}^*\mathbf{x} + \kappa^*\mathbf{p} + \bar{\mathbf{u}}^*.$$

It should be clear that the boundedness of $\delta\mathbf{u}$ on $\mathscr{P}$ assumed in (8.10) implies that $\delta\mathbf{u}^*$ is bounded on $\mathscr{P}$. This in turn, forces $\overline{W}^*$ to be the zero tensor. Now, if $\mathscr{S}_1$ is nontrivial, the periodicity of the original problem forces $\delta\mathbf{u}^*$, and, thus, κ^* and $\bar{\mathbf{u}}^*$, to vanish on $\mathscr{S}_1$. Thus, $\mathbf{u}^*$ is periodic. Uniqueness Theorem 6.1 can now be applied, showing that $\delta^n E = E^*$ vanishes on $\mathscr{B}$.

Suppose, now, that 2iv) and not 1) holds. Obviously 2iv) also forces κ^* to zero. Since a traction problem is being considered and $\bar{\mathbf{u}}^*$ arises from a rigid displacement, it may be assumed that $\bar{\mathbf{u}}^* = \mathbf{0}$. Again, uniqueness Theorem 6.1 applies showing that $\delta^n E = E^*$ vanishes on $\mathscr{B}$. Thus, $\delta^n E$ is periodic.

The proof of the theorem is completed in the obvious manner, that is, the arguments from the previous two paragraphs are repeated successively for $\delta^{n-1}E$, $\delta^{n-2}E$, etc. □

Two corollaries to Theorem 8.3 will be given. The first is the expected "uniqueness theorem". The second is a repeat of Theorem 8.3 combined with the observation (via Theorem 4.3) that both (8.10) and assumption 2 are satisfied whenever $\mathbf{u}$ satisfies

$$\mathbf{u}(\mathbf{x}) = O(|\mathbf{x}|) \quad \text{as } |\mathbf{x}| \to \infty \text{ on } \mathscr{B}.$$

Proofs of these two corollaries will be left to the reader.

THEOREM 8.4. *Let $\mathscr{B}$ be a periodic exterior body with positive definite elasticity. Suppose $(\mathbf{u}, E, S)$ is the difference between two solutions to the same slightly periodic boundary value problem such that, on any given $\mathscr{P}$*

$$\mathbf{u}(\mathbf{x}) = O(1) \quad \textit{as } |\mathbf{x}| \to \infty \textit{ on } \mathscr{P}$$

and, for some nonnegative integer, n, either

$$\delta^n\mathbf{u}(\mathbf{x}) = O\left(|\mathbf{x}|^{1/2}\right) \quad \textit{as } |\mathbf{x}| \to \infty \textit{ on } \mathscr{B}$$

or

$$\delta^n \nabla E(\mathbf{x}) = O\left(|\mathbf{x}|^{1/2}\right) \quad \textit{as } |\mathbf{x}| \to \infty \textit{ on } \mathscr{B}.$$

If, either

1) $\mathcal{S}_1$, *the surface on which the displacement is prescribed, is nontrivial or*

2) *the problem is a traction problem and*

$$E_{11}(\mathbf{x}) = o(1) \quad as\ |\mathbf{x}| \to \infty\ on\ \mathcal{P},$$

then E and S vanish on $\mathcal{B}$ and **u** *is a rigid displacement.*

THEOREM 8.5. *Let $\mathcal{B}$ be a periodic exterior body with positive definite elasticity. Suppose $(\mathbf{u}, E, S)$ is a solution to a periodic boundary value problem on $\mathcal{B}$ such that*

$$\mathbf{u}(\mathbf{x}) = O(|\mathbf{x}|) \quad as\ |\mathbf{x}| \to \infty\ on\ \mathcal{B}$$

and, for some nonnegative integer, n, either

$$\delta^n \mathbf{u}(\mathbf{x}) = O\left(|\mathbf{x}|^{1/2}\right) \quad as\ |\mathbf{x}| \to \infty\ on\ \mathcal{B}$$

or

$$\delta^n \nabla E(\mathbf{x}) = O\left(|\mathbf{x}|^{1/2}\right) \quad as\ |\mathbf{x}| \to \infty\ on\ \mathcal{B}.$$

Then E is periodic with $\overline{W} = 0$.

9. Applications involving periodic fractional spaces. The proofs of the results discussed in the previous section can be fairly easily adapted to cover analogous problems on periodic fractional spaces. The only serious modification results from the fact that on fractional spaces $r(\mathbf{x})$ cannot, in general, be considered as equivalent (for large $r(\mathbf{x})$) to the distance from **x** to the boundary of $\mathcal{B}$. As a result, to insure that uniqueness Theorem 5.1 can be applied to some $\delta^m \mathbf{u}$, it becomes expedient to replace to bounds involving $(1+|x|)^{1/2}$ with bounds involving $(1+|x|)^0$. This is not always necessary, and, following Theorem 9.5 will be a brief discussion on relaxing these particular bounds.

Throughout this section it is to be understood that ρ and $\rho(\mathbf{x})$ both denote the distance from any given **x** in $\mathcal{B}$ to the boundary of $\mathcal{B}$, and, as in the previous section, x denotes $\mathbf{x} \cdot \mathbf{e}^1$.

The first five theorems of this section are the direct analogues of the corresponding theorems from the previous section. No further discussion of their proofs will be made.

THEOREM 9.1. *Let $\mathcal{B}$ be a periodic fractional space with positive definite elasticity. Suppose $(\mathbf{u}, E, S)$ is a solution to a periodic boundary value problem on $\mathcal{B}$ which satisfies all of the following:*

i) *For some n*

$$\delta E(\mathbf{x}) = O(\rho^n) \quad as\ \rho(\mathbf{x}) \to \infty.$$

ii) *For each* **x** *in $\mathcal{B}$*

$$\delta E(\mathbf{x} + k\mathbf{p}) = o(1) \quad as\ k \to \infty\ on\ \mathbb{N}.$$

iii) *Given any $R > 0$, there is a constant, A_R, such that*

$$|\delta^3 \mathbf{u}(\mathbf{x})| \leqq A_R$$

whenever **x** *is in $\mathcal{B}$ and $\rho(\mathbf{x}) \leqq R$.*

Then E is periodic on $\mathcal{B}$.

THEOREM 9.2. *Let $\mathscr{B}$ be a periodic fractional space with positive definite elasticity. Suppose* $(\mathbf{u}, E, S)$ *is the difference between two solutions to the same slightly periodic boundary value problem on $\mathscr{B}$ which satisfies both of the following*:

i) *For some n*

$$\delta E(\mathbf{x}) = O(\rho^n) \quad \text{as } \rho(\mathbf{x}) \to \infty.$$

ii) *Given any* $R > 0$, *there is a constant*, A_R, *such that*

$$|\delta^3 \mathbf{u}(\mathbf{x})| \leqq A_R$$

whenever $\mathbf{x}$ *is a point in $\mathscr{B}$ with* $\rho(\mathbf{x}) \leqq R$.

If, in addition, $(\mathbf{u}, E, S)$ *satisfies any of the following conditions, then E and S vanish on $\mathscr{B}$ and* $\mathbf{u}$ *is a rigid displacement.*

1. *For each* $\mathbf{x}$ *in* $\mathscr{B}$

$$S(\mathbf{x} + k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \text{ on } \mathbb{N}.$$

2. *For each* $\mathbf{x}$ *in* $\mathscr{B}$

$$\nabla S(\mathbf{x} + k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \text{ on } \mathbb{N}$$

and

$$S(\mathbf{x}) = o(1) \quad \text{as } \rho(\mathbf{x}) \to \infty.$$

3. *For each* $\mathbf{x}$ *in* $\mathscr{B}$

$$\nabla S(\mathbf{x} + k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \text{ on } \mathbb{N}$$

and there are three points on $\mathscr{S}_2$, *the surface on which* $\mathbf{s}$ *is prescribed, such that the normals to* $\mathscr{S}_2$ *at these points span all of* $\mathbb{R}^3$.

4. *For some nonnegative integer, n, and each* $\mathbf{x}$ *in* $\mathscr{B}$

$$\nabla^n \mathbf{u}(\mathbf{x} + k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \quad \text{on } \mathbb{N}$$

and $\mathscr{S}_2$, *the surface on which* $\mathbf{u}$ *is prescribed, is nondegenerate* (*i.e., the projection of* $\mathscr{S}_2$ *onto the* X_2-X_3 *plane contains a relatively open subset of the plane*).

THEOREM 9.3. *Let $\mathscr{B}$ be a periodic fractional space with positive definite elasticity. Suppose* $(\mathbf{u}, E, S)$ *is the solution to a periodic boundary value problem on $\mathscr{B}$ such that, on any given* $\mathscr{P}$,

$$\mathbf{u}(\mathbf{x} + \mathbf{p}) - \mathbf{u}(\mathbf{x}) = O(1) \quad \text{as } |\mathbf{x}| \to \infty \text{ on } \mathscr{P}$$

and, for some nonnegative integer, n, either

$$\delta^n \mathbf{u}(\mathbf{x}) = O(1) \text{ as } |\mathbf{x}| \to \infty \text{ on } \mathscr{B}$$

or

$$\delta^n \nabla E(\mathbf{x}) = O(1) \quad \text{as } |\mathbf{x}| \to \infty \text{ on } \mathscr{B}.$$

Assume, also, that either

1. $\mathscr{S}_1$, *the surface on which* $\mathbf{u}$ *is prescribed, is nontrivial, or*
2. *the problem is a traction problem with* $(\mathbf{u}, E, S)$ *satisfying one of the following*:

i) $\delta u_1(\mathbf{x} + \mathbf{p}) - \delta u_1(\mathbf{x}) = o(1)$ *as* $|\mathbf{x}| \to \infty$ *on* $\mathscr{P}$;

ii) $E_{11}(\mathbf{x} + \mathbf{p}) - E_{11}(\mathbf{x}) = o(1)$ *as* $|\mathbf{x}| \to \infty$ *on* $\mathscr{P}$;

iii) *there is a nonincreasing function, $f(r)$, which approaches zero as r approaches infinity, such that*

$$|E_{11}(\mathbf{x})| \leqq [\rho + 1 + |x|] f(\rho)$$

for all $\mathbf{x}$ *in* $\mathscr{B}$;

iv) *there is a nonincreasing function, $f(r)$, which approaches zero as r approaches infinity, such that*

$$|u_1(\mathbf{x})| \leqq [\rho^2 + (1 + |x|)\rho] f(\rho)$$

for all $\mathbf{x}$ *in* $\mathscr{B}$.

Then E is periodic on $\mathscr{B}$ with $\overline{W} = 0$.

THEOREM 9.4. *Let $\mathscr{B}$ be a periodic fractional space with positive definite elasticity. Suppose $(\mathbf{u}, E, S)$ is the difference between two solutions to the same slightly periodic boundary value problem on $\mathscr{B}$ such that, on any given $\mathscr{P}$,*

$$\mathbf{u}(\mathbf{x}) = O(1) \quad \text{as } |\mathbf{x}| \to \infty \text{ on } \mathscr{P}$$

and, for some nonnegative integer, n, either

$$\delta^n \mathbf{u}(\mathbf{x}) = O(1) \quad \text{as } |\mathbf{x}| \to \infty \text{ on } \mathscr{B}$$

or

$$\delta^n \nabla E(\mathbf{x}) = O(1) \quad \text{as } |\mathbf{x}| \to \infty \text{ on } \mathscr{B}.$$

If either

1) $\mathscr{S}_1$, *the surface on which the displacement is prescribed, is nontrivial or*

2) *the problem is a traction problem and*

$$E_{11}(\mathbf{x}) = o(1) \quad \text{as } |\mathbf{x}| \to \infty \text{ on } \mathscr{P}$$

then E and S vanish on $\mathscr{B}$ and $\mathbf{u}$ *is a rigid displacement.*

THEOREM 9.5. *Let $\mathscr{B}$ be a periodic fractional space with positive definite elasticity. Suppose $(\mathbf{u}, E, S)$ is a solution to a periodic boundary value problem on $\mathscr{B}$ such that*

$$\mathbf{u}(\mathbf{x}) = O(|\mathbf{x}|) \quad \text{as } |\mathbf{x}| \to \infty \text{ on } \mathscr{B}$$

and, for some nonnegative integer, n, either

$$\delta^n \mathbf{u}(\mathbf{x}) = O(1) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{B}$$

or

$$\delta^n \nabla E(\mathbf{x}) = O(1) \quad \text{as } |\mathbf{x}| \to \infty \text{ on } \mathscr{B}.$$

Then E is periodic with $\overline{W} = 0$.

As pointed out at the beginning of this section, the theorems in this section are slightly weaker than the analogous theorems in the previous section. This is, at least partly, due to the fact that the bounds developed in §4 behave poorly near the boundaries of the domains. Duffin [1], however, has shown that solutions to certain problems in elastostatics can be continued across planar portions of the boundary. The extension is also an elastic state but is now defined on a larger body, $\mathscr{B}^*$, which contains the original body. For example, if $\mathscr{B}$ is a half space, $\mathscr{B}^*$ may be taken to be the entire space. Use of these reflexion principles of Duffin have led to several uniqueness results for problems involving half spaces (see Knops and Payne [7, Chap. 6]). They also lead to the following result.

THEOREM 9.6. *Suppose that $\mathscr{B}$ is a periodic fractional space with positive definite elasticity. Suppose, also, that outside of some cylinder of finite radius centered about the X_1-axis, the boundary of $\mathscr{B}$ consists of a pair of half planes both parallel to the X_1-axis. Let such a body be termed a "periodic perfect fractional space". Then all of the theorems in* §8 *remain true if the phrase "periodic exterior body" is replaced by the phrase "periodic perfect fractional space".*

The proof of this theorem shall be left to the dedicated reader.

10. The displacement problem. It has been shown (see Howell [3] and [5], respectively) that Theorems 5.1 and 6.1 remain true if the problem being considered is a displacement problem on a homogeneous, isotropic body with strongly elliptic elasticity. Thus, the results of §§8 and 9 can easily be extended to cover periodic and slightly periodic displacement problems on homogeneous, isotropic bodies with strongly elliptic elasticity.

11. Comparison with half-space results. To a certain extent the uniqueness results in §9 can be viewed as extensions of the known results for half-space traction and displacement problems (see, Knops and Payne [7, Chap. 6]). In §9 the bodies (periodic fractional spaces) are more general and the class of problems considered (slightly periodic general boundary value problems) are, also, more general. However, via Duffin's reflexion principle [1], half-space traction and displacement problems can be treated as if the functions were defined on all of space. One can then prove the desired uniqueness using properties of biharmonic functions defined on all of space (for example, one could use Corollary 4.4 of this paper) with little regard as to whether or not the elasticity field is positive definite or strongly elliptic.

The asymptotic behavior assumed in the half-space problems, however, should be compared to the behavior assumed in Theorem 9.2 under either assumption 3 or 4. By assuming that the body was not a perfect half-space it was possible to show uniqueness under weaker assumptions concerning the asymptotic behavior of the elastic state. Interestingly, the sort of asymptotic behavior assumed in this theorem is definitely not sufficient to insure uniqueness in half-space problems. For the traction problem on the half-space $\{\mathbf{x}\colon x_2>0\}$ an appropriate counterexample would be any elastic state $(\mathbf{u}, E, S)$ in which S is constant and

$$S_{i2}=0 \quad \text{for } i=1,2,3.$$

For the displacement problem on the same half-space one can use

$$\mathbf{u}(x,y,z)=y\mathbf{v}^0$$

as a counterexample where $\mathbf{v}^0$ is any fixed nonzero vector.

Acknowledgment. A major portion of this paper was prepared under the support of an Incentive Fund Research Grant from the University of Alabama in Huntsville. The author is grateful for this assistance.

REFERENCES

[1] R. J. DUFFIN, *Analytic continuation in elasticity*, J. Rational Mech. Anal., 5 (1956), pp. 939–949.

[2] M. E. GURTIN AND E. STERNBERG, *Theorems in linear elastostatics for exterior domains*, Arch. Rational Mech. Anal., 8 (1961), pp. 99–119.

[3] K. B. Howell, *Uniqueness in linear elastostatics for problems involving unbounded bodies*, J. Elasticity, 10 (1980), pp. 407–427.

[4] ______, *Periodic and "slightly" periodic boundary value problems in elastostatics on bodies bounded in all but one direction*, J. Elasticity, 11 (1981), pp. 293–316.

[5] ______, *Periodic and "slightly" periodic boundary value problems in elastostatic on bodies unbounded in several directions*, Int. J. Engrg. Sci., 20 (1982), pp. 455–481.

[6] ______, *Asymptotic behavior of periodic strain states*, this Journal, 16 (1985), to appear.

[7] R. J. Knops and L. E. Payne, *Uniqueness Theorems in Linear Elasticity*, Springer-Verlag, Berlin, 1971.

[8] N. I. Muskhelishvili, *Some Basic Problems of the Mathematical Theory of Elasticity*, J. R. M. Radok, trans., Noordhoff, Groningen, 1953.

[9] M. Nicolesco, *Les functions polyharmoniques*, Actualités Sci. Ind., No. 331, 1936.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

014

COMPLETENESS OF THE EIGENFUNCTIONS OF THE EQUILATERAL TRIANGLE*

MARK A. PINSKY†

Abstract. It is proved that the eigenfunctions obtained by multiple reflection are complete: the generalized Fourier series of a smooth function coverges absolutely to the given function.

Recently there has been a revival of interest in the detailed structure of the eigenvalues and eigenfunctions of the equilateral triangle [2], [3]. While the existence of a complete set of eigenfunctions may be deduced from general theorems on self-adjoint elliptic operators, it does not follow that a concrete set of eigenfunctions is in fact complete. The purpose of this note is to prove that the eigenfunctions obtained in [3] by multiple reflection form a complete set. This problem is more difficult than the corresponding problem for a square because in the present case the associated parallelogram covers the triangle eighteen times, in contrast to the fourfold covering in the case of a square.

Let D be the equilateral triangle

$$D=\{(x,y): 0<y<x\sqrt{3}, 0<y<\sqrt{3}(1-x)\}.$$

Let Z^2 be the integer lattice

$$Z^2=\{(m,n): m=0, \pm 1, \pm 2, \cdots; n=0, \pm 1, \pm 2, \cdots\},$$

and let G be the group of transformations of Z^2 generated by the operations S_1, S_2 where

$$S_1: (m,n)\to(m,m-n),$$
$$S_2: (m,n)\to(n-m,n).$$

This group has six elements, including the identity and the transformations S_1S_2, S_2S_1 and $S_1S_2S_1$ $(=S_2S_1S_2)$. Under the action of this group the lattice Z^2 splits into orbits which are equivalence classes of lattice points which are transformed into one another by the group operations. Every orbit is of the form

$$\mathscr{S}=\{(m,n),(m,m-n),(-n,m-n),(-n,-m),(n-m,-m),(n-m,n)\}.$$

The union of these orbits is the entire lattice Z^2.

In general an orbit may have six elements or fewer. For the construction of eigenfunctions we need a certain class of orbits which are defined as follows.

DEFINITION. $\mathscr{S}$ is a *special orbit* if the following conditions hold:

(i) $m\neq 2n$ for all $(m,n)\in\mathscr{S}$;

(ii) $n\neq 2m$ for all $(m,n)\in\mathscr{S}$;

(iii) $m+n$ is a multiple of 3 for all $(m,n)\in\mathscr{S}$.

*Received by the editors October 26, 1983, and in revised form January 25, 1984. This research was supported by the National Science Foundation.

† Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

We denote these orbits by $(\mathcal{S}_k)$. In [3] it was shown that to each special orbit $\mathcal{S}_k$ there corresponds an eigenfunction with zero boundary conditions, i.e. a function ψ satisfying

$$\begin{aligned} &\Delta\psi+\lambda\psi=0 && \text{in } D,\\ &\psi=0 && \text{on the boundary of } D, \end{aligned}$$

where the eigenvalue is given by $\lambda=(16\pi^2/27)(m^2+n^2-mn)$. The eigenfunction is given by the formula

$$(1) \qquad \psi_k=\sum_{(m,n)\in\mathcal{S}_k}\pm\exp\left(\frac{2\pi i}{3}\right)\left(mx+\frac{(2n-m)y}{\sqrt{3}}\right),$$

where the $\pm$ signs are chosen according to the following ordering of elements in the orbit

$$(2) \qquad \begin{tikzcd} & (m,n) \arrow[dr, "S_1"] & \\ (n-m,n) \arrow[ur, "S_2"] & & (m,m-n) \arrow[d, "S_2"] \\ (n-m,-m) \arrow[u, "S_1"] & & (-n,m-n) \arrow[dl, "S_1"] \\ & (-n,-m) \arrow[ul, "S_2"] & \end{tikzcd}$$

Each transition induces a change of sign in the coefficient of the exponential in (1). We can now state the result.

THEOREM. *Let f be a smooth real-valued function with compact support in the equilateral triangle D. Then*

$$(3) \qquad f=\sum_{k=1}^{\infty}a_k\psi_k$$

for suitably chosen coefficients (a_k).

Proof. Let $\tilde{D}$ be the parallelogram and R_1, R_2, R_3 the reflection operators with translation operators T_1, T_2 defined as follows:

$$\begin{aligned} &\tilde{D}=\{(x,y):0<y<3\sqrt{3}/2, y/\sqrt{3}<x<3+y/\sqrt{3}\},\\ &R_1:\ (x,y)\to(x,-y),\\ &R_2:\ (x,y)\to\left(-\tfrac{1}{2}x+\tfrac{1}{2}y\sqrt{3}, \tfrac{1}{2}x\sqrt{3}+\tfrac{1}{2}y\right),\\ &R_3:\ (x,y)\to\left(\tfrac{1}{2}3-\tfrac{1}{2}x-\tfrac{1}{2}y\sqrt{3}, \tfrac{1}{2}y+\tfrac{1}{2}\sqrt{3}-\tfrac{1}{2}x\sqrt{3}\right),\\ &T_1:\ (x,y)\to(x+3,y),\\ &T_2:\ (x,y)\to\left(x+\tfrac{1}{2}3, y+\tfrac{1}{2}3\sqrt{3}\right). \end{aligned}$$

The parallelogram $\tilde{D}$ covers the triangle D eighteen times as shown in Fig. 1. Given f on D which vanishes on the boundary, there is a unique $\tilde{f}$ on $\mathbb{R}^2$ such that

$$\begin{aligned} &T_1\tilde{f}=\tilde{f}, \quad T_2\tilde{f}=\tilde{f},\\ &R_1\tilde{f}=-\tilde{f}, \quad R_2\tilde{f}=-\tilde{f}, \quad R_3\tilde{f}=-\tilde{f},\\ &\tilde{f}|_D=f. \end{aligned}$$

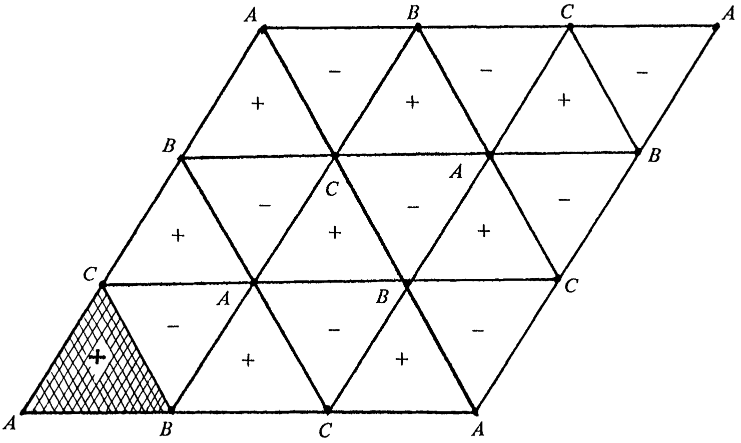


FIG. 1. *The basic triangle (shaded) and the covering parallelogram. Corresponding vertices are labeled A, B, C. A function is reflected according to the ± signs as indicated.*

The function $\tilde{f}$ may be obtained by multiple reflection and periodic extension as indicated in Fig. 1. Now $\tilde{f}$ is also smooth and may be written as an absolutely convergent trigonometric series

$$\tilde{f}=\sum_{(m,n)\in Z^2} A_{mn}\exp\left(\frac{2\pi i}{3}\right)\left(mx+\frac{(2n-m)y}{\sqrt{3}}\right). \tag{4}$$

To see this, note that the transformation $x'=x-y/\sqrt{3}$, $y'=2y/\sqrt{3}$ carries the parallelogram $\tilde{D}$ to the square $0<x'<3$, $0<y'<3$ for which the complete set of eigenfunctions is known to be [4, p. 300]

$$\begin{aligned}\exp\left(\frac{2\pi i}{3}\right)(mx'+ny')&=\exp\left(\frac{2\pi i}{3}\right)\left(\frac{m(x-y)}{\sqrt{3}}+n\frac{2y}{\sqrt{3}}\right)\\&=\exp\left(\frac{2\pi i}{3}\right)\left(mx+\frac{(2n-m)y}{\sqrt{3}}\right).\end{aligned}$$

Applying the reflection operators and following the computation of [3, p. 822], we have

$$R_1f=\sum_{(m,n)\in Z^2} A_{m,m-n}\exp\left(\frac{2\pi i}{3}\right)\left(mx+\frac{(2n-m)y}{\sqrt{3}}\right),$$

$$R_2f=\sum_{(m,n)\in Z^2} A_{n-m,n}\exp\left(\frac{2\pi i}{3}\right)\left(mx+\frac{(2n-m)y}{\sqrt{3}}\right),$$

$$R_3f=\sum_{(m,n)\in Z^2} A_{-n,-m}\exp\left(-\frac{2\pi i}{3}\right)(m+n)\exp\left(\frac{2\pi i}{3}\right)\left(mx+\frac{(2n-m)y}{\sqrt{3}}\right).$$

By the uniqueness of the representation (4) we have

$$R_1\tilde{f} = -\tilde{f} \Rightarrow A_{m,m-n} = -A_{mn}, \tag{5}$$

$$R_2\tilde{f} = -\tilde{f} \Rightarrow A_{n-m,n} = -A_{mn}, \tag{6}$$

$$R_3\tilde{f} = -\tilde{f} \Rightarrow A_{-n,-m} = -A_{mn}\exp\left(\frac{2\pi i}{3}\right)(m+n). \tag{7}$$

We now rewrite the absolutely convergent sum (4) as a sum over the orbits. Within each orbit A_{mn} has the same absolute value, from (5) and (6). Furthermore (7) shows that A_{mn} is zero unless $m+n$ is a multiple of three. We claim further that $m \neq 2n$, $n \neq 2m$ for each (mn). Indeed if this is not the case, then the orbit must have the form

$$\mathscr{S} = \{(2n,n),(2n,n),(-n,n),(-n,-2n),(-n,-2n),(-n,n)\}.$$

Again using properties (5) and (6) we see that the sum over this orbit is zero. Therefore we conclude that *we may restrict the sum to the special orbits* $\mathscr{S}_k$. The proof is complete.

The above methods may be used to prove completeness of the eigenfunctions of the fundamental domains of general crystallographic groups, as considered by Bérard [1]. The details are entirely similar to the case of the equilateral triangle treated here.

Acknowledgment. I would like to thank Gilbert Strang for a helpful initial conversation.

REFERENCES

[1] P. Bérard, *Spectres et groupes cristallographiques* I: *domaines euclidiens*, Inventiones Mathematica, 58 (1980), pp. 179–199.

[2] S. S. Lee and S. H. Crandall, *The eigenmodes of an equilateral triangle*, in Mélanges Theodore Vogel, Universite Libre de Bruxelles, 1978, pp. 255–268.

[3] M. Pinsky, *Eigenvalues of an equilateral triangle*, this Journal, 11 (1980), pp. 819–827.

[4] A. Zygmund, *Trigonometrical Series*, Cambridge Univ. Press, Cambridge, 1959.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

HYPERGEOMETRIC FUNCTIONS OF SCALAR MATRIX ARGUMENT ARE EXPRESSIBLE IN TERMS OF CLASSICAL HYPERGEOMETRIC FUNCTIONS*

RAMESHWAR D. GUPTA† AND DONALD ST. P. RICHARDS‡

Abstract. It is shown that the hypergeometric function of $m \times m$ scalar matrix argument (cf. Herz, Annals of Math., 61 (1955), pp. 474–523) may be expressed as the Pfaffian of a matrix whose entries are evaluated in terms of classical hypergeometric functions. Applications, in multivariate statistical theory, are made to the distributions of eigenvalues of various random matrices.

AMS-MOS subject classifications (1980). Primary 33A30, 62H10

Key words. hypergeometric function of matrix argument, Pfaffian, distribution theory, random eigenvalues

1. Introduction. The hypergeometric functions of $m \times m$ argument, introduced by Herz (1955), have found widespread applications in multivariate statistical theory (Muirhead (1982)) and analytic number theory (Shimura (1982)). The great importance of these functions has led to series expansions in terms of zonal polynomials (Constantine (1963)), and also to relations with the classical hypergeometric functions (Herz (1955), Muirhead (1975), Koornwinder and Sprinkhuizen-Kuyper (1978), Gupta and Richards (1982)). However, zonal polynomial series generally converge somewhat slowly, while connections with the classical counterparts are only available for $m = 2$.

In this article, we use the results of de Bruijn (1955) to show that Herz's hypergeometric function ${}_2F_1(\alpha,\beta;\gamma;rI_m)$, of $m \times m$ scalar matrix argument, is related to the Pfaffian of a matrix whose entries are expressible in terms of the classical hypergeometric functions. Applications, in multivariate analysis, are made to the distributions of random eigenvalues of two matrix statistics.

2. Preliminaries. Throughout, $R > 0$ (or $0 < R$) will mean that R is a positive definite $m \times m$ symmetric matrix; $R > S$ will mean that $R - S > 0$.

Herz (1955) defines the hypergeometric function ${}_2F_1(\alpha,\beta;\gamma;R)$ through the integral formula

$$(2.1) \qquad {}_2F_1(\alpha,\beta;\gamma;R) = \frac{1}{B_m(\alpha,\gamma-\alpha)} \int_0^I |S|^{\alpha-p} |I-S|^{\gamma-\alpha-p} |I-RS|^{-\beta} \, dS,$$

for complex α,β,γ with $\mathrm{Re}\,\alpha > p-1, \mathrm{Re}\,\beta > p-1, \mathrm{Re}(\gamma-\alpha) > p-1$, and $0 < R < I$. Here, $p = (m+1)/2$, I is the $m \times m$ identity matrix, and with $S = (s_{ij})$, $dS = \prod_{i \leq j}^m ds_{ij}$ is Lebesgue measure. $\int_0^I$ denotes integration over $\{S: 0 < S < I\}$, and the multivariate beta

*Received by the editors October 4, 1983. This work was supported by the Natural Science and Engineering Research Council of Canada under grant A4850.

† Division of Computer Science, Engineering and Mathematics, University of New Brunswick, St. John, New Brunswick, Canada EZL 4L5.

‡ Departments of Statistics and Mathematics, University of North Carolina, Chapel Hill, North Carolina 27514 and Departments of Mathematics and Statistics, University of Wyoming, Laramie, Wyoming 82071.

function $B_m(\cdot,\cdot)$ may be defined as

$$B_m(\alpha,\beta)=\frac{\Gamma_m(\alpha)\Gamma_m(\beta)}{\Gamma_m(\alpha+\beta)},$$

where

$$\Gamma_m(\alpha)=\pi^{m(m-1)/4}\Gamma(\alpha)\Gamma\left(\alpha-\frac{1}{2}\right)\cdots\Gamma\left(\alpha-\frac{1}{2}(m-1)\right),\quad \operatorname{Re}\alpha>p-1,$$

is the multivariate gamma function.

Let μ be a Borel measure on a possibly infinite interval (a,b) of the real line, and let $\phi_1,\phi_2,\cdots,\phi_n$ be members of $L^1(\mu)$. The results of de Bruijn (1955) pertain to the multiple integral

$$\Omega=\int_{a<x_m<\cdots<x_1<b}\cdots\int |(\phi_i(x_j))|\,d\mu(x_1)\cdots d\mu(x_m), \tag{2.2}$$

where $|\cdot|$ denotes the determinant. Introducing the signature function

$$E(x_1,\cdots,x_m)=\prod_{i<j}\operatorname{sgn}(x_j-x_i),\qquad x_1,\cdots,x_m\in(a,b),$$

into the integrand of (2.2), it follows from a symmetry argument that

$$\begin{aligned}\Omega&=\frac{1}{m!}\int_a^b\cdots\int_a^b E(x_1,\cdots,x_m)|(\phi_i(x_j))|\,d\mu(x_1)\cdots d\mu(x_m)\\ &=\int_a^b\cdots\int_a^b E(x_1,\cdots,x_m)\phi_1(x_1)\cdots\phi_m(x_m)\,d\mu(x_1)\cdots d\mu(x_m).\end{aligned}\tag{2.3}$$

Expanding the signature function in the form

$$E(x_1,\cdots,x_m)=\frac{1}{2^n n!}\sum_{j_1=1}^{m}\cdots\sum_{j_m=1}^{m}E(j_1,\cdots,j_m)E(x_{j_1},x_{j_2})\cdots E(x_{j_{2n-1}},x_{j_{2n}}),$$

where $n=[m/2]$, the greatest integer not exceeding $m/2$, and integrating termwise in (2.3), we find that Ω may be written as the Pfaffian of a certain skew-symmetric $m\times m$ matrix $A=(a_{ij})$. We recall (cf. Weyl (1946)) that for any $m\times m$ skew-symmetric matrix $A=(a_{ij})$, the *Pfaffian* of A may be defined by

$$\operatorname{Pf}(A)=\frac{1}{2^n n!}\sum_{j_1=1}^{m}\cdots\sum_{j_m=1}^{m}a_{j_1j_2}\cdots a_{j_{2n-1},j_{2n}}E(j_1,j_2,\cdots,j_m),$$

where $n=[m/2]$. When m is even, a well-known result is that $[\operatorname{Pf}(A)]^2=|A|$; if m is odd, we can use the result that $\operatorname{Pf}(A)=\operatorname{Pf}(A_+)$, where

$$A_+=\begin{pmatrix}0 & a_{12} & \cdot & \cdot & \cdot & a_{1m} & 1\\ a_{21} & 0 & \cdot & \cdot & \cdot & a_{2m} & 1\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ a_{m1} & a_{m2} & \cdot & \cdot & \cdot & 0 & 1\\ -1 & -1 & \cdot & \cdot & \cdot & -1 & 0\end{pmatrix}$$

to again express $[\operatorname{Pf}(A)]^2$ as a determinant.

Consequently $\Omega = \mathrm{Pf}(A)$, with $A = (a_{ij})$ being $m \times m$ and skew-symmetric. If m is even, $m = 2n$, then for $1 \leqq i \neq j \leqq 2n$,

$$(2.4) \qquad a_{ij} = -a_{ji} = \int_a^b \int_a^b \phi_i(x)\phi_j(y)\operatorname{sgn}(y-x)\,d\mu(x)\,d\mu(y).$$

If m is odd, $m = 2n+1$, then in addition to (2.4) holding for $1 \leqq i \neq j \leqq 2n+1$,

$$(2.5) \qquad a_{i,2n+2} = -a_{2n+2,i} = \int_a^b \phi_i(x)\,d\mu(x), \qquad 1 \leqq i \leqq 2n+1,$$

and $a_{2n+2,2n+2} = 0$.

All details behind (2.4) and (2.5) are given by de Bruijn (1955).

3. Hypergeometric functions of scalar matrix argument. In (2.1), we set $R = rI$, $0 < r < 1$, and make the substitutions $S = H\Delta H^{-1}$, where H is $m \times m$ and orthogonal, and $\Delta = \mathrm{diag}(s_1, \cdots, s_m)$ contains the characteristic roots of S. We may assume that $s_1 > s_2 > \cdots > s_m$, since the manifold corresponding to possible equality of any subset of the $\{s_i\}_{i=1}^m$ is of lower dimension, and is therefore of measure zero. The Jacobian of the transformation is given in Constantine (1963) and Muirhead (1982), from which we readily obtain

$$(3.1) \quad B_m(\alpha, \gamma-\alpha)_2F_1(\alpha, \beta; \gamma; rI) = c_m \underset{1 > s_1 > \cdots > s_n > 0}{\int \cdots \int} \prod_{i<j}^{m} (s_i - s_j)\,d\mu(s_1) \cdots d\mu(s_m),$$

where $d\mu(x) = x^{\alpha-p}(1-x)^{\gamma-\alpha-p}(1-rx)^{-\beta}dx$, $0 < x < 1$, and $c_m = \pi^{m^2/2}/\Gamma_m(m/2)$. The well-known Vandermonde formula

$$\prod_{i<j}^{m} (s_i - s_j) = \left|\left(s_j^{m-i}\right)\right|$$

applied to (3.1) shows that (3.1) is of the form (2.2). It then follows from the results outlined in §2 that the integral in (3.1) equals $\mathrm{PF}(A)$, where $A = (a_{ij})$ is $m \times m$ skew-symmetric with

$$(3.2) \qquad a_{ij} = -a_{ji} = \int_0^1 \int_0^1 x^{m-i} y^{m-j} \operatorname{sgn}(y-x)\,d\mu(x)\,d\mu(y),$$

for $1 \leqq i \neq j \leqq m$ if m is even, and with the modification in (2.5) if m is odd. Writing a_{ij} as a difference of two integrals,

$$(3.3) \qquad a_{ij} = \left(\int_0^1 \int_0^y - \int_0^1 \int_y^1\right) x^{m-i} y^{m-j}\,d\mu(x)\,d\mu(y),$$

it is straightforward to verify that the *sum* of the two integrals in (3.3) equals

$$(3.4) \qquad B_1(\alpha+p-i, \gamma-\alpha-p+1) B_1(\alpha+p-j, \gamma-\alpha-p+1) \\ \times {}_2F_1(\beta, \alpha+p-i; \gamma-i+1; r)\ {}_2F_1(\beta, \alpha+p-j; \gamma-j+1; r).$$

Therefore, it suffices to calculate the first integral in (3.3), viz.

$$(3.5) \\ \int_0^1 \int_0^y x^{\alpha+p-i-1} y^{\alpha+p-j-1} (1-x)^{\gamma-\alpha-p} (1-y)^{\gamma-\alpha-p} (1-rx)^{-\beta} (1-ry)^{-\beta}\,dx\,dy.$$

Writing

$$(1-rx)^{-\beta}=\sum_{k=0}^{\infty}\frac{(\beta)_k}{k!}r^k x^k,$$

where $(\beta)_k=\Gamma(k+\beta)/\Gamma(\beta)$, $k=0,1,\cdots$, and integrating termwise shows that the inner integral in (3.5) equals

$$\sum_{k=0}^{\infty}\frac{(\beta)_k}{k!}r^k\int_0^y x^{\alpha+k+p-i-1}(1-x)^{\gamma-\alpha-p}dx$$

$$=y^{\alpha+p-i}\sum_{k=0}^{\infty}\frac{(\beta)_k r^k y^k}{k!(\alpha+k+p-i)}{}_2F_1(\alpha+k+p-i,-\gamma+\alpha+p;\alpha+k+p-i+1;y).$$

Next, we expand this last ${}_2F_1$ and again integrate termwise to deduce that the integral in (3.5) equals

$$(3.6)\quad \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\frac{(\beta)_k(-\gamma+\alpha+p)_l\Gamma(m+2\alpha+k+l+1-i-j)\Gamma(\gamma-\alpha-p+1)}{k!l!(\alpha+k+l+p-i)\Gamma(\alpha+\gamma+k+l+p+1-i-j)}$$

$$\times r^k\,{}_2F_1(\beta,m+2\alpha+k+l+1-i-j;\alpha+\gamma+k+l+p+1-i-j;r).$$

Consequently, $a_{ij}(1\leqq i\neq j\leqq m)$ is the difference between twice the expression in (3.6) and that in (3.4). If m is odd, $m=2n+1$, then (2.5) leads to

$$(3.7)\quad a_{i,2n+2}=-a_{2n+2,i}=\frac{\Gamma(\alpha+p-i)\Gamma(\gamma-\alpha-p+1)}{\Gamma(\gamma-i+1)}{}_2F_1(\beta,\alpha+p-i;\gamma-i+1;r),$$

for $i=1,\cdots,2n+1$.

At this point, we address some remarks to the computability of the matrix A and its Pfaffian. First, since A is skew-symmetric, only $\frac{1}{2}m(m-1)$ of the entries a_{ij} need to be computed. In addition, numerical studies indicate that the series in (3.6) converge fairly quickly for various choices of the parameters. Preliminary work seems to indicate that for small values of m, it is more efficient to compute the generalized hypergeometric function from the formula involving $\mathrm{Pf}(A)$ or $|A|$ than through the use of partitional sums.

To conclude this section we remark that from the results obtained in (3.6) and (3.7), we can use confluence relations to obtain similar results for the confluent hypergeometric and Bessel functions of Herz (1955) and Muirhead (1970).

4. Applications to latent root distributions. In multivariate statistical theory, the distributions of several test criteria defined in terms of the latent roots of random matrices may be expressed in terms of hypergeometric functions of scalar matrix argument. Let U_1 and U_2 be two $m\times m$ independent random matrices having Wishart distributions $W(n_1,\Sigma)$ and $W(n_2,\Sigma)$ respectively, where n_1, $n_2\geqq m$. The matrix $U=(U_1+U_2)^{-1/2}U_1(U_1+U_2)^{-1/2}$ is called the generalized B statistic. Sugiyama (1967) shows that $\Pr(\lambda_1<r)$, the cumulative distribution function of λ_1, the largest latent root of U, is proportional to $r^{mn_1/2}{}_2F_1(-\frac{1}{2}n_2+p,\frac{1}{2}n_2;\frac{1}{2}n_1+p;rI_m)$, $0<r<1$. Consequently, the results of the previous section lead to computable representations for the distribution function of λ_1. Similar applications may be made to problems discussed by Hayakawa (1967), Muirhead (1975), Gupta and Richards (1982), among others.

We also provide some explicit details for a problem discussed by James (1975). Let S be a real $m\times m$ symmetric matrix, and

$$F^{(m)}(r)=\int_{-\infty<S<rI}\exp\left(-\frac{1}{4}trS^2\right)dS,\qquad -\infty<r<\infty.$$

$F^{(m)}(r)$ is related to the distribution of the largest latent root of a random symmetric matrix whose components are normally distributed. As before, we set $S=H\Delta H^{-1}$ and obtain

$$F^{(m)}(r)=c_m\underset{r>s_1>\cdots>s_m>-\infty}{\int\cdots\int}\prod_{i<j}^{m}(s_i-s_j)\exp\left(-\frac{1}{4}\sum_{i=1}^{m}s_i^2\right)ds_1\cdots ds_m.$$

Thus, de Bruijn's results show that $F^{(m)}(r)=c_m\operatorname{Pf}(A)$, where $A=(a_{ij})$ is skew-symmetric $m\times m$, with

$$(4.1)\qquad a_{ij}=-a_{ji}=\int_{-\infty}^{r}\int_{-\infty}^{r}x^{m-i}y^{m-j}\operatorname{sgn}(y-x)\exp\left(-\frac{1}{4}x^2-\frac{1}{4}y^2\right)dx\,dy,$$

for $1\leqq i\neq j\leqq 2[\frac{1}{2}m]$, and with the usual modifications if m is odd. As before, we write a_{ij} as the difference of two integrals,

$$(4.2)\qquad a_{ij}=\left(\int_{-\infty}^{r}\int_{-\infty}^{y}-\int_{-\infty}^{r}\int_{y}^{r}\right)x^{m-i}y^{m-j}\exp\left(-\frac{(x^2+y^2)}{4}\right)dx\,dy.$$

The sum of the two integrals in (4.2) equals $F_i(r)F_j(r)$, where

$$F_i(r)=\int_{-\infty}^{r}x^{m-i}\exp\left(-\frac{1}{4}x^2\right)dx,\qquad -\infty<r<\infty.$$

With the usual notation for the incomplete gamma functions (Erdélyi et al. (1953, p. 266)), it may be shown that for $i=1,2,\cdots,m$,

$$F_i(r)=\begin{cases}(-1)^{m-i+1}2^{m-i}\Gamma\left(\dfrac{m-i+1}{2},\dfrac{r^2}{4}\right), & r\leqq 0,\\[2ex] (-1)^{m-i+1}2^{m-i}\Gamma\left(\dfrac{m-i+1}{2}\right)+2^{m-i}\gamma\left(\dfrac{(m-i+1)}{2},\dfrac{r^2}{4}\right), & r>0.\end{cases}$$

Hence, as before, we need only compute the first integral in (4.2), viz.

$$(4.3)\qquad \int_{-\infty}^{r}\int_{-\infty}^{y}x^{m-i}y^{m-j}\exp\left(-\frac{(x^2+y^2)}{4}\right)dx\,dy.$$

First, we assume $r\leqq 0$. Replacing (x,y) by $(-2s^{1/2},-2t^{1/2})$, we find that (4.3) equals

$$(4.4)\qquad (-1)^{i+j}2^{2m-i-j}\int_{r^2/4}^{\infty}\int_{t}^{\infty}s^{(m-i-1)/2}t^{(m-j-1)/2}e^{-s-t}ds\,dt.$$

The inner integral in (4.4) equals

$$\Gamma\left(\frac{m-i+1}{2}\right)-\int_0^t s^{(m-i-1)/2}e^{-s}ds$$

$$=\Gamma\left(\frac{m-i+1}{2}\right)-\frac{2t^{(m-i+1)/2}}{m-i+1}{}_1F_1\left(\frac{m-i+1}{2};\frac{m-i+3}{2};-t\right).$$

Expanding the ${}_1F_1$ and integrating termwise, it follows that (4.4) equals

(4.5)

$$(-1)^{i+j}2^{2m-i-j}\left[\Gamma\left(\frac{m-i+1}{2}\right)\Gamma\left(\frac{m-j+1}{2},\frac{r^2}{4}\right)\right.$$
$$\left.-\frac{2}{m-i+1}\sum_{k=0}^{\infty}\frac{(-1)^k((m-i+1)/2)_k}{k!((m-i+3)/2)_k}\Gamma\left(k+m-\frac{i+j}{2}+1,\frac{r^2}{4}\right)\right].$$

To evaluate (4.3) when $r>0$, we split the region $\{-\infty<x<y<r\}$ into the disjoint union of $R_1=\{-\infty<x<y\leqq 0\}$, $R_2=\{-\infty<x<0,0<y<r\}$ and $R_3=\{0<x<y<r\}$. Letting $\phi_i(x)=x^{m-i}\exp(-x^2/4)$, $x\in\mathbb{R}$, it follows from (4.5) that

$$\iint_{R_1}\phi_i(x)\phi_j(y)\,dx\,dy=(-1)^{i+j}2^{2m-i-j}\Gamma\left(\frac{m-i+1}{2}\right)\Gamma\left(\frac{m-j+1}{2}\right).$$

Trivially,

$$\iint_{R_2}\phi_i(x)\phi_j(y)\,dx\,dy=(-1)^{m-i+1}2^{2m-i-j}\Gamma\left(\frac{m-i+1}{2}\right)\gamma\left(\frac{m-j+1}{2},\frac{r^2}{4}\right),$$

while the methods used earlier will also prove that

$$\iint_{R_3}\phi_i(x)\phi_j(y)\,dx\,dy=\frac{2^{2m-i-j}}{m-i+1}\sum_{k=0}^{\infty}\frac{(-1)^k((m-i+1)/2)_k}{k!((m+3-i)/2)_k}\gamma\left(k+m-\frac{i+j}{2}+1,\frac{r^2}{4}\right).$$

Summing the last three expressions evaluates (4.3) for $r>0$, from which (4.2) can be computed. These formulae lead to explicit results for the function $F^{(m)}(r)$. When $m=2$, some numerical work indicates that the series derived here are adequate for computations purposes; these matters will be considered more extensively elsewhere. We should finally note that for m odd, similar expessions for the a_{ij} are obtained above, since then,

$$a_{i,2n+2}=\int_{-\infty}^{r}\phi_i(x)\,dx\equiv F_i(r).$$

Acknowledgments. The suspicion that the results given here would be valid, originated from conversations with Professor R. Askey; we are grateful to him for drawing our attention to de Bruijn's paper.

REFERENCES

[1] N. G. DE BRUIJN (1955), *On some multiple integrals involving determinants*, J. Indian Math. Soc., 19, pp. 133–151.

[2] A. G. CONSTANTINE (1963), *Some non-central distribution problems in multivariate analysis*, Ann. Math. Statist., 34, pp. 1270–1285.

[3] A. Erdélyi, W. Magnus, F. Oberhettinger and F. G. Tricomi (1953), *Higher Transcendental Functions*, Vol. I, McGraw-Hill, New York.

[4] R. D. Gupta and D. St. P. Richards (1982), *Series expansions for a hypergeometric function of matrix argument with applications*, Austral. J. Statist., 24, pp. 216–220.

[5] T. Hayakawa (1967), *On the distribution of the maximum latent root of a positive definite symmetric matrix*, Ann. Inst. Statist. Math., 19, pp. 1–17; correction 21 (1969), p. 221.

[6] C. S. Herz (1955), *Bessel functions of matrix argument*, Ann. of Math., 61, pp. 474–523.

[7] A. T. James (1975), *Special functions of matrix and single argument in statistics*, in Theory and Application of Special Functions, R. A. Askey, ed., Academic Press, New York, pp. 497–520.

[8] T. Koornwinder and I. Sprinkhuizen-Kuyper (1978), *Hypergeometric functions of* 2×2 *matrix argument are expressible in terms of Appell's functions* F_4, Proc. Amer. Math. Soc., 70, pp. 39–42.

[9] R. J. Muirhead (1970), *Asymptotic distributions of some multivariate tests*, Ann. Math. Statist., 41, pp. 1002–1010.

[10] ______ (1975), *Expressions for some hypergeometric functions of matrix argument with applications*, J. Multivariate Anal., 5, pp. 283–293.

[11] ______ (1982), *Aspects of Multivariate Statistical Theory*, John Wiley, New York.

[12] G. Shimura (1982), *Confluent hypergeometric functions on tube domains*, Math. Ann., 260, pp. 269–302.

[13] T. Sugiyama (1967), *Distribution of the largest latent root and the smallest latent root of the generalized B statistic and F statistic in multivariate analysis*, Ann. Math. Statist., 38, pp. 1152–1159.

[14] H. Weyl (1946), *The Classical Groups*, Princeton Univ. Press, Princeton, NJ.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

016

AN INFINITE SERIES WITH PRODUCTS OF JACOBI POLYNOMIALS AND JACOBI FUNCTIONS OF THE SECOND KIND*

MIZAN RAHMAN† AND MIHR J. SHAH‡

Abstract. Let $C_n^\lambda(x)$ and $D_n^\lambda(x)$ be the ultraspherical polynomial and ultraspherical function of the second kind, respectively. Askey, Koornwinder and Rahman showed that the integral $\int_{-1}^1 D_n^\lambda(x)C_m^\lambda(x)C_l^\lambda(x)(1-x^2)^{2\lambda-1}\,dx$ vanishes when $l+m+n$ is odd and $|l-m|<n<l+m$. We find a dual of this result, namely, that

$$\sum_{n=0}^{\infty}(n+\lambda)\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}C_n^\lambda(\cos\beta)C_n^\lambda(\cos\gamma)D_n^\lambda(\cos\alpha)$$

vanishes when $|\beta-\gamma|<\alpha<\beta+\gamma$. We evaluate this sum for $\alpha<|\beta-\gamma|$ and $\alpha>\beta+\gamma$, and apply Hsü's limiting technique to give a derivation of the Bessel function integral $\int_0^\infty x^{1+\nu}Y_\nu(ax)J_\nu(bx)J_\nu(cx)\,dx$ that has recently been evaluated by Askey, Koornwinder and Rahman by a different method.

1. Introduction. Integrals and sums of products of classical orthogonal polynomials have been a matter of curiosity for a long time (see [20] for an impressive list of integral formulas). Many of these formulas have also proved very useful. During the last fifteen years or so Askey and some of his co-workers [1]–[5], [7]–[9], [11] have pointed out some interesting applications of such integral and sum formulas. To mention one of the many such applications, Gasper [22], [23] proved that the linearization coefficients in

$$\frac{P_m^{(\alpha,\beta)}(x)}{P_m^{(\alpha,\beta)}(1)}\frac{P_n^{(\alpha,\beta)}(x)}{P_n^{(\alpha,\beta)}(1)}=\sum_{k=|n-m|}^{n+m}g(k,m,n;\alpha,\beta)\frac{P_k^{(\alpha,\beta)}(x)}{P_k^{(\alpha,\beta)}(1)},\tag{1.1}$$

where $P_n^{(\alpha,\beta)}(x)$ is the Jacobi polynomial defined by

$$P_n^{(\alpha,\beta)}(x)=\frac{(\alpha+1)_n}{n!}\,{}_2F_1\left[\begin{matrix}-n,\,n+\alpha+\beta+1\\ \alpha+1\end{matrix};\ \frac{1-x}{2}\right],\tag{1.2}$$

are nonnegative for all nonnegative (k,m,n) if and only if (α,β) belongs to a certain set. Gasper showed that the nonnegativity of these coefficients is all that is required to obtain a convolution structure, a Wiener–Lévy theorem and the positivity of a generalized translation operator for Jacobi polynomials. Orthogonality of the Jacobi polynomials implies that the coefficient $g(k,m,n;\alpha,\beta)$ in (1.1) is a positive multiple of the integral

$$\int_{-1}^{1}(1-x)^\alpha(1+x)^\beta P_k^{(\alpha,\beta)}(x)P_m^{(\alpha,\beta)}(x)P_n^{(\alpha,\beta)}(x)\,dx.\tag{1.3}$$

* Received by the editors July 18, 1983, and in revised form December 10, 1983.

† Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6. The work of this author was supported by the Natural Science and Engineering Research Council under grant A6197.

‡ Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6. This is part of the material submitted to the Department of Mathematics and Statistics at Carleton University as partial fulfillment towards a Ph.D. degree in Mathematics.

Recently Rahman [28] was able to evaluate this integral as a single series with nonnegative terms in the case $\alpha \geq \beta > -1$ and $\alpha+\beta+1 \geq 0$.

As a dual to the linearization problem, Gasper [24] showed that the kernel $K(x,y,z;\alpha,\beta)$ in

$$(1.4) \qquad \frac{P_n^{(\alpha,\beta)}(x) P_n^{(\alpha,\beta)}(y)}{P_n^{(\alpha,\beta)}(1) P_n^{(\alpha,\beta)}(1)} = \int_{-1}^{1} K(x,y,z;\alpha,\beta) \frac{P_n^{(\alpha,\beta)}(z)}{P_n^{(\alpha,\beta)}(1)} (1-z)^{\alpha}(1+z)^{\beta} dz$$

is nonnegative on the set

$$(1.5) \qquad U = \left\{(\alpha,\beta) | \alpha \geq \beta \geq -\tfrac{1}{2}, \alpha > -\tfrac{1}{2}\right\}.$$

Orthogonality implies

$$(1.6) \qquad K(x,y,z;\alpha,\beta) = \sum_{n=0}^{\infty} h_n^{(\alpha,\beta)} P_n^{(\alpha,\beta)}(x) P_n^{(\alpha,\beta)}(y) P_n^{(\alpha,\beta)}(z) \frac{n!}{(\alpha+1)_n},$$

where the normalization constant $h_n^{(\alpha,\beta)}$ is given by

$$(1.7) \qquad h_n^{(\alpha,\beta)} = \left\{ \int_{-1}^{1} (1-x)^{\alpha}(1+x)^{\beta} \left[P_n^{(\alpha,\beta)}(x)\right]^2 dx \right\}^{-1}$$
$$= \frac{(2n+\alpha+\beta+1)\Gamma(n+\alpha+\beta+1)n!}{2^{\alpha+\beta+1}\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}, \qquad \mathrm{Re}\,\alpha, \beta > -1.$$

Gasper [24], [25] used the nonnegativity of this kernel to construct a convolution structure for Jacobi series and to prove the positivity of a generalized translation operator.

Special functions and many of their integral and sum formulas have perhaps been more widely used by physicists than mathematicians. To mention only one example that is relevant to this work, the Clebsch–Gordon coefficients with zero magnetic quantum numbers have the following representation [13], [17]

$$(1.8) \qquad \left(C_{k0m0}^{n0}\right)^2 = \frac{2n+1}{2} \int_{-1}^{1} P_k(x) P_m(x) P_n(x)\, dx$$

where $P_n(x) = P_n^{(0,0)}(x)$ is the Legendre polynomial of degree n. Obviously, the integral in (1.8) is the same as in (1.3) for $\alpha=\beta=0$. From the symmetry and orthogonality properties of the Legendre polynomials it is clear that the integral in (1.8) vanishes if $k+m+n$ is odd or the triangle inequality

$$(1.9) \qquad |n-m| \leq k \leq m+n$$

is not satisfied. These inequalities must also be satisfied for the nonvanishing of the integral (1.3).

In trying to classify the various fluctuation modes in an investigation of the stability properties of some special solutions of the $O(n)$ nonlinear 2-dimensional σ-model Din and Zakrzewski [14] found that the classification depended crucially on the vanishing of the sum

$$\sum_{k=|n-m|}^{n+m} \frac{2k+1}{k(k+1)-m(m+1)} \left(C_{k0m0}^{n0}\right)^2.$$

Din showed that this vanishing follows from

$$\int_{-1}^{1} Q_k(x)P_m(x)P_n(x)\,dx=0 \tag{1.10}$$

where $k+m+n$ is odd, $|n-m|<k<m+n$, and $Q_k(x)$ is an appropriately defined Legendre function [19], [29] of the second kind on the cut $-1<x<1$. Then he showed [13] that (1.10) holds. This is a very interesting and somewhat surprising result because the integral (1.10) vanishes on the set complementary to the set where (1.8) vanishes. Din did not compute its value when k,m,n do not satisfy (1.9). The integral (1.10) obviously vanishes when $k+m+n$ is even and Askey [6] evaluated it when $k+m+n$ is odd and either $k<|n-m|$ or $k>n+m$.

Dougall's [15] generalization of the integral in (1.8) to ultraspherical polynomials is well known:

$$\int_{-1}^{1} C_k^\lambda(x)C_m^\lambda(x)C_n^\lambda(x)(1-x^2)^{\lambda-1/2}dx \tag{1.11}$$

$$=\frac{\Gamma(\frac{1}{2})\Gamma(\lambda+\frac{1}{2})}{\Gamma(\lambda+1)}\,\frac{(\lambda)_{s-k}(\lambda)_{s-m}(\lambda)_{s-n}(2\lambda)_s}{(s-k)!(s-m)!(s-n)!(\lambda+1)_s},$$

where $k+m+n=2s$ is even and $|n-m|\le k\le n+m$, and zero otherwise. Here $C_n^\lambda(x)$ is the ultraspherical polynomial that can be defined by

$$C_n^\lambda(\cos\theta)=\sum_{k=0}^{n}\frac{(\lambda)_k(\lambda)_{n-k}}{k!(n-k)!}\cos(n-2k)\theta. \tag{1.12}$$

Recently Askey, Koornwinder and Rahman [10] found a generalization of Din's integral. They showed that

$$I(k,m,n)=\int_{-1}^{1} D_k^\lambda(x)C_m^\lambda(x)C_n^\lambda(x)(1-x^2)^{2\lambda-1}dx \tag{1.13}$$

vanishes when (i) $k+m+n$ is even, and (ii) $k+m+n$ is odd and $|n-m|<k<n+m$, where $D_k^\lambda(x)$ is the ultraspherical function of the second kind defined, on the cut $-1<x<1$, by

$$(1-x^2)^{\lambda-1/2}D_k^\lambda(x)=\frac{2\Gamma(\lambda+\frac{1}{2})(2\lambda)_k}{\Gamma(\frac{1}{2})\Gamma(\lambda+1)k!}\sum_{l=0}^{\infty}\frac{(1-\lambda)_l(1)_{l+k}}{l!(\lambda+1)_{l+k}}\cos(k+2l+1)\theta, \tag{1.14}$$

$x=\cos\theta$. They also showed that

(1.15)

$$I(m+n+1+2k,m,n)$$

$$=-\left[\frac{\Gamma(\lambda+\frac{1}{2})}{\Gamma(\lambda+1)}\right]^2\frac{(2\lambda)_m(2\lambda)_n(2\lambda)_{m+n+1+2k}}{m!n!(m+n+1+2k)!}\,\frac{(1-\lambda)_k(\lambda)_{m+n+1+k}(m+k)!(n+k)!}{k!(2\lambda)_{m+n+1+k}(\lambda+1)_{m+k}(\lambda+1)_{n+k}},$$

$k=0,1,2,\cdots$, while

(1.16)

$$I(m+n+1-2k,m,n)$$
$$=-\left[\frac{\Gamma(\lambda+\frac{1}{2})}{\Gamma(\lambda+1)}\right]^2\frac{(2\lambda)_m(2\lambda)_n(2\lambda)_{m+n+1-2k}}{m!n!(m+n+1-2k)!}\,\frac{(-\lambda)_{k-m}(\lambda)_{m+n+1-k}(n-k)!\Gamma(k)}{(\lambda)_k(2\lambda)_{m+n+1-k}(\lambda+1)_{n-k}\Gamma(k-m)},$$

where $0\le m\le n$ and $k=1,2,\cdots,m,m+1,\cdots,[(m+n+1)/2]$.

The purpose of this paper is to find duals of (1.13), (1.15) and (1.16) in the same sense as the dual to Dougall's integral (1.11) is Dougall's infinite sum [14]:

$$(1.17)\qquad \sum_{n=0}^{\infty}(n+\lambda)\left\{\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\right\}^2 C_n^\lambda(\cos\alpha)C_n^\lambda(\cos\beta)C_n^\lambda(\cos\gamma)$$
$$=2^{-2\lambda}\pi\{\Gamma(\lambda)\}^{-4}(\sin\alpha\sin\beta\sin\gamma)^{1-2\lambda}$$
$$\cdot\left\{\sin\frac{\alpha+\beta+\gamma}{2}\sin\frac{\beta+\gamma-\alpha}{2}\sin\frac{\gamma+\alpha-\beta}{2}\sin\frac{\alpha+\beta-\gamma}{2}\right\}^{\lambda-1}$$

if $0<\alpha,\beta,\gamma<\pi$, $0<\operatorname{Re}\lambda$, and a triangle can be drawn with sides α,β,γ, assuming that the sum of any two of them is less than or equal to π; and, 0 otherwise.

We show that the dual to (1.13), (1.15) and (1.16) is

(1.18)

$$\sum_{n=0}^{\infty}(n+\lambda)\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}C_n^\lambda(\cos\beta)C_n^\lambda(\cos\gamma)D_n^\lambda(\cos\alpha)$$
$$=\frac{2^{2-2\lambda}}{\Gamma^2(\lambda)}\begin{cases}0 & \text{if } |\beta-\gamma|<\alpha<\beta+\gamma\le\pi,\\ 2^{-2\lambda-1}\sin\pi\lambda(-16D)^{-\lambda}, & \\ & \alpha<|\beta-\gamma|, \text{ or } \pi<\beta+\gamma<2\pi \text{ and } \alpha+\beta+\gamma<2\pi,\\ -2^{-2\lambda-1}\sin\pi\lambda(-16D)^{-\lambda} & \text{if } \alpha>\beta+\gamma\end{cases}$$

where $-\frac{1}{2}<\operatorname{Re}\lambda<1$ and $0<\alpha,\beta,\gamma<\pi$ and

$$(1.19)\qquad -16D=\sin\frac{\alpha+\beta+\gamma}{2}\sin\frac{\alpha+\beta-\gamma}{2}\sin\frac{\beta+\gamma-\alpha}{2}\sin\frac{\beta-\gamma-\alpha}{2}.$$

Hsü [26] showed that Sonine's integral formula [30, p. 411]

(1.20)

$$\int_0^\infty J_\nu(ax)J_\nu(bx)J_\nu(cx)x^{1-\nu}dx$$
$$=\begin{cases}0 & \text{if } a,b,c \text{ are not sides of a triangle,}\\ \dfrac{2^{\nu-1}\Delta^{\nu-1/2}}{\sqrt{\pi}\,(abc)^\nu\Gamma(\nu+\frac{1}{2})} & \text{if } a,b,c \text{ are sides of a triangle of area } \Delta^{1/2},\end{cases}$$

follows as limiting cases of both (1.11) and (1.17), provided $\mathrm{Re}\,\nu > -\frac{1}{2}$. $J_\nu(z)$ is the Bessel function that is related to the ultraspherical polynomials through Mehler's formula

$$\lim_{n\to\infty} n^{-2\nu} C_n^{\nu+1/2}\left(\cos\frac{z}{n}\right) = \frac{\sqrt{\pi}}{\Gamma\left(\nu+\frac{1}{2}\right)}(2z)^{-\nu}J_\nu(z). \tag{1.21}$$

One can show that a similar limiting formula holds for $D_n^\lambda(z)$:

$$\lim_{n\to\infty} n^{-2\nu} D_n^{\nu+1/2}\left(\cos\frac{z}{n}\right) = -\frac{\sqrt{\pi}}{\Gamma\left(\nu+\frac{1}{2}\right)}(2z)^{-\nu}Y_\nu(z), \tag{1.22}$$

where $Y_\nu(z)$ is the Bessel function of the second kind. It is possible to use (1.21) and (1.22) in (1.13) and obtain a formula complementary to (1.20) in the same sense as (1.13) is complementary to (1.11), by using Hsü's limiting procedure. However, because of the simplicity of Bessel functions, Askey, Koornwinder and Rahman [10] preferred a direct computation to show that

$$\int_0^\infty Y_\nu(ax)J_\nu(bx)J_\nu(cx)x^{1+\nu}dx = \begin{cases} 0 & \text{if } |b-c|<a<b+c, \\ -\dfrac{2^{-\nu-1}(-\Delta)^{-\nu-1/2}}{\sqrt{\pi}\,\Gamma\left(\frac{1}{2}-\nu\right)(abc)^{-\nu}} & \text{if } a<|b-c|, \\ \dfrac{2^{-\nu-1}(-\Delta)^{-\nu-1/2}}{\sqrt{\pi}\,\Gamma\left(\frac{1}{2}-\nu\right)(abc)^{-\nu}} & \text{if } b+c<a. \end{cases} \tag{1.23}$$

In (1.20) and (1.23) Δ is given by the same algebraic expression, namely,

$$\Delta = \left((b+c)^2-a^2\right)\left(a^2-(b-c)^2\right)/16. \tag{1.24}$$

However, in (1.20) it is positive and represents the square of the area of the plane triangle of sides a, b, c, whereas in (1.23) Δ is negative because a, b, c do not satisfy the triangle inequality. The complementary nature of the two formulas is vividly displayed by the striking similarity of the two expressions on the right hand sides of (1.20) and (1.23).

In §2 we first introduce a dual to the integral in (1.13) and carry out the computations in §3 that lead to (1.18). In §4 we give an alternative derivation of (1.23) by applying Hsü's limiting procedure to (1.18).

2. The dual kernel. The essential ingredients of this work are contained in Rahman [27] where he gave a more accessible proof and extensions of the following formula of Feldheim [21]:

$$\begin{aligned} &F_4\left(\alpha_1,\alpha_2;\alpha+1,\beta+1;\ \frac{\rho(1-x)(1-y)}{4},\ \frac{\rho(1+x)(1+y)}{4}\right) \\ &= \sum_{n=0}^{\infty}\frac{(\alpha+1)_n(\alpha+\beta+1)_n}{n!(\beta+1)_n(\alpha+\beta+1)_{2n}}(\alpha_1)_n(\alpha_2)_n\rho^n\,{}_2F_1\left[\begin{matrix}\alpha_1+n,\ \alpha_2+n\\ \alpha+\beta+2+2n\end{matrix};\ \rho\right] \\ &\quad\cdot\frac{P_n^{(\alpha,\beta)}(x)}{P_n^{(\alpha,\beta)}(1)}\frac{P_n^{(\alpha,\beta)}(y)}{P_n^{(\alpha,\beta)}(1)} \end{aligned} \tag{2.1}$$

where the parameters $\alpha, \beta, \alpha_1, \alpha_2, \rho$ are restricted by the requirement of convergence of the infinite series involved, and F_4 is an Appell function [12], [18] defined by

$$(2.2) \qquad F_4(\alpha_1, \alpha_2; \beta_1, \beta_2; u, v) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\alpha_1)_{m+n}(\alpha_2)_{m+n}}{m!n!(\beta_1)_m(\beta_2)_n} u^m v^n.$$

Jacobi functions of the second kind, $Q_n^{(\alpha,\beta)}(z)$, defined for all z in the complex plane cut along the segment $(-1,1)$, are given by [19], [29]

$$(2.3) \qquad Q_n^{(\alpha,\beta)}(z) = 2^{n+\alpha+\beta} \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(2n+\alpha+\beta+2)}$$

$$\cdot (z-1)^{-\alpha-n-1}(z+1)^{-\beta} {}_2F_1\left[\begin{matrix} n+1, n+\alpha+1 \\ 2n+\alpha+\beta+2 \end{matrix}; \frac{2}{1-z}\right],$$

$$= 2^{n+\alpha+\beta} \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(2n+\alpha+\beta+2)}$$

$$\cdot (z-1)^{-\alpha}(z+1)^{-n-\beta-1} {}_2F_1\left[\begin{matrix} n+1, n+\beta+1 \\ 2n+\alpha+\beta+2 \end{matrix}; \frac{2}{1+z}\right].$$

This can be used to define $Q_n^{(\alpha,\beta)}(x)$ on the cut and to compute $P_n^{(\alpha,\beta)}(x)$ there using Durand's formulas [16]:

$$(2.4) \qquad P_n^{(\alpha,\beta)}(x) = \frac{i}{\pi} \lim_{\varepsilon \to 0} \left[e^{i\pi\alpha} Q_n^{(\alpha,\beta)}(x+i\varepsilon) - e^{-i\pi\alpha} Q_n^{(\alpha,\beta)}(x-i\varepsilon) \right],$$

$$(2.5) \quad Q_n^{(\alpha,\beta)}(x) = \frac{1}{2} \lim_{\varepsilon \to 0} \left[e^{i\pi\alpha} Q_n^{(\alpha,\beta)}(x+i\varepsilon) + e^{-i\pi\alpha} Q_n^{(\alpha,\beta)}(x-i\varepsilon) \right], \qquad -1 < x < 1.$$

In [27] Rahman put $\alpha_1 = 1$, $\alpha_2 = \beta + 1$, $\rho = 2/(1+z)$ in (2.1) and used (2.4) to give an alternative proof of Gasper's result [24]

(2.6)

$$K(x, y, z; \alpha, \beta)$$

$$= \begin{cases} 0 \quad \text{if } f < |g-h|, \\[2mm] \dfrac{\Gamma(\alpha+1)(1-h^2)^{-\alpha} f^{-2\alpha}(f^2-g^2-h^2)^{\alpha-\beta-1}}{\Gamma(\alpha-\beta)\Gamma(\beta+1)2^{\alpha+\beta+1}} \\[2mm] \qquad \cdot {}_2F_1\left[\begin{matrix} \dfrac{\beta-\alpha+1}{2}, \dfrac{\beta-\alpha+2}{2} \\ \beta+1 \end{matrix}; G^{-2}\right], \qquad h < f-g, \ \mathrm{Re}(\alpha-\beta) > 0, \\[2mm] \dfrac{\Gamma(\alpha+1)(1-h^2)^{-\alpha} f^{-2\alpha}(gh)^{\alpha-\beta-1}(1-G^2)^{\alpha-1/2}}{\Gamma(\alpha+\frac{1}{2})\Gamma(\frac{1}{2})2^{\alpha+\beta+2}} \\[2mm] \qquad \cdot {}_2F_1\left[\begin{matrix} \alpha-\beta, \alpha+\beta \\ \alpha+\frac{1}{2} \end{matrix}; \tfrac{1}{2}(1-G)\right], \qquad |f-g| < h < f+g, \ \mathrm{Re}\,\alpha > -\tfrac{1}{2}, \end{cases}$$

where $K(x,y,z;\alpha,\beta)$ is defined in (1.6), and

$$\text{(2.7)}\qquad \begin{matrix} f=\sin\phi\sin\psi, & g=\cos\phi\cos\psi, & h=\cos\theta, \\ x=\cos 2\phi, & y=\cos 2\psi, & z=\cos 2\theta, \end{matrix} \quad 0<\theta,\phi,\psi<\frac{\pi}{2},$$

$$\text{(2.8)}\qquad G=\frac{g^2+h^2-f^2}{2gh}.$$

One could have used (2.5) in (2.1) with the above choices of α_1 and α_2 to obtain a different kernel, but it is not the right kernel because it does not vanish in the triangular region $|f-g|<h<f+g$, analogous to the vanishing of the integral (1.13), unless $\alpha=\beta=0$. The reason is similar to the fact that the appropriate weight factor in (1.13) is $(1-x^2)^{2\lambda-1}$ and not $(1-x^2)^{\lambda-1/2}$, as one might expect. As it turns out, the appropriate dual to (1.13) is obtained by making the following identification of the parameters in (2.1):

$$\text{(2.9)}\qquad \alpha_1=\alpha+\beta+1,\quad \alpha_2=\alpha+1,\quad \rho=\frac{2}{1+z}.$$

Thus we have, on using the obvious symmetry of F_4,

$$\text{(2.10)}\qquad F_4\left(\alpha+\beta+1,\alpha+1;\beta+1,\alpha+1;\ \frac{(1+x)(1+y)}{2(1+z)},\ \frac{(1-x)(1-y)}{2(1+z)}\right)$$
$$=\sum_{n=0}^{\infty}\frac{n!(\alpha+\beta+1)_n}{(\beta+1)_n(\alpha+\beta+1)_{2n}}(\alpha+\beta+1)_n\left(\frac{2}{1+z}\right)^n$$
$$\cdot {}_2F_1\left[\begin{matrix}\alpha+\beta+1+n,\alpha+n+1\\ \alpha+\beta+2+2n\end{matrix};\ \frac{2}{1+z}\right]P_n^{(\alpha,\beta)}(x)P_n^{(\alpha,\beta)}(y).$$

Using [18, 2.1.4(23), p. 64], the second formula in (2.3), and (1.7), we get

$$\text{(2.11)}\quad F_4\left(\alpha+\beta+1,\alpha+1;\beta+1,\alpha+1;\ \frac{(1+x)(1+y)}{2(1+z)},\ \frac{(1-x)(1-y)}{2(1+z)}\right)$$
$$=2(z+1)^{\alpha+\beta+1}\sum_{n=0}^{\infty}h_n^{(\alpha,\beta)}\frac{(\alpha+\beta+1)_n}{(\beta+1)_n}P_n^{(\alpha,\beta)}(x)P_n^{(\alpha,\beta)}(y)Q_n^{(\alpha,\beta)}(z),$$

where $-1\le x,y\le 1$ and z is in the cut-plane.

Using (2.5) we then have, for $-1\le x,y\le 1$ and $-1<z<1$,

(2.12)
$$L(x,y,z;\alpha,\beta)\equiv\sum_{n=0}^{\infty}h_n^{(\alpha,\beta)}\frac{(\alpha+\beta+1)_n}{(\beta+1)_n}P_n^{(\alpha,\beta)}(x)P_n^{(\alpha,\beta)}(y)Q_n^{(\alpha,\beta)}(z)$$
$$=\frac{1}{4}(z+1)^{-\alpha-\beta-1}\lim_{\varepsilon\to 0}\left[e^{i\pi\alpha}A(x,y,z+i\varepsilon)+e^{-i\pi\alpha}A(x,y,z-i\varepsilon)\right],$$

where, for abbreviation, we write

$$\text{(2.13)}\quad A(x,y,z)=F_4\left(\alpha+\beta+1,\alpha+1;\beta+1,\alpha+1;\ \frac{(1+x)(1+y)}{2(1+z)},\ \frac{(1-x)(1-y)}{2(1+z)}\right).$$

There are some convergence difficulties in both (2.11) and (2.12). The series on the l.h.s. of (2.11) may diverge if z, but neither x nor y, is close to -1. One can use an analytic continuation [18, 5.11(9), p. 240] to remove the difficulty with the F_4 function. However, the series on the r.h.s. of (2.12) definitely diverges for $\operatorname{Re}\alpha \geq \frac{1}{2}$. So our results are necessarily restricted to $-1 < \operatorname{Re}\alpha < \frac{1}{2}$.

3. Computation of $L(x,y,z;\alpha,\beta)$. The Appell function in (2.13) is absolutely convergent if

$$\left|\frac{(1+x)(1+y)}{2(1+z)}\right|^{1/2} + \left|\frac{(1-x)(1-y)}{2(1+z)}\right|^{1/2} < 1. \tag{3.1}$$

With $\theta, \phi, \psi, f, g, h$ defined as in (2.7) this condition implies

$$f+g<h. \tag{3.2}$$

When this inequality is satisfied, $A(x,y,z)$ is single-valued on the real axis of the z-plane, so we simply have

(3.3)

$$L(x,y,z;\alpha,\beta) = \frac{1}{2}\cos\pi\alpha(2h^2)^{-\alpha-\beta-1}F_4(\alpha+\beta+1,\alpha+1;\beta+1,\alpha+1;\ g^2/h^2, f^2/h^2).$$

When (3.1) is not satisfied, $A(x,y,z)$ may not converge, but by [12, 20(v) p. 102] it can be expressed in terms of a ${}_2F_1$ which admits many analytic continuations. Let

$$\frac{g^2}{h^2} = -\frac{s}{(1-s)(1-t)}, \qquad \frac{f^2}{h^2} = -\frac{t}{(1-s)(1-t)}. \tag{3.4}$$

Solving for s and t we get

$$s = \frac{f^2+g^2-h^2 \pm 4\sqrt{-D}}{2f^2}, \qquad t = \frac{f^2+g^2-h^2 \pm 4\sqrt{-D}}{2g^2}, \tag{3.5}$$

where

$$D = \left((f+g)^2 - h^2\right)\left(h^2 - (f-g)^2\right)/16 \tag{3.6}$$

is the same as Δ in (1.2) with a,b,c replaced by h,f and g, respectively, and equals the square of the area of a plane triangle of sides f,g,h whenever $D>0$. However, if $h<|f-g|$ or $h>f+g, D$ is negative and, by (3.5), s,t are both real. If f,g,h are real and positive, as they are whenever $0<\theta,\phi,\psi<\pi/2$, and satisfy the triangle inequality $|f-g|<h<f+g$ it follows from (3.5) that $\operatorname{Im}s \neq 0$, $\operatorname{Im}t \neq 0$.

We could express D in terms of θ,ϕ,ψ as follows:

$$-64D = (z-h_+)(z-h_-) \tag{3.7}$$

where $h_\pm = \cos 2(\phi \pm \psi)$.

As a function of $z, \sqrt{-D}$ has a branch-point at h_+ and another at h_-. In the z-plane cut along the real axis from $\min(h_\pm)$ to $\max(h_\pm)$, $\sqrt{-D}$ is single-valued and so, for $|\operatorname{Re}z| > h_\pm$ we may choose either of the two signs in (3.5). We choose the following branch:

$$s = \frac{f^2+g^2-h^2-4\sqrt{-D}}{2f^2}, \qquad t = \frac{f^2+g^2-h^2-4\sqrt{-D}}{2g^2}. \tag{3.8}$$

Let us introduce two auxiliary parameters F and G:

$$F=\frac{f^2+g^2-h^2}{2fg},\qquad G=\frac{g^2+h^2-f^2}{2gh}. \tag{3.9}$$

Of course, G is the same as in (2.8). One can easily check the following identities

$$f^2g^2(F^2-1)=-4D=g^2h^2(G^2-1), \tag{3.10}$$

$$s=\frac{g}{f}\left(F-\sqrt{F^2-1}\right)=\frac{g^2}{f^2}\left(1-\frac{h}{g}\left(G+\sqrt{G^2-1}\right)\right), \tag{3.11}$$

$$t=\frac{f}{g}\left(F-\sqrt{F^2-1}\right)=1-\frac{h}{g}\left(G+\sqrt{G^2-1}\right). \tag{3.12}$$

Whenever s,t are real it follows from (3.4) that

$$\begin{array}{lll} \text{either} & \text{(i)} & s<0,\, t<0, \\ \text{or} & \text{(ii)} & 0<s<1,\, t>1, \\ \text{or} & \text{(iii)} & 0<t<1,\, s>1. \end{array} \tag{3.13}$$

Let us now consider two separate cases.

Case 1. $D<0$. This is the case when f,g,h do not satisfy the triangle inequality, that is, either $h<|f-g|$ or $f+g<h$. By Bailey [12, 20(v), p. 102]

$$\begin{aligned} &F_4\left(\alpha+\beta+1,\alpha+1;\beta+1,\alpha+1;\frac{-s}{(1-s)(1-t)},\frac{-t}{(1-s)(1-t)}\right) \\ &\quad=(1-t)^{\alpha+\beta+1}{}_2F_1\left[\begin{matrix}\alpha+\beta+1,\alpha+1\\ \beta+1\end{matrix};\frac{s(t-1)}{1-s}\right]. \end{aligned} \tag{3.14}$$

Denoting $v=s(t-1)/(1-s)$ and using the quadratic transformation [18, 2.11(34), p. 113]

$${}_2F_1\left[\begin{matrix}\alpha_1,\alpha_2\\ \alpha_1-\alpha_2+1\end{matrix};v\right]=(1+v)^{-\alpha_1}{}_2F_1\left[\begin{matrix}\frac{\alpha_1}{2},\frac{\alpha_1+1}{2}\\ \alpha_1-\alpha_2+1\end{matrix};\frac{4v}{(1+v)^2}\right], \tag{3.15}$$

we have

$$\begin{aligned} &F_4\left(\alpha+\beta+1,\alpha+1;\beta+1,\alpha+1;\frac{-s}{(1-s)(1-t)},\frac{-t}{(1-s)(1-t)}\right) \\ &\quad=\left(\frac{1-t}{1+v}\right)^{\alpha+\beta+1}{}_2F_1\left[\begin{matrix}\frac{\alpha+\beta+1}{2},\frac{\alpha+\beta+2}{2}\\ \beta+1\end{matrix};\frac{4v}{(1+v)^2}\right]. \end{aligned} \tag{3.16}$$

Let us first assume $f+g<h$ for which the kernel in terms of the Appell function is given in (3.3). Using (3.16) we shall now express it as a ${}_2F_1$ function. From the definition of F and G it follows that $F<-1$, $G>1$ and hence $s<0$, $t<0$, $v>0$. Using (3.9), (3.11) and (3.12) one can easily show that

$$\frac{1-t}{1+v}=\frac{h}{2g}G^{-1}\quad\text{and}\quad\frac{4v}{(1+v)^2}=G^{-2}. \tag{3.17}$$

Thus, it follows from (3.3), (3.16) and (3.17) that

$$L(x,y,z;\alpha,\beta)=2^{-\alpha-\beta-2}\cos\pi\alpha\left(g^2+h^2-f^2\right)^{-\alpha-\beta-1} \cdot {}_2F_1\left[\begin{matrix}\frac{\alpha+\beta+1}{2},\ \frac{\alpha+\beta+2}{2} \\ \beta+1\end{matrix};G^{-2}\right], \tag{3.18}$$

$-1<\operatorname{Re}\alpha<\frac{1}{2}$. For $\alpha=\beta$ this leads to

$$L(x,y,z;\beta,\beta)=2^{-6\beta-4}(-D)^{-(\beta+1/2)}\cos\pi\beta,\qquad f+g<h,\quad -1<\beta<\tfrac{1}{2}. \tag{3.19}$$

Let us now consider the case $h<|f-g|$. First, assume that $f<g$ so that $h<g-f$. In this case $F>1$, $G>1$ and consequently $0<t<1$ and $s>1$ while $v>0$. Hence the r.h.s. of (3.16) is single-valued with the result that the kernel is given by the same formula (3.18).

We finally look at the case $g<f$ with $h<f-g$. Here $F>1$ and $G<-1$ so that $0<s<1$, $t>1$ and $v>0$. The r.h.s. of (3.16) is double-valued in this case, so we take $\arg(1-t)=-\pi$ in the upper half-plane and $\arg(1-t)=\pi$ in the lower half-plane. Then from (2.12), (3.16) and (3.17) we get

$$L(x,y,z;\alpha,\beta)=2^{-\alpha-\beta-2}\cos\pi(\beta+1)\left(f^2-g^2-h^2\right)^{-\alpha-\beta-1} \cdot {}_2F_1\left[\begin{matrix}\frac{\alpha+\beta+1}{2},\ \frac{\alpha+\beta+2}{2} \\ \beta+1\end{matrix};G^{-2}\right], \tag{3.20}$$

$-1<\operatorname{Re}\alpha<\frac{1}{2}$. In the ultraspherical case $\alpha=\beta$ this reduces to

$$L(x,y,z;\beta,\beta)=-2^{-6\beta-4}(-\Delta)^{-(\beta+1/2)}\cos\pi\beta,\qquad h<f-g,\quad -1<\operatorname{Re}\beta<\tfrac{1}{2}. \tag{3.21}$$

Case 2. $D>0$. In this case f,g,h do form the sides of a plane triangle so that

$$|f-g|<h<f+g. \tag{3.22}$$

As a function of z, $\sqrt{-D}$ is purely imaginary and double-valued across the cut $\min(h_{\pm})<z<\max(h_{\pm})$ and we take the branch $i\sqrt{D}$ in the upper half-plane and $-i\sqrt{D}$ in the lower half-plane. So, in the upper half-plane

$$s=\frac{f^2+g^2-h^2-4i\sqrt{D}}{2f^2},\qquad t=\frac{f^2+g^2-h^2-4i\sqrt{D}}{2g^2}, \tag{3.23}$$

while in the lower half-plane

$$s=\frac{f^2+g^2-h^2+4i\sqrt{D}}{2f^2},\qquad t=\frac{f^2+g^2-h^2+4i\sqrt{D}}{2g^2}. \tag{3.24}$$

In this case $-1<F,\ G<1$. Also, from (3.12)

$$t=1-\frac{hG}{g}-\frac{hG}{g}\sqrt{1-G^{-2}}\,.$$

In order that this agrees with the choice of branches in (3.23) and (3.24) we need to choose $\arg(1-G^{-2})=\pi$ in the upper half and $-\pi$ in the lower half-plane. So

$$\pm\pi=\arg(1-G^{-2})=\arg(-G^{-2}) \tag{3.25}$$

in the upper and lower half-planes, respectively. From (2.13), (3.16) and (3.17) we have, by [18, 2.9(34), p. 107]

(3.26)

$$A(x,y,z)=\left(\frac{h}{2gG}\right)^{\alpha+\beta+1}{}_2F_1\left[\begin{matrix}\frac{\alpha+\beta+1}{2},\ \frac{\alpha+\beta+2}{2}\\ \beta+1\end{matrix};G^{-2}\right]$$

$$=\left(\frac{h}{2gG}\right)^{\alpha+\beta+1}\left\{\frac{\Gamma(\beta+1)\Gamma(\frac{1}{2})(-G^{-2})^{-(\alpha+\beta+1)/2}}{\Gamma\left(\frac{\alpha+\beta+2}{2}\right)\Gamma\left(\frac{\beta-\alpha+1}{2}\right)}\,{}_2F_1\left[\begin{matrix}\frac{\alpha+\beta+1}{2},\ \frac{\alpha-\beta+1}{2}\\ \frac{1}{2}\end{matrix};G^2\right]\right.$$

$$\left.+\frac{\Gamma(\beta+1)\Gamma(-\frac{1}{2})(-G^{-2})^{-(\alpha+\beta+2)/2}}{\Gamma\left(\frac{\alpha+\beta+1}{2}\right)\Gamma\left(\frac{\beta-\alpha}{2}\right)}\,{}_2F_1\left[\begin{matrix}\frac{\alpha+\beta+2}{2},\ \frac{\alpha-\beta+2}{2}\\ \frac{3}{2}\end{matrix};G^2\right]\right\}.$$

Hence

(3.27)

$$\lim_{\varepsilon\to 0}\left[e^{i\pi\alpha}A(x,y,z+i\varepsilon)+e^{-i\pi\alpha}A(x,y,z-i\varepsilon)\right]$$

$$=2\left(\frac{h}{2g}\right)^{\alpha+\beta+1}\frac{\Gamma(\beta+1)\Gamma(\frac{1}{2})}{\Gamma\left(\frac{\alpha+\beta+2}{2}\right)\Gamma\left(\frac{\beta-\alpha+1}{2}\right)}\cos\frac{\alpha-\beta-1}{2}\pi$$

$$\cdot{}_2F_1\left[\begin{matrix}\frac{\alpha+\beta+1}{2},\ \frac{\alpha-\beta+1}{2}\\ \frac{1}{2}\end{matrix};G^2\right]$$

$$-2\left(\frac{h}{2g}\right)^{\alpha+\beta+1}G\frac{\Gamma(\beta+1)\Gamma(\frac{1}{2})}{\Gamma\left(\frac{\alpha+\beta+1}{2}\right)\Gamma\left(\frac{\beta-\alpha}{2}\right)}\cos\frac{\alpha-\beta}{2}\pi$$

$$\cdot{}_2F_1\left[\begin{matrix}\frac{\alpha+\beta+2}{2},\ \frac{\alpha-\beta+2}{2}\\ \frac{3}{2}\end{matrix};G^2\right].$$

Thus, in the region $|f-g|<h<f+g$,

(3.28)

$$L(x,y,z;\alpha,\beta)=\frac{1}{2}(4gh)^{-\alpha-\beta-1}$$

$$\cdot\left\{\frac{\Gamma(\beta+1)\Gamma(\frac{1}{2})}{\Gamma\left(\frac{\alpha+\beta+2}{2}\right)\Gamma\left(\frac{\beta-\alpha+1}{2}\right)}\sin\frac{\alpha-\beta}{2}\pi\,{}_2F_1\left[\begin{matrix}\frac{\alpha+\beta+1}{2},\ \frac{\alpha-\beta+1}{2}\\ \frac{1}{2}\end{matrix};G^2\right]\right.$$

$$\left.-\frac{\Gamma(\beta+1)\Gamma(-\frac{1}{2})G}{\Gamma\left(\frac{\alpha+\beta+1}{2}\right)\Gamma\left(\frac{\beta-\alpha}{2}\right)}\cos\frac{\alpha-\beta}{2}\pi\,{}_2F_1\left[\begin{matrix}\frac{\alpha+\beta+2}{2},\ \frac{\alpha-\beta+2}{2}\\ \frac{3}{2}\end{matrix};G^2\right]\right\}.$$

Using the reflection formula for the gamma function

$$\Gamma(z)\Gamma(1-z)=\frac{\pi}{\sin\pi z},\qquad 0<\operatorname{Re}z<1 \tag{3.29}$$

and [18, 2.11(3), p. 111], this can be further simplified to

(3.30)

$$L(x,y,z;\alpha,\beta)=-\frac{1}{2}(4gh)^{-\alpha-\beta-1}\frac{\pi\Gamma(\beta+1)}{\Gamma(\alpha+\frac{3}{2})\Gamma\left(\frac{\beta-\alpha}{2}\right)\Gamma\left(\frac{\beta-\alpha+1}{2}\right)}$$

$$\cdot{}_2F_1\left[\begin{matrix}\alpha+\beta+1,\alpha-\beta+1\\ \alpha+\frac{3}{2}\end{matrix};\frac{1}{2}(1-G)\right]$$

$$=-2^{-2\alpha-3}(2gh)^{-\alpha-\beta-1}\frac{\sqrt{\pi}\,\Gamma(\beta+1)}{\Gamma(\alpha+\frac{3}{2})\Gamma(\beta-\alpha)}\,{}_2F_1\left[\begin{matrix}\alpha+\beta+1,\alpha-\beta+1\\ \alpha+\frac{3}{2}\end{matrix};\frac{1}{2}(1-G)\right],$$

$-1<\operatorname{Re}\alpha<\frac{1}{2}$, with

$$L(x,y,z;\beta,\beta)=0\quad\text{when } |f-g|<h<f+g,\quad -1<\operatorname{Re}\beta<\tfrac{1}{2}. \tag{3.31}$$

In closing this section let us summarize the results on the ultraspherical polynomials. Noting that

$$\begin{aligned}C_n^\lambda(x)&=\frac{(2\lambda)_n}{(\lambda+\frac{1}{2})_n}P_n^{(\lambda-1/2,\lambda-1/2)}(x),\qquad -1\le x\le 1,\\ D_n^\lambda(x)&=\frac{2}{\pi}\frac{(2\lambda)_n}{(\lambda+\frac{1}{2})_n}Q_n^{(\lambda-1/2,\lambda-1/2)}(x),\qquad -1<x<1,\\ h_n^{(\lambda-1/2,\lambda-1/2)}&=2^{1-2\lambda}(n+\lambda)\frac{\Gamma(n+2\lambda)n!}{\Gamma^2(n+\lambda+\frac{1}{2})},\end{aligned} \tag{3.32}$$

formulas (3.19), (3.21) and (3.31) can be written as

(3.33)

$$\sum_{n=0}^{\infty}(n+\lambda)\frac{\Gamma(n+1)}{\Gamma(2\lambda+n)}C_n^\lambda(\cos 2\phi)C_n^\lambda(\cos 2\psi)D_n^\lambda(\cos 2\theta)$$

$$=\frac{2^{2\lambda}}{\pi}\left[\frac{\Gamma(\lambda+\frac{1}{2})}{\Gamma(2\lambda)}\right]^2\begin{cases}0 & \text{if } |\cos(\phi+\psi)|<\cos\theta<\cos(\phi-\psi),\\ 2^{-6\lambda-1}\sin\pi\lambda(-D)^{-\lambda} & \text{if } \cos(\phi-\psi)<\cos\theta \text{ or}\\ & \quad 0<\cos\theta<\cos(\phi+\psi),\\ -2^{-6\lambda-1}\sin\pi\lambda(-D)^{-\lambda} & \text{if } 0<\cos\theta<-\cos(\phi+\psi),\end{cases}$$

where D is defined in (3.6), but can be given explicitly in terms of θ,ϕ,ψ as

$$(3.34)\qquad 16D=\sin(\theta+\phi+\psi)\sin(\theta+\phi-\psi)\sin(\theta+\psi-\phi)\sin(\phi+\psi-\theta).$$

Convergence of the series on the l.h.s. of (3.33) requires $-\frac{1}{2}<\operatorname{Re}\lambda<1$ except for some exceptional values of θ,ϕ and ψ.

In order that we may apply Hsü's [26] limiting procedure directly to our formulas we now replace θ,ϕ,ψ by $\alpha/2$, $(\pi-\beta)/2$, $(\pi-\gamma)/2$, respectively, in (3.33), use the duplication formula for the gamma function, and the symmetry property of the ultraspherical polynomials, namely, $C_n^\lambda(-x)=(-1)^n C_n^\lambda(x)$. Formula (3.33) then leads to (1.18) while (3.34) becomes (1.19).

4. Derivation of (1.23). Following Hsü we now set

$$(4.1)\qquad \alpha=ah,\quad \beta=bh,\quad \gamma=ch$$

in (1.18), where a,b,c,h are arbitrary positive numbers. If $\max(a,b,c)>(a+b+c)/2$, that is, a,b,c do not form the sides of a triangle then α,β,γ will not form the sides of a triangle. Obviously, this is the case that applies to (1.18). For small h

$$(4.2)\quad 2^{2-2\lambda}\{\Gamma(\lambda)\}^{-2}\cdot 2^{-2\lambda-1}\sin\pi\lambda(-16D)^{-\lambda}$$
$$\simeq 2^{-4\nu-1}\{\Gamma(\nu+\tfrac{1}{2})\}^{-2}\cos\pi\nu(-\Delta)^{-\nu-1/2}h^{-4\nu-2}$$

where $\lambda=\nu+\frac{1}{2}$, and Δ is the same as defined in (1.24).

Let us split the sum on the l.h.s. of (4.1) into two ranges: $0\le n\le[\omega/h]$ and $1+[\omega/h]\le n$, where ω is an arbitrary positive number independent of h. From (1.21) and (1.22) we have, for $nz=O(1)$,

$$(4.3)\quad \begin{aligned} n^{-2\nu}C_n^{\nu+1/2}(\cos z)&=\frac{\sqrt{\pi}}{\Gamma(\nu+\frac{1}{2})}(2nz)^{-\nu}J_\nu(nz)+\varepsilon_n,\\ n^{-2\nu}D_n^{\nu+1/2}(\cos z)&=-\frac{\sqrt{\pi}}{\Gamma(\nu+\frac{1}{2})}(2nz)^{-\nu}Y_\nu(nz)+\varepsilon_n',\end{aligned}$$

where $\varepsilon_n \to 0$, $\varepsilon'_n \to 0$ as $n \to \infty$. Hence

$$(4.4)\qquad n^{-6\nu}C_n^{\nu+1/2}(\cos\beta)C_n^{\nu+1/2}(\cos\gamma)D_n^{\nu+1/2}(\cos\alpha)$$
$$= -\pi^{3/2}\{\Gamma(\nu+\tfrac{1}{2})\}^{-3}(2nh)^{-3\nu}(abc)^{-\nu}J_\nu(bnh)J_\nu(cnh)Y_\nu(anh)+\eta_n$$

where $\eta_n \to 0$ as $n \to \infty$. Also

$$(4.5)\qquad (n+\nu+\tfrac{1}{2})\frac{\Gamma(n+1)}{\Gamma(n+2\nu+1)} = n^{1-2\nu}(1+\eta'_n)$$

where $\eta'_n = O(n^{-1})$ as $n \to \infty$.

Thus the summation over the first range contributes

(4.6)

$$h^{4\nu+2}\sum_{n=0}^{[\omega/h]}(n+\nu+\tfrac{1}{2})\frac{\Gamma(n+1)}{\Gamma(n+2\nu+1)}C_n^{\nu+1/2}(\cos\beta)C_n^{\nu+1/2}(\cos\gamma)D_n^{\nu+1/2}(\cos\alpha)$$
$$= -\pi^{3/2}\{\Gamma(\nu+\tfrac{1}{2})\}^{-3}(8abc)^{-\nu}h\sum_{n=0}^{[\omega/h]}(nh)^{1+\nu}J_\nu(bnh)J_\nu(cnh)Y_\nu(anh)(1+\eta'_n)$$
$$+h\sum_{n=0}^{[\omega/h]}(nh)^{4\nu+1}\eta_n(1+\eta'_n).$$

The principal part of (4.6), namely,

$$(4.7)\quad h\sum_{n=0}^{[\omega/h]}(nh)^{1+\nu}J_\nu(bnh)J_\nu(cnh)Y_\nu(anh) \to \int_0^\omega x^{1+\nu}J_\nu(bx)J_\nu(cx)Y_\nu(ax)\,dx$$

as $h \to 0$, provided the integral converges which, for fixed ω, requires that $\operatorname{Re}\nu > -\frac{1}{2}$. Since $(nh)^{1+\nu}J_\nu(bnh)J_\nu(cnh)Y_\nu(anh) = (nh)^{1+2\nu}O(1)$, the remainder part clearly approaches 0 as $h \to 0$ if $\operatorname{Re}\nu > -\frac{1}{2}$.

In order to deal with the second range of n for the sum in (1.18) we need the following asymptotic formulas

(4.8)

$$C_n^{\nu+1/2}(\cos\theta) = \frac{2}{\Gamma(\nu+\frac{1}{2})}(2\sin\theta)^{-\nu-1/2}n^{\nu-1/2}\left[\cos\{(n+\nu+\tfrac{1}{2})\theta - \tfrac{1}{2}\nu\pi - \tfrac{1}{2}\pi\} + R\right],$$
$$|R| < C(n\sin\theta)^{-1}$$

and

$$(4.9)\quad D_n^{\nu+1/2}(\cos\theta) = \frac{2}{\Gamma(\nu+\frac{1}{2})}(2\sin\theta)^{-\nu-1/2}\left[\cos\{(n+\nu+\tfrac{1}{2})\theta - \tfrac{1}{2}\nu\pi + \tfrac{1}{4}\pi\} + R'\right],$$
$$|R'| < C'(n\sin\theta)^{-1},$$

where $0 < \theta < \pi$, $-\frac{1}{2} < \nu < \frac{1}{2}$ and C, C' are independent of n and θ. The first of these formulas is due to Stieltjes [26, Equation (4.2)]; see also [29, Equation (8.21.18), p. 198]. The second formula (4.9) does not seem to be explicitly given in the literature, but can be easily deduced by using the known asymptotic formula for Legendre functions of the second kind [18, 3.9.1(1), p. 162] and its relation with the ultraspherical function. Now

(4.10)

$$h^{4\nu+2}n^{1-2\nu}(1+\eta_n')C_n^{\nu+1/2}(\cos\beta)C_n^{\nu+1/2}(\cos\gamma)D_n^{\nu+1/2}(\cos\alpha)$$

$$=\left\{\frac{2^{1/2}-\nu}{\Gamma(\nu+\frac{1}{2})}\right\}^3 h^{4\nu+2}(\sin\alpha\sin\beta\sin\gamma)^{-\nu-1/2}n^{\nu-1/2}\left\{\cos\left(N\beta-\frac{\nu\pi}{2}-\frac{\pi}{4}\right)+\rho_1\right\}$$

$$\cdot\left\{\cos\left(N\gamma-\frac{\nu\pi}{2}-\frac{\pi}{4}\right)+\rho_2\right\}\left\{\cos\left(N\alpha-\frac{\nu\pi}{2}+\frac{\pi}{4}\right)+\rho_3\right\}$$

$$=\left\{\frac{2^{1/2-\nu}}{\Gamma(\nu+\frac{1}{2})}\right\}^3 h^{4\nu+2}(\sin\alpha\sin\beta\sin\gamma)^{-\nu-1/2}n^{\nu-1/2}$$

$$\cdot\left\{\cos\left(N\beta-\frac{\nu\pi}{2}-\frac{\pi}{4}\right)\cos\left(N\gamma-\frac{\nu\pi}{2}-\frac{\pi}{4}\right)\cos\left(N\alpha-\frac{\nu\pi}{2}+\frac{\pi}{4}\right)+\rho\right\},$$

$$N=n+\nu+\tfrac{1}{2},$$

where ρ_1,ρ_2,ρ_3 and consequently ρ have the same type of bounds as R and R' have in (4.8) and (4.9). Using $\eta_n'=O(n^{-1})$ we find that $\rho=(nh)^{-1}O(1)$.

Since

$$4\cos\left(N\beta-\frac{\nu\pi}{2}-\frac{\pi}{4}\right)\cos\left(N\gamma-\frac{\nu\pi}{2}-\frac{\pi}{4}\right)\cos\left(N\alpha-\frac{\nu\pi}{2}+\frac{\pi}{4}\right)$$

$$=\cos\left\{N(\beta+\gamma+\alpha)-\frac{3\nu\pi}{2}-\frac{\pi}{4}\right\}+\cos\left\{N(\beta+\nu-\alpha)-\frac{\nu\pi}{2}-\frac{3\pi}{4}\right\}$$

$$+\cos\left\{N(\beta-\gamma+\alpha)-\frac{\nu\pi}{2}+\frac{\pi}{4}\right\}+\cos\left\{N(\gamma-\beta+\alpha)-\frac{\nu\pi}{2}+\frac{\pi}{4}\right\},$$

the principal term in the summation over (4.10) on the second range reduces to an expression of the type

$$p^{\nu+1/2}\sum_{n=1+[\omega/h]}^{\infty} n^{\nu-1/2}\cos(nq+\tau) \tag{4.11}$$

where p and q are $\sim h$ and τ is a fixed constant. The bound for (4.11) is

$$p^{\nu+1/2}(\omega h^{-1})^{\nu-1/2}q^{-1}=\omega^{\nu-1/2}O(1). \tag{4.12}$$

The contribution of the remainder term is

$$p^{\nu+1/2}\sum_{n=1+[\omega/h]}^{\infty} n^{\nu-1/2}(nh)^{-1}=O(1)h^{\nu-1/2}\sum_{n=1+[\omega/h]}^{\infty} n^{\nu-3/2}=\omega^{\nu-1/2}O(1). \tag{4.13}$$

Therefore, as $h \to 0$,

(4.14)

$$h^{4\nu+2}\sum_{n=0}^{\infty}\left(n+\nu+\tfrac{1}{2}\right)\frac{\Gamma(n+1)}{\Gamma(n+2\nu+1)}C_n^{\nu+1/2}(\cos\beta)C_n^{\nu+1/2}(\cos\gamma)D_n^{\nu+1/2}(\cos\alpha)$$

$$= -\pi^{3/2}\left\{\Gamma\left(\nu+\tfrac{1}{2}\right)\right\}^{-3}(8abc)^{-\nu}\int_0^{\omega}x^{1+\nu}Y_\nu(ax)J_\nu(bx)J_\nu(cx)\,dx+o(1)+\omega^{\nu-1/2}O(1)$$

provided $-\frac{1}{2}<\nu<\frac{1}{2}$.

Using (1.18), (4.1), (4.2) and (4.14) we then have, for $\omega\to\infty$,

(4.15)

$$\int_0^{\infty}x^{1+\nu}Y_\nu(ax)J_\nu(bx)J_\nu(cx)\,dx$$

$$=\begin{cases}0 \quad \text{if } |b-c|<a<b+c,\\ \dfrac{2^{-\nu-1}\Gamma(\nu+\frac{1}{2})\cos\pi\nu}{\pi^{3/2}(abc)^{-\nu}}(-\Delta)^{-\nu-1/2} & \text{if } a>b+c,\\ -\dfrac{2^{-\nu-1}\Gamma(\nu+\frac{1}{2})\cos\pi\nu}{\pi^{3/2}(abc)^{-\nu}}(-\Delta)^{-\nu-1/2} & \text{if } a<|b-c|.\end{cases}$$

This immediately leads to (1.23) when we use $\Gamma(\frac{1}{2}-\nu)\Gamma(\frac{1}{2}+\nu)=\pi\sec\pi\nu$.

Although the asymptotic formulas (4.8) and (4.9) that we had to use to derive (4.12) and (4.13) are valid for real ν and $-\frac{1}{2}<\nu<\frac{1}{2}$, by analytic continuation, equation (4.15) obviously holds true also for complex ν with $-\frac{1}{2}<\operatorname{Re}\nu<\frac{1}{2}$.

REFERENCES

[1] R. Askey, *Orthogonal polynomials and positivity*, in Studies in Applied Mathematics 6, Special Functions and Wave Propagation, D. Ludwig and F. W. J. Olver, eds., Society for Industrial and Applied Mathematics, Philadelphia, 1970, pp. 64–85.

[2] ______, *Linearization of the product of orthogonal polynomials*, in Problems in Analysis, R. Gunning, ed., Princeton Univ. Press, Princeton, NJ, 1970, pp. 223–228.

[3] ______, *Positive Jacobi polynomial sums*, Tohoku Math. J., 24 (1972), pp. 109–119.

[4] ______, *Summability of Jacobi series*, Trans. Amer. Math. Soc., 179 (1973), pp. 71–84.

[5] ______, *Orthogonal Polynomials and Special Functions*, CBMS Regional Conference Series in Applied Mathematics, 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[6] ______, *An integral of products of Legendre functions and a Clebsch–Gordon sum*, Lett. Math. Phys., 6 (1982), pp. 299–302.

[7] R. Askey and J. Fitch, *Integral representatio for Jacobi polynomials and some applications*, J. Math. Anal. Appl., 26 (1969), pp. 411–437.

[8] R. Askey and G. Gasper, *Linearization of the product of Jacobi polynomials* III, Canad. J. Math., 23 (1971), pp. 332–338.

[9] ______, *Jacobi polynomial expansions of Jacobi polynomials with nonnegative coefficients*, Proc. Camb. Phil. Soc., 70 (1971), pp. 243–255.

[10] R. Askey, T. H. Koornwinder and Mizan Rahman, *An integral of products of ultraspherical functions and a q-extension*, submitted.

[11] R. Askey and S. Wainger, *A convolution structure for Jacobi series*, Amer. J. Math., 91 (1969), pp. 463–485.

[12] W. N. Bailey, *Generalized Hypergeometric Series*, Stechert-Hafner Service Agency, New York and London, 1964.

[13] A. M. DIN, *A simple sum formula for Clebsch–Gordon coefficients*, Lett. Math. Phys., 5 (1981), pp. 207–211.

[14] A. M. DIN AND W. J. ZAHRZEWSKI, *Stability properties of classical solutions to nonlinear σ-models*, Nucl. Phys. B168 (1980), pp. 173–180.

[15] J. DOUGALL, *A theorem of Sonine in Bessel functions, with two extensions to spherical harmonies*, Proc. Edin. Math. Soc., 37 (1919), pp. 33–47.

[16] L. DURAND, *Product formulas and Nicholson-type integrals for Jacobi functions* I, *Summary of results*, this Journal, 9 (1978), pp. 76–86.

[17] A. R. EDMONDS, *Angular Momentum in Quantum Mechanics*, 2nd edition, Princeton Univ. Press, Princeton, NJ, 1960.

[18] A. ERDÉLYI, ed., *Higher Transcendental Functions, Vol.* I, McGraw-Hill, New York, 1953.

[19] ———, *Higher Transcendental Functions, Vol.* II, McGraw-Hill, New York, 1953.

[20] ———, *Tables of Integral Transforms, Vol.* II, McGraw-Hill, New York, 1954.

[21] E. FELDHEIM, *Contributions à la théorie des polynômes de Jacobi*, Mat. Fiz. Lapok, 48 (1941), pp. 453–504. (In Hungarian, French summary.)

[22] G. GASPER, *Linearization of the product of Jacobi polynomials* I, Canad. J. Math., 22 (1970), pp. 171–175.

[23] ———, *Linearization of the product of Jacobi polynomials* II, Canad. J. Math., 22 (1970), pp. 582–593.

[24] ———, *Positivity and the convolution structure for Jacobi series*, Ann. Math., 93 (1971), pp. 112–118.

[25] ———, *Banach algebras for Jacobi series and positivity of a kernel*, Ann. Math., 95 (1972), pp. 261–280.

[26] H.-Y. HSÜ, *Certain integrals and infinite series involving ultraspherical polynomials and Bessel functions*, Duke Math. J., 4 (1938), pp. 374–383.

[27] MIZAN RAHMAN, *On a generalization of the Poisson kernel for Jacobi polynomials*, this Journal, 8 (1977), pp. 1014–1031.

[28] ———, *A nonnegative representation of the linearization coefficient of the product of Jacobi polynomials*, Canad. J. Math., 33 (1981), pp. 915–928.

[29] G. SZEGÖ, *Orthogonal Polynomials*, AMS Colloquium Publications 23, 4th edition, American Mathematical Society, Providence, RI, 1975.

[30] G. N. WATSON, *Theory of Bessel Functions*, 2nd edition, Cambridge Univ. Press, Cambridge, 1944.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

017

MULTIPLE HYPERGEOMETRIC FUNCTIONS AND SIMPLE LIE GROUPS SL AND Sp*

JAN HRABOWSKI[†]

Abstract. With any multiple hypergeometric series one can associate a Lie group called its group of symmetries. This article is devoted to those series which group of symmetries contains a subgroup isomorphic to the special linear group **SL** or to the sympletic group **Sp**. The work owes much to the group-theoretic analysis of the Lauricella series F_D in Miller [J. Math. Phys., 13 (1972), pp. 1393–1399], *Symmetry and the Separation of Variables*, Addison-Wesley, Reading, MA, 1977.

1. Introduction. We assume that the reader is familiar with the concept of a Lie group/algebra of symmetries of a system of differential equations (see [7] for the general definition and [6] for the special case of linear systems in one unknown to which we restrict ourselves here). We will also use freely a standard terminology of the Lie theory (see e.g. [1]).

The theme of this article has its origin in the following observations:

(i) *With any hypergeometric series*[1] *one can associate a constant coefficient system, termed canonical* [4].

(ii) *The canonical system of the Lauricella series F_D admits a Lie algebra of symmetries of classical type A* [5], [6].

These observations led W. Miller, Jr. to conclude that "...the use of **SL**-symmetry (...) will lead inevitably to the remarkable function F_D." [5]. While this emphasis on F_D seems to us premature we agree that series with **SL** and other classical groups of symmetry with their rich transformation properties and differential recurrence relations are indeed remarkable.

The classification of these series is carried out in §§2–3. Some of their elementary properties are studied in §§4–6. In §7 we present an interplay between canonical systems and hypergeometric series which is equivalent to that in [4].

1.1. Notation. **N**, **Z**, **Q**, **R**, **C** denote the sets of natural, integral, rational, real and complex numbers. $\mathfrak{S}_n$ is the symmetric group of an n-element set. Lie algebras are defined over **C**. Functions and coordinate systems are complex valued. A family (f.ex. $(x_i)_{i \in I}$) is often denoted by the corresponding bold face character (**x** in this case). Also, $\mathbf{1} = (1, 1, \cdots)$. δ_{ij} is the Kronecker delta ($=1$ if $i=j$ and 0 otherwise). The family $(\delta_{ij})_j$ is denoted by $\mathbf{1}_i$.

If H is a group acting on a group N then $H \ltimes N$ denotes their semidirect product.

2. Canonical systems. Our approach to canonical systems, while equivalent to, differs from that in [4]. Let Ω be a finite set of generators of a free abelian group F. Let $(x_\omega)_{\omega \in \Omega}$ be a local coordinate system indexed by Ω and ∂_{x_ω} be the corresponding partial derivatives. A monomial $\prod_\omega \partial_{x_\omega}^{n_\omega}$ will be abbreviated to $\partial_{\mathbf{x}}^{\mathbf{n}}$.

DEFINITION. The *canonical system of* Ω (or of Ω, F) is the system

$$\partial_{\mathbf{x}}^{\mathbf{n}^+} - \partial_{\mathbf{x}}^{\mathbf{n}^-} \tag{2.1}$$

*Received by the editors March 8, 1983, and in final revised form October 9, 1984.

[†] 385 Grand Street, New York, New York 10002.

[1] A formal power series $\sum a_{\mathbf{n}} z_1^{n_1} \cdots z_m^{n_m}$ is called *hypergeometric* if $a(n_1, \cdots, n_i + 1, \cdots, n_m)/a_{\mathbf{n}} = P_i(\mathbf{n})$ where P_i is a rational function in m variables, $1 \leqq i \leqq m$.

where $\mathbf{n}=\mathbf{n}^+-\mathbf{n}^-$, $\mathbf{n}^\pm \geqq 0$, runs through the group $\rho(\Omega)=\{\mathbf{n}\in\mathbf{Z}^\Omega;\ \Sigma_\omega n_\omega \omega=0\}$ of $\mathbf{Z}$-linear dependence relations in Ω.

Open problem. Find a finite set of generators of the ideal generated by (2.1).

3. Canonical systems with symmetry Lie algebra of classical type. Any linear, constant coefficient system, in particular (2.1), is invariant under the abelian Lie algebra spanned by ∂_{x_ω}, $\omega\in\Omega$, and the trivial symmetry 1. Let $\mathfrak{h}$ denote the abelian Lie algebra of those symmetries of (2.1) which have the form

$$H=\sum_{\omega\in\Omega} a_\omega x_\omega \partial_{x_\omega}, \qquad a_\omega \in \mathbf{C}. \tag{3.1}$$

It is easy to check that $H\in\mathfrak{h}$ if and only if $\Sigma_\omega a_\omega n_\omega=0$ for all $\mathbf{n}\in\rho(\Omega)$. Note that

$$\mathfrak{h}^* \cong \mathbf{C}\otimes_{\mathbf{Z}} F. \tag{3.2}$$

The canonical systems (2.1) which interest us here have the following property (S):

(S_1) *The system is invariant under a simple Lie algebra* $\mathfrak{g}$.

(S_2) $\mathfrak{h}\oplus\mathbf{C}.1$ *is a Cartan subalgebra of* $\mathfrak{g}\oplus\mathbf{C}.1$.

(It follows that $\tilde{\mathfrak{h}}=(\mathfrak{h}\oplus\mathbf{C})\cap\mathfrak{g}$ is a Cartan subalgebra of $\mathfrak{g}$). In view of (3.2) we may identify Ω with a set of generators of $\mathfrak{h}^*$ (or $\tilde{\mathfrak{h}}^*$). Let $R\subset\tilde{\mathfrak{h}}^*$ be the root system of $(\mathfrak{g},\tilde{\mathfrak{h}})$. Since ∂_{x_ω} is an ω-root vector we have $\Omega\subset R$.

LEMMA. *If* (S) *holds then*

(a) $(\Omega+\Omega)\cap(R\cup\{0\})=\varnothing$,

(b) $(\mathbf{C}\Omega+\mathbf{C}\Omega)\cap R\subset\mathbf{C}\Omega$.

Proof. If $A\subset R$ then $\mathfrak{g}^A$ will denote the vector space spanned by the α-root vectors, $\alpha\in A$. The condition (a) means that $\mathfrak{g}^\Omega=\operatorname{span}_{\mathbf{C}}(\partial_{x_\omega})$ is an abelian subalgebra [1, Ch. 7, no. 1.3, Prop. 10]. Let $\mathfrak{n}$ be the *isotropy subalgebra* of $\mathfrak{g}$ relative to the point $\mathbf{x}=\mathbf{0}$. ($X\in\mathfrak{g}$ is in $\mathfrak{n}$ if its first order term vanishes at $\mathbf{x}=\mathbf{0}$.) Since $\tilde{\mathfrak{h}}\subset\mathfrak{n}$ we have $\mathfrak{n}=\mathfrak{g}^{\tilde\Omega}\oplus\tilde{\mathfrak{h}}$, $(\tilde\Omega+\tilde\Omega)\cap R\subset\tilde\Omega$, [1, Ch. 8, no. 3.1, Prop. 1]. On the other hand $\mathfrak{n}\oplus\mathfrak{g}^\Omega=\mathfrak{g}$ hence $\tilde\Omega=\mathbf{C}\Omega$. Q.E.D.

Conditions (a), (b) imply that $\mathbf{C}\Omega$ is a *parabolic subset* of R [1, Ch. 8, no. 3.4]. Parabolic subsets of classical type (A, B, C or D) have been listed in [1, Ch. 8, §13(III)]. Those which satisfy (a), (b) are listed in Table 1. We use standard realizations of classical root systems R in a vector space with basis (ε_i) [1, Ch. 6, §4 or Ch. 8, §13].

TABLE 1

Sets Ω of classical type.

Type	R	Ω	card(Ω)
A_l[2]	$\varepsilon_i-\varepsilon_j;\ i,j\in\{1,\cdots,p,-q,\cdots,-1\}$ $i\neq j,\ p+q=l+1,\ 1\leqq p\leqq q$	$\Omega^A_{l,p}=\{\omega_{ij}=\varepsilon_i-\varepsilon_{-j};$ $1\leqq i\leqq p,\ 1\leqq j\leqq q\}$	pq
B_l	$\pm\varepsilon_i,\ \pm\varepsilon_j\pm\varepsilon_k;\ 1\leqq i,j,k\leqq l, j\neq k$	$\Omega^B_l=\{\omega_{\pm i}=\varepsilon_1\pm\varepsilon_{i+1};$ $1\leqq i\leqq l-1\}\cup\{\omega_0=\varepsilon_1\}$	$2l-1$
C_l	$\pm 2\varepsilon_i,\ \pm\varepsilon_j\pm\varepsilon_k;\ 1\leqq i,j,k\leqq l,$ $j\neq k$	$\Omega^C_l=\{\omega_{ij}=\omega_{ji}=\ \varepsilon_i+\varepsilon_j;$ $1\leqq i,j\leqq l\}$	$l(l+1)/2$
D_l $l\geqq 3$	$\pm\varepsilon_i\pm\varepsilon_j;\ 1\leqq i\neq j\leqq l$	$\Omega^D_{l,1}=\{\omega_{\pm i}=\varepsilon_1\pm\varepsilon_{i+1};$ $1\leqq i\leqq l-1\}$	$2l-2$
		$\Omega^D_l=\{\omega_{ij}=\varepsilon_i+\varepsilon_j;\ i\neq j\}$	$l(l-1)/2$

[2] There are $[l/2]$ distinct subsets Ω in the root system of type A_l. It is convenient to use a different indexing of R in each case.

Between these sets we have the following isomorphisms:

$$\Omega^A_{1,1} \cong \Omega^B_1 \cong \Omega^C_1,\ \Omega^B_2 \cong \Omega^C_2,\ \Omega^A_{3,1} \cong \Omega^D_3,\ \Omega^A_{3,2} \cong \Omega^D_{3,1},\ \Omega^D_4 \cong \Omega^D_{4,1}. \tag{3.3}$$

Note that $\rho(\Omega^A_{l,1}) = \{0\}$ which means that the canonical systems are trivial in this case.

Table 2 lists canonical systems of the sets Ω in Table 1. Only second order elements of (2.1), which suffice to generate the ideal, are given. If $\Omega = \{\omega_i\}_{i\in I}$ then the coordinate system (x_{ω_i}) will be abbreviated to (x_i).

TABLE 2

Canonical systems of classical type

Ω	Canonical system
$\Omega^A_{l,p}$	$\partial_{x_{ar}}\partial_{x_{bs}} - \partial_{x_{as}}\partial_{x_{br}}$; $1 \leqq a, b \leqq p, 1 \leqq r, s \leqq q$
Ω^B_l	$\partial_{x_i}\partial_{x_{-i}} - \partial^2_{x_0}$; $1 \leqq i \leqq l-1$
Ω^C_l	Same as $\Omega^A_{l,p}$ but with $1 \leqq a, b, r, s \leqq l$. Note that $x_{ij} = x_{ji}$
$\Omega^D_{l,1}$	$\partial_{x_i}\partial_{x_{-i}} - \partial_{x_j}\partial_{x_{-j}}$; $1 \leqq i, j \leqq l-1$
Ω^D_l	Same as Ω^C_l but with $a \neq r, s$ and $b \neq r, s$

Examples. Sets $\Omega^A_{l,2}$ lead to functions named (unfortunately) F_D by Lauricella [6]. In the particular case $l = 3$ F_D is the ordinary hypergeometric function ${}_2F_1$ with canonical equation

$$\partial_{x_{11}}\partial_{x_{22}} - \partial_{x_{12}}\partial_{x_{21}}. \tag{3.4}$$

$\Omega^A_{5,3}$ leads to the quadruple function K_{16} of Exton [3]. Ω^C_2 leads to the Gogenbauer function [2] with canonical equation

$$\partial_{x_{11}}\partial_{x_{22}} - \partial^2_{x_{12}}. \tag{3.5}$$

Sets $\Omega^D_{l,1}$ lead to functions named (unfortunately) F_C by Lauricella.

It remains to determine which systems in Table 2 satisfy (S). One can easily show that any transformation which leaves invariant the system of Ω^B_l or $\Omega^D_{l,1}$ has to permute the lines $\mathbf{C}.\partial_{x_\omega}$, $\omega \in \Omega$. This implies that these systems have no symmetries other than those mentioned at the beginning of this section. In view of the last isomorphism in (3.3) the system of Ω^D_l has to be excluded as well. It is known that $\Omega^A_{l,2}$ and Ω^C_2 satisfy (S) [5], [6]. It turns out that all systems of type A and C satisfy (S).

3.1. Verification that systems of type A and C satisfy (S). The verification will consist of 3 steps:

(I) Choosing a basis of the (abstract) Lie algebra $\mathfrak{g}$.

(II) Finding all Lie algebras of first order differential operators isomorphic to $\mathfrak{g}$ such that $\mathfrak{g}^\omega$ is spanned by ∂_{x_ω} and $\mathfrak{g}^{C\Omega}$ is isotropic at $\mathbf{x} = \mathbf{0}$. These algebras depend on a complex parameter c (c.f. §§6.1–2).

(III) Determining the parameter c for which $\mathfrak{g}$ is an algebra of symmetries. (The straightforward calculations of this step are omitted and only the result is stated).

Type A. (I) Let $(\hat{E}_{ij})$ be the canonical basis of $\mathfrak{gl}(l+1)$ and (E_{ij}) be its projection on $\mathfrak{sl}(l+1)$. Thus, $\Sigma E_{ii} = 0$. The nonzero brackets are

$$\begin{aligned} [E_{ij}, E_{jk}] &= E_{ik} \qquad (i \neq k), \\ [E_{ij}, E_{ji}] &= E_{ii} - E_{jj}. \end{aligned} \tag{3.1.1}$$

(II) We use $(1,\cdots,p,-1,\cdots,-q)$ to index the canonical basis of $\mathbf{C}^{l+1}$.

$$\tag{3.1.2}\begin{aligned}
&E_{i,-j}=\partial_{x_{ij}},\\
&E_{ij}=-\sum_k x_{jk}\partial_{x_{ik}}+c_+\delta_{ij},\ c_+=cq/(l+1),\\
&E_{-i,-j}=\sum_k x_{ki}\partial_{x_{kj}}+c_-\delta_{ij},\ c_-=-cp/(l+1),\\
&E_{-i,j}=-\sum_{k,r} x_{ki}x_{jr}\partial_{x_{kr}}+cx_{ji},\ c=c_+-c_-.
\end{aligned}$$

(III) $c=-1$.

Type C. (I) $\mathfrak{g}\cong\mathfrak{sp}(2l)$ has a basis $\{X_{\pm i,\pm j};\ 1\leqq i,j\leqq l\}$. The nonzero brackets are

$$\tag{3.1.3}\begin{aligned}
&[X_{ij},X_{-j,-k}]=(1+\delta_{ij})(1+\delta_{jk})X_{i,-k}\quad\text{if } i\neq k,\\
&[X_{ij},X_{-j,-i}]=X_{i,-i}-X_{j,-j},\\
&[X_{i,-j},X_{jk}]=(1+\delta_{jk})X_{ik},\\
&[X_{-i,-j},X_{j,-k}]=(1+\delta_{ij})X_{-i,-k},\\
&X_{i,-j}\text{ combine as }E_{ij}\text{ in (3.1.1).}
\end{aligned}$$

(II)

$$\tag{3.1.4}\begin{aligned}
&X_{ij}=\partial_{x_{ij}},\\
&X_{i,-j}=-\sum_k(1+\delta_{kj})x_{jk}\partial_{x_{ik}}+c\delta_{ij},\\
&X_{-i,-j}=-\sum_{k,r}(1+\delta_{ik})(1+\delta_{jr})x_{ik}x_{jr}\partial_{x_{kr}}+2c(1+\delta_{ij})x_{ij}.
\end{aligned}$$

(III) $c=-\frac{1}{2}$.

4. Hypergeometric system. Let us recall the concept of *separation of variables* relative to an abelian Lie algebra [6]. Let $\mathcal{S}$ be a system of linear differential operators on $\mathbf{C}^n$ and $\mathfrak{h}$ be an abelian Lie algebra of symmetries of $\mathcal{S}(1\notin\mathfrak{h})$. The separation of $\mathcal{S}$ relative to $\mathfrak{h}$ is the following system of operators on the product space $\mathfrak{h}^*\times\mathbf{C}^n$:

$$\tag{4.1}\begin{aligned}
&\text{(a)}\quad 1\otimes\mathcal{S},\\
&\text{(b)}\quad 1\otimes H-h_H\otimes 1,\qquad H\in\mathfrak{h},
\end{aligned}$$

where h_H is the (linear) function on $\mathfrak{h}^*$ defined by $h_H(\alpha)=\langle H,\alpha\rangle$. Note that the system (4.1) is linear not only over constants but also over functions defined on $\mathfrak{h}^*$.

DEFINITION. A *hypergeometric system* of Ω is the separation of the canonical system (2.1) relative to an algebra $\tilde{\mathfrak{h}}=\{H+\langle H,\alpha_0\rangle;\ H\in\mathfrak{h}\}$, where $\mathfrak{h}$ is defined by (3.1) and $\alpha_0\in\mathfrak{h}^*$.

Open problem. Find the dimension of the space of solutions of a hypergeometric system.

Remark. There is no essential loss of generality in setting $\alpha_0=0$. Let τ_{α_0}; $\mathfrak{h}^*\to\mathfrak{h}^*$ be the translation $\alpha\mapsto\alpha+\alpha_0$. If a function f on $\mathfrak{h}^*\times\mathbf{C}^n$ satisfies the separation of $\mathcal{S}$ relative to $\mathfrak{h}$ then $f^{\tau_{-\alpha_0}}$ satisfies the separation relative to $\tilde{\mathfrak{h}}$. However for the canonical systems of $\Omega^A_{l,p}$ and Ω^C_l we choose $\tilde{\mathfrak{h}}$ to be a Cartan subalgebra of the simple Lie algebra of symmetries.

5. Recurrence relations. Consider the separation (4.1) of a system $\mathscr{S}$ relative to $\mathfrak{h}$. If $\omega \in \mathfrak{h}^*$ then $\tau_\omega\colon \mathfrak{h}^* \to \mathfrak{h}^*$ will denote the translation $\tau_\omega(\alpha) = \alpha + \omega$. τ_ω transforms functions on $\mathfrak{h}^*$ and, in particular,

$$\tau_\omega \cdot h_H = h_H + \langle \omega, H \rangle, \qquad H \in \mathfrak{h}. \tag{5.1}$$

LEMMA. *Let $\omega \in \mathfrak{h}^*$ and X_ω be a symmetry of $\mathscr{S}$ which is also an ω-root vector. Then $\tau_\omega^{-1} \otimes X_\omega$ leaves invariant the separated system* (4.1).

Proof. Since X_ω is a symmetry of $\mathscr{S}$, $\tau_\omega^{-1} \otimes X_\omega$ is a symmetry of $1 \otimes \mathscr{S}$. If f is a function on $\mathfrak{h}^*$ then

$$(h_H \circ \tau_{-\omega} - \tau_{-\omega} \circ h_H) \cdot f = h_H \cdot (f^{\tau_{-\omega}}) - (h_H - \langle \omega, H \rangle) f^{\tau_{-\omega}} = \langle \omega, H \rangle f^{\tau_{-\omega}}.$$

Hence,

$$[1 \otimes H - h_H \otimes 1, \tau_{-\omega} \otimes X_\omega] = \langle \omega, H \rangle \tau_{-\omega} \otimes X_\omega - [h_H, \tau_{-\omega}] \otimes X_\omega = 0. \quad \text{Q.E.D.}$$

Let $\mathfrak{g}$ be a Lie algebra of symmetries of $\mathscr{S}$ which admits the root decomposition $\mathfrak{g} = \oplus_{\omega \in \mathfrak{h}^*} \mathfrak{g}^\omega$ relative to $\mathfrak{h}$. Then the map

$$\sum_\omega X_\omega \mapsto \sum_\omega \tau_\omega^{-1} \otimes X_\omega, \qquad X_\omega \in \mathfrak{g}^\omega, \tag{5.2}$$

is a homomorphism of $\mathfrak{g}$ into the Lie algebra of $\mathbf{C}$-linear endomorphisms on the space of functions defined on $\mathfrak{h}^* \times \mathbf{C}^n$. The endomorphisms $\tau_\omega^{-1} \otimes X_\omega$ are called *recurrences* of the separated system. Note that a recurrence $\tau_\omega^{-1} \otimes X_\omega$ is τ_ω^{-1}-semilinear.

Example. Consider the system F_C ($\Omega_{l,1}^D$ in Table 2). A basis of $\mathfrak{h}$ is given by

$$H_i = x_i \partial_{x_i} - x_{-i} \partial_{x_{-i}}, \qquad i = 1, \cdots, l-1, \quad H_l = \sum_{\pm i} x_{\pm i} \partial_{x_{\pm i}}. \tag{5.3}$$

If $R_{\pm i}$ is the recurrence corresponding to $\partial_{x_{\pm i}}$ and h_i denotes h_{H_i},

$$R_{\pm i} \cdot f(\mathbf{h}, \mathbf{x}) = \partial_{x_{\pm i}} f(\cdots, h_i \pm 1, \cdots, h_l + 1, \mathbf{x}). \tag{5.4}$$

6. Transformations. Let $\mathscr{S}$ be a linear system in the variables (x_i) and $\mathfrak{h}$ be an abelian Lie algebra of symmetries $(1 \notin \mathfrak{h})$. Let $\sigma_\mathbf{x}$ be a transformation of $\mathbf{C}^n$ which preserves $\mathscr{S}$ and $\mathfrak{h}$. Let $\sigma_\mathfrak{h}$ be the corresponding transformation of $\mathfrak{h}^*$. It is clear that the transformation $(\sigma_\mathfrak{h}, \sigma_\mathbf{x})$ of $\mathfrak{h}^* \times \mathbf{C}^n$ is a symmetry of the separated system $\mathscr{S}_\mathfrak{h}$.

Let Ω be a set of generators of a free $\mathbf{Z}$-module. Let $A(\Omega)$ be the group of *automorphisms* of Ω, i.e., the group of permutations of Ω induced by $\mathbf{Z}$-linear transformations. If $\sigma \in A(\Omega)$ then

$$\sigma_\mathbf{x}(x_\omega) = x_{\sigma(\omega)} \tag{6.1}$$

is an $\mathfrak{h}$-preserving symmetry of the canonical system of Ω.

Example. Consider the canonical system of the Example in §5. Let S_0 be the group $\{1, -1\}$ (the 0-sphere). $A(\Omega_{l,1}^D) = \mathfrak{S}_{l-1} \ltimes \mathrm{S}_0^{l-1}$, where

$$\begin{aligned} \sigma \cdot \omega_{\pm i} &= \omega_{\pm \sigma(i)}, & \sigma &\in \mathfrak{S}_{l-1}, \\ \boldsymbol{\varepsilon} \cdot \omega_{\pm i} &= \omega_{\pm \varepsilon_i i}, & \boldsymbol{\varepsilon} &= (\varepsilon_i) \in \mathrm{S}_0^{l-1}. \end{aligned} \tag{6.2}$$

The corresponding transformation of the hypergeometric system are

$$\begin{aligned} &\boldsymbol{\sigma}\cdot f(\mathbf{h},\mathbf{x})=f\big(h_{\sigma(1)},\cdots,h_{\sigma(l-1)},h_l,x_{\sigma(1)},\cdots,x_{-\sigma(l-1)}\big),\\ &\boldsymbol{\varepsilon}\cdot f(\mathbf{h},\mathbf{x})=f\Big(\varepsilon_1 h_1,\cdots,\varepsilon_{l-1}h_{l-1},h_l,x_{\varepsilon_1 1},\cdots,x_{-\varepsilon_{l-1}(l-1)}\Big).\end{aligned} \tag{6.3}$$

Let $\Omega=\Omega^A_{l,p}$ or Ω^C_l and R be the root system containing Ω.

PROPOSITION. *Let $\sigma\in A(R)$. It is possible to find a transformation $\sigma_{\mathbf{x}}$ of $\mathbf{C}^\Omega$ which induces σ on R if and only if the Lie algebra $\mathfrak{g}^{\sigma\Omega}$ acts transitively in the $\mathbf{x}$-space. $\sigma_{\mathbf{x}}$ is then an $\tilde{\mathfrak{h}}$-preserving symmetry of the canonical system.*

Let G be the group of transformation obtained by exponentiating (3.1.2) or (3.1.4). If σ is induced by the adjoint action of $g\in G$ on $\mathfrak{h}^*$, i.e., if σ is in the Weyl group $W(R)$ then we may take $g=\sigma_{\mathbf{x}}$. Since $W(R)$ is equal to $A(R)$ for type C and is of index 2 in $A(R)$ for type A the Proposition is of interest only in connection with the system $\Omega^A_{l,p}$ and the permutation $\sigma(\omega)=-\omega$. In this case $\sigma_{\mathbf{x}}$ exists if and only if $p=(l+1)/2$.

In the following we confine ourselves to selected examples of $\sigma_{\mathbf{x}}$. We take advantage of the well-known homogeneous spaces of the classical groups in question.

6.1. Transformations of $\Omega^A_{l,p}$. Let $L_{p,l+1}$, be the space of $(p,l+1)$-matrices of rank p. Let (ξ_{ij}) be the canonical coordinates of $L_{p,l+1}$. The group $\mathbf{GL}(p,\mathbf{C})$ acts on $L_{p,l+1}$ by the left matrix multiplication. For $c\in\mathbf{C}$ let $\mathcal{O}^{(c)}$ be the space of functions f on $L_{p,l+1}$ which satisfy

$$f(u\cdot a)=\det(u)^c f(a),\qquad u\in\mathbf{GL}(p,\mathbf{C}). \tag{6.1.1}$$

In particular, $\mathcal{O}^{(0)}$ is the field of functions on $\mathbf{GL}(p,\mathbf{C})\backslash L_{p,l+1}$, i.e., on the set of p-dimensional subspaces of $\mathbf{C}^{l+1}$. $\mathcal{O}^{(c)}$ is a 1-dimensional space over $\mathcal{O}^{(0)}$. We use $(1,\cdots,p,-1,\cdots,-q)$ to index the columns in $L_{p,l+1}$. Let $X=(\xi_{ij})$ and $X_\pm$ be the submatrix consisting of columns with indices >0 (resp. <0). Then $\det(X_+)^c\in\mathcal{O}^{(c)}$ and the entries of $\mathbf{x}=(X_+)^{-1}X_-$ form a coordinate system of $\mathcal{O}^{(0)}$. The group $\mathbf{GL}(l+1,\mathbf{C})$ acts on $L_{p,l+1}$ by the right matrix multiplication which induces an action on $\mathcal{O}^{(c)}$. Using $\mathcal{O}^{(c)}=\mathcal{O}^{(0)}\det(X_+)^c$ we obtain a 1-parameter family of multiplier representations on the $\mathbf{x}$-space. Passing to the Lie algebra $\mathfrak{sl}(l+1,\mathbf{C})$ we obtain the family (3.1.2). Thus the canonical system of $\Omega^A_{l,p}$ should be viewed as a system of operators on $\mathcal{O}^{(-1)}$.

$A(R)=\mathbf{S}_0\times W(R)$, $W(R)=\mathfrak{S}_{l+1}$. Let $h_i=h_{E_{ii}}$. The action of $A(R)$ on $\tilde{\mathfrak{h}}$ is given by

$$(\varepsilon,\sigma)_{\tilde{\mathfrak{h}}}\cdot h_i=\varepsilon h_{\sigma(i)}. \tag{6.1.2}$$

We may identify $W(R)$ with the group of permutation matrices in $\mathbf{GL}(l+1,\mathbf{C})$. This group acts on $L_{p,l+1}$ by permuting the columns of $X=(\xi_{ij})$. We identify $\mathfrak{S}_p\times\mathfrak{S}_q$ with the subgroup of $W(R)$ which permutes the columns in X_+ and X_-.

Let us apply $W(R)$ to $\det(X_+)$. If $\sigma=(\sigma',\sigma'')\in\mathfrak{S}_p\times\mathfrak{S}_q$ then

$$\sigma(\det(X_+))/\det(X_+)=\operatorname{sgn}(\sigma'). \tag{6.1.3}$$

Any $\sigma\in\mathfrak{S}_{l+1}$ is $\mathfrak{S}_p\times\mathfrak{S}_q$-conjugated to a permutation σ_r such that

$$\sigma_r(i)=i,\ 1\leq i\leq r,\qquad \sigma_r(r+k)=-k,\qquad 1\leq k\leq p-r. \tag{6.1.4}$$

Then

$$\sigma_r(\det(X_+))/\det(X_+)=\det\big(X_+^{-1}\sigma_r(X_+)\big)=\det\begin{bmatrix} x_{r+1,1},\cdots,x_{r+1,p-r}\\ x_{p,1},\cdots,x_{p,p-r}\end{bmatrix}. \tag{6.1.5}$$

Let us apply $\mathfrak{S}_{l+1}$ to $\mathcal{O}^{(0)}$. For $\sigma=(\sigma',\sigma'')\in\mathfrak{S}_p\times\mathfrak{S}_q$

$$\sigma_{\mathbf{x}}\cdot x_{ij}=x_{\sigma'(i)\sigma''(j)}. \tag{6.1.6}$$

Let now σ_r be the permutation defined by (6.1.4) and

$$\sigma_r(-k)=r+k, 1\leqq k\leqq p-r;\ \sigma_r(-k)=-k, p-r+1\leqq k\leqq 9. \tag{6.1.7}$$

Then

$$(\sigma_r)_{\mathbf{x}}\cdot\mathbf{x}=\left[\begin{array}{c|c} \dfrac{1_r}{0} & \mathbf{x}_{(1,p-r)} \end{array}\right]^{-1}\left[\begin{array}{c|c} \dfrac{0}{1_{p-r}} & \mathbf{x}_{(p-r+1,q)} \end{array}\right], \tag{6.1.8}$$

where 1_k is the (k,k)-identity matrix and $\mathbf{x}_{(a,b)}$ is the submatrix of $\mathbf{x}$ consisting of columns in the interval (a,b). For example,

$$(\sigma_{p-1})_{\mathbf{x}}\cdot\mathbf{x}=\left(x_{p1}^{-1}\cdot 1_p\right)\begin{bmatrix} -x_{11} & & & & & \\ \vdots & & & & & \\ -x_{i1} & & \cdots & \begin{vmatrix} x_{ij} & x_{pj} \\ x_{i1} & x_{p1}\end{vmatrix} & \cdots & \\ \vdots & & & & & \\ -x_{p-1,1} & & & & & \\ 1 & x_{p2} & \cdots & x_{pj} & \cdots & x_{pq} \end{bmatrix}, \tag{6.1.9}$$

$$(\sigma_0)_{\mathbf{x}}\cdot\mathbf{x}=\left[\mathbf{x}_{(1,p)}^{-1}\,\vdots\,\mathbf{x}_{(1,p)}^{-1}\mathbf{x}_{(p+1,q)}\right]. \tag{6.1.10}$$

Now let $p=q$. The action of $\mathbf{S}_0$ is determined by the action of the element $(-1,\sigma_0)\in A(\Omega_{l,p}^A)$ given by

$$(-1,\sigma_0)\cdot\det(X_+)=\det(X_+), \qquad (-1,\sigma_0)_{\mathbf{x}}\cdot\mathbf{x}={}^t\mathbf{x}, \tag{6.1.11}$$

where ${}^t\mathbf{x}$ is the transpose matrix of $\mathbf{x}$.

6.2. Transformations of Ω_l^C. Let Φ be the antisymmetric bilinear form

$$\Phi(\mathbf{x},\mathbf{y})=\sum_{i=1}^{l}\begin{vmatrix} x_i & y_i \\ x_{-i} & y_{-i}\end{vmatrix} \tag{6.2.1}$$

on $\mathbf{C}^{2l}$ (we use $\pm 1,\cdots,\pm l$ for the index set). Let N_{2l} be the subset of $L_{l,2l}$ consisting of matrices which rows are pairwise Φ-orthogonal. We define $\mathcal{O}^{(c)}$, ξ_{ij} and x_{ij} by restricting to N_{2l} the functions defined in §6.1. The Φ-orthogonality implies

$$x_{ij}=x_{ji}. \tag{6.2.2}$$

$\mathcal{O}^{(0)}$ is the field of functions on the set $\mathbf{GL}(l,\mathbf{C})\backslash N_{2l}$ of l-dimensional isotropic subspaces of $\mathbf{C}^{2l}$. The symplectic group $\mathbf{Sp}(\Phi)$ acts on N_{2l} by the right matrix multiplication. Passing to the action of its Lie algebra $\mathfrak{sp}(\Phi)$ on $\mathcal{O}^{(c)}$ we obtain a 1-parameter family of multiplier representations on $\mathcal{O}^{(0)}$ given by (3.1.4). The canonical system of Ω_l^C should be viewed as a set of operators on $\mathcal{O}^{(-1/2)}$.

$A(R)=W(R)=\mathfrak{S}_l\ltimes\mathbf{S}_0^l$. Let $h_i=h_{X_{i,-i}}$. The action of $A(R)$ on $\tilde{\mathfrak{h}}$ is given by

$$\sigma_{\tilde{\mathfrak{h}}}\cdot h_i=h_{\sigma(i)}, \quad \sigma\in\mathfrak{S}_p, \qquad \varepsilon_{\tilde{\mathfrak{h}}}\cdot h_i=\varepsilon_i h_i, \quad \varepsilon\in S_0^l. \tag{6.2.3}$$

There exists a subgroup of $\mathbf{Sp}(\Phi)$ isomorphic to $\mathfrak{S}_l \ltimes \{1, i, -1, -i\}^l$ which induces $W(R)$ on $\tilde{\mathfrak{h}}\cdot$ Let M: $\mathbf{C} \to \mathrm{End}(\mathbf{C}^2)$ be the $\mathbf{R}$-algebra homomorphism defined by $M(i) = \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$. Then

$$\text{(6.2.4)} \qquad \boldsymbol{\varepsilon}\begin{bmatrix} \xi_{jk} \\ \xi_{j,-k} \end{bmatrix} = M(\varepsilon_k)\begin{bmatrix} \xi_{jk} \\ \xi_{j,-k} \end{bmatrix}, \qquad \boldsymbol{\varepsilon} \in \{1, i, -1, -i\}^l, \quad k = 1, \cdots, l,$$

$$\sigma \cdot \xi_{i,\pm j} = \xi_{i,\pm\sigma(j)}, \qquad \sigma \in \mathfrak{S}_l.$$

In particular, the action of $\mathfrak{S}_l$ on $\mathcal{O}^{(c)}$ is the restriction to the diagonal subgroup of the action of $\mathfrak{S}_l \times \mathfrak{S}_l$ described in Sec. 6.1. Thus,

$$\text{(6.2.5)} \qquad \sigma(\det(X_+))/\det(X_+) = \mathrm{sgn}(\sigma),$$

$$\text{(6.2.6)} \qquad \sigma_{\mathbf{x}} \cdot x_{ij} = x_{\sigma(i)\sigma(j)}.$$

Consider now the action of $\mathbf{S}_0^l$. Any element of $\mathbf{S}_0^l$ is $\mathfrak{S}_l$-conjugated to the element $\boldsymbol{\varepsilon}_r$ defined by

$$\text{(6.2.7)} \qquad (\boldsymbol{\varepsilon}_r)_j = 1, \quad 1 \leqq j \leqq r, \qquad (\boldsymbol{\varepsilon}_r)_j = -1, \quad r+1 \leqq j \leqq l.$$

Arguing as in (6.1.5)–(6.1.8) we obtain

$$\text{(6.2.8)} \qquad (\boldsymbol{\varepsilon}_r^{1/2} \cdot \det(X_+))/\det(X_+) = \det\begin{bmatrix} x_{r+1,r+1} & ,\cdots, & x_{r+1,l} \\ x_{l,r+1} & ,\cdots, & x_{l,l} \end{bmatrix},$$

$$\text{(6.2.9)} \qquad (\boldsymbol{\varepsilon}_r^{1/2})_{\mathbf{x}} \cdot \mathbf{x} = \left[\begin{array}{c|c} \begin{array}{c} 1_r \\ \hline 0 \end{array} & \mathbf{x}_{(r+1,l)} \end{array}\right]^{-1} \left[\begin{array}{c|c} \mathbf{x}_{(1,r)} & \begin{array}{c} 0 \\ \hline -1_{l-r} \end{array} \end{array}\right].$$

For example,

$$\text{(6.2.10)} \qquad (\boldsymbol{\varepsilon}_{l-1}^{1/2})_{\mathbf{x}} \cdot \mathbf{x} = (x_{ll}^{-1} \cdot 1_l)\begin{bmatrix} & \vdots & & \vdots \\ \cdots & \begin{vmatrix} x_{ij} & x_{il} \\ x_{lj} & x_{ll} \end{vmatrix} & \cdots & x_{il} \\ & \vdots & & \vdots \\ \cdots & x_{lj} & \cdots & -1 \end{bmatrix},$$

$$\text{(6.2.11)} \qquad (\boldsymbol{\varepsilon}_0^{1/2})_{\mathbf{x}} \cdot \mathbf{x} = -\mathbf{x}^{-1}.$$

7. Hypergeometric series. Let Ω be a set of generators of a free $\mathbf{Z}$-module, $\mathfrak{h}$ be the Lie algebra defined by (3.1) and $\tau_\omega : \mathfrak{h}^* \to \mathfrak{h}^*$ be the translation $\alpha \mapsto \alpha + \omega$.

LEMMA 1. *If a function F on $\mathfrak{h}^* \times \mathbf{C}^\Omega$ satisfies the recurrences*

$$\text{(7.1)} \qquad \tau_\omega^{-1} \otimes \partial_{x_\omega} \cdot F = F, \qquad \omega \in \Omega,$$

then F satisfies the canonical system of Ω.

Proof. Let $\mathbf{n} = \mathbf{n}^+ - \mathbf{n}^- \in \rho(\Omega)$. If F satisfies (7.1) then

$$(\tau^{-1} \otimes \partial_{\mathbf{x}})^{\mathbf{n}^\pm} \cdot F = \tau_{\Sigma n_\omega^\pm \omega}^{-1} \otimes \partial_{\mathbf{x}}^{\mathbf{n}^\pm} \cdot F = F.$$

Since $\Sigma n_\omega^+ \omega = \Sigma n_\omega^- \omega$ we obtain $\partial_{\mathbf{x}}^{\mathbf{n}^+} \cdot F = \partial_{\mathbf{x}}^{\mathbf{n}^-} \cdot F$. Q.E.D.

Let $(\mathbf{1}_\omega)$ be the canonical basis of $\mathbf{C}^\Omega$ and π: $\mathbf{C}^\Omega \to \mathfrak{h}^*$ be defined by $\pi(\mathbf{1}_\omega) = \omega$.

LEMMA 2. *Let* $\mathbf{a}=(a_\omega)$ *be a map of* $\mathfrak{h}^*$ *into* $\mathbf{C}^\Omega$. *The function* $\mathbf{x}^\mathbf{a}=\prod_\omega x_\omega^{a_\omega}$ *satisfies the separation equations*

$$(7.2)\qquad (H-h_H)\cdot\mathbf{x}^\mathbf{a}=0,\qquad H\in\mathfrak{h},$$

if and only if

$$(7.3)\qquad \pi\circ\mathbf{a}=-\mathrm{id}.$$

Proof. Let $H=-\Sigma h_\omega x_\omega\partial_{x_\omega}\in\mathfrak{h}$. Thus, $h_\omega=h_H(\omega)$. We have

$$H\cdot\mathbf{x}^\mathbf{a}=-\left(\sum_\omega h_\omega a_\omega\right)\mathbf{x}^\mathbf{a}=h_H\left(-\sum_\omega a_\omega\omega\right)\mathbf{x}^\mathbf{a}=(-h_H\circ\pi\circ a)\mathbf{x}^\mathbf{a}.\qquad \text{Q.E.D.}$$

Let $\Delta_\omega=\tau_\omega^{-1}-\mathbf{1}_\omega$, $\omega\in\mathfrak{h}^*$. Thus Δ_ω is an operator defined on functions of $\mathfrak{h}^*$ into $\mathbf{C}^\Omega$ by $\Delta_\omega\cdot\varphi(\alpha)=\varphi(\alpha-\omega)-\mathbf{1}_\omega$. Δ_ω is invertible and $\Delta_\omega^{-1}=\tau_\omega+\mathbf{1}_\omega$. We write $\Delta^\mathbf{n}$ instead of $\prod_\omega\Delta_\omega^{n_\omega}$, $\mathbf{n}=(n_\omega)\in\mathbf{Z}^\Omega$. It is easy to see that Δ_ω permutes the set of functions $\mathbf{a}$ satisfying (7.3). Let Γ be the gamma function [2], $\Gamma(\mathbf{x})$ be $\prod_\omega\Gamma(x_\omega)$ and $\mathbf{z}!$ be $\Gamma(\mathbf{x}+\mathbf{1})$.

DEFINITION. Let $\mathbf{a}$: $\mathfrak{h}^*\to\mathbf{C}^\Omega$ satisfy (7.3) and $A=\{\Delta^\mathbf{n}\cdot\mathbf{a};\ \mathbf{n}\in\mathbf{Z}^\Omega\}$. The *hypergeometric series* of Ω, $\mathbf{a}$ is the formal expression

$$(7.4)\qquad F_\mathbf{a}^\Omega=\sum_{\mathbf{b}\in A}\mathbf{x}^\mathbf{b}/\Gamma(\mathbf{1}+\mathbf{b}).$$

PROPOSITION. $F_\mathbf{a}^\Omega$ *satisfies the hypergeometric system of* Ω, $\mathfrak{h}$. *Furthermore it satisfies the recurrences* (7.1).

Proof. In view of Lemmas 1–2 it is enough to prove the second assertion. Let $\gamma(\mathbf{x})$ denote $1/\Gamma(\mathbf{1}+\mathbf{x})$. γ satisfies the functional equations

$$(7.5)\qquad \gamma(\mathbf{x}-\mathbf{1}_\omega)=x_\omega\gamma(\mathbf{x}).$$

Applying the recurrence to each term of the series we have

$$\tau_\omega^{-1}\otimes\partial_{x_\omega}\cdot\gamma(\mathbf{b})\mathbf{x}^b=b_\omega^{\tau-\omega}\gamma(\mathbf{b}^{\tau-\omega})\mathbf{x}^{\Delta_\omega\mathbf{b}}=\gamma(\Delta_\omega\mathbf{b})\mathbf{x}^{\Delta_\omega\mathbf{b}}.\qquad \text{Q.E.D.}$$

In particular, let $B\subset\Omega$ be a basis of $\mathfrak{h}^*$ and $\mathbf{a}_B:\mathfrak{h}^*\to\mathbf{C}^\Omega$ be the linear function defined by $\mathbf{a}_B(\beta)=-\mathbf{1}_\beta$, $\beta\in B$. We write F_B^Ω instead of $F_{a_B}^\Omega$. Let $(\varepsilon_{\beta\omega})$ be the matrix over $\mathbf{Q}$ defined by

$$(7.6)\qquad \omega=\sum_{\beta\in B}\varepsilon_{\beta\omega}\beta,\qquad \omega\in\complement B.$$

Let $\alpha=\Sigma_\beta h_\beta\beta\in\mathfrak{h}^*$. Then

$$\begin{aligned}(\Delta^\mathbf{n}\cdot\mathbf{a}_B)(\alpha)&=\mathbf{a}_B\left(\alpha-\sum_\omega n_\omega\omega\right)-\sum_\omega n_\omega\mathbf{1}_\omega=\mathbf{a}_B(\alpha)-\sum_\omega n_\omega(\mathbf{a}_B(\omega)+\mathbf{1}_\omega)\\&=-\sum_\beta h_\beta\mathbf{1}_\beta-\sum_{\omega\in\complement B}n_\omega\left(\mathbf{1}_\omega-\sum_{\beta\in B}\varepsilon_{\beta\omega}\mathbf{1}_\beta\right)\\&=\sum_\beta\left(\sum_{\omega\in\complement B}n_\omega\varepsilon_{\beta\omega}-h_\beta\right)\mathbf{1}_\beta-\sum_{\omega\in\complement B}n_\omega\mathbf{1}_\omega.\end{aligned}$$

Let us introduce new variables

$$u_\beta = x_\beta, \quad \beta \in B, \qquad z_\omega = x_\omega \prod_\beta x_\beta^{-\varepsilon_{\beta\omega}}, \quad \omega \in \mathrm{C}B. \tag{7.7}$$

Then

$$\begin{aligned} F_B^\Omega &= \mathbf{u}^{-\mathbf{h}} \sum_{\mathbf{n}\in\mathbf{N}} \mathrm{C}B\mathbf{z}^{\mathbf{n}}/\mathbf{n}!\Gamma(\mathbf{1}-\mathbf{h}-\boldsymbol{\varepsilon}\mathbf{n}) \\ &= (\mathbf{u}^{-\mathbf{h}}/\Gamma(\mathbf{1}-\mathbf{h}))\sum_{\mathbf{n}} \mathbf{z}^{\mathbf{n}}/\mathbf{n}!(\mathbf{1}-\mathbf{h};\ -\boldsymbol{\varepsilon}\cdot\mathbf{n}), \end{aligned} \tag{7.8}$$

where (**x**; **n**) is the *Pochhammer symbol* $\Gamma(\mathbf{x}+\mathbf{n})/\Gamma(\mathbf{x})$.

7.1. Example.

$$\Omega_2^C;\ \omega_{11}+\omega_{22}=2\omega_{12}.$$

We are separating the wave equation (3.5) with the Lie algebra $\tilde{\mathfrak{h}}$ spanned by

$$\begin{aligned} X_{1,-1} &= -2x_{11}\partial_{x_{11}} - x_{12}\partial_{x_{12}} - \frac{1}{2}, \\ X_{2,-2} &= -2x_{22}\partial_{x_{22}} - x_{12}\partial_{x_{12}} - \frac{1}{2}. \end{aligned} \tag{7.1.1}$$

There are 2 nonequivalent bases B in Ω_2^C:

$$a = \{\omega_{11}, \omega_{22}\} \quad\text{and}\quad b = \{\omega_{11}, \omega_{12}\}. \tag{7.1.2}$$

The affine functions of the variable $\alpha \in \tilde{\mathfrak{h}}^*$ which appear in the series are related by

$$a_{11}\omega_{11} + a_{22}\omega_{22} = b_{11}\omega_{11} + b_{12}\omega_{12}. \tag{7.1.3}$$

Hence,

$$a_{11} = b_{11} + \frac{1}{2}b_{12}, \qquad a_{22} = \tfrac{1}{2}b_{12}. \tag{7.1.4}$$

The relations between these functions and the functions h_i defined in §6.2 are obtained from (7.1.1) by replacing $X_{i,-i}$ by h_i and $x_{ij}\partial_{x_{ij}}$ by $-b_{ij}$ if $\omega_{ij} \in B$ and by 0 otherwise. For example,

$$a_{ii} = \frac{1}{2}h_i + \frac{1}{4}. \tag{7.1.5}$$

We have

(7.1.6)

$$\begin{aligned} F_a^{C_2} &= (\mathbf{x}^{-\mathbf{a}}/\Gamma(\mathbf{1}-\mathbf{a}))\sum_n z_a^n/n!\left(\mathbf{1}-\mathbf{a};\ -\frac{1}{2}n, -\frac{1}{2}n\right), \qquad z_a = x_{12}(x_{11}x_{22})^{-1/2}, \\ F_b^{C_2} &= (\mathbf{x}^{-\mathbf{b}}/\Gamma(\mathbf{1}-\mathbf{b}))\sum_n z_b^n/n!(\mathbf{1}-\mathbf{b};\ n, -2n), \quad z_b = z_a^{-2}. \end{aligned}$$

The transformation $x_{ii} \mapsto x_{ii}$, $x_{12} \mapsto -x_{12}$ is a symmetry of the hypergeometric system. Hence $F_a^{C_2}(-z_a)$ and therefore the even component $pF_a^{C_2}$ of $F_a^{C_2}$ also satisfy the hypergeometric system (but not the recurrences of §5). We have

$$pF_a^{C_2}=(\mathbf{x}^{-\mathbf{a}}/\Gamma(\mathbf{1}-\mathbf{a}))\sum_n z_a^{2n}/(2n)!(\mathbf{1}-\mathbf{a};\ -n,-n). \tag{7.1.7}$$

Using the well-known identities [2]

$$\begin{aligned}&(x;\ -n)=(-1)^n(1-x;n)^{-1},\\&(2x;2n)=2^{2n}(x;n)\left(x+\frac{1}{2};n\right),\quad \text{in particular,}\\&(2n)!=(1;2n)=2^{2n}\left(\frac{1}{2};n\right)n!,\end{aligned} \tag{7.1.8}$$

we may express these functions in terms of ${}_2F_1$:

$$\begin{aligned}&pF_a^{C_2}=(\mathbf{x}^{-\mathbf{a}}/\Gamma(\mathbf{1}-\mathbf{a}))\ {}_2F_1\left(a_{11},a_{22},\frac{1}{2},z_a^2/4\right)\\&F_b^{C_2}=(\mathbf{x}^{-\mathbf{b}}/\Gamma(\mathbf{1}-\mathbf{b}))\ {}_2F_1\left(\frac{1}{2}b_{12},\frac{1}{2}b_{12}+\frac{1}{2},1-b_{11},4z_b\right).\end{aligned} \tag{7.1.9}$$

REFERENCES

[1] N. Bourbaki, *Groupes et algèbres de Lie*, Masson, Paris 1981.

[2] A. Erdelyi et al., *Higher Transcendental Functions*, McGraw–Hill, New York, 1951.

[3] H. Exton, *Multiple Hypergeometric Functions and Applications*, Ellis Harwood, Chichester, 1976.

[4] E. G. Kalnins et. al., *The Lie theory of two-variable hypergeometric functions*, Stud. Appl. Math., 62 (1980), pp. 143–173.

[5] W. Miller, Jr., *Lie theory and the Lauricella functions F_D*, J. Math. Phys., 13 (1972), pp. 1393–1399.

[6] ______, *Symmetry and Separation of Variables*, Addison–Wesley, Reading, MA, 1977.

[7] L. W. Ovsyannikov, *Group Theory of Differential Equations*, Nauka, Moscow, 1978. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

SOME EXPLICIT PADÉ APPROXIMANTS FOR THE FUNCTION Φ'/Φ AND A RELATED QUADRATURE FORMULA INVOLVING BESSEL FUNCTIONS*

JET WIMP†

Abstract. In this paper we determined in closed form the $[n|n]$ Padé approximant for the logarithmic derivative of the confluent hypergeometric function of the first kind, and also an explicit formula for the error.

We next show how the recurrence defining the numerators and denominators of the approximants can be used to deduce a certain discrete orthogonality relationship. A consequence of this is a discrete orthogonality relation for the Bessel function of the first kind and an exact quadrature formula involving this function.

1. Introduction. In [3], W. A. Fair applied the τ-method to the differential equation,

$$zy'(z)+(c-z)y(z)+zy^2(z)-a=0, \qquad y(0):=a/c, \tag{1.1}$$

satisfied by

$$y(z):=u'(z)/u(z), \qquad u(z):=\Phi(a,c;z), \tag{1.2}$$

to determine the $[n|n]$ Padé approximant to $y(z)$. (Here Φ is the usual notation for the confluent hypergeometric function. In this, as well as all other functions, we will follow the notation of [2]. For an account of Fair's work, see [5, v. 2, 10.4].)

Let $A_n(z)$, $B_n(z)$ denote the numerator and denominator approximants respectively. It was found that both A_n, B_n satisfy the three-term recurrence

$$y_{n+2}(z)+(a_nz+b_n)y_{n+1}+c_nz^2y_n=0, \qquad n=0,1,2,\cdots, \tag{1.3}$$

$$\begin{aligned} a_n&=-(2a-c)/(2n+c+2)(2n+c+4),\\ b_n&=-1,\\ c_n&=-(n+a+1)(n+c-a+1)/(2n+c+1)_3(2n+c+2). \end{aligned} \tag{1.4}$$

The initial conditions were

$$\begin{aligned} A_0&=\frac{a}{c}, & A_1&=\frac{a}{c}+\frac{a(a+1)z}{(c)_3},\\ B_0&=1, & B_1&=1+\frac{(2a-c)z}{c(c+2)}. \end{aligned} \tag{1.5}$$

Only in the case $c=2a=2\nu+1$ (the Bessel function case) could closed form expressions be determined for A_n, B_n.

* Received by the editors March 17, 1983 and in revised form October 8, 1983. This week was supported in part by NATO Research grant 027.81.

† Department of Mathematical Sciences, Drexel University, Philadelphia, Pennsylvania 19104.

In the first part of this paper, we derived closed form expressions for A_n, B_n for all a,c as well as an explicit formula for the error, $y-A_n/B_n$.

In the second part, we show how the recurrence (1.3) leads naturally to a certain orthogonality relation which contains a discrete orthogonality relation for the Bessel function $J_{2n+\nu}(x)$ as a special case.

Our approach is, roughly, that used by Askey and Wimp in [1] to derive closed form Padé approximants to $\Psi(a+1,c;z)/\Psi(a,c;z)$ and depends on a study of the recurrence *associated* with (1.3), i.e., the equation obtained when a_n, b_n, c_n are replaced by $a_{n+\delta}$, $b_{n+\delta}$, $c_{n+\delta}$ respectively.

2. The Padé approximants. We start off with a specialization of a recurrence for a generalized hypergeometric function given by the present author, which can be found in [5, v. 2, 12.2]. We let $r=0$, $s=1$, make an obvious identification of parameters, an equally obvious change of dependent variable, and replace n by $n+\delta$. The recurrence becomes

$$(2.1)\qquad g_{n+2}-g_{n+1}\left\{z+\frac{(2a-c)}{(2n+2\delta+c+2)(2n+2\delta+c+4)}\right\}$$
$$-g_n\frac{(n+\delta+a+1)(n+\delta+c-a+1)}{(2n+2\delta+c+1)_3(2n+2\delta+c+2)}=0.$$

Note this is the associated equation for the equation satisfied by $z^nA_n(1/z)$, $z^nB_n(1/z)$. As shown in the cited reference, the equation, under appropriate conditions, has a basis of solutions

$$(2.2)\qquad g_n^1\equiv g_n^1(z,\delta):=z^n\Phi\left(-a-n-\delta,\,-c-2n-2\delta;\,-\frac{1}{z}\right),$$
$$g_n^2\equiv g_n^2(z,\delta):=(-z)^{-n}\frac{\Gamma(n+\delta+a+1)\Gamma(n+\delta+c+1-a)}{\Gamma(2n+2\delta+c+1)\Gamma(2n+2\delta+c+2)}$$
$$\cdot\Phi\left(n+\delta+c+1-a,2n+2\delta+c+2;\,-\frac{1}{z}\right).$$

(In what follows, we assume that a,n,c,d are such that these functions exist and are linearly independent. If this is not the case, straightforward redefinitions and/or limit processes may always be invoked to assure the existence and linear independence of the functions. The volumes [5] contain many examples of these procedures.)

We wish to develop a *polynomial* basis of solutions of the recurrence (2.1). As is common, we shall use the recurrence for $n=-1$ as well as $n=0,1,2,\cdots,$. Our first polynomial solution will be denoted $p_n\equiv p_n(z,\delta)$ and satisfies the conditions

$$(2.3)\qquad p_{-1}=0,\qquad p_0=1.$$

Note that $p_{n-1}(z,\delta+1)$ is also a solution of the recurrence and satisfies

$$(2.4)\qquad p_{-2}(z,\delta+1):=\frac{(2\delta+c-1)_3(2\delta+c)}{(a+\delta)(\delta+c-a)},$$
$$p_{-1}(z,\delta+1)=0.$$

A little algebra shows we may write

$$p_n = \frac{\Delta_n}{\Delta_0}, \qquad \Delta_n \equiv \Delta_n(z,\delta) := \left(g_n^2 g_{-1}^1 - g_n^1 g_{-1}^2\right). \tag{2.5}$$

We next wish to determine Δ_0 explicitly. Here, and in what follows, three contiguous relations for $\Phi(\alpha,\beta;x)$ will be useful (see [2, v. 1, 6.4]).

$$\Phi(\alpha+1,\beta+1;x) = \frac{\beta}{\alpha}\Phi'(\alpha,\beta;x), \tag{2.6}$$

$$\Phi(\alpha+1,\beta+2;x) = \frac{-\beta(\beta+1)}{x\alpha(\beta-\alpha)}\{\alpha\Phi(\alpha,\beta;x) - \beta\Phi'(\alpha,\beta;x)\}, \tag{2.7}$$

$$\Phi(\alpha,\beta-1;x) = \frac{x}{(\beta-1)}\Phi'(\alpha,\beta;x) + \Phi(\alpha,\beta;x). \tag{2.8}$$

Through the use of (2.6) and (2.8), Δ_0 may be written in terms of cross products of confluent functions with parameters $(-a-\delta, 1-c-2\delta)$ and $(\delta+c-a, 2\delta+c+1)$. These functions are a basis of solutions of the confluent hypergeometric differential equation, and thus Δ_0 is nothing more than a known Wronskian, (see [2, v. 1, 6.7(6)]). Evaluating the Wronskian gives

$$\Delta_0 = \frac{z e^{-1/z}\Gamma(\delta+a)\Gamma(\delta+c-a)}{\Gamma(2\delta+c-1)\Gamma(2\delta+c)}. \tag{2.9}$$

Referring back to (2.5) we see that p_n may be written in two parts

$$g_n^2 g_{-1}^1/\Delta_0, \qquad -g_n^1 g_{-1}^2/\Delta_0, \tag{2.10}$$

and, after using Kummer's transformation on g_{-1}^1, g_{-1}^2, that the first consists of a term $O(z^{-n-2})$ as $z \to \infty$. Thus p_n must be the polynomial part of the second part. To find this, we use on $g_n^1 g_{-1}^2/\Delta_0$ the formula

$$\Phi(a,c;x)\Phi(b,d;-x) = \sum_{k=0}^{\infty} \frac{(a)_k x^k}{(c)_k k!} {}_3F_2\left(\begin{matrix} -k, 1-k-c, b \\ 1-k-a, d \end{matrix} \middle| 1 \right). \tag{2.11}$$

The result is

(2.12)

$$p_n(z,\delta) = \sum_{k=0}^{n} \frac{(-a-n-\delta)_k(-1)^k z^{n-k}}{(-c-2n-2\delta)_k k!} {}_3F_2\left(\begin{matrix} -k, 2n+2\delta+c+1-k, \delta+a \\ n+\delta+a+1-k, 2\delta+c \end{matrix} \middle| 1 \right).$$

This construction can be made to yield a rational approximant (see [2, v. 2, 10.5]) which is the $[n-1|n]$ Padé approximant to a certain function. This is not quite the Padé element we seek, but what we want is easily obtainable from the analysis to follow.

Observe that the expressions (2.2) reveal that

$$g_{n-1}^1(z,\delta+1) = \frac{g_n^1(z,\delta)}{z}, \qquad g_{n-1}^2(z,\delta+1) = -z g_n^2(z,\delta), \tag{2.13}$$

and from the recurrence (2.1) we get

(2.14)
$$g_1^j(z,\delta)=\left[z+\frac{(2a-c)}{(2\delta+c)(2\delta+c+2)}\right]g_0^j(z,\delta)+g_{-1}^j(z,\delta)\frac{(\delta+a)(\delta+c-a)}{(2\delta+c-1)_3(2\delta+c)},$$
$$j=1,2.$$

Using these and performing a lot of algebra gives

$$\begin{aligned}\text{(2.15)}\qquad \frac{p_{n-1}(z,\delta+1)}{p_n(z,\delta)}&=\frac{M\left(g_n^2g_0^1-g_0^2g_n^1\right)}{\left(g_n^2g_{-1}^1-g_n^1g_{-1}^2\right)}\\&=\frac{\Phi(c+1-a,c+2;\,-1/z)}{z\Phi(c-a,c;\,-1/z)}-\frac{Mr_n\Delta_0}{g_{-1}^2\left(r_ng_{-1}^1-g_{-1}^2\right)},\end{aligned}$$

where

$$\text{(2.16)}\qquad M:=\frac{-(2\delta+c-1)_3(2\delta+c)}{(\delta+a)(\delta+c-a)},\qquad r_n:=r_n(z,\delta)=g_n^2/g_n^1.$$

We now write $z^nA_n(1/z)$, $z^nB_n(1/z)$ as linear combinations of $p_n(z,0)$, $p_{n-1}(z,1)$. We have

$$\text{(2.17)}\qquad \begin{aligned}z^nA_n\left(\frac{1}{z}\right)&=\frac{a}{c}p_n(z,0)+\frac{a(c-a)p_{n-1}(z,1)}{c^2(c+1)},\\ z^nB_n\left(\frac{1}{z}\right)&=p_n(z,0).\end{aligned}$$

Employing generously the properties of the Φ functions and (2.15), (2.16) shows

$$\text{(2.18)}\qquad \begin{aligned}f_n:=\frac{A_n(z)}{B_n(z)}&=\frac{a}{c}+\frac{(c-a)p_{n-1}(z^{-1},1)}{c^2(c+1)p_n(z^{-1},0)}=y(z)+R_n(z),\\ y(z)&=\frac{\Phi'(a,c;z)}{\Phi(a,c;z)},\end{aligned}$$

$$\text{(2.19)}\qquad R_n=\frac{ze^z\Gamma^2(c)r_n(z^{-1},0)}{\Gamma(a)\Gamma(c-a)\Phi^2(a,c;z)\left(r_n(z^{-1},0)/r_{-1}(z^{-1},0)-1\right)}.$$

This demonstrates the incredibly rapid convergence of the approximations, since $r_n(z,0)=O((z/2)^{2n}n^{2a-1/2-c}/(2n)!)$, $n\to\infty$. Also, $R_n=O(z^{2n+1})$ as $z\to 0$, demonstrating the claimed Padé character of the approximants. (Of course, as pointed out in [5], we must stay away from the zeros of $\Phi(a,c;z)$.)

3. An orthogonality relation and a quadrature formula involving Bessel functions. It is a well-known fact that if a sequence of polynomials $\{p_n(z)\}$ satisfies a three-term recurrence

$$\text{(3.1)}\qquad y_{n+2}+(a_n+zb_n)y_{n+1}+c_ny_n=0,\qquad n=-1,0,1,2,\cdots,$$
$$p_{-1}=0,\qquad p_0=1,$$

and

$$\lambda_n := c_n/b_nb_{n-1} > 0, \qquad n=1,2,\cdots, \tag{3.2}$$

then there is a distribution function $\psi \in \Psi^*$ such that $\{p_n\}$ is an orthogonal set on $(-\infty,\infty)$ with respect to $d\psi$. (Here, Ψ^* is the class of bounded nondecreasing functions on $(-\infty,\infty)$ with an infinite number of points of increase whose moments exist.)

Obviously the recurrence (2.1) doesn't qualify as it stands, but some minor changes produce a recurrence that does generate a system of orthogonal polynomials (with respect to a discrete distribution).

Let

$$z \to iz, \quad a := \frac{c}{2} + i\kappa, \quad g_n := i^n w_n. \tag{3.3}$$

We have

$$\begin{aligned} w_{n+2} - w_{n+1}\left\{ z + \frac{2\kappa}{(2n+2\delta+c+2)(2n+2\delta+c+4)} \right\} \\ + w_n \frac{\left(n+\delta+\frac{c}{2}+1\right)^2 + \kappa^2}{(2n+2\delta+c+1)_3(2n+2\delta+c+2)} = 0. \end{aligned} \tag{3.4}$$

As long as $c > -1$, this recurrence satisfies (3.2).

It is further known (see [6]) that if λ_n is bounded, then a ratio of appropriately defined polynomials generated from the recurrence will converge to the Stieltjes transform of the distribution function. What this means in our case is that

$$\lim_{n\to\infty} \frac{ip_{n-1}(iz,\delta+1)}{p_n(iz,\delta)} := h(z) = \int_{-\infty}^{\infty} \frac{d\psi}{z-t}, \qquad z \notin \operatorname{Supp}\psi. \tag{3.5}$$

Also, ψ has, in this case, compact support. The computations in the previous section give $h(z)$. Without loss of generality, we may put $\delta = 0$, and we find

$$\begin{aligned} h(z) &= m(z)/d(z), \\ m(z) &:= \Phi\left(\frac{c}{2}+1-i\kappa, c+2; \frac{i}{z}\right), \\ d(z) &:= z\Phi\left(\frac{c}{2}-i\kappa, c; \frac{i}{z}\right). \end{aligned} \tag{3.6}$$

Note that Kummer's transformation shows that $\bar{h}(z) = h(z)$ for z real; hence $h(z)$ is real for z real, as it should be. Also, the only singularities of $h(1/z)$ in the finite plane are poles. These are real and occur at the zero of the denominator. Denote the zeros of $d(1/z)$ by z_k in increasing order, $k = \pm 1, \pm 2, \pm 3, \cdots, z_1$ denoting the smallest positive zero.

The relation (3.5) may be inverted to find ψ by using a complex inversion formula given in Stone, [8, p. 163]. Without loss of generality we take $\psi(-\infty) = 0$. Then

$$\psi(t) = \frac{1}{2\pi i} \int_{\Gamma_t} h(z)\,dz, \tag{3.7}$$

where Γ_t is the rectangle $[-\infty - i\varepsilon, t - i\varepsilon, t + i\varepsilon, -\infty + i\varepsilon]$.

The integral is easily evaluated by using known contiguous relations for Φ, see [2, v. 1, p. 253 (8), (9); p. 252 (1)], and the argument theorem. We find

$$\psi(t)=K\sum_{z_k^{-1}\text{ interior to }\Gamma_t} z_k^{-2},$$

(3.8)
$$\psi(t)=\begin{cases}0, & -\infty<t<z_{-1}^{-1},\\ K\displaystyle\sum_{-\infty<z_k^{-1}<t} z_k^{-2}, & K=c^2(c+1)\Big/\left(\dfrac{c^2}{4}+\kappa^2\right),\end{cases}$$

i.e., ψ has a jump of z_k^{-2} at z_k^{-1}.

ψ is bounded. This follows on considering the differential equation

(3.9)
$$y''+\left[\frac{1}{4}+\frac{\kappa}{z}+\frac{c(2-c)}{4z}\right]y=0,$$

satisfied by

(3.10)
$$y=z^{c/2}e^{-iz/2}\Phi\left(\frac{c}{2}-i\kappa,c;iz\right),$$

and applying a result of Tricomi (see [9, p. 104 and the subsequence analysis of Bessel functions]). We find that

(3.11)
$$|z_{k-1}-z_k|=2\pi+o(1),\qquad k\to\pm\infty.$$

Thus the series $\Sigma_{-\infty}^{\infty}z_k^{-2}$ converges. The polynomials

(3.12)
$$w_n(z):=i^{-n}p_n(iz,0)$$
$$=\sum_{k=0}^{n}\frac{i^{-k}z^{n-k}\left(-\frac{c}{z}-i\kappa-n\right)_k}{(-2n-c)_k k!}\,{}_3F_2\left(\left.\begin{matrix}-k,2n+c+1-k,i\kappa+\frac{c}{2}\\ n+\frac{c}{2}+i\kappa+1-k,c\end{matrix}\right|1\right)$$

are orthogonal with respect to this distribution, i.e.,

(3.13)
$$\sum_{k=-\infty}^{\infty}{}' z_k^{-2}w_n\left(z_k^{-1}\right)w_m\left(z_k^{-1}\right)=\sigma_n\delta_{mn}.$$

(The prime means the term $k=0$ is to be deleted.)

σ_n can be determined from the formula $c_n=b_n\sigma_{n+1}/b_{n-1}\sigma_n$. Thus

(3.14)
$$\sigma_n=\frac{\sigma_0|(1+c/2+i\kappa)_n|^2}{(c+1)_{2n}(c+2)_{2n}}.$$

To determine σ_0 we use the representation (3.5), i.e.,

$$\frac{zm(z)}{d(z)}=K\sum_{-\infty}^{\infty}{}'z_k^{-2}\frac{z}{\left(z-z_k^{-1}\right)}.$$

Letting $z \to \infty$ shows

$$1 = K \sum_{-\infty}^{\infty}{}' z_k^{-2} = \sigma_0.$$

It is not obvious that w_n as given by (3.12) is a polynomial with real coefficients. When $\kappa = 0$, a representation may be found which does not have this defect. Then the ${}_3F_2$ in (3.12) can be summed by Watson's formula [2, v. 1, 4.4(6)]. This corresponds to the Bessel function case, which we will next investigate. Since for $\kappa = 0$ Bessel functions are Φ functions (3.6) with double the argument, we replace z everywhere it occurs by $x/2$ and let $c = 1 + 2\nu$, $\nu > -\frac{1}{2}$. Making the substitution

$$\text{(3.15)} \qquad u_n := 2^{2n}(\nu+1)_n w_n,$$

gives the difference equation

$$\text{(3.16)} \qquad u_{n+2} - 2(n+\nu+2)x u_{n+1} + u_n = 0,$$

whose solutions we recognize to be $J_{\nu+n+1}(x^{-1})$, $Y_{\nu+n+1}(x^{-1})$. The corresponding polynomial solutions, $q_n(x)$, with initial values $q_{-1} = 0$, $q_0 = 1$, are, as determined from (3.12), Lommel polynomials,

$$\text{(3.17)} \qquad q_n(x) = R_{n,\nu+1}\left(\frac{1}{x}\right) = \sum_{k=0}^{n/2} \frac{(-1)^k (n-k)!\Gamma(\nu+n+1-k)(2x)^{n-2k}}{k!(n-2k)!\Gamma(\nu+k+1)},$$

see [2, v. 2, 5.2(26)]. (Empty sums are to be interpreted as 0.)

Formula (3.13) shows that Lommel polynomials are orthogonal with respect to a discrete distribution, i.e.,

$$\text{(3.18)} \qquad \sum_{k=-\infty}^{\infty}{}' x_k^{-2} R_{n,\nu+1}(x_k) R_{m,\nu+1}(x_k) = \tau_n \delta_{mn},$$

$$\text{(3.19)} \qquad \tau_n := 1/2(n+\nu+1),$$

x_k denoting the zeros of $x^{-\nu}J_\nu(x)$. Since $q_n(-x) = (-1)^n q_n(x)$ we can write

$$\text{(3.20)} \qquad \left[1 + (-1)^{m+n}\right] \sum_{k=1}^{\infty} x_k^{-2} R_{n,\nu+1}(x_k) R_{m,\nu+1}(x_k) = \tau_n \delta_{mn}.$$

The relation (3.18) was obtained in a different way by Schwartz [7], and generalized to basic Bessel functions by Ismail [4].

Of course (3.19) implies that when a function has a formal expansion in Lommel polynomials,

$$\text{(3.21)} \qquad f(x) = \sum_{n=0}^{\infty} A_n R_{2n,\nu+1}(x).$$

The coefficients A_n can be determined by applying the orthogonality relationship (3.19). However the surprising implications of the orthogonality are for functions on $[0, \infty)$ possessing a Neumann series,

$$\text{(3.22)} \qquad f(x) = \sum_{n=0}^{\infty} B_n J_{2n+\nu+1}(x).$$

We know

$$\left\{ R_{n,\nu+1}(x) = \frac{-\pi x}{2}\left[Y_{n+\nu+1}(x)J_\nu(x) - J_{n+\nu+1}(x)Y_\nu(x)\right]. \right. \tag{3.23}$$

Substituting this in (3.18) gives a *discrete* orthogonality relation for the functions $\{J_{2n+\nu+1}\}$,

$$\sum_{k=1}^{\infty} J_{2n+\nu+1}(x_k)J_{2m+\nu+1}(x_k)Y_\nu^2(x_k) = \frac{\delta_{mn}}{\pi^2(2n+\nu+1)}, \tag{3.24}$$

and this provides a closed form expression for the coefficients B_n,

$$B_n = \pi^2(2n+\nu+1)\sum_{k=1}^{\infty} f(x_k)J_{2n+\nu+1}(x_k)Y_\nu^2(x_k). \tag{3.25}$$

However, it is known [2, v. 2, 7.10.1(8)] that these same coefficients possess an integral representation and equating the two gives a *quadrature formula*:

$$\int_0^\infty t^{-1}f(t)J_{2n+\nu+1}(t)\,dt = \frac{\pi^2}{2}\sum_{k=1}^{\infty} f(x_k)\cdot J_{2n+\nu+1}(x_k)Y_\nu^2(x_k). \tag{3.26}$$

Note when f is bounded, the series on the right converges since $\psi \in \Psi^*$. At first glance this formula seems very surprising, since the quadrature formula is *exact*. It must be borne in mind, though, that the formula is true only for functions possessing a convergent Neumann series of the form (3.22), and that class of functions is, unhappily, rather small.

Using a result for $\nu=0$ given in Watson [10, p. 533] allows us to deduce the following representation theorem.

THEOREM. *Let* $\nu=0$, $n=0,1,2,\cdots$. *Let* $f \in L_1[0,\infty) \cap C^1[0,\infty)$ *and satisfy the integral equation*

$$2f'(t) = \int_0^\infty \frac{J_1(u)}{u}\left[f(u+t) - f(u-t)\right] dt, \qquad t>0. \tag{3.27}$$

Then the quadrature formula (3.26) *is valid when the right-hand side converges.*

The condition (3.27) on F is necessary and sufficient for the existence and convergence of the series (3.22), so the class of functions for which (3.26) holds cannot be enlarged in any substantial way. That class, however, does include a number of commonly occurring transcendental functions, see the tables in [5, v. 2, 9.4]. For arbitrary ν, a more complicated set of conditions has been given by Wilkins, [11].

It is not clear whether the same sort of analysis can be applied to the more general orthogonal system (3.13). The major stumbling block is that when $k \neq 0$ the differential equation (3.9) cannot be transformed into a differential equation of Sturm–Liouville type in a way that allows an integral representation analogous to the left-hand side of (3.26) to be derived.

REFERENCES

[1] R. ASKEY AND J. WIMP, *Associated Laguerre and Hermite polynomials*, Proc. Roy Soc. Edinburgh, 95 A (1983), pp. 1–23.

[2] A. ERDÉLYI, et al, *Higher Transcendental Functions*, 3 volumes, McGraw-Hill, New York, 1954.

[3] W. G. FAIR, *Padé approximation to the solution of the Ricatti equation*, Math Comp., 18 (1964), pp. 627–634.

[4] M. E. H. ISMAIL, *The zeros of basic Bessel functions, the functions $J_{\nu+ax}(x)$ and associated orthogonal polynomials*, J. Math. Anal. Appl., 86 (1982), pp. 1–19.

[5] Y. L. LUKE, *The Special Functions And Their Approximations*, 2 volumes, Academic Press, New York, 1969.

[6] O. PERRON, *Lehre von den Kettenbruchen*, Chelsea, New York, 1950.

[7] H. M. SCHWARTZ, *A class of continued fractions*, Duke Math. J., 6 (1940), pp. 48–65.

[8] M. H. STONE, *Linear Transformation in Hilbert Space*, Am rican Mathematical Society, New York, 1932.

[9] F. A. TRICOMI, *Differential Equations*, Blackie and Son, London, 1961.

[10] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Cambridge Univ. Press, London, 1962.

[11] J. E. WILKINS, JR., *Neumann series of Bessel functions*, Trans. Amer. Math. Soc., 64 (1948), pp. 359–385; 69 (1950), pp. 55–65.

SIAM J. MATH. ANAL.
Vol. 16, No. 4, July 1985

019

ASYMPTOTIC EXPANSIONS OF MELLIN TRANSFORMS AND ANALOGUES OF WATSON'S LEMMA*

AVRAM SIDI†

Abstract. In this paper the asymptotic behavior of the Mellin transform $\hat{f}(x)=\int_0^\infty t^{x-1}f(t)\,dt$ of $f(t)$, for $x\to+\infty$, is analyzed. In particular, it is shown for certain classes of functions $u_k(t)$, $k=0,1,\cdots$, that form asymptotic sequences for $t\to+\infty$, that if $f(t)\sim\sum_{k=0}^\infty A_k u_k(t)$ as $t\to+\infty$, then $\hat{f}(x)\sim\sum_{k=0}^\infty A_k\hat{u}_k(x)$ as $x\to+\infty$. In this sense the results of this paper are analogues of Watson's lemma for Laplace integrals. Several illustrative examples involving summation of everywhere divergent moment series and special functions are appended.

1. Introduction. Let $f(t)$ be a function that is locally integrable for $0<t<+\infty$ such that, for some real constant σ, $t^{\sigma-1}f(t)$ is absolutely integrable in any finite interval of the form $[0,a]$, and

$$f(t)=O(t^{-\mu})\quad\text{as } t\to+\infty,\quad \text{any } \mu>0. \tag{1.1}$$

Then the Mellin transform $\hat{f}(x)$ of $f(t)$, defined by

$$\hat{f}(x)=\int_0^\infty t^{x-1}f(t)\,dt, \tag{1.2}$$

exists for all sufficiently large x.

The purpose of this work is to give an asymptotic analysis of $\hat{f}(x)$ for $x\to+\infty$. Surprisingly, this problem does not seem to have received much attention. Doetsch [2, Vol. 2, Chap. 5] has considered the problem of analytic continuation of the Mellin transform beyond the strip in which its integral representation converges, and has obtained results on the singularity structure of it. Riekstinš [8] has considered the asymptotic expansion of the inverse Mellin transform. Wagner [11] has obtained some Tauberian theorems for Mellin transforms. Handlesman and Lew [3], [4], [5] have used techniques involving the Mellin transform for obtaining asymptotic expansions for other integral transforms.

The results of this work can be summarized in an informal way as follows. Consider the sequence of functions $\{u_0(t),u_1(t),\cdots\}$ and the sequence of the corresponding Mellin transforms $\{\hat{u}_0(x),\hat{u}_1(x),\cdots\}$. Assume that

$$\lim_{t\to+\infty}\frac{u_q(t)}{u_k(t)}=0,\qquad q>k,\quad k=0,1,\cdots, \tag{1.3}$$

and

$$\lim_{x\to+\infty}\frac{\hat{u}_q(x)}{\hat{u}_k(x)}=0,\qquad q>k,\quad k=0,1,\cdots, \tag{1.4}$$

* Received by the editors September 7, 1982, and in revised form October 18, 1983.

† Computer Science Department, Technion–Israel Institute of Technology, Haifa, Israel. The research for this paper was done while the author was a National Research Council NASA Research Associate at Lewis Research Center, Cleveland, Ohio 44135.

i.e., that $\{u_0(t), u_1(t), \cdots\}$ and $\{\hat{u}_0(x), \hat{u}_1(x), \cdots\}$ are asymptotic sequences as $t \to +\infty$ and $x \to +\infty$ respectively. If the function $f(t)$ has an asymptotic expansion of the form

$$(1.5) \qquad f(t) \sim \sum_{k=0}^{\infty} A_k u_k(t) \quad \text{as } t \to +\infty,$$

then, for some choices of the $u_k(t)$, $\hat{f}(x)$ has the asymptotic expansion

$$(1.6) \qquad \hat{f}(x) \sim \sum_{k=0}^{\infty} A_k \hat{u}_k(x) \quad \text{as } x \to +\infty,$$

i.e., the asymptotic expansion of $\hat{f}(x)$ is that obtained by taking the Mellin transform of the right-hand side of (1.5) term by term.

In the sense of (1.5) and (1.6) the results of the present work are analogues of Watson's lemma for Laplace transforms. Recall that essentially Watson's lemma concerns sequences $\{w_0(t), w_1(t), \cdots\}$, where $w_k(t) = t^{\gamma_k}$, $k = 0, 1, \cdots$, and $-1 < \operatorname{Re}\gamma_0 < \operatorname{Re}\gamma_1 < \cdots$, and states that if the function $g(t)$ has an asymptotic expansion of the form $g(t) \sim \sum_{k=0}^{\infty} B_k w_k(t)$ as $t \to 0+$, then the Laplace transform $\bar{g}(s) = \int_0^{\infty} e^{-st} g(t)\,dt$ of $g(t)$ has the asymptotic expansion $\bar{g}(s) \sim \sum_{k=0}^{\infty} B_k \bar{w}_k(s)$ as $s \to +\infty$. For Watson's lemma and its generalizations see Olver [7].

The use of the asymptotic expansion of $f(t)$ as $t \to +\infty$ to obtain that of $\hat{f}(x)$ as $x \to +\infty$ can be heuristically justified as follows. Consider the integral

$$(1.7) \qquad I(b;x) = \int_0^b t^{x-1} f(t)\,dt, \qquad 0 < b < +\infty.$$

Making the change of variable of integration $\xi = \log(b/t)$, (1.7) becomes

$$(1.8) \qquad I(b;x) = b^x \int_0^{\infty} e^{-x\xi} f(be^{-\xi})\,d\xi.$$

Now the asymptotic expansion of $I(b;x)$ for $x \to +\infty$ can be obtained by expanding $f(be^{-\xi})$ asymptotically for $\xi \to 0+$ and applying Watson's lemma or its generalizations. But expanding $f(be^{-\xi})$ for $\xi \to 0+$ is equivalent to expanding $f(t)$ for $t \to b-$. This and the fact that $\hat{f}(x) = I(+\infty; x)$ suggest that one should consider expanding $f(t)$ for $t \to +\infty$ in order to analyze the asymptotic behavior of $\hat{f}(x)$ for $x \to +\infty$.

Finally note that by making the change of variable of integration $t = e^{-\eta}$, (1.2) becomes

$$(1.9) \qquad \hat{f}(x) = \int_{-\infty}^{+\infty} e^{-x\eta} f(e^{-\eta})\,d\eta,$$

which is a two-sided Laplace transform. Hence, our results carry over to such transforms naturally. This point has been noted by various authors.

The main results of this work are given in the next section. These results are illustrated in §3 with examples that involve the summation of everywhere divergent moment series and some special functions.

2. Main results. Let the function $f(t)$ be as in the first paragraph of §1. In Theorems 2.1 and 2.2 of the present section we show that for some choices of the functions $u_k(t)$, $k = 0, 1, \cdots$, the sequences $\{u_0(t), u_1(t), \cdots\}$ and $\{\hat{u}_0(x), \hat{u}_1(x), \cdots\}$ are asymptotic sequences as $t \to +\infty$ and $x \to +\infty$ respectively, and that (1.5) implies (1.6). In the proofs of our results we make use of the following simple observations.

LEMMA 2.1. *For any fixed $T>0$*

$$\int_0^T t^{x-1}f(t)\,dt = O(T^x) \quad as\ x\to+\infty. \tag{2.1}$$

Proof. (2.1) is a consequence of the assumption that, for some real constant σ, $t^{\sigma-1}f(t)$ is absolutely integrable in any finite interval of the form $[0,a]$. □

LEMMA 2.2. *Assume $\{u_0(t),u_1(t),\cdots\}$ is an asymptotic sequence as $t\to+\infty$ and $f(t)$ has the asymptotic expansion in* (1.5). *Set*

$$r_n(t)=f(t)-\sum_{k=0}^{n-1}A_ku_k(t), \qquad n=1,2,\cdots. \tag{2.2}$$

Then for each positive integer n, there exist positive constants K and T that depend only on n, such that

$$|r_n(t)|\le K|u_n(t)|, \qquad t\ge T. \tag{2.3}$$

Proof. (2.3) follows from the fact that $\lim_{t\to+\infty}[r_n(t)/u_n(t)]=A_n$, which in turn is a consequence of $r_n(t)=A_nu_n(t)+O(u_{n+1}(t))$ as $t\to+\infty$ and (1.3). □

THEOREM 2.1. *Let*

$$u_k(t)=t^{-\lambda_k}\exp(-\alpha_kt^{\beta_k}), \qquad k=0,1,\cdots, \tag{2.4}$$

where

$$\lambda_k\ real,\quad \mathrm{Re}\,\alpha_k>0,\quad k=0,1,\cdots,\quad 0<\beta_0\le\beta_1\le\beta_2\le\cdots, \tag{2.5}$$

(2.6) *when $k<q$, $\beta_k=\beta_q$ implies either one of the four combinations* (a *and* c), (a *and* d), (b *and* c), *and* (b *and* d), *with*
a) $\mathrm{Re}\,\alpha_k<\mathrm{Re}\,\alpha_q$, b) $\mathrm{Re}\,\alpha_k=\mathrm{Re}\,\alpha_q$ *and* $\lambda_k<\lambda_q$,
c) $|\alpha_k|<|\alpha_q|$, d) $|\alpha_k|=|\alpha_q|$ *and* $\lambda_k<\lambda_q$,

and no restrictions are imposed on λ_k and α_k when $\beta_k<\beta_q$. Then $\{u_0(t),u_1(t),\cdots\}$ and $\{\hat{u}_0(x),\hat{u}_1(x),\cdots\}$ are asymptotic sequences as $t\to+\infty$ and $x\to+\infty$ respectively. If, for any nonnegative integer n, there exists an integer $N>n$, such that

(2.7)
either a) $\beta_n<\beta_N$
or b) $\beta_n=\beta_N$ *and* $|\alpha_n|<\mathrm{Re}\,\alpha_N$,
or c) $\beta_n=\beta_N$, $|\alpha_n|=\mathrm{Re}\,\alpha_N$ *and* $\lambda_n\leqq\lambda_N$

then (1.5) *implies* (1.6).

Proof. The first part of the theorem is a direct consequence of (2.4)–(2.6),

$$\hat{u}_k(x)=\beta_k^{-1}\alpha_k^{-(x-\lambda_k)/\beta_k}\Gamma\left(\frac{x-\lambda_k}{\beta_k}\right), \tag{2.8}$$

and Stirling's formula for the gamma function.

For the second part of the theorem it is sufficient to show that, for each positive integer n,

$$\hat{r}_n(x)=O(\hat{u}_n(x)) \quad \text{as } x\to+\infty, \tag{2.9}$$

where $r_n(t)$ has been defined in (2.2). For a given positive integer n, let N be as in the statement of the theorem. Then by Lemma 2.2 there exist positive constants K and T

that depend only on N, hence only on n, for which $|r_N(t)|\leq K|u_N(t)|$ when $t\geq T$. Now for sufficiently large x

$$(2.10)\qquad \hat{r}_N(x)=\int_0^T t^{x-1}f(t)\,dt-\sum_{k=0}^{N-1}A_k\int_0^T t^{x-1}u_k(t)\,dt+\int_T^\infty t^{x-1}r_N(t)\,dt.$$

Each one of the integrals $\int_0^T t^{x-1}f(t)\,dt$ and $\int_0^T t^{x-1}u_k(t)\,dt$, $k=0,1,\cdots,N-1$, is $O(T^x)$ as $x\to+\infty$, by Lemma 2.1. Furthermore,

$$(2.11)\qquad \left|\int_T^\infty t^{x-1}r_N(t)\,dt\right|\leq\int_T^\infty t^{x-1}|r_N(t)|\,dt\leq K\int_T^\infty t^{x-1}|u_N(t)|\,dt$$
$$<K\int_0^\infty t^{x-1}|u_N(t)|\,dt=K\beta_N^{-1}(\operatorname{Re}\alpha_N)^{-(x-\lambda_N)/\beta_N}\Gamma\left(\frac{x-\lambda_N}{\beta_N}\right).$$

Invoking now (2.7), (2.11) can be replaced by

$$(2.12)\qquad \int_T^\infty t^{x-1}r_N(t)\,dt=O(\hat{u}_n(x))\quad\text{as } x\to+\infty.$$

Thus (2.10) becomes

$$(2.13)\qquad \hat{r}_N(x)=O(T^x)+O(\hat{u}_n(x))=O(\hat{u}_n(x))\quad\text{as } x\to+\infty,$$

by (2.8) and Stirling's formula. But

$$(2.14)\qquad \hat{r}_n(x)=\hat{r}_N(x)+\sum_{k=n}^{N-1}A_k\hat{u}_k(x),$$

and $\hat{u}_k(x)=O(\hat{u}_n(x))$ as $x\to+\infty$ for each $k\geq n$ since $\{\hat{u}_0(x),\hat{u}_1(x),\cdots\}$ is an asymptotic sequence as $x\to+\infty$. Combining this and (2.13) in (2.14), (2.9) follows. This completes the proof of the theorem. □

The special case of (1.5) with $u_k(t)$ as given by (2.4) in Theorem 2.1, such that $\alpha_k=\alpha_{k+1}$ and $\beta_k=\beta_{k+1}$ for all $k=0,1,\cdots$, is of importance, and we turn to this case in Theorems 2.2 and 2.3 below. We first note that for this special case $u_k(t)$ is of the form

$$(2.15)\qquad u_k(t)=t^{-\lambda_k}\exp(-\alpha t^\beta),\qquad k=0,1,\cdots,$$

and, consequently

$$(2.16)\qquad \hat{u}_k(x)=\beta^{-1}\alpha^{-(x-\lambda_k)/\beta}\Gamma\left(\frac{x-\lambda_k}{\beta}\right),\qquad k=0,1,\cdots.$$

THEOREM 2.2. *Let $u_k(t)$ and $\hat{u}_k(x)$ be as in* (2.15) *and* (2.16), *where α,β and λ_k are all real, and*

$$(2.17)\qquad \alpha>0,\beta>0,\qquad \lambda_0<\lambda_1<\lambda_2<\cdots.$$

Then $\{u_0(t),u_1(t),\cdots\}$ and $\{\hat{u}_0(x),\hat{u}_1(x),\cdots\}$ are asymptotic sequences as $t\to+\infty$ and $x\to+\infty$ respectively, and (1.5) *implies* (1.6).

Proof. We observe that, with the present $u_k(t)$, all the conditions of Theorem 2.1 are satisfied with $\alpha_k=\alpha$, $\beta_k=\beta$, $k=0,1,\cdots$, and with $N=n+1$ for each nonnegative integer n. Therefore, Theorem 2.1 holds. This proves the theorem. □

If α in Theorem 2.2 is not real, then the proof of this theorem is no longer valid since (2.7) is not satisfied for any $N>n$. However, different arguments establish essentially the same result when α is complex with positive real part if $f(t)$ satisfies further conditions in the complex t-plane. This is given in Theorem 2.3 below.

THEOREM 2.3. *Let $u_k(t)$ and $\hat{u}_k(x)$ be as in* (2.15) *and* (2.16), *where α is now complex, and*

$$\operatorname{Re}\alpha>0, \quad \beta>0, \quad \lambda_0<\lambda_1<\lambda_2<\cdots. \tag{2.18}$$

Then $\{u_0(t),u_1(t),\cdots\}$ and $\{\hat{u}_0(x),\hat{u}_1(x),\cdots\}$ are asymptotic sequences as $t\to\infty$ (along any path in the complex t-plane possibly cut along the negative real axis) and $x\to+\infty$ respectively. Denote $\theta=\arg t$, $\omega=\arg\alpha$, $\theta_1=\min(0,-\omega/\beta)$, and $\theta_2=\max(0,-\omega/\beta)$. Assume that for some $T_0\geq 0$ and $\delta>0$ the function $f(t)$ is analytic in the set $D=\{t:|t|\geq T_0,\theta\in S\}$, where $S=(\theta_1-\delta,\theta_2+\delta)$, and that

$$f(t)\sim\sum_{k=0}^{\infty}A_k u_k(t) \quad as\ t\to\infty,\ \theta\in S. \tag{2.19}$$

If, in addition, for sufficiently large x

$$\lim_{R\to\infty}\int_{L(R)}t^{x-1}f(t)\,dt=0, \tag{2.20}$$

where $L(\rho)=\{t:t=\rho e^{i\theta},\theta$ goes from 0 to $-\omega/\beta\}$, then (1.6) *holds.*

Proof. As in Theorem 2.1, the first part of the present theorem is a direct consequence of (2.15), (2.16), (2.18), and Stirling's formula.

To prove that (1.6) holds we proceed as follows. Since $f(t)$ is analytic in D and satisfies (2.20), we can write

$$\hat{f}(x)=\left(\int_0^{T_0}+\int_{L(T_0)}+\int_C\right)t^{x-1}f(t)\,dt, \tag{2.21}$$

where $C=\{t:t=\rho e^{-i\omega/\beta}$, ρ goes from T_0 to $+\infty\}$. By Lemma 2.1

$$\int_0^{T_0}t^{x-1}f(t)\,dt=O(T_0^x) \quad \text{as } x\to+\infty. \tag{2.22}$$

Similarly, by analyticity properties of $f(t)$,

$$\left|\int_{L(T_0)}t^{x-1}f(t)\,dt\right|\leq\frac{|\omega|}{\beta}\left(\max_{\theta_1\leq\theta\leq\theta_2}|f(T_0e^{i\theta})|\right)T_0^x. \tag{2.23}$$

Now the integral along C can be reexpressed as

$$\int_C t^{x-1}f(t)\,dt=e^{-i\omega x/\beta}\int_{T_0}^{\infty}\rho^{x-1}f(\rho e^{-i\omega/\beta})\,d\rho, \tag{2.24}$$

and the function $F(\rho)=f(\rho e^{-i\omega/\beta})$ satisfies

$$F(\rho)\sim\exp(-|\alpha|\rho^\beta)\sum_{k=0}^{\infty}A_k\exp\left(\frac{i\omega\lambda_k}{\beta}\right)\rho^{-\lambda_k} \quad \text{as } \rho\to\infty \tag{2.25}$$

by (2.19). Thus the function $F_1(\rho)=H(\rho-T_0)F(\rho)$, where $H(y)$ is the Heaviside unit function, satisfies all the requirements of Theorem 2.2. Consequently

(2.26)

$$\int_{T_0}^{\infty}\rho^{x-1}F(\rho)\,d\rho=\hat{F}_1(\rho)=\sum_{k=0}^{n-1}A_k\exp\left(\frac{i\omega\lambda_k}{\beta}\right)\hat{v}_k(x)+O(\hat{v}_n(x))\quad \text{as } x\to+\infty,$$

where

$$\hat{v}_k(x)=\beta^{-1}|\alpha|^{-(x-\lambda_k)/\beta}\Gamma\left(\frac{x-\lambda_k}{\beta}\right),\qquad k=0,1,\cdots. \tag{2.27}$$

Combining (2.22)–(2.27) in (2.21), (1.6) follows. □

Remark. In Theorem 2.3, if we assume that (2.19) holds *uniformly* for $\theta\in S$, then (2.20) is automatically satisfied. To see this observe that, under this condition, there exist positive constants K and $T>T_0$ independent of t, such that

$$|f(t)|\le K|u_0(t)|,\qquad t\ge T,\quad \theta\in S. \tag{2.28}$$

Thus, for $R\ge T$,

$$\begin{aligned}\left|\int_{L(R)}t^{x-1}f(t)\,dt\right|&\le K\left|\int_{L(R)}|t|^{x-1}|u_0(t)||dt|\right|\\&=KR^{x-\lambda_0}\int_{\theta_1}^{\theta_2}\exp\left[-|\alpha|R^{\beta}\cos(\omega+\beta\theta)\right]d\theta.\end{aligned} \tag{2.29}$$

But for $\theta\in[\theta_1,\theta_2]$ we have $|\omega+\beta\theta|\le|\omega|<\pi/2$, thus $\cos(\omega+\beta\theta)\ge\cos\omega>0$. Consequently

$$\left|\int_{L(R)}t^{x-1}f(t)\,dt\right|\le K\frac{|\omega|}{\beta}R^{x-\lambda_0}\exp(-|\alpha|R^{\beta}\cos\omega), \tag{2.30}$$

and (2.20) follows by letting $R\to\infty$.

The corollary below gives a reformulation or rearrangement of the asymptotic expansion in (1.6) when $u_k(t)$ are as in Theorem 2.2 or Theorem 2.3 and $(\lambda_{k+1}-\lambda_k)/\beta$ is a fixed rational number for all $k=0,1,\cdots$. The form of the asymptotic expansion that is given by this corollary is more familiar and revealing than (1.6) itself, and we make extensive use of it in Examples 2–5.

COROLLARY. *Let p and q be two positive relatively prime integers, and let*

$$\lambda_{k+1}-\lambda_k=\frac{p}{q}\beta,\qquad k=0,1,\cdots, \tag{2.31}$$

in Theorem 2.2 *or Theorem* 2.3. *Then there exist constants* B_j, $j=0,1,\cdots$, *such that, for any positive integer n*,

$$\hat{f}(x)=\alpha^{-x/\beta}\Gamma\left(\frac{x-\lambda_0}{\beta}\right)\left[\sum_{j=0}^{n-1}\frac{B_j}{x^{j/q}}+O\left(\frac{1}{x^{n/q}}\right)\right]\quad \textit{as } x\to+\infty, \tag{2.32}$$

with $B_0=A_0\beta^{-1}\alpha^{\lambda_0/\beta}$.

Proof. Starting with (2.9), we have, for any positive integer n,

$$\hat{f}(x)=\hat{u}_0(x)\left[\sum_{k=0}^{n-1} A_k \frac{\hat{u}_k(x)}{\hat{u}_0(x)}+O\left(\frac{\hat{u}_n(x)}{\hat{u}_0(x)}\right)\right] \quad \text{as } x\to+\infty. \tag{2.33}$$

Making use of (2.31) and the formula (see [1, formula 6.1.47, p. 257])

$$z^{b-a}\frac{\Gamma(z+a)}{\Gamma(z+b)}\sim 1+\sum_{j=1}^{\infty}\frac{c_j}{z^j} \quad \text{as } z\to\infty, \tag{2.34}$$

where c_j are constants independent of z, we have from (2.16)

$$\begin{aligned}\frac{\hat{u}_k(x)}{\hat{u}_0(x)}&=\alpha^{(\lambda_k-\lambda_0)/\beta}\frac{\Gamma((x-\lambda_k)/\beta)}{\Gamma((x-\lambda_0)/\beta)}\\ &\sim x^{-kp/q}\sum_{j=0}^{\infty}\frac{d_{kj}}{x^j} \quad \text{as } x\to+\infty,\end{aligned} \tag{2.35}$$

where d_{kj} are constants independent of x. From (2.35) it also follows that

$$\frac{\hat{u}_n(x)}{\hat{u}_0(x)}=O(x^{-np/q}) \quad \text{as } x\to+\infty. \tag{2.36}$$

Combining (2.35) and (2.36) in (2.33), (2.32) follows. □

Finally we can introduce integral powers of $\log t$ in the functions $u_k(t)$ and still retain Theorems 2.1–2.3. The only additional results that one needs for proving this are

$$\int_0^\infty t^{x-1}(\log t)^m\exp(-\alpha t^\beta)\,dt=\left(-\frac{\partial}{\partial x}\right)^m\left[\beta^{-1}\alpha^{-x/\beta}\Gamma\left(\frac{x}{\beta}\right)\right] \tag{2.37}$$

and the asymptotic behavior of the psi function and its derivatives (see [1, formulas 6.3.18, 6.4.11, pp. 259–260]). We shall not pursue this further, as the results and the techniques for proving them are now obvious.

Note also that all the results of this section hold true if the integral $\int_0^\infty t^{x-1}f(t)\,dt$ is replaced by $\int_a^\infty t^{x-1}f(t)\,dt$ for any $a>0$, as the proofs depend solely on the asymptotic behavior of $f(t)$ for $t\to+\infty$ or $t\to\infty$ in a sector in the complex t-plane. We have already used this in the proof of Theorem 2.3, and shall use it in some of the examples in the next section.

3. Examples. We shall illustrate the results of the previous section by several examples. The first example is a straightforward application of Theorem 2.1. The second example arises in applying the T-transformation of Levin [6] to the partial sums of the everywhere divergent moment series

$$H(z)\sim\sum_{i=1}^{\infty}\frac{\mu_i}{z^i} \quad \text{as } z\to\infty, \tag{3.1}$$

where

$$H(z)=\int_0^\infty\frac{w(t)}{z-t}\,dt \tag{3.2}$$

and

$$(3.3)\qquad \mu_i=\int_0^\infty w(t)t^{i-1}dt,\qquad i=1,2,\cdots.$$

Ultimately, one is interested in the asymptotic behavior of the partial sum $\sum_{i=1}^{r-1}\mu_i/z^i$ as $r\to\infty$. It is easy to show that

$$(3.4)\qquad \sum_{i=1}^{r-1}\frac{\mu_i}{z^i}=H(z)-\frac{1}{z^r}\int_0^\infty\frac{w(t)t^{r-1}}{1-t/z}dt.$$

The integral on the right is simply a Mellin transform, and the problem is to find its asymptotic expansion as $r\to\infty$. This integral is actually related to the converging factor for the series in (3.1). In [10] the cases $w(t)=t^\gamma e^{-t}$, $\gamma>-1$, and $w(t)=t^\gamma E_m(t)$, $\gamma>-1$, $\gamma+m>0$, where $E_m(t)$ is the exponential integral, were considered. The results of [10] were used in [9] in the derivation of new numerical quadrature formulas for infinite range integrals with $w(x)$ above as the weight functions. For further details see [9], [10]. Finally, the rest of the examples deal with some special functions when their orders tend to infinity.

Example 1.

$$(3.5)\qquad I(x)=\int_0^\infty\frac{t^{x-1}e^{-ct}}{1-ze^{-t}}dt,\qquad \operatorname{Re}c>0.$$

Here $f(t)=e^{-ct}/(1-ze^{-t})$ satisfies all the requirements of Theorem 2.1 with $A_k=z^k$, $\lambda_k=0$, $\alpha_k=c+k$, $\beta_k=1$, $k=0,1,\cdots$. Hence

$$(3.6)\qquad I(x)\sim\Gamma(x)\sum_{k=0}^\infty\frac{z^k}{(c+k)^x}\quad\text{as } x\to+\infty.$$

It is worth noting that for $|z|<1$ this series converges and $\sim$ can be replaced by $=$. For $|z|>1$, however, the series diverges, but by Theorem 2.1, it represents $I(x)$ asymptotically as $x\to\infty$. A special case of this example is $I(x)=\Gamma(x)\zeta(x)$, where $\zeta(x)$ is the Riemann Zeta function, and is obtained by setting $c=1$, $z=1$.

Example 2.

$$(3.7)\qquad I(x)=\int_0^\infty\frac{t^{x-1}w(t)}{z-t}dt,\qquad z\notin[0,\infty).$$

We assume that

$$(3.8)\qquad w(t)\sim e^{-t}\sum_{k=0}^\infty\frac{c_k}{t^{k+\sigma}}\quad\text{as } t\to+\infty.$$

Therefore,

$$(3.9)\qquad f(t)=\frac{w(t)}{z-t}\sim e^{-t}\sum_{k=0}^\infty\frac{c_k}{t^{k+\sigma+1}}\sum_{j=0}^\infty\frac{z^j}{t^j}=e^{-t}\sum_{k=0}^\infty\frac{A_k(z)}{t^{k+\sigma+1}}\quad\text{as } t\to+\infty,$$

where $A_k(z)$ are polynomials in z. Thus $f(t)$ satisfies all the conditions of Theorem 2.2 and the corollary, with $\alpha=1$, $\beta=1$, $\lambda_k=k+\sigma+1$, $k=0,1,\cdots$. Hence

$$(3.10)\qquad I(x)\sim\Gamma(x-\sigma-1)\sum_{j=0}^{\infty}\frac{B_j(z)}{x^j}\quad \text{as } x\to+\infty,$$

where $B_j(z)$ are polynomials in z, and are independent of x. The $B_j(z)$ can be determined in terms of the c_k and z, but we shall not go into this. Furthermore, this expansion is valid for all $z\notin[0,\infty)$.

For the cases (a) $w(t)=t^\gamma e^{-t}$, and (b) $w(t)=t^\gamma E_m(t)$, where $E_m(t)=\int_1^\infty e^{-ty}/y^m dy$ is the exponential integral, (3.8) holds. For (a) (3.8) holds with $\sigma=-\gamma$ and $c_0=1$, $c_k=0$, $k\geq 1$. For (b) (3.8) holds with $\sigma=-\gamma+1$ and $c_k=(-1)^k(m)_k$, $k=0,1,\cdots$, where $(m)_k$ is the Pochhamer symbol, see [1, formula 5.1.51, p. 231]. Using entirely different techniques, in [10], (3.10) was shown to be valid for all $z\notin[0,\infty)$ for case (a) and for z with $\operatorname{Re} z<0$ for case (b). It is now obvious that (3.10) is valid for all $z\notin[0,\infty)$ as long as $w(t)$ satisfies (3.8).

Example 3. Asymptotic expansion of $K_\nu(z)$ as $\nu\to+\infty$.

Here $\operatorname{Re} z>0$ and $K_\nu(z)$ is the modified Bessel function of the second kind of order ν and has the integral representation, see [1, formula 9.6.23, p. 376],

$$(3.11)\qquad K_\nu(z)=\frac{\pi^{1/2}(z/2)^\nu}{\Gamma(\nu+1/2)}\int_1^\infty e^{-zt}(t^2-1)^{\nu-1/2}dt.$$

Making the change of variable or integration $t^2-1=\xi^2$, we have

$$(3.12)\qquad \begin{aligned} q_\nu(z)&=\int_1^\infty e^{-zt}(t^2-1)^{\nu-1/2}dt\\ &=\int_0^\infty \exp\left[-z(1+\xi^2)^{1/2}\right]\frac{\xi^{2\nu}}{(1+\xi^2)^{1/2}}d\xi.\end{aligned}$$

Now

$$(3.13)\qquad \frac{\exp\left[-z(1+\xi^2)^{1/2}\right]}{(1+\xi^2)^{1/2}}=\frac{e^{-z\xi}}{\xi}\left\{\frac{\exp\left(z\left[\xi-(1+\xi^2)^{1/2}\right]\right)}{(1+1/\xi^2)^{1/2}}\right\}.$$

It is not difficult to see that the term inside the curly brackets has a convergent expansion of the form

$$(3.14)\qquad q(\xi;z)=\frac{\exp\left(z\left[\xi-(1+\xi^2)^{1/2}\right]\right)}{(1+1/\xi^2)^{1/2}}=\sum_{k=0}^{\infty}\frac{A_k(z)}{\xi^k},\qquad \xi>1$$

with $A_k(z)$ being polynomials in z and $A_0(z)=1$. Therefore, the integrand of (3.12) satisfies all the conditions of Theorem 2.3 and the corollary, with $x=2\nu$, $\alpha=z$, $\beta=1$, $\lambda_k=k$, $k=0,1,\cdots$. Consequently

$$(3.15)\qquad q_\nu(z)\sim z^{-2\nu}\Gamma(2\nu)\sum_{j=0}^{\infty}\frac{B_j(z)}{\nu^j}\quad \text{as } \nu\to+\infty,$$

where the $B_j(z)$ are polynomials in z and $B_0(z)=1$. By (3.15) and the duplication formula for the gamma function, see [1, formula 6.1.18, p. 256], we obtain

$$K_\nu(z) \sim \frac{1}{2}\left(\frac{2}{z}\right)^\nu \Gamma(\nu)\left[1+\sum_{j=1}^{\infty}\frac{B_j(z)}{\nu^j}\right] \quad \text{as } \nu \to +\infty, \tag{3.16}$$

which, for integer ν, has also been derived in [12] by analyzing the power series of $K_\nu(z)$ for small z.

Example 4. Asymptotic expansion of $Y_\nu(z)$ as $\nu \to +\infty$.

Here $\operatorname{Re} z>0$ and $Y_\nu(z)$ is the Bessel function of the second kind of order ν and has the integral representation, see [1, formula 9.1.22, p. 360],

$$Y_\nu(z)=\frac{1}{\pi}\int_0^\pi \sin(z\sin\theta-\nu\theta)\,d\theta-\frac{1}{\pi}\int_0^\infty\left\{e^{\nu t}+e^{-\nu t}\cos(\nu\pi)\right\}e^{-z\sinh t}\,dt. \tag{3.17}$$

Integrating by parts once, we see that the first integral is $O(\nu^{-1})$ as $\nu \to +\infty$. In the second integral we have two contributions:

$$\begin{aligned} I_1&=-\frac{1}{\pi}\int_0^\infty e^{\nu t}e^{-z\sinh t}\,dt,\\ I_2&=-\frac{1}{\pi}\cos(\nu\pi)\int_0^\infty e^{-\nu t}e^{-z\sinh t}\,dt. \end{aligned} \tag{3.18}$$

Using Watson's lemma, we can show that $I_2 \sim \cos(\nu\pi)\sum_{j=1}^{\infty} b_j(z)/\nu^j$ as $\nu \to +\infty$, with $b_j(z)$ being independent of ν. In I_1 we make the change of variable of integration $e^t=\xi$, and obtain

$$I_1=-\frac{1}{\pi}\int_1^\infty \xi^{\nu-1}e^{-z\xi/2}e^{-z/(2\xi)}\,d\xi. \tag{3.19}$$

We can now apply Theorem 2.3 and the corollary, and obtain

$$I_1 \sim -\frac{1}{\pi}\left(\frac{z}{2}\right)^{-\nu}\Gamma(\nu)\left\{1+\sum_{j=1}^{\infty}\frac{c_j(z)}{\nu^j}\right\} \quad \text{as } \nu \to +\infty, \tag{3.20}$$

where the $c_j(z)$ are independent of ν. Consequently

$$Y_\nu(z) \sim -\frac{1}{\pi}\left(\frac{2}{z}\right)^\nu \Gamma(\nu)\left[1+O(\nu^{-1})\right] \quad \text{as } \nu \to +\infty. \tag{3.21}$$

Example 5. Asymptotic expansion of $H_\nu(z)-Y_\nu(z)$ as $\nu \to +\infty$.

Here $\operatorname{Re} z>0$ and $H_\nu(z)$ is the Struve function of order ν. We have, see [1, formula 12.1.18, p. 496],

$$H_\nu(z)-Y_\nu(z)=\frac{2(z/2)^\nu}{\pi^{1/2}\Gamma(\nu+1/2)}\int_0^\infty e^{-zt}(1+t^2)^{\nu-1/2}\,dt. \tag{3.22}$$

Making the change of variable of integration $1+t^2=\xi^2$, we have

$$q_\nu(z)=\int_0^\infty e^{-zt}(1+t^2)^{\nu-1/2}\,dt=\int_1^\infty \exp\left[-z(\xi^2-1)^{1/2}\right]\frac{\xi^{2\nu}}{(\xi^2-1)^{1/2}}\,d\xi. \tag{3.23}$$

As in Example 3,

$$\frac{\exp\left[-z(\xi^2-1)^{1/2}\right]}{(\xi^2-1)^{1/2}}=\frac{e^{-z\xi}}{\xi}\sum_{k=0}^{\infty}\frac{A_k(z)}{\xi^k},\qquad \xi>1, \tag{3.24}$$

where the $A_k(z)$ are polynomials in z and $A_0(z)=1$. Following Example 3, we obtain

$$H_\nu(z)-Y_\nu(z)\sim\frac{1}{\pi}\left(\frac{2}{z}\right)^{\nu}\Gamma(\nu)\left[1+O(\nu^{-1})\right]\quad \text{as } \nu\to+\infty. \tag{3.25}$$

Acknowledgments. The author wishes to thank Dr. John S. Lew of IBM T. J. Watson Research Center, Yorktown Heights, NY, for his kindness in locating the relevant literature, and the anonymous referee for his suggestions, which helped to improve the presentation of the material in the original version of this paper.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, Applied Mathematics Series 55, Government Printing Office, Washington, DC, 1964.

[2] G. Doetsch, *Handbuch der Laplace-Transformation*, Birkhäuser Verlag, Basel and Stuttgart, 1972.

[3] 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.

[4] ______, *Asymptotic expansion of Laplace transforms near the origin*, this Journal, 1 (1970), pp. 118–130.

[5] ______, *Asymptotic expansion of a class of integral transforms with algebraically dominated kernels*, J. Math. Anal. Appl., 35 (1971), pp. 405–433.

[6] D. Levin, *Development of nonlinear transformations for improving convergence of sequences*, Internat. J. Comput. Math., B3 (1973), pp. 371–388.

[7] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[8] E. Riekstinš, *Asymptotic Expansions of Integrals*, Izdat. Zinatne, Riga, 1977. (In Russian.) For a review of Volume 2 of this work see Mathematical Reviews 57 #3715.

[9] A. Sidi, *Numerical quadrature rules for some infinite range integrals*, Math. Comp., 38 (1982), pp. 127–142.

[10] ______, *Converging factors for some asymptotic moment series that arise in numerical quadrature*, J. Austral. Math. Soc. (Series B), 24 (1982), pp. 223–233.

[11] E. Wagner, *Taubersche Sätze für die Mellin-Transformation*, Wiss. Z. Univ. Halle, XV 66M, H.4, pp. 617–624.

[12] J. Wimp, *Recent developments in recursive computation* in Studies in Applied Mathematics 6, Special Functions and Wave Propagation, D. Ludwig and F. W. J. Olver, eds., Society for Industrial and Applied Mathematics, Philadelphia, 1970, pp. 110–123.

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

THE EXISTENCE OF HOMOCLINIC ORBITS AND THE METHOD OF MELNIKOV FOR SYSTEMS IN R^n*

JOSEPH GRUENDLER[†]

Abstract. We consider a periodically forced dynamical system possessing a small parameter, in arbitrary dimension. When the parameter is zero the system is autonomous with an explicitly known homoclinic orbit; we develop a criterion for this homoclinic orbit to persist for small, nonzero values of the parameter. The theory is applied to an example arising from a magnetized spherical pendulum.

The theory is a generalization to arbitrary dimension of the method of Melnikov. The example is a generalization to R^4 of a system in R^2 considered by Holmes.

1. Introduction. There has been considerable recent interest in the problem of chaotic solutions to deterministic systems. Some basic references to this subject are [1], [10] and [17]. One approach to predicting the onset of chaos is the use of perturbation theory and the method of Melnikov to detect transverse homoclinic orbits.

In his original work, [13], Melnikov considered the case of an analytic system in R^2. Recently, Holmes [9] and Sanders [18] have reduced the differentiability requirements to C^2 and made the proof more geometric. See also [5] and [4]. In [11], Holmes and Marsden consider the case of integrable Hamiltonian systems in higher dimension. The purpose of this work is to extend Melnikov's method to the non-Hamiltonian case of C^2 and arbitrary (finite) dimension. We also provide a pair of examples in R^4.

To motivate the general definitions in this section, consider a movable block with a smooth, curved surface as shown in Fig. 1 and a mass sliding on this surface. For our unperturbed system we assume that the block is at rest and that the mass slides with no damping. If the mass starts with a small velocity in the vicinity of point A, the motion will be periodic without passing through point B, and similarly for motions near point A'.

If the mass starts at a point higher than point B, the motion will be periodic with each cycle passing twice through point B.

Thus, in the phase portrait, points A and A' with zero velocity are centers while point B with zero velocity is an unstable equilibrium. Passing through this unstable equilibrium is a separatrix in the shape of a figure eight separating the two types of motion.

If S represents the horizontal displacement measured from point B, the differential equation has the form $\ddot{S}=F(S)$. In phase space, $x=(S,\dot{S})$, this takes the form $\dot{x}=f(x)$ where the origin is a saddle equilibrium and where there exists a homoclinic orbit.

We now add two perturbations. The first is viscous damping. In addition, we move the block with a sinusoidal motion. If the coordinate S moves with the block, we experience a sinusoidal inertial force. The differential equation becomes

$$\ddot{S}=F(S)-\varepsilon_1\dot{S}+\varepsilon_2\cos\omega t.$$

In phase space this takes the form

*Received by the editors May 4, 1984.

[†] Department of Mathematics, North Carolina A & T State University, Greensboro, North Carolina 27411.

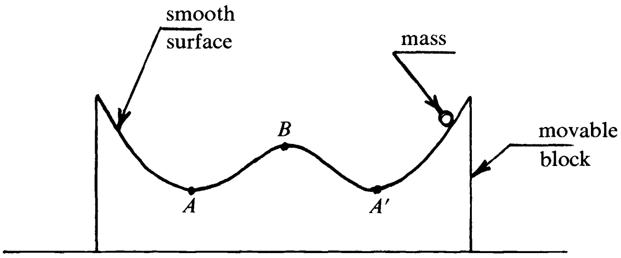


FIG. 1

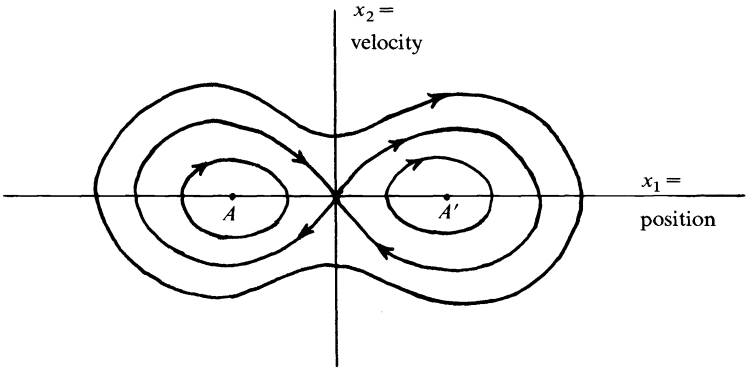


FIG. 2

$$(1.1) \qquad \dot{x} = f(x) - \varepsilon_1 Ax + \varepsilon_2 u \cos \omega t, \quad \text{where } A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad u = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

In this work, ω will denote angular frequency (radians per second). The circular frequency (cycles per second) will then be $\omega/2\pi$ and the period $2\pi/\omega$.

Numerical examples show that for certain values of $\varepsilon = (\varepsilon_1, \varepsilon_2)$ the solutions to equations like (1.1) exhibit a chaotic nature. See [10], [14], [15] and [16]. There is a way to see why such behavior should be expected. Fix the damping coefficient, ε_1, and vary ε_2, the amplitude of the applied force. When ε_2 is small the mass will oscillate with frequency ω near A or A'. When ε_2 is large the mass will again oscillate with frequency ω but now with large amplitude so that each cycle passes through point B twice.

For certain intermediate values of ε_2 the motion will alternate between a few cycles near A and a few cycles near A' with the numbers of each set of cycles apparently random.

Numerical experiments suggest that the onset of this chaotic behavior involves the appearance of homoclinic orbits in the perturbed system. This is reasonable since the existence of transverse homoclinic orbits is known to produce exotic dynamics (e.g., the existence of horseshoes). One approach, then, to studying the chaotic solutions to (1.1) is to study the homoclinic orbits of the perturbed system. The first step in this, determining the values $\varepsilon = (\varepsilon_1, \varepsilon_2)$ for which such orbits exist, is the problem to be considered here. We first state the problem precisely.

We introduce the notation B_δ to denote an open ball of radius δ centered at the origin in some Euclidean space. For future reference we now describe the dynamical system which we will study.

DEFINITION 1.1. A *perturbed homoclinic system* (PHS) is a dynamical system

$$\dot{x}(t)=f(x(t))+h(x(t),t,\varepsilon), \tag{1.2}$$

where f is a C^2 vector field on R^n and h is a C^2 function h: $R^n \times R \times B_\delta \to R^n$ for some $B_\delta \subset R^N$.

The function h is periodic in t with frequency ω and satisfies $h(0,t,\varepsilon)=h(x,t,0)=0$. The unperturbed system, $\dot{x}=f(x)$, has a saddle equilibrium at the origin and possesses at least one homoclinic orbit.

We could consider f,h to be defined on some open set $U \subset R^n$ containing the origin. We use all of R^n for convenience and because this covers the applications.

Notice that (1.1) does not satisfy $h(0,t,\varepsilon)=0$. This can always be fixed up. Since $x=0$ is an equilibrium for the unperturbed system, for sufficiently small $\|\varepsilon\|$, (1.1) has a periodic orbit which goes to zero as $\varepsilon \to 0$. We can then subtract this orbit from x.

Let us denote by W^s, $W^u \subset R^n$ the stable and unstable manifolds, respectively, of the origin for the unperturbed system $x=f(\dot{x})$ obtained by setting $\varepsilon=0$ in (1.2) and write $d_s=\dim(W^s)$, $d_u=\dim(W^u)$. The perturbed system is nonautonomous but as it is periodic in t we can consider the flow of (1.2) as autonomous on $R^n \times S^1$. Let $t \to \phi(t,x,\xi,\varepsilon)$ be the solution to (1.2) satisfying $\phi(\xi,x,\xi,\varepsilon)=x$ and let $t \to \bar{t}$ be the projection $R \to S^1$ obtained by reducing t mod$2\pi/\omega$. Then ϕ induces an orbit in $R^n \times S^1$ given by $t \to (\phi(t,x,\xi,\varepsilon),\bar{t})$.

The fixed point $x=0$ in R^n now becomes a periodic orbit $t \to (0,\bar{t})$. We denote the stable and unstable manifolds of this orbit by

$$\hat{W}_\varepsilon^s=\left\{(x,\bar{\xi}) \in R^n \times S^1 \,\middle|\, \lim_{t \to +\infty} \phi(t,x,\xi,\varepsilon)=0\right\},$$

$$\hat{W}_\varepsilon^u=\left\{(x,\bar{\xi}) \in R^n \times S^1 \,\middle|\, \lim_{t \to -\infty} \phi(t,x,\xi,t)=0\right\}.$$

We will find it convenient to look at sections of these manifolds. We define

$$W_\varepsilon^s(\bar{\xi})=\left\{x \in R^n \,\middle|\, \lim_{t \to \infty} \phi(t,x,\xi,\varepsilon)=0\right\},$$

$$W_\varepsilon^u(\bar{\xi})=\left\{x \in R^n \,\middle|\, \lim_{t \to -\infty} \phi(t,x,\xi,\varepsilon)=0\right\}.$$

Notice that making the identification $R^n=R^n \times \{\bar{\xi}\}$ we can think of the above as

$$W_\varepsilon^s(\bar{\xi})=\hat{W}_\varepsilon^s \cap (R^n \times \{\bar{\xi}\}), \qquad W_\varepsilon^u(\bar{\xi})=\hat{W}_\varepsilon^u \cap (R^n \times \{\bar{\xi}\}).$$

With this notation we get

$$W^s=W_0^s(\bar{\xi}) \quad \text{for all } \xi, \qquad W^u=W_0^u(\bar{\xi}) \quad \text{for all } \xi.$$

By hypothesis, W^s and W^u intersect. The question becomes: when do $\hat{W}_\varepsilon^s$ and $\hat{W}_\varepsilon^u$ do so?

DEFINITION 1.2. A *Melnikov function* for a PHS is a C^1 map Δ: $S^1 \times B_\delta \to R^{d_b}$ for some $B_\delta \subset R^N$ given by $(\bar{\xi},\varepsilon) \to \Delta(\bar{\xi},\varepsilon)$ with $d_b=\dim(T_PW^s \cap T_PW^u)=\operatorname{codim}(T_PW^s+T_PW^u)$, where $P \in W^s \cap W^u$. The function Δ has the property that $\Delta(\bar{\xi}^*,\varepsilon^*)=0$ if and only if the PHS has a homoclinic orbit when $\varepsilon=\varepsilon^*$.

We give a brief description of the Melnikov function to be developed. Let $\lambda_1, \lambda_2, \cdots, \lambda_n$ denote the eigenvalues of $Df(0)$ and assume for convenience that these are distinct. Let $t \to \gamma(t)$ be a homoclinic orbit for the unperturbed system and consider the variational equation $\dot{u}(t) = Df(\gamma(t))u(t)$. We show in §2 that this equation has a fundamental solution $\{\psi^{(1)}, \cdots, \psi^{(n)}\}$ with the property that each $\psi^{(i)}$ is asymptotic to some distinct $e^{\lambda_j t}$ as $t \to \infty$ and each $\psi^{(i)}$ is asymptotic to some distinct $e^{\lambda_k t}$ as $t \to -\infty$. Define an index set I by $i \in I$ if and only if $\lim_{t\to\infty} \psi^{(i)}(t) = \lim_{t\to-\infty} \psi^{(i)}(t) = \infty$. It follows from stable manifold theory that the order of I is d_b (see [6]).

Now form the function $D(t) = \det(\psi^{(1)}(t), \cdots, \psi^{(n)}(t)) \exp(-\int_0^t (\nabla \cdot f)(\gamma(s)))$ and let $K_{ij}(t, \bar{\xi})$ denote the result of replacing $\psi^{(i)}(t)$ in D by $h^{(j)}(\gamma(t), t + \bar{\xi})$ where $h^{(j)}(x, t) = \partial h / \partial \varepsilon_j(x, t, 0)$. Next, define

$$\Delta_{ij}(\bar{\xi}) = -\int_{-\infty}^{\infty} K_{ij}(t, \bar{\xi})\, dt, \qquad i \in I.$$

When a mild restriction on the K_{ij} is satisfied (Definition 3.1) the conditions for a homoclinic clinic become (Theorem 3.2):

$$\sum_{j=1}^{N} \Delta_{ij}(\bar{\xi})\varepsilon_j = 0, \qquad i \in I.$$

In §4 we provide a condition for the homoclinic orbit to be transverse (Theorem 4.1). When $d_b = 1$ this condition becomes

$$\sum_{k=1}^{N} \frac{\partial \Delta_{ik}}{\partial \bar{\xi}}(\bar{\xi}^*)\varepsilon_k^* \neq 0, \qquad I = \{i\}.$$

The reader who wishes to see an example of these formulas before their derivation can begin with §6.

We conclude this section by stating some standard results used later. For proofs of the next two theorems see [6].

THEOREM 1.3. *Let $\dot{x} = f(x) + h(x, t, \varepsilon)$ be a* PHS. *Let $P_1 \in W^s$, fix $\bar{\xi}$ and let Π^+ be a plane of dimension d_u in R^n transverse to W^s at P_1. Then, for sufficiently small $\|\varepsilon\|$, Π^+ meets $W^s_\varepsilon(\bar{\xi})$ in a point $q^+(\bar{\xi}, \varepsilon)$, where q^+ is C^1 in $(\bar{\xi}, \varepsilon)$ and $q^+(\bar{\xi}, \varepsilon) \to 0$ as $\varepsilon \to 0$.*

Similarly, a point $P_2 \in W^u$ and a plane Π^- of dimension d_s and transverse to W^u at P_2 define a point $q^-(\bar{\xi}, \varepsilon)$.

The general idea, now, is to find q^- and q^+ from a common point $P \in W^s \cap W^u$ and construct a function, Δ, using the requirement that $\Delta = 0$ when $q^+ = q^-$. The trick is to do this in such a way that Δ is easily obtainable from the perturbation h. The first step toward this end is to show that the orbits through q^+ and q^- stay near the orbit through P. Thus, in the following theorem one can think of $\nu(\bar{\xi}, \varepsilon)$ as $q^{\pm}(\bar{\xi}, \varepsilon)$. The reason for the slightly more abstract statement is to allow for some additional applications of the result.

In this next theorem we introduce some notation to be used throughout the work. We use M or M_i to denote a continuous, positive function on some $B_\delta \subset R^N$ satisfying $M(0) = 0$.

THEOREM 1.4. *Let*

$$\dot{x} = f(x) + h(x, t, \varepsilon) \tag{1.3}$$

be a PHS. *Let $B_{\delta_2} \subset R^N$, ν: $S^1 \times \bar{B}_{\delta_2} \to R^n$ be a C^1 function denoted $(\bar{\xi}, \varepsilon) \to \nu(\bar{\xi}, \varepsilon)$ with $\nu(\bar{\xi}, \varepsilon) \in W^s_\varepsilon(\bar{\xi})$. Let $D\nu$ represent the derivative of ν with respect to ε and let $t \to \gamma(t, \bar{\xi}, \varepsilon)$*

denote the solution to (1.3) *satisfying* $\gamma(\xi,\bar{\xi},\varepsilon)=\nu(\bar{\xi},\varepsilon)$. *Then*

a) γ *is* C^1 *in* ε *at* $\varepsilon=0$.

b) *Let* $D\gamma$ *denote the derivative of* γ *with respect to* ε. *Then* $t\to D\gamma(t,\bar{\xi},0)$ *satisfies the nonhomogeneous variational equation*

$$\dot{U}(t)=Df(\gamma(t,\bar{\xi},0))U(t)+D_1h(\gamma(t,\bar{\xi},0),t,0)U(t)+D_3h(\gamma(t,\bar{\xi},0),t,0),$$
$$U(\xi)=D\nu(\bar{\xi},0).$$

c) *Define* R_γ *by*

$$\gamma(t,\bar{\xi},\varepsilon)=\gamma(t,\bar{\xi},0)+D\gamma(t,\bar{\xi},0)\varepsilon+R_\gamma(t,\bar{\xi},\varepsilon).$$

Then there exist a constant $\alpha>0$ *and a function* M *such that* $\|R_\gamma(t,\bar{\xi},\varepsilon)\|\leqq\|\varepsilon\|M(\varepsilon)e^{-\alpha(t-\xi)}$ *for all* $t\geqq\xi$.

2. The variational equation. To motivate this section, we give a brief outline of Melnikov's original construction. Let $\dot{x}=f(x)+h(x,t,\varepsilon)$ be a PHS with $n=2$. Then $W^s\cap W^u$ consists of one, or possibly two, homoclinic orbits say $t\to\gamma(t)$. Let $P=\gamma(0)$.

Now $\dot{\gamma}(0)$ is a vector tangent to the orbit γ at P and $t\to\dot{\gamma}(t)$ is a solution to the variational equation $\dot{u}(t)=Df(\gamma(t))u(t)$. If $t\to\psi(t)$ is any solution to the variational equation independent from $\dot{\gamma}$, then $\psi(0)$ is a vector transverse to the orbit γ at P.

By Theorem 1.3 the line through P and along $\psi(0)$ determines $q^+(\bar{\xi},\varepsilon)$ and $q^-(\bar{\xi},\varepsilon)$. Melnikov now defines $\Delta(\bar{\xi},\varepsilon)=\Omega(q^+(\bar{\xi},\varepsilon)-q^-(\bar{\xi},\varepsilon),\dot{\gamma}(0))$ for some nondegenerate two form Ω. The fact that $\dot{\gamma}$ satisfies the variational equation is used in a nice way to get an expression in terms of $D_3h(x,t,0)$ for the part of Δ linear in ε.

The purpose of this section is to generalize the fundamental solution $\{\psi,\dot{\gamma}\}$.

We consider the system

$$\dot{x}(t)=f(x(t)), \tag{2.1}$$

where f is a C^1 vector field on R^n, the origin is a saddle equilibrium, and there exists a homoclinic orbit.

Let W^s, W^u denote, respectively, the stable and unstable manifolds of the origin and let $P\in W^s\cap W^u$. We denote $d_s=\dim(W^s)$, $d_u=\dim(W^u)$, and $d_b=\dim(T_pW^s\cap T_pW^u)$. Let $t\to\gamma(t)$ be the solution to (2.1) satisfying $\gamma(0)=P$.

The variational equation along γ is

$$\dot{u}(t)=Df(\gamma(t))u(t). \tag{2.2}$$

We seek a fundamental solution, $\{\psi^{(1)},\cdots,\psi^{(n)}\}$, to (2.2) possessing a number of properties. First, we require $\psi^{(n)}(t)=\dot{\gamma}(t)$.

We next require that the initial vectors, $\psi^{(i)}(0)$, span certain vector spaces. This requirement is summarized in Fig. 3. Note that the first d_u of these span a plane transverse to W^s while another d_s of them span a plane transverse to W^u.

The next requirement involves the exponential behavior of $\psi^{(i)}(t)$ as $t\to\pm\infty$. Let us write $\psi^{(i)}(t)\sim t^ae^{bt}v$ as $t\to+\infty$ to mean $\lim_{t\to\infty}\psi^{(i)}(t)t^{-a}e^{-bt}=v\in R^n$. Similarly for $t\to-\infty$. We now require that there exist a numbering, $\{\lambda_1,\cdots,\lambda_n\}$, according to algebraic multiplicity of the eigenvalues of $Df(0)$ such that $\psi^{(i)}(t)\sim t^{k_i}e^{\lambda_it}v^{(i)}$ as $t\to+\infty$ for some positive integers k_i and some vectors $v^{(i)}$. We also require that there be a permutation, σ, on n symbols and vectors $\bar{v}^{(i)}$ such that $\psi^{(i)}(t)\sim t^{k_{\sigma(i)}}e^{\lambda_{\sigma(i)}t}\bar{v}^{(i)}$ as $t\to-\infty$.

Our final requirement involves the signs of $\text{Re}(\lambda_i)$ and $\text{Re}(\lambda_{\sigma(i)})$. These requirements are summarized in Fig. 4.

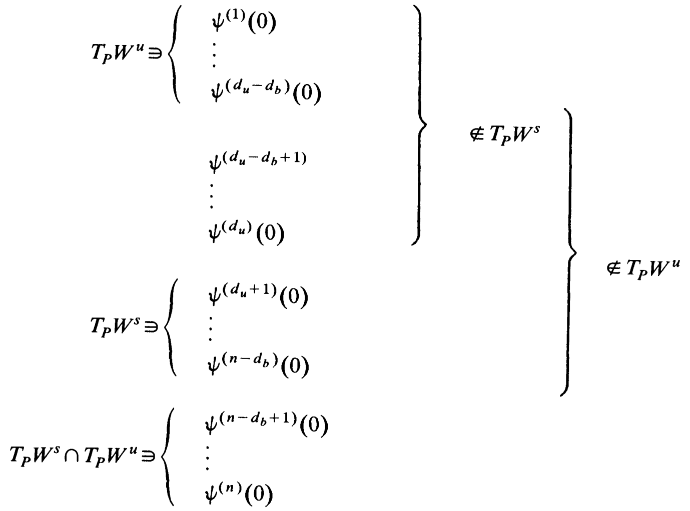


FIG. 3

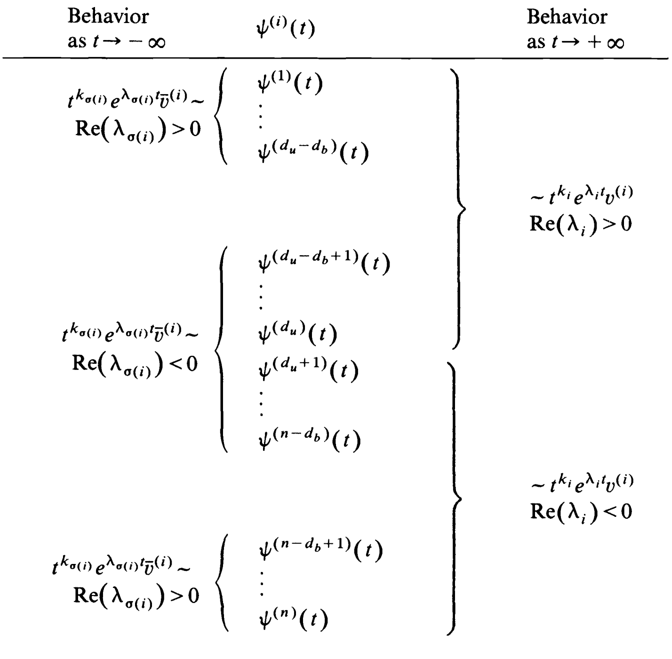

Behavior as $t \to -\infty$
$\psi^{(i)}(t)$
Behavior as $t \to +\infty$
$t^{k_{\sigma(i)}}e^{\lambda_{\sigma(i)}t}\bar{v}^{(i)} \sim$ $\mathrm{Re}(\lambda_{\sigma(i)})>0$
$\psi^{(1)}(t)$
$\psi^{(d_u-d_b)}(t)$
$\sim t^{k_i}e^{\lambda_i t}v^{(i)}$ $\mathrm{Re}(\lambda_i)>0$
$\psi^{(d_u-d_b+1)}(t)$
$t^{k_{\sigma(i)}}e^{\lambda_{\sigma(i)}t}\bar{v}^{(i)} \sim$ $\mathrm{Re}(\lambda_{\sigma(i)})<0$
$\psi^{(d_u)}(t)$
$\psi^{(d_u+1)}(t)$
$\psi^{(n-d_b)}(t)$
$\sim t^{k_i}e^{\lambda_i t}v^{(i)}$ $\mathrm{Re}(\lambda_i)<0$
$t^{k_{\sigma(i)}}e^{\lambda_{\sigma(i)}t}\bar{v}^{(i)} \sim$ $\mathrm{Re}(\lambda_{\sigma(i)})>0$
$\psi^{(n-d_b+1)}(t)$
$\psi^{(n)}(t)$

FIG. 4

Notice that for any n-form Ω we have

$$\Omega(\psi^{(1)}(t),\cdots,\psi^{(n)}(t)) \sim \Omega(v^{(1)},\cdots,v^{(n)})t^{k_1+\cdots+k_n}e^{tr(Df(0))t} \quad \text{as } t \to +\infty$$

and

$$\Omega(\psi^{(1)}(t),\cdots,\psi^{(n)}(t)) \sim \Omega(\bar{v}^{(1)},\cdots,\bar{v}^{(n)})t^{k_1+\cdots k_n}e^{tr(Df(0))t} \quad \text{as } t \to -\infty.$$

To prove the existence of the desired solutions to (2.2) begin with standard results concerning the asymptotic behavior of linear systems e.g. [2, Chap. 3, §8]. First apply these results at $+\infty$ and $-\infty$ separately and then splice them together. The fact that d_b

solutions must decay in both directions follows from a standard fact in stable manifold theory. A detailed proof of the following result can be found in [6].

THEOREM 2.1. *Equation* (2.2) *has a fundamental set of solutions* $\{\psi^{(1)},\cdots,\psi^{(n)}\}$ *satisfying the conditions summarized in Figs.* 3 *and* 4.

3. Definition of Δ.

We are now prepared to construct our Melnikov function. The original idea is due to Melnikov [13]. For a modernized treatment of the two dimensional case see Holmes [9], Sanders [18], Greenspan [4] and Greenspan and Holmes [5]. The essential new feature in the present, higher dimensional case is the use of the special solutions, given in §2, for the variational equation. We begin with some definitions and a statement of the main result.

Let

$$\dot{x}=f(x)+h(x,t,\varepsilon) \tag{3.1}$$

be a PHS with $x\in R^n$, $\varepsilon\in R^N$ and denote $h^{(j)}(x,t)=(\partial h/\partial\varepsilon_j)(x,t,0), j=1,2,\cdots,N$. Let $P\in W^s\cap W^u$ and let $t\rightarrow\gamma(t)$ be the solution to $\dot{x}=f(x)$ satisfying $\gamma(0)=P$.

The variational equation along γ is

$$\dot{u}(t)=Df(\gamma(t))u(t). \tag{3.2}$$

Let $\{\psi^{(1)},\cdots,\psi^{(n)}\}$ be solutions to (3.2) as described in §2. We can assume that $\det(\psi^{(1)}(0),\cdots,\psi^{(n)}(0))=1$. Note that the vectors $\{\psi^{(1)}(t),\cdots,\psi^{(d_u)}(t)\}$ are transverse to W^s at $\gamma(t)$ while the vectors $\{\psi^{(d_u-d_b+1)}(t),\cdots,\psi^{(n-d_b)}(t)\}$ are transverse to W^u as $\gamma(t)$.

We now define $K_{ij}(t,\bar{\xi})$ for $1\leqq i\leqq n-d_b$, $1\leqq j\leqq N$ to be the function obtained by substituting $h^{(j)}(\gamma(t),t+\xi)$ for $\psi^{(i)}(t)$ in the expression $\Omega(\psi^{(1)}(t),\cdots,\psi^{(n)}(t))\exp(-\int_0^t(\nabla\cdot f)(\gamma(s))\,ds)$. The function $K_{ij}(t,\bar{\xi})$ represents the projection onto the direction of $\psi^{(i)}(t)$ of the ε_j-linear part of the vector field h evaluated along γ.

DEFINITION 3.1. The equation (3.1) will be said to have a *transverse perturbation* if, given $\xi\in R$ and $\varepsilon\in R^N$, $\varepsilon\neq 0$, there exists a $t\in R$ and integers p,q with $1\leqq p\leqq d_u$ and $d_u-d_b+1\leqq q\leqq n-d_b$ such that

$$\sum_{j=1}^{N} K_{pj}(t,\bar{\xi})\varepsilon_j\neq 0 \quad\text{and}\quad \sum_{j=1}^{N} K_{qj}(t,\bar{\xi})\varepsilon_j\neq 0.$$

The derivation later in this section provides criteria for (3.1) to have a homoclinic orbit without the assumption that the perturbation is transverse. However, because this assumption is not too restrictive and is satisfied by the examples to follow it will be made throughout.

To measure the separation of $W_\varepsilon^s(\bar{\xi})$ and $W_\varepsilon^u(\bar{\xi})$ we define

$$\Delta_{ij}(\bar{\zeta})=-\int_{-\infty}^{\infty}K_{ij}(t,\bar{\xi})\,dt$$

for $a_u-d_b+1\leqq i\leqq d_u$ and $1\leqq j\leqq N$. Note that the range of i is such that in the integrand, K_{ij}, the $\psi^{(i)}$ omitted (and replaced by $h^{(j)}$) is one which exhibits exponential growth at both $+\infty$ and $-\infty$. We will show that $\Delta_{ij}(\bar{\xi})$ is the coefficient of ε_j in the ε-linear part of the distance between $W_\varepsilon^s(\bar{\xi})$ and $W_\varepsilon^u(\bar{\xi})$ along $\psi^{(i)}(0)$.

We now assume that $N\geqq d_b$, define a $d_b\times N$ matrix $A(\xi)=[a_{ij}]$ by $a_{ij}=\Delta_{i+d_u-d_b,j}$ and let $\bar{A}(\xi)$ denote the first d_b columns of $A(\xi)$. Note that, after a possible renumbering of the ε_j, when the rank of $A(\bar{\xi})$ is d_b we have $\det(\bar{A}(\bar{\xi}))\neq 0$.

The main result of this section is given in Theorem 3.2. This result is a corollary of Theorem 3.3.

THEOREM 3.2. *Suppose that* $N \geqq d_b$, *that* (3.1) *has a transverse perturbation, that* $\bar{\xi}^*$, ε^* *satisfy* $\sum_{j=1}^{N} \Delta_{ij}(\bar{\xi}^*)\varepsilon_j^* = 0$ *and that* $\det(\bar{A}(\bar{\xi}^*)) \neq 0$. *Then there exist an open interval* $J \subset R$, *containing zero and a* C^1 *map* $\alpha\colon J \to R^{d_b}$ *with* $\alpha(0) = (\varepsilon_1^*, \cdots, \varepsilon_{d_b}^*)$ *such that* (3.1) *has a homoclinic orbit when* $\varepsilon = s(\alpha_1(s), \cdots, \alpha_{d_b}(s), \varepsilon_{d_b+1}^*, \cdots, \varepsilon_N^*)$.

The rest of this section consists of a proof of the preceding result. We begin by stating two standard facts for future reference. First, if A is any linear transformation on R^n and $v^{(1)}, \cdots, v^{(n)}$ any n vectors in R^n then

$$\sum_{i=1}^{n} \Omega(v^{(1)}, \cdots, v^{(i-1)}, Av^{(i)}, v^{(i+1)}, \cdots, v^{(n)}) = \operatorname{Tr}(A)\Omega(v^{(1)}, \cdots, v^{(n)}). \tag{3.3}$$

Second, since each $\psi^{(i)}$ is a solution to $\dot{u} = Df(\gamma(t))u$ and $\Omega(\psi^{(1)}(0), \cdots, \psi^{(n)}(0)) = 1$,

$$\Omega(\psi^{(1)}(t), \cdots, \psi^{(n)}(t)) = \exp\left(\int_0^t (\nabla \cdot f)(\gamma(s))\, ds\right). \tag{3.4}$$

To follow the stable manifold under the perturbation we need a general plane (i.e., affine linear subspace) transverse to W^s at an arbitrary point, say $\gamma(t_0)$, on γ. To get a general set of vectors transverse to W^s at $\gamma(t_0)$ we define

$$u^{(i)+} = \psi^{(i)}(t_0) + \sum_{k=d_u+1}^{n} a_{ki}^{+}\psi^{(k)}(t_0), \quad 1 \leqq i \leqq d_u \tag{3.5}$$

for arbitrary scalars a_{ki}^{+}. We now let $\Pi^{+}(t_0, a^{+})$ denote the plane through $\gamma(t_0)$ spanned by the $u^{(i)+}$. Similarly, we define

$$u^{(i)-} = \psi^{(i)}(t_0) + \sum_{k \in I_u} a_{ki}^{-}\psi^{(k)}(t_0), \qquad d_u - d_b + 1 \leqq i \leqq n - d_b, \tag{3.6}$$

where $I_u = \{1, 2, \cdots, d_u - d_b\} \cup \{n - d_b + 1, \cdots, n\}$ and let $\Pi^{-}(t_0, a^{-})$ denote the plane, transverse to W^u at $\gamma(t_0)$, spanned by the $u^{(i)-}$.

From Theorem 1.3 the planes $\Pi^{+}(t_0)$ and $\Pi^{-}(t_0)$ determine, respectively, points

$$q^{+}(t_0, \bar{\xi}, \varepsilon, a^{+}) \in \Pi^{+}(t_0, a^{+}) \cap W_\varepsilon^s(\bar{t}_0 + \bar{\xi}), \quad q^{-}(t_0, \bar{\xi}, \varepsilon, a^{-}) \in \Pi^{-}(t_0, a^{-}) \cap W_\varepsilon^u(\bar{t}_0, \bar{\xi}).$$

Then (3.1) has a homoclinic solution $t \to \gamma(t, \varepsilon)$ with $\gamma(t, 0) = \gamma(t)$ if and only if $q^{+} = q^{-}$ for some t_0, $\bar{\xi}, a^{\pm}$.

We now define Δ^{+} and Δ^{-} by

$$q^{+}(t_0, \bar{\xi}, \varepsilon, a^{+}) - \gamma(t_0) = \sum_{i=1}^{d_u} \Delta_i^{+}(t_0, \bar{\xi}, \varepsilon, a^{+})u^{(i)+}, \tag{3.7}$$

$$q^{-}(t_0, \bar{\xi}, \varepsilon, a^{-}) - \gamma(t_0) = \sum_{i=d_u-d_b+1}^{n-d_b} \Delta_i^{-}(t_0, \bar{\xi}, \varepsilon, a^{-})u^{(i)-}. \tag{3.8}$$

If we define $\Delta_i = \Delta_i^+ - \Delta_i^-$ for $d_u - d_p + 1 \leqq i \leqq d_u$, the condition $q^+ - q^- = 0$ becomes

$$F_i(t_0, \bar{\xi}, \varepsilon, a^+, a^-) = \Delta_i^+(t_0, \bar{\xi}, \varepsilon, a^+) - \sum_{k=d_u-d_b+1}^{n-d_b} a_{ik}^- \Delta_k^-(t_0, \bar{\xi}, \varepsilon, a^-) = 0, \qquad 1 \leqq i \leqq d_u - d_b, \tag{3.9}$$

$$F_i(t_0, \bar{\xi}, \varepsilon, a^+, a^-) = \Delta_i(t_0, \bar{\xi}, \varepsilon, a^+, a^-) = 0, \qquad d_u - d_b + 1 \leqq i \leqq d_u, \tag{3.10}$$

$$F_i(t_0, \bar{\xi}, \varepsilon, a^+, a^-) = -\Delta_i^-(t_0, \bar{\xi}, \varepsilon, a^-) + \sum_{k=1}^{d_u} a_{ik}^+ \Delta_k^+(t_0, \bar{\xi}, \varepsilon, a^+) = 0, \qquad d_u + 1 \leqq i \leqq n - d_b, \tag{3.11}$$

$$\sum_{k=1}^{d_u} a_{ik}^+ \Delta_k^+(t_0, \bar{\xi}, \varepsilon, a^+) - \sum_{k=d_u-d_b+1}^{n-d_b} a_{ik}^- \Delta_k^-(t_0, \bar{\xi}, \varepsilon, a^-) = 0, \qquad n - d_b + 1 \leqq i \leqq n. \tag{3.12}$$

Notice that (3.12) can be satisfied by taking $a_{ij}^+ = a_{ij}^- = 0$ whenever $n - d_b + 1 \leqq i \leqq n$.

From Theorem 1.3 the $\Delta_i^\pm$ are C^1 functions of ε so that we can write

$$\Delta_i^\pm(t_0, \bar{\xi}, \varepsilon, a^\pm) = \sum_{j=1}^{N} \Delta_{ij}^\pm(t_0, \bar{\xi}) \varepsilon_j + R_i^\pm(t_0, \bar{\xi}, \varepsilon, a^\pm), \tag{3.13}$$

where we use $+$ for $1 \leqq i \leqq d_u$ and $-$ for $d_u - d_b + 1 \leqq i \leqq n - d_b$ and $\|R_i^\pm(t_0, \bar{\xi}, \varepsilon, a^\pm)\| < \|\varepsilon\| M_i^\pm(\varepsilon, a^\pm)$. In a similar way we write

$$\Delta_i(t_0, \bar{\xi}, \varepsilon, a^+, a^-) = \sum_{i=1}^{N} \Delta_{ij}(\bar{\xi}) \varepsilon_j + R_i(t_0, \bar{\xi}, \varepsilon, a^+, a^-)$$

for $d_u - d_b + 1 \leqq i \leqq d_u$. We will obtain explicit formulas for the $\Delta_{ij}^\pm$ and then, from the definiton of Δ_i, obtain $\Delta_{ij} = \Delta_{ij}^+ - \Delta_{ij}^-$.

Let $t \to \gamma^\pm(t, t_0, \bar{\xi}, \varepsilon, a^\pm)$ denote the solutions to (3.1) satisfying $\gamma^\pm(t_0 + \xi, t_0, \bar{\xi}, \varepsilon, a^\pm) = q^\pm(t_0, \bar{\xi}, \varepsilon, a^\pm)$. Then, from Theorem 1.4, we have

$$\gamma^\pm(t, t_0, \bar{\xi}, \varepsilon, a^\pm) = \gamma(t - \xi) + \sum_{j=1}^{N} v^{(j)\pm}(t, t_0, \bar{\xi}, a^\pm) + R^\pm(t, t_0, \bar{\xi}, \varepsilon, a^\pm), \tag{3.14}$$

where $\|R^\pm(t, t_0, \bar{\xi}, \varepsilon, a^\pm)\| \leqq \|\varepsilon\| M^\pm(\varepsilon, a^\pm) e^{\mp \alpha(t - t_0 - \xi)}$ and the $v^{(j)\pm}$ satisfy

$$\dot{v}^{(j)\pm}(t, t_0, \bar{\xi}, a^\pm) = Df(\gamma(t - \xi)) v^{(j)\pm}(t, t_0, \bar{\xi}, a^\pm) + h^{(j)}(\gamma(t - \xi), t). \tag{3.15}$$

Substituting $t = t_0 + \xi$ in (3.14), we get

$$q^\pm(t_0, \bar{\xi}, \varepsilon, a^\pm) = \gamma(t_0) + \sum_{j=1}^{N} v^{(j)\pm}(t_0 + \xi, t_0, \bar{\xi}, a^\pm) \varepsilon_j + R^\pm(t_0 + \xi, t_0, \bar{\xi}, \varepsilon, a^\pm). \tag{3.16}$$

Using (3.4)–(3.8) we see that

$$\begin{aligned} &\Delta_i^\pm(t_0, \bar{\xi}, \varepsilon, a^\pm) \\ &\quad = \Omega\big(\psi^{(1)}(t_0), \cdots, \psi^{(i-1)}(t_0), q^\pm(t_0, \bar{\xi}, \varepsilon, a^\pm) - \gamma(t_0), \psi^{(i+1)}(t_0), \cdots, \psi^{(n)}(t_0)\big) \\ &\quad\quad \times \exp\left(-\int_0^{t_0} (\nabla \cdot f)(\gamma(t))\, dt\right), \end{aligned}$$

where the ranges of + and − are as in (3.13). Substituting (3.16) into this result and comparing with (3.13) we see that

$$\Delta^{\pm}_{ij}(t_0,\bar{\xi})=\Omega\big(\psi^{(1)}(t_0),\cdots,\psi^{(i-1)}(t_0),v^{(j)\pm}(t_0+\xi,t_0,\bar{\xi},a^{\pm}),\psi^{(i+1)}(t_0),\cdots,\psi^{(n)}(t_0)\big)$$
$$\times\exp\left(-\int_0^{t_0}(\nabla\cdot f)(\gamma(t))\,dt\right),$$

where, again, the ranges of + and − are as in (3.13).

We now define

$$\phi^{+}_{ij}(t)=\Omega\big(\psi^{(1)}(t),\cdots,\psi^{(i-1)}(t),v^{(j)+}(t+\xi,t_0,\bar{\xi},a^{\pm}),\psi^{(i+1)}(t),\cdots,\psi^{(n)}(t)\big) \tag{3.17}$$

so that

$$\Delta^{+}_{ij}(t_0,\bar{\xi})=\phi^{+}_{ij}(t_0)\exp\left(-\int_0^{t_0}(\nabla\cdot f)(\gamma(t))\,dt\right). \tag{3.18}$$

Differentiating (3.17), substituting (3.15), and using (3.3), we get

$$\dot{\phi}^{+}_{ij}(t)=(\nabla\cdot f)(\gamma(t))\phi^{+}_{ij}(t)+K_{ij}(t,\bar{\xi})\exp\left(\int_0^{t}(\nabla\cdot f)(\gamma(s))\,ds\right).$$

This is a first order, linear, ordinary differential equation for ϕ^{+}_{ij}. Using an integrating factor we get

$$\phi^{+}_{ij}(t)\exp\left(-\int_0^{t}(\nabla\cdot f)(\gamma(s))\,ds\right) \tag{3.19}$$
$$=\phi^{+}_{ij}(t_0)\exp\left(-\int_0^{t_0}(\nabla\cdot f)(\gamma(s))\,ds\right)+\int_{t_0}^{t}K_{ij}(s,\bar{\xi})\,ds.$$

We wish to let $t\to\infty$ in this equation. We claim that the left-hand side goes to zero. This follows from the following observations:

i) In (3.19) $1\leqq i\leqq d_u$ so that $\mathrm{Re}(\lambda_i)>0$.

ii) From (3.14), $v^{(j)}(t,t_0,\bar{\xi},a^{\pm})\to 0$ as $t\to\infty$.

iii) Using (ii) and (3.12) as $t\to\infty$ the asymptotic behavior of $\phi^{+}_{ij}(t)$ is a factor which goes to zero multiplied by $\exp((\lambda_1+\cdots+\lambda_{i-1}+\lambda_{i+1}+\cdots+\lambda_n)t)$.

iv) As $t\to\infty$, the asymptotic behavior of $\exp(-\int_0^t(\nabla\cdot f)(\gamma(s))\,ds)$ is $\exp(-(\lambda_1+\cdots+\lambda_n)t)$.

A similar argument shows that $K_{ij}(t,\bar{\xi})$ is dominated by $e^{-\lambda_i t}$ as $t\to\infty$ so that this function is integrable from t_0 to ∞.

Now, letting $t\to\infty$ in (3.19) and using (3.18), we get

$$\Delta^{+}_{ij}(t_0,\bar{\xi})=-\int_{t_0}^{\infty}K_{ij}(t,\bar{\xi})\,dt,\qquad 1\leqq i\leqq d_u.$$

By a similar argument we get

$$\Delta^{-}_{ij}(t_0,\bar{\xi})=\int_{-\infty}^{t_0}K_{ij}(t,\bar{\xi})\,dt,\qquad d_u-d_b+1\leqq i\leqq n-d_b.$$

Finally, combining these formulas yields

$$\Delta_{ij}(\bar{\xi})=-\int_{-\infty}^{\infty}K_{ij}(t,\bar{\xi})\,dt,\qquad d_u-d_b+1\leqq i\leqq d_u.$$

We see now that the ε-linear part of the equation $F(t_0,\bar{\xi},\varepsilon,a^+,a^-)=0$ becomes

$$(3.20)\qquad \bar{F}_i\left(t_0,\bar{\xi},\varepsilon,a^-\right)=\sum_{j=1}^{N}\left[\Delta_{ij}^+\left(t_0,\bar{\xi}\right)-\sum_{k=d_u-d_b+1}^{n-d_b}a_{ik}^-\Delta_{kj}^-\left(t_0,\bar{\xi}\right)\right]\varepsilon_j=0,$$

$$1\leqq i\leqq d_u-d_b,$$

$$(3.21)\qquad \bar{F}_i(\bar{\xi},\varepsilon)=\sum_{j=1}^{N}\Delta_{ij}(\bar{\xi})\varepsilon_j=0,\qquad d_u-d_b+1\leqq i\leqq d_u,$$

$$(3.22)\qquad \bar{F}_i\left(t_0,\bar{\xi},\varepsilon,a^+\right)=\sum_{j=1}^{N}\left[-\Delta_{ij}^-\left(t_0,\bar{\xi}\right)+\sum_{k=1}^{d_u}a_{ik}^+\Delta_{kj}^+\left(t_0,\bar{\xi}\right)\right]\varepsilon_j=0,$$

$$d_u+1\leqq i\leqq n-d_b.$$

Suppose now that we have a solution for (3.21) i.e., we have a $\bar{\xi}^*$ and ε^* such that $\Sigma_{j=1}^N\Delta_{ij}(\bar{\xi}^*)\varepsilon_j^*=0$. Assuming that (3.1) has a transverse pertrubation we can find, for ξ^* and ε^*, t_0, p and q such that $\Sigma_{j=1}^N K_{pj}(t_0,\xi^*)\varepsilon_j^*\neq 0$. If we let $\phi(t)=\Sigma_{j=1}^N\Delta_{pj}^+(t,\bar{\xi}^*)\varepsilon_j^*$, we have $\dot{\phi}(t_0)=\Sigma_{j=1}^N K_{pj}(t_0,\bar{\xi}^*)\varepsilon_j\neq 0$ so $\phi(t)\neq 0$ for all $t\neq t_0$ sufficiently near t_0. Using this argument again we see that there exists a t_0^* at or near t_0 such that $\Sigma_{j=1}^N\Delta_{pj}^+(t_0^*,\bar{\xi}^*)\varepsilon_j^*\neq 0$ and $\Sigma_{j=1}^N\Delta_{qj}^-(t_0^*,\bar{\xi}^*)\varepsilon_j^*\neq 0$. Now, for $d_u+1\leqq i\leqq n-d_b$ define

$$x_{i-d_u}^*=\frac{\Sigma_{j=1}^N\Delta_{ij}^-\left(t_0^*,\bar{\xi}^*\right)\varepsilon_j^*}{\Sigma_{j=1}^N\Delta_{pj}^+\left(t_0^*,\bar{\xi}^*\right)\varepsilon_j^*}$$

and let $b_{ip}^+=x_{i-d_u}^*$, $b_{ij}^+=0$ for all other $1\leqq j\leqq d_u$.

Similarly, for $1\leqq i\leqq d_u-d_b$ define

$$z_i^*=\frac{\Sigma_{j=1}^N\Delta_{ij}^+\left(t_0^*,\bar{\xi}^*\right)\varepsilon_j^*}{\Sigma_{j=1}^N\Delta_{qj}^-\left(t_0^*,\bar{\xi}^*\right)\varepsilon_j^*}$$

and let $b_{iq}^-=z_i^*$, $b_{ij}^-=0$ for all other $d_u-d_b+1\leqq j\leqq n-d_b$. Then $\bar{F}(t_0^*,\bar{\xi}^*,\varepsilon^*,b^+,b^-)=0$. The following result shows that if $\mathrm{rank}(A(\bar{\xi}^*))=d_b$, then $F(t_0^*,\bar{\xi}^*,\varepsilon,a^+,a^-)=0$ for nearby values.

THEOREM 3.3. *Use the notation above, assume that* (3.1) *has a transverse perturbation, and that* $\mathrm{rank}(A(\bar{\xi}^*))=d_b$. *Then there exist an open interval* $J\subset R$ *containing zero and* C^1 *maps* $\alpha\colon J\to R^{d_b}$, $\alpha^+\colon J\to R^{d_s-d_b}\times R^{d_u}$, $\alpha^-\colon J\to R^{d_u-d_b}\times R^{d_s}$ *such that* $\alpha(0)=(\varepsilon_1^*,\cdots,\varepsilon_{d_b}^*)$, $\alpha^+(0)=b^+$, $\alpha^-(0)=b^-$ *and* $F(t_0^*,\bar{\xi}^*,\ s(\alpha_1(s),\cdots,\alpha_{d_b}(s),\varepsilon_{d_b+1}^*,\cdots,\varepsilon_N^*),\ \alpha^+(s),\alpha^-(s))=0$.

Proof. Let $\bar{A}(\bar{\xi}^*)$ denote the first d_b columns of $A(\bar{\xi}^*)$. By renumbering the ε_i if necessary we can assume that $\det(\bar{A}(\bar{\xi}^*))\neq 0$. For a sufficiently small interval $J\subset R$ containing zero and $B_\delta\subset R^{d_b}$ containing $(\varepsilon_1^*,\cdots,\varepsilon_{d_b}^*)$ we define a C^1 map $\phi\colon J\times R^{d_s-d_b}\times B_\delta\times R^{d_u-d_b}\to R^{n-d_b}$ as follows: first, for $(s,x,y,z)\in J\times R^{d_s-d_b}\times B_\delta\times R^{d_u-d_b}$ introduce the notation

$$e(s,y)=\left(sy_1,\cdots,sy_{d_b},s\varepsilon_{d_b+1}^*,\cdots,s\varepsilon_N^*\right);$$

for $d_u+1\leqq i\leqq n-d_b$ let

$$\beta_{ip}^+(x)=x_i,\qquad \beta_{ij}^+(x)=0\quad\text{for } 1\leqq j\leqq d_u;$$

for $1 \leqq i \leqq d_u - d_b$ let

$$\beta^-_{iq}(z) = z_i, \qquad \beta^-_{ij}(z) = 0 \quad \text{for } d_u - d_b + 1 \leqq j \leqq n - d_b.$$

Now define

$$\phi(s,x,y,z) = \frac{1}{s} F\big(t_0^*, \bar{\xi}^*, e(s,y), \beta^+(x), \beta^-(z)\big).$$

This is valid for $s \neq 0$. We extend ϕ in a C^1 way to $s = 0$ by defining

$$\phi(0,x,y,z) = \bar{F}\Big(t_0^*, \bar{\xi}^*, \big(y_1, \cdots, y_{d_b}, \varepsilon^*_{d_b+1}, \cdots, \varepsilon^*_N\big), \beta^+(x), \beta^-(z)\Big).$$

Then $\phi(0, x^*, (\varepsilon_1^*, \cdots, \varepsilon^*_{d_b}), z^*) = 0$. Furthermore, the derivative of ϕ with respect to (x,y,z) at $(0, x^*, (\varepsilon_1^*, \cdots, \varepsilon^*_{d_b}), z^*)$ has a matrix representation of the form

$$\left(\sum_{j=1}^{N} \Delta^+_{pj}\big(t_0^*, \bar{\xi}^*\big)\varepsilon_j^*\right)^{d_s - d_b} \left(\sum_{j=1}^{N} \Delta^-_{qj}\big(t_0^*, \bar{\xi}^*\big)\varepsilon_j^*\right)^{d_u - d_b} \times \left[\begin{array}{c:c:c} I & * & 0 \\ \hdashline 0 & \bar{A}(\bar{\xi}^*) & 0 \\ \hdashline 0 & * & I \end{array}\right].$$

The result now follows from the implicit function theorem. □

4. Transversality condition. The construction in §3 is sufficient for locating a homoclinic point of the perturbed system. An additional construction is required to determine when this point is one of transverse intersection of the stable and unstable manifolds.

Let

$$\dot{x} = f(x) + h(x, t, \varepsilon) \tag{4.1}$$

be a PHS with $t \to \gamma(t)$ a homoclinic orbit for $\dot{x} = f(x)$ and let $\{\psi^{(1)}, \cdots, \psi^{(n)}\}$ be as in Theorem 2.1 and $u^{(i)\pm}$ as in §3. Now let U^+ be a coordinate neighborhood on W^s centered at $\gamma(t_0)$, denote coordinates in U^+ by $s^+ = (s_i^+)$, $i \in \{d_u + 1, \cdots, n\}$, and let $P(s^+)$ denote the point with coordinates s^+. We can assume that $P(0) = \gamma(t_0)$ and $(\partial P/\partial s_i^+)(0) = \psi^{(i)}(t_0)$.

Let $\Pi^+(t_0, s^+)$ be the plane through $P(s^+)$ generated by $\{u^{(1)+}, \cdots, u^{(d_u)+}\}$. By choosing U^+ small enough we can assume that $\Pi^+(t_0, s^+)$ is transverse to W^s at $P(s^+)$ so, for sufficiently small ε, we have a uniquely defined point $q^+(t_0, \bar{\xi}, \varepsilon, a^+, s^+) \in \Pi^+(t_0, s^+) \cap W^s_\varepsilon(\bar{\xi})$ with $q^+ \to P(s^+)$ as $\varepsilon \to 0$.

We define Δ_i^+ by

$$q^+\big(t_0, \bar{\xi}, \varepsilon, a^+, s^+\big) = P(s^+) + \sum_{i=1}^{d_u} \Delta_i^+\big(t_0, \bar{\xi}, \varepsilon, a^+, s^+\big) u^{(i)^+}. \tag{4.2}$$

Following the development in the preceding section we let $t \to \gamma^+(t, s^+)$ denote the solution to $\dot{x} = f(x)$ with $\gamma^+(t_0, s^+) = P(s^+)$, write

$$\Delta_i^+\big(t_0, \bar{\xi}, \varepsilon, a^+, s^+\big) = \sum_{j=1}^{N} \Delta^+_{ij}\big(t_0, \bar{\xi}, s^+\big)\varepsilon_j + R_i^+\big(t_0, \bar{\xi}, \varepsilon, a^+, s^+\big), \tag{4.3}$$

and derive

$$\Delta^+_{ij}\big(t_0, \bar{\xi}, s^+\big) = -\int_{t_0}^{\infty} K^+_{ij}\big(t, \bar{\xi}, s^+\big)\, dt,$$

where the expression for K_{ij}^+ is obtained from K_{ij} of the preceding section by replacing $\gamma(t)$ by $\gamma^+(t,s^+)$.

In a similar way we define a neighborhood, U^-, of $\gamma(t_0)$ on W^u, establish coordinates $s^-=(s_i^-)$, $i\in I_u=\{1,\cdots,d_u-d_b\}\cup\{n-d_b+1,\cdots,n\}$ and define $P(s^-)$, $\Pi^-(t_0,s^-)$, $q^-(t_0,\bar{\xi},\varepsilon,a^-,s^-)$, $\Delta_i^-(t_0,\bar{\xi},\varepsilon,a^-,s^-)$, $\Delta_{ij}^-(t_0,\bar{\xi},s^-)$ in an analogous way.

We now define a $d_b\times d_b$ matrix $B(t_0,\bar{\xi},\varepsilon)=[b_{ij}]$, by

$$b_{i1}=\sum_{k=1}^{N}\frac{\partial\Delta_{i+d_u-d_b,k}}{\partial\bar{\xi}}(t_0,\bar{\xi},0)\varepsilon_k$$

and

$$b_{ij}=\sum_{k=1}^{N}\left[\frac{\partial\Delta^+_{i+d_u-d_b,k}}{\partial s^+_{j+n-d_b-1}}(t_0,\bar{\xi},0)-\frac{\partial\Delta^-_{i+d_u-d_b,k}}{\partial s^-_{j+n-d_b-1}}(t_0,\bar{\xi},0)\right]\varepsilon_k$$

for $2\leqq j\leqq d_b$, $1\leqq i\leqq d_b$. Suppose, now, that (4.1) has a homoclinic orbit for $\varepsilon=\varepsilon^*$. Then, if ε^* is sufficiently small, there exist t_0, $\bar{\xi}^*$, b^+, b^- such that $q^+(t_0^*,\bar{\xi}^*,\varepsilon^*,b^+,0)=q^-(t_0^*,\bar{\xi}^*,\varepsilon^*,b^-,0)$ by the previous section. Let $q\in W^s_{\varepsilon^*}(\bar{\xi}^*)\cap W^u_{\varepsilon^*}(\bar{\xi}^*)$ denote the common value $q^+=q^-$ and let $\hat{q}=(q,\bar{\xi}^*)\in R^n\times S^1$ so that $\hat{q}\in\hat{W}^s_{\varepsilon^*}\cap\hat{W}^u_{\varepsilon^*}$.

THEOREM 4.1. *With the notation of the preceding paragraph, $\hat{q}$ is a point of transverse intersection of $\hat{W}^s_{\varepsilon^*}$ and $\hat{W}^u_{\varepsilon^*}$ if and only if* $\det(B(t_0^*,\bar{\xi}^*,\varepsilon^*))\neq 0$ *when $\|\varepsilon^*\|$ is sufficiently small.*

Proof. Since $(q^+(t_0^*,\bar{\xi}^*,\varepsilon^*,b^+,s^+),\bar{\xi}^*)\in\hat{W}^s_{\varepsilon^*}$ for all s^+,

$$\left(\frac{\partial q^+}{\partial s_j^+}(t_0^*,\bar{\xi}^*,\varepsilon^*,b^+,0),0\right)\in T_{\hat{q}}\hat{W}^s_{\varepsilon^*},\qquad j\in\{d_u+1,\cdots,n-1\},\tag{4.4}$$

and

$$\left(\frac{\partial q^+}{\partial\bar{\xi}}(t_0^*,\bar{\xi}^*,\varepsilon^*,b^+,0),1\right)\in T_{\hat{q}}\hat{W}^s_{\varepsilon^*}.\tag{4.5}$$

Similarly,

$$\left(\frac{\partial q^-}{\partial s_j^-}(t_0^*,\bar{\xi}^*,\varepsilon^*,b^-,0),0\right)\in T_{\hat{q}}\hat{W}^u_{\varepsilon^*},\qquad j\in I_u, j\neq n,\tag{4.6}$$

and

$$\left(\frac{\partial q^-}{\partial\bar{\xi}}(t_0^*,\bar{\xi}^*,\varepsilon^*,b^-,0),1\right)\in T_{\hat{q}}\hat{W}^u_{\varepsilon^*}.\tag{4.7}$$

In the notation of the previous section the solution to (4.1) for $\varepsilon=\varepsilon^*$ passing through q at $t=t_0^*+\xi^*$ is $t\to\gamma^{\pm}(t,t_0^*,\bar{\xi}^*,\varepsilon^*,b^{\pm})$. Then $t\to(\gamma^{\pm}(t,t_0^*,\bar{\xi}^*,\varepsilon^*,b^{\pm}),t)$ is a homoclinic orbit in $R^n\times S^1$ passing through $\hat{q}$ at $t=t_0^*+\xi^*$ so that

$$\left(\dot{\gamma}^{\pm}(t^*+\xi^*,t_0^*,\bar{\xi}^*,\varepsilon^*,b^{\pm}),1\right)\in T_{\hat{q}}\hat{W}^s_{\varepsilon^*}\cap T_{\hat{q}}\hat{W}^u_{\varepsilon^*}.\tag{4.8}$$

Recall that, as $\varepsilon\to 0$, $\dot{\gamma}^{\pm}(t,t_0,\bar{\xi},\varepsilon,a^{\pm})\to\dot{\gamma}(t-\xi)=\psi^{(n)}(t-\xi)$.

Combining (4.2) and (4.3) we get

(4.9)

$$\frac{\partial q^+}{\partial s_j^+}\left(t_0^*,\bar{\xi}^*,\varepsilon^*,b^+,0\right)$$

$$=\psi^{(j)}\left(t_0^*\right)+\sum_{k=1}^{N}\sum_{i=1}^{d_u}\frac{\partial\Delta_{ik}^+}{\partial s_j^+}\left(t_0^*,\bar{\xi}^*,0\right)\left[\psi^{(i)}\left(t_0^*\right)+\sum_{r=d_u+1}^{n=d_b}b_{ri}^+\psi^{(r)}\left(t_0^*\right)\right]\varepsilon_k^*+o(\|\varepsilon\|)$$

and

$$\frac{\partial q^+}{\partial \bar{\xi}}\left(t_0^*,\bar{\xi}^*,\varepsilon^*,b^+,0\right)$$

$$=\sum_{k=1}^{N}\sum_{i=1}^{d_u}\frac{\partial\Delta_{ik}^+}{\partial \bar{\xi}}\left(t_0^*,\bar{\xi}^*,0\right)\left[\psi^{(i)}\left(t_0^*\right)+\sum_{r=d_u+1}^{n-d_b}b_{ri}^+\psi^{(r)}\left(t_0^*\right)\right]\varepsilon_k^*+o(\|\varepsilon\|).$$

The formulas are valid for $j\in\{d_u+1,\cdots,n-1\}$. Similar formulas can be obtained for $\partial q^-/\partial s_j$ valid for $j\in I_u, j\neq n$ and for $\partial q^-/\partial\bar{\xi}$.

From (4.9) we see that the d_s+1 vectors in (4.4), (4.5) and (4.8) are, for sufficiently small ε, linearly independent in $T_{\hat{q}}\hat{W}^s_{\varepsilon^*}$. Similarly, the d_u+1 vectors in (4.6), (4.7) and (4.8) are, for sufficiently small ε, linearly independent in $T_{\hat{q}}\hat{W}^u_{\varepsilon^*}$. Thus, $\hat{q}$ is a point of transverse intersection of these manifolds if and only if the vectors in (4.4)–(4.8) are linearly independent, i.e., if and only if the following determinant is nonzero:

$$D\left(t_0^*,\bar{\xi}^*,\varepsilon^*\right)$$

$$=\det\Bigg(\overbrace{\cdots\left(\frac{\partial q^-}{\partial s_j^-},0\right),\cdots,}^{j\in I_u,j\neq n}\left(\frac{\partial q^-}{\partial\bar{\xi}},1\right),\underbrace{\cdots,\left(\frac{\partial q^+}{\partial s_j^+},0\right),\cdots,}_{d_u+1\leqq j\leqq n-1}\left(\frac{\partial q^+}{\partial\bar{\xi}},1\right),\left(\dot{\gamma}^{\pm},1\right)\Bigg).$$

We expand this determinant by minors of the last row using (4.9) and the analogous results. We find that the minor for the $(n+1)$, $(n+1)$ entry is zero. The two nonzero minors yield:

$$D\left(t_0^*,\bar{\xi}^*,\varepsilon^*\right)=-\det\left(\psi^{(1)}\left(t_0^*\right),\cdots,\psi^{(n)}\left(t_0^*\right)\right)\left|\begin{array}{c:c:c} I & 0 & 0\\ \hdashline 0 & B & 0\\ \hdashline 0 & 0 & I\end{array}\right|+o\left(\|\varepsilon\|^{d_b}\right)$$

$$=-\det\left(B\left(t_0^*,\bar{\xi}^*,\varepsilon^*\right)\right)\exp\left(\int_0^{t_0}(\nabla\cdot f)(\gamma(s))\,ds\right)+o\left(\|\varepsilon\|^{d_b}\right).\qquad\square$$

Note that when $d_b=1$ the condition for transversality reduces to

$$\sum_{k=1}^{N}\frac{\partial\Delta_{d_uk}}{\partial\bar{\xi}}(\bar{\xi})\varepsilon_k\neq 0.$$

5. The special case of a manifold of homoclinic orbits. We wish to consider the case where $W^s\cap W^u$ has a connected component which is a manifold. This will occur, for example, if one of W^s, W^u contains the other. Let

$$\dot{x}=f(x)+h(x,t,\varepsilon)\tag{5.1}$$

be a PHS. Let W^s, W^u denote respectively the stable and unstable manifolds of the origin for the unperturbed system $\dot{x}=f(x)$ and denote $d_s=\dim(W^s)$, $d_u=\dim(W^u)$. Assume $W^s \cap W^u$ has a connected component, W^B, with a manifold structure. Let $\dim(W^B)=d_b$. We assume $d_b \geqq 2$.

We establish a coordinate system on W^B. Let (U,ϕ) be a local chart on W^B with $0 \in U$, ϕ: $U \to R^{d_b}$ a homeomorphism onto its image with $\phi(0)=0$. We can assume $S^{d_b-1} \subset \mathrm{Im}(\phi)$. Define $\Sigma^{d_b-1}=\phi^{-1}(s^{d_b-1})$. If $q \in W^B$ we define $(s_1(q),\cdots,s_{d_b-1}(q))$ as follows:

Let $t \to \beta_q(t)$ be the solution to $\dot{x}=f(x)$ with $\beta_q(0)=q$ and choose t_q so that $\beta_q(t_q) \in \Sigma^{n-1}$. Using Hartman's theorem, we can assume that U is small enough that t_q is unique and that the orbit β_q meets Σ^{d_b-1} transversely. Now let $(s_1(q),\cdots,s_{d_b-1}(q))$ be the spherical coordinates of $\phi(\beta_q(t_q)) \in S^{d_b-1}$. We will refer to $(s_1(q),\cdots,s_{d_b-1}(q))$ as the B-coordinates of q.

Notice that it is now possible to check the perturbation of every orbit in W^B for homoclinic points by checking the orbit through an arbitrary point on Σ^{d_b-1}. It is also possible to combine the constructions in §2 and §4.

Let $P \in \Sigma^{d_b-1}$ be arbitrary and denote the B-coordinates of P by $\theta=(\theta_1,\cdots,\theta_{d_b-1})$. Let $t \to \gamma(t,s)$ be the solution to $\dot{x}=f(x)$ such that $\gamma(0,s) \in \Sigma^{d_b-1}$ and has B-coordinates $s=(s_1,\cdots,s_{d_b-1})$. Then $\gamma(0,\theta)=P$.

We now consider the variational equation

$$\dot{u}(t,\theta)=Df(\gamma(t,\theta))u(t,\theta). \tag{5.2}$$

Let $t \to \psi^{(j)}(t,\theta)$, $1 \leqq j \leqq n$, be the solutions to (5.2) as in Theorem 2.1. Then $\psi^{(n)}(t,\theta)=\dot{\gamma}(t,\theta)$.

In the present special case it is sometimes possible to specify $\psi^{(j)}$ other than $\psi^{(n)}$. Notice that from Theorem 1.4 γ is C^1 in s and that the functions $t \to (\partial\gamma/\partial s_i)(t,\theta)$, $1 \leqq i \leqq d_b-1$ are solutions to (5.2). Also, from stable manifold theory (see [6])

$$\lim_{t \to +\infty} \frac{\partial\gamma}{\partial s_i}(t,\theta)=\lim_{t \to -\infty} \frac{\partial\gamma}{\partial s_i}(t,\theta)=0.$$

Some or all of $\partial\gamma/\partial s_i$ can be used for $\psi^{(j)}$, $n-d_b+1 \leqq j \leqq n-1$ according to the following theorem. The proof of this theorem follows from slight modification of the proof of Theorem 2.1.

THEOREM 5.1. *Let* $\{\lambda_1,\cdots,\lambda_n\}$ *be a numbering of the eigenvalues of* $Df(0)$ *according to algebraic multiplicity. Let*

$$\lim_{t \to \infty} \frac{\partial\gamma}{\partial s_j}(t,\theta)\exp(-\lambda_{\alpha(j)}t)t^{-p_j}=v^{(j)},$$

and

$$\lim_{t \to \infty} \frac{\partial\gamma}{\partial s_j}(t,\theta)\exp(-\lambda_{\beta(j)}t)t^{-q_j}=\bar{v}^{(j)}$$

with $v^{(j)}$, $\bar{v}^{(j)}$ *nonzero vectors in* R^n *and* $\lambda_{\alpha(j)}<0<\lambda_{\beta(j)}$, $1 \leqq j \leqq d_b-1$. *If the pairs* $(\alpha(j),p_j)$ *are all distinct and the pairs* $(\beta(j),q_j)$ *are all distinct, then we can take*

$$\psi^{(j)}(t,\theta)=\frac{\partial\gamma}{\partial s_{j-n+d_b}}(t,\theta), \qquad n-d_b+1 \leqq j \leqq n-1$$

in Theorem 2.1.

We now carry out, at P, the construction of §3. Computationally, one replaces $\gamma(t)$ with $\gamma(t,\theta)$ to get, in particular, $K_{ij}(t,\bar{\xi},\theta)$, $\Delta^{\pm}_{ij}(\bar{\xi},\theta)$ and $\Delta_{ij}(\bar{\xi},\theta)=\Delta^{+}_{ij}(\bar{\xi},\theta)-\Delta^{-}_{ij}(\bar{\xi},\theta)$. Definition 3.1 must now be replaced by the following:

DEFINITION 5.2. Equation (5.1) will be said to have a *uniformly transverse perturbation* if, given $\xi\in R$ and $\varepsilon\in R^N$ there exists a $t\in R$ and integers p,q with $1\leqq p\leqq d_u$ and $d_u-d_b+1\leqq q\leqq n-d_b$ such that

$$\sum_{j=1}^{N} K_{pj}(t,\bar{\xi},\theta)\varepsilon_j\neq 0 \text{ and } \sum_{j=1}^{N} K_{qj}(t,\bar{\xi},\theta)\varepsilon_j\neq 0 \quad \text{for all } \theta\in S^{d_b-1}.$$

The linearized conditions for a homoclinic orbit become

$$\sum_{j=1}^{N}\Delta_{ij}(\bar{\xi},\theta)\varepsilon_j=0, \qquad d_u-d_b\leqq i\leqq d_u. \tag{5.3}$$

We showed in Theorem 3.3 via the implicit function theorem that when (5.3) is satisfied the exact conditions for a homoclinic orbit can be satisfied with a slight change in d_b of the ε_j's. In the present situation, if we can solve (5.3) for $\bar{\xi},\theta$ and ε it is possible to satisfy the exact conditions for a homoclinic orbit by a slight modification of $\bar{\xi}$ and θ without changing ε.

We define a new $d_b\times d_b$ matrix $C(\bar{\xi},\theta,\varepsilon)=[c_{ij}]$ by

$$c_{i1}=\sum_{k=1}^{N}\frac{\partial\Delta_{i+d_u-d_b,k}}{\partial\bar{\xi}}(\bar{\xi},\theta)\varepsilon_k, \qquad c_{ij}=\sum_{k=1}^{N}\frac{\partial\Delta_{i-d_u+d_b,k}}{\partial\theta_{j-1}}(\bar{\xi},\theta)\varepsilon_k, \quad 2\leqq j\leqq d_b$$

for $1\leqq i\leqq d_b$. After proving the appropriate modification of Theorem 3.3 one gets the following version of Theorem 3.2.

THEOREM 5.3. *Suppose that* (5.1) *has a uniformly transverse perturbation, that* $\bar{\xi}^*$, θ^* *and* ε^* *satisfy* (5.3), *and* $\det(C(\bar{\xi}^*,\theta^*,\varepsilon^*))\neq 0$. *Then there exists an open interval* $J\subset R$ *containing the origin such that* (5.1) *has a homoclinic orbit when* $\varepsilon=s\varepsilon^*$ *for all* $s\in J$.

Let us now assume that we are able to take

$$\psi^{(j)}=\frac{\partial\gamma}{\partial s_{j-n+d_b}}, \qquad n-d_b+1\leqq j\leqq n-1. \tag{5.4}$$

This provides a nice unification of the results in the two preceding sections for we can replace the coordinates $s^{+}_{i+n-d_b}$ and $s^{-}_{i+n-d_b}$ in §4 with the common value θ_i as introduced in this section. This leads to the following version of Theorem 4.1.

THEOREM 5.4. *Suppose that* (5.1) *has a uniformly transverse perturbation, that* (5.4) *holds, and that* $\bar{\xi}^*$, θ^*, ε^* *satisfy* (5.3). *Then the homoclinic orbit predicted by these values is transverse if and only if* $\det(C(\bar{\xi}^*,\theta^*,\varepsilon^*))\neq 0$.

A further specialization occurs when $W^s=W^u$. In this case $d_s=d_u=d_b=n/2$, the planes Π^+ and Π^- agree and the $a^{\pm}_{ij}$ do not appear. The condition of a transverse intersection is not needed and the conditions for a homoclinic orbit are always

$$\sum_{j=1}^{N}\Delta_{ij}(\bar{\xi},\theta)\varepsilon_j=0, \qquad 1\leqq i\leqq\frac{n}{2}.$$

6. Application to a damped, magnetized spherical pendulum. As an application of our theory we consider a spherical pendulum. We assume that the pendulum bob is magnetized, that we have a second magnet fixed vertically below the pendulum support, and that these magnets are arranged to repel each other. Further, we allow for a fixed amount of damping, with damping constant c, along one direction of motion. The preceding constitutes our unperturbed system.

We incorporate three perturbation terms. The first perturbation is to replace the single fixed magnet with two magnets separated by a distance ε_1. The second perturbation is radially symmetric damping with damping coefficient ε_2. The third perturbation is an externally applied force of the form $\varepsilon \cos \omega t$ applied along an arbitrary horizontal direction. This force appears as two components with independent amplitudes ε_3 and ε_4.

The equations for this system can be derived by standard means. For details see [6]. The resulting equations are:

$$\begin{aligned}\ddot{x}&=x-2x(x^2+y^2)-3\varepsilon_1 x-\varepsilon_2\dot{x}+\varepsilon_3\cos\omega t,\\ \ddot{y}&=y-2y(x^2+y^2)-c\dot{y}-\varepsilon_1 y-\varepsilon_2\dot{y}+\varepsilon_4\cos\omega t.\end{aligned}$$

An equation of this type with slightly different coefficients has been studied numerically by Moon [14]. When $y=\dot{y}=0$ we get the equation studied by Holmes in [10].

Now the unperturbed equations have an equilibrium at $(x,\dot{x},y,\dot{y})=0$ but the perturbed equations do not. Accordingly we seek a periodic solution of the form

$$x_s(t,\varepsilon)=\sum_{i=1}^{4}\varepsilon_i w^{(i)}(t)+o(\|\varepsilon\|),\qquad y_s(t,\varepsilon)=\sum_{i=1}^{4}\varepsilon_i w^{(i)}(t)+o(\|\varepsilon\|).$$

By substituting into the differential equations we get

$$x_s(t,\varepsilon)=\frac{-\varepsilon_3}{\omega^2+1}\cos\omega t+o(\|\varepsilon\|),\qquad y_s(t,\varepsilon)=\frac{-\varepsilon_4}{\omega^2+1}\cos\omega t+o(\|\varepsilon\|).$$

We now replace x by $x-(\varepsilon_3/\omega^2+1)\cos\omega t+o(\|\varepsilon\|)$, y by $y-(\varepsilon_4/\omega^2+1)\cos\omega t+o(\|\varepsilon d\|)$, to get

$$\begin{aligned}\ddot{x}&=x-2x(x^2+y^2)-3\varepsilon_1 x-\varepsilon_2\dot{x}\\ &\quad+\frac{2\varepsilon_3}{\omega^2+1}(3x^2+y^2)\cos\omega t+\frac{4\varepsilon_4}{\omega^2+1}xy\cos\omega t+o(\|\varepsilon\|),\\ \ddot{y}&=y-2y(x^2+y^2)-c\dot{y}-\varepsilon_1 y-\varepsilon_2\dot{y}\\ &\quad+\frac{4\varepsilon_3}{\omega^2+1}xy\cos\omega t+\frac{2\varepsilon_4}{\omega^2+1}(x^2+3y^2)\cos\omega t+o(\|\varepsilon\|).\end{aligned}\tag{6.1}$$

These equations take the first order form

$$\begin{aligned}\dot{u}_1&=u_2,\\ \dot{u}_2&=u_1-2u_1(u_1^2+u_3^2)-3\varepsilon_1 u_1-\varepsilon_2 u_2\\ &\quad+\frac{2\varepsilon_3}{\omega^2+1}(3u_1^2+u_3^2)\cos\omega t+\frac{4\varepsilon_4}{\omega^2+1}u_1u_3\cos\omega t+o(\|\varepsilon\|),\\ \dot{u}_3&=u_4,\\ \dot{u}_4&=u_3-2u_3(u_1^2+u_3^2)-cu_4-\varepsilon_1 u_3-\varepsilon_1 u_4\\ &\quad+\frac{4\varepsilon_3}{\omega^2+1}u_1u_3\cos\omega t+\frac{2\varepsilon_4}{\omega^2+1}(u_1^2+3u_3^2)\cos\omega t+o(\|\varepsilon\|).\end{aligned}\tag{6.2}$$

The unperturbed equations are

$$
\begin{aligned}
&\dot{u}_1=f_1(u)=u_2,\\
&\dot{u}_2=f_2(u)=u_1-2u_1(u_1^2+u_3^2),\\
&\dot{u}_3=f_3(u)=u_4,\\
&\dot{u}_4=f_4(u)=u_3-2u_3(u_1^2+u_3^2)-2cu_4,
\end{aligned}
\tag{6.3}
$$

from which we can compute

$$
Df(u)=\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1-6u_1^2-2u_3^2 & 0 & -4u_1u_3 & 0 \\ 0 & 0 & 0 & 1 \\ -4u_3u_1 & 0 & 1-2u_1^2-6u_3^2 & -2c \end{bmatrix}
$$

and

$$
Df(0)=\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & -2c \end{bmatrix}.
$$

The eigenvalues of this matrix are ± 1 and $-c\pm\alpha$ where $\alpha=\sqrt{c^2+1}$. Thus $d_s=d_u=2$. Since one direction of motion is damped, we expect $d_b=1$. A homoclinic solution for (6.3) is given by $\gamma(t)=(r(t),\dot{r}(t),0,0)$ where $r(t)=\operatorname{sech} t$; see [10]. As a point on γ, we have $P=\gamma(0)=(1,0,0,0)\in W^s\cap W^u$. The variational equation along this solution, $\dot{\psi}=Df(\gamma(t))\psi$, becomes

$$
\dot{\psi}_1=\psi_2,\quad \dot{\psi}_2=(1-6r^2)\psi_1,\quad \dot{\psi}_3=\psi_4,\quad \dot{\psi}_4=(1-2r^2)\psi_3-2c\psi_4. \tag{6.4}
$$

These uncouple into a pair of second order equations

$$
\ddot{\psi}_1=(1-6r^2)\psi_1, \tag{6.5}
$$

$$
\ddot{\psi}_3=(1-2r^2)\psi_3-2c\dot{\psi}_3. \tag{6.6}
$$

We seek four solutions to (6.4) as described in §2. We first take

$$
\psi^{(4)}(t)=\dot{\gamma}(t)=(\dot{r}(t),\ddot{r}(t),0,0). \tag{6.7}
$$

We find that, as $t\to+\infty$, $\psi^{(4)}$ is asymptotic to e^{-t}. Thus we set $\lambda_4=1$. Since, as $t\to-\infty$, $\psi^{(4)}$ is asymptotic to e^t, $\lambda_{\sigma(4)}=1$.

From (6.7) we get a solution to (6.5), namely $\psi_1(t)=\dot{r}(t)$. By the method of variation of parameters we can get a second solution to (6.5). Substituting $\psi_1(t)=Q(t)\dot{r}(t)$ into (6.5) we find

$$
Q(t)=-\tfrac{3}{2}t-\tfrac{1}{2}\sinh t\cosh t+\coth t.
$$

The arbitrary constant multiplying Q has been chosen so that

$$
\begin{vmatrix} Q\dot{r} & \dot{r} \\ (Q\dot{r})^{\cdot} & \ddot{r} \end{vmatrix}=1.
$$

We find that $Q(t)\dot{r}(t)$ is asymptotic to e^t at $+\infty$ and e^{-t} at $-\infty$. Thus we set $\lambda_2=1$, $\lambda_{\sigma(2)}=-1$ and $\psi^{(2)}(t)=((Q\dot{r})(t),(Q\dot{r})^{\cdot}(t),0,0)$.

We now must find $\psi^{(1)}$ asymptotic to $e^{(-c+\alpha)t}$ at $\pm\infty$ and $\psi^{(3)}$ asymptotic to $e^{(-c-\alpha)t}$ at $\pm\infty$. These solutions will come from (6.6).

We seek a solution to (6.6) of the form $\psi_3(t)=e^{-ct}w(t)$. Substituting we get

$$\ddot{w}-(\alpha^2-2r^2)w=0. \tag{6.8}$$

Note that if $t\to w(t)$ is one solution to this equation, $t\to w(-t)$ will be another.

We seek a solution to (6.8) of the form $w(t)=\phi(t)\operatorname{sech}^{\alpha}t$. Substituting yields

$$\ddot{\phi}-(2\alpha\tanh t)\dot{\phi}-\left[(\alpha^2+\alpha-2)\operatorname{sech}^2 t\right]\phi=0.$$

We now make a change in the independent variable by letting $x=\operatorname{sech}^2 t$. This produces

$$x(1-x)\frac{d^2\phi}{dx^2}+\left[(\alpha+1)-\left(\alpha+\frac{3}{2}\right)x\right]\frac{d\phi}{dx}-\left(\frac{\alpha}{2}+1\right)\left(\frac{\alpha}{2}-\frac{1}{2}\right)\phi=0.$$

This is a hypergeometric equation. Using the second form of the solution given in [7, 15.5.3], we get

$$\phi=(\tanh t)F\left(\frac{\alpha}{2},\frac{\alpha}{2}+\frac{3}{2};\alpha+1;x\right).$$

Using formula [7, 15.2.24], substituting $a=\alpha/2$, $b=\alpha/2+\frac{1}{2}$, $c=\alpha+1$, $z=x=\operatorname{sech}^2 t$, and combining with [7, 15.1.13] and [7, 15.1.14], we get

$$\phi(t)=\tanh t\left[\frac{2^{\alpha}}{\alpha+1}\frac{\alpha+\tanh t}{(1+\tanh t)^{\alpha}\tanh t}\right]=\frac{2^{\alpha}}{\alpha+1}\frac{\alpha+\tanh t}{(1+\tanh t)^{\alpha}}.$$

Dropping the constant we can take

$$w(t)=\frac{(\alpha+\tanh t)\operatorname{sech}^{\alpha}t}{(1+\tanh t)^{\alpha}}=\frac{(\alpha+\tanh t)(1-\tanh t)^{\alpha}}{\operatorname{sech}^{\alpha}t}.$$

Note that $w(t)$ is asymptotic to $e^{\alpha t}$ as $t\to\pm\infty$. We now take as our two solutions to (6.6)

$$v_1(t)=e^{-ct}w(-t),\qquad v_2(t)=e^{-ct}w(t).$$

For another method of solving (6.6) see [19]. We now define

$$\psi^{(1)}(t)=\frac{1}{c\sqrt{2\alpha}}(0,0,v_1(t),\dot{v}_1(t)),\qquad \psi^{(3)}(t)=\frac{1}{c\sqrt{2\alpha}}(0,0,v_2(t),\dot{v}_2(t)).$$

The arbitrary constant has been chosen so that $\det(\psi^{(1)}(0),\psi^{(2)}(0),\psi^{(3)}(0),\psi^{(4)}(0))=1$. From stable manifold theory we have (see [6])

$$\begin{aligned}&\psi^{(1)}(0)\notin T_PW^s,\ \in T_PW^u, &&\psi^{(3)}(0)\in T_PW^s,\ \notin T_PW^u\\ &\psi^{(2)}(0)\notin T_PW^s,\ \notin T_PW^u, &&\psi^{(4)}(0)\in T_PW^s\cap T_PW^u.\end{aligned}$$

This confirms that $d_b=1$.

At this point we can compute the K_{ij}. For details see [6]. The results are

$$\begin{aligned}
&K_{11}=K_{12}=K_{13}=0,\\
&K_{14}(t,\bar{\xi})=\frac{2}{c\sqrt{2\alpha}\,(\omega^2+1)}\,\frac{e^{ct}(\alpha+\tanh t)\operatorname{sech}^{\alpha+2}t}{(1+\tanh t)^{\alpha}}\cos\omega(t+\xi),\\
&K_{21}(t,\bar{\xi})=-3\operatorname{sech}^2 t\tanh t,\\
&K_{22}(t,\bar{\xi})=\operatorname{sech}^2 t\tanh^2 t,\\
&K_{23}(t,\bar{\xi})=\frac{6}{\omega^2+1}\operatorname{sech}^3 t\tanh t\cos\omega(t+\xi),\\
&K_{24}(t,\bar{\xi})=0,\\
&K_{31}=K_{32}=K_{33}=0,\\
&K_{34}(t,\bar{\xi})=\frac{-2}{c\sqrt{2\alpha}\,(\omega^2+1)}\,\frac{e^{ct}(\alpha-\tanh t)\operatorname{sech}^{\alpha+2}t}{(1-\tanh t)^{\alpha}}\cos\omega(t+\xi).
\end{aligned}$$

From these results one can easily verify that the system has a transverse perturbation. Thus, we need only calculate $\Delta_{2j}(\bar{\xi})$ for $1\leqq j\leqq 4$. For details see [6]. The results are

$$\Delta_{21}=\Delta_{24}=0,\quad \Delta_{22}=-\frac{2}{3},\quad \Delta_{23}(\bar{\xi})=\frac{\pi\omega\sin\omega\xi}{\cos(\pi\omega/2)}.$$

The condition for a homoclinic orbit now becomes

$$\sum_{j=1}^{4}\Delta_{2j}(\bar{\xi})\varepsilon_j=-\frac{2}{3}\varepsilon_2+\left[\frac{\pi\omega\sin\omega\xi}{\cosh(\pi\omega/2)}\right]\varepsilon_3=0.$$

This is the same result as obtained by Holmes in [10]. The analysis here shows that the effect of the damping is to project the problem onto the x-$\dot{x}$ plane when ε is small. However, in the next example we will see that new homoclinic orbits will appear when ε is large enough (compared to the damping coefficient c). A weakness of the present method is that it cannot detect homoclinic orbits which appear in this way.

7. Application to an undamped, magnetized spherical pendulum. The results of §6 are not valid when $c=0$, which case we consider here. This case will utilize the theory of §5.

The case $c=0$ represents a pendulum in which the unperturbed system is undamped and radially symmetric. The viscous damping perturbation is also radially symmetric but the ε_1 term, representing the separation of the fixed magnets, destroys the radial symmetry and leaves two planes of symmetry. The two forcing terms can be thought of as components of a single force applied at some arbitrary angle.

The equations of motion become

$$\begin{aligned}
\ddot{x}&=x-2x(x^2+y^2)-3\varepsilon_1 x-\varepsilon_2\dot{x}\\
&\quad+\frac{2\varepsilon_3}{\omega^2+1}(3x^2+y^2)\cos\omega t+\frac{4\varepsilon_4}{\omega^2+1}xy\cos\omega t+o(\|\varepsilon\|),\\
\ddot{y}&=y-2y(x^2+y^2)-\varepsilon_1 y-\varepsilon_2\dot{y}\\
&\quad+\frac{4\varepsilon_3}{\omega^2+1}xy\cos\omega t+\frac{2\varepsilon_4}{\omega^2+1}(x^2+3y^2)\cos\omega t+o(\|\varepsilon\|).
\end{aligned}\tag{7.1}$$

The unperturbed equations are

$$(7.2)\qquad \ddot{x}=x-2x(x^2+y^2),\qquad \ddot{y}=y-2y(x^2+y^2).$$

From the previous section we see that the eigenvalues for the equilibrium at $(x,\dot{x},y,\dot{y})=0$ consist of double eigenvalues at $\lambda=\pm1$. Thus $d_s=d_u=2$. Because of the symmetry we expect $d_b=2$.

We now transform (7.2) to polar coordinates by introducing $x=r\cos\theta$, $y=\mathrm{r}\sin\theta$ to get

$$(7.3)\qquad 2\dot{r}\dot{\theta}+r\ddot{\theta}=0,$$

$$(7.4)\qquad \ddot{r}-r\dot{\theta}^2=r-2r^3.$$

Equation (7.3) can be integrated to yield

$$r^2\dot{\theta}=\text{constant},$$

which expresses conservation of angular momentum.

In polar coordinates the saddle equilibrium at the origin is located at $r=\dot{r}=0$. Thus, any orbit on W^s or W^u must satisfy $r\to0$ which requires, by conservation of angular momentum, $\dot{\theta}\equiv0$. Hence, W^u and $W$are each a subset of $S_1=\{(r,\dot{r},\theta,\dot{\theta})\,|\,\dot{\theta}=0\}$.

Furthermore, suppose $t\to(r(t),\dot{r}(t),\theta(t),\dot{\theta}(t))$ is any orbit with $\dot{\theta}(0)=0$, $r(0)\neq0$. Then $r(t)^2\dot{\theta}(t)=r(0)^2\dot{\theta}(0)=0$ so that $\dot{\theta}(t)\equiv0$. Hence, S_1 is invariant.

It is easily verified that a Hamiltonian for (7.2) is given by

$$H(x,\dot{x},y,\dot{y})=\frac{1}{2}(\dot{x}^2+\dot{y}^2)-\frac{1}{2}(x^2+y^2)+\frac{1}{2}(x^2+y^2)^2.$$

In polar coordinates this becomes

$$\overline{H}(r,\dot{r},\theta,\dot{\theta})=\frac{1}{2}(\dot{r}^2+r^2\dot{\theta}^2)-\frac{1}{2}r^2+\frac{1}{2}r^4.$$

Notice that $\overline{H}(0,0,\theta,\dot{\theta})=0$. This means that W^s and W^u are subsets of the level surface $S_2=\{(r,\dot{r},\theta,\dot{\theta})\,|\,\overline{H}(r,\dot{r},\theta,\dot{\theta})=0\}$.

Now let $t\to(r(t),\dot{r}(t),\theta(t),\dot{\theta}(t))$ be any orbit lying on $S_1\cap S_2$. Then $\dot{\theta}(t)=0$ and

$$(7.5)\qquad \frac{1}{2}\dot{r}^2-\frac{1}{2}r^2+\frac{1}{2}r^4=0.$$

Note that this requires $r\leqq1$.

Integrating (7.5) we get

$$r(t)=\operatorname{sgn}(r(0))\operatorname{sech}\left[t+\operatorname{sech}^{-1}(|r(0)|)\right].$$

Since $\lim_{t\to+\infty}r(t)=\lim_{t\to-\infty}r(t)=0$,

$$W^s=W^u=S_1\cap S_2=\left\{(r,\dot{r},\theta,\dot{\theta})|\dot{r}^2=r^2(1-r^2),0\leqq\theta\leqq2\pi,\dot{\theta}=0\right\}.$$

This verifies that $d_b=2$.

We now utilize the formulation of §5. To establish B-coordinates on $W^B=W^s\cap W^u$ we choose $0<\delta<1$ and define $U=\{(r,\dot{r},\theta,\dot{\theta})\in W^s\,||r|>\delta\}$. We can then establish a coordinate chart (U,ϕ), where ϕ: $U\to R^2$ is given by

$$\phi_1(r,\dot{r},\theta,\dot{\theta})=\cos\theta\operatorname{sgn}(r)\exp\left[-\operatorname{sgn}(\dot{r})\operatorname{sech}^{-1}|r|\right],$$

$$\phi_2(r,\dot{r},\theta,\dot{\theta})=\sin\theta\operatorname{sgn}(r)\exp\left[-\operatorname{sgn}(\dot{r})\operatorname{sech}^{-1}|r|\right].$$

Since $\Sigma^1=\phi^{-1}(S^1)=\{(r,\dot{r},\theta,\dot{\theta})|r=1,\dot{r}=0,0\leqq\theta\leqq 2\pi,\dot{\theta}=0\}$ and since θ is constant on homoclinic orbits, $s(r,\dot{r},\theta,\dot{\theta})=s_1(r,\dot{r},\theta,\dot{\theta})=\theta$. Let $P\in\Sigma^1$ be arbitrary and denote the B-coordinates of P by θ.

We now introduce rotated coordinates

$$\bar{x}=x\cos\theta+y\sin\theta,\qquad \bar{y}=-x\sin\theta+y\cos\theta.$$

We substitute these relations into (7.1) and then define $u_1=\bar{x}$, $u_2=\dot{\bar{x}}$, $u_3=\bar{y}$, $u_4=\dot{\bar{y}}$. The result is

$$\begin{aligned}
&\dot{u}_1=u_2,\\
&\dot{u}_2=u_1-2u_1\left(u_1^2+u_3^2\right)\\
&\quad -\varepsilon_1\left[u_1(3\cos^2\theta+\sin^2\theta)-2u_3\sin\theta\cos\theta\right]-\varepsilon_2u_2\\
&\quad +\frac{2\varepsilon_3}{\omega^2+1}\left[3u_1^2\cos\theta-2u_1u_3\sin\theta+u_3^2\cos\theta\right]\cos\omega t\\
&\quad +\frac{2\varepsilon_4}{\omega^2+1}\left[3u_1^2\sin\theta+2u_1u_3\cos\theta+u_3^2\sin\theta\right]\cos\omega t+o\left(\|\varepsilon\|\right),\\
&\dot{u}_3=u_4,\\
&\dot{u}_4=u_3-2u_3\left(u_1^2+u_3^2\right)\\
&\quad -\varepsilon_1\left[-2u_1\sin\theta\cos\theta+u_3(3\sin^2\theta+\cos^2\theta)\right]-\varepsilon_2u_4\\
&\quad +\frac{2\varepsilon_3}{\omega^2+1}\left[-u_1^2\sin\theta+2u_1u_3\cos\theta-3u_3^2\sin\theta\right]\cos\omega t\\
&\quad +\frac{2\varepsilon_4}{\omega^2+1}\left[u_1^2\cos\theta+2u_1u_3\sin\theta+3u_3^2\cos\theta\right]\cos\omega t+o\left(\|\varepsilon\|\right).
\end{aligned}\tag{7.6}$$

The unperturbed equations are

$$\begin{aligned}
&\dot{u}_1=f_1(u)=u_2, &&\dot{u}_3=f_3(u)=u_4,\\
&\dot{u}_2=f_2(u)=u_1-2u_1\left(u_1^2+u_3^2\right), &&\dot{u}_4=f_4(u)=u_3-2u_3\left(u_1^2+u_3^2\right),
\end{aligned}$$

and we have

$$Df(u)=\begin{bmatrix} 0 & 1 & 0 & 0\\ 1-6u_1^2-2u_3^2 & 0 & -4u_1u_3 & 0\\ 0 & 0 & 0 & 1\\ -4u_3u_1 & 0 & 1-2u_1^2-6u_3^2 & 0\end{bmatrix}.$$

Notice that the unperturbed equations are invariant under this coordinate change but that in the u coordinate system P has coordinates $(1,0,0,0)$ so that the B-coordinate of P is 0.

A solution $t\to\gamma(t,s)$ to the unperturbed system with $\gamma(0,s)=(\cos s,0,\sin s,0)$ is given by $\gamma(t,s)=(r(t)\cos s,\dot{r}(t)\cos s,r(t)\sin s,\dot{r}(t)\sin s)$, where $r(t)=\operatorname{sech} t$.

The variational equation $\dot{\psi}(t,0)=Df(\gamma(t,0))\psi(t,0)$ takes the form

$$\begin{aligned}
&\dot{\psi}_1=\psi_2, &&\dot{\psi}_3=\psi_4,\\
&\dot{\psi}_2=(1-6r^2)\psi_1, &&\dot{\psi}_4=(1-2r^2)\psi_3.
\end{aligned}\tag{7.7}$$

These uncouple into a pair of second order equations:

$$\ddot{\psi}_1=(1-6r^2)\psi_1, \tag{7.8}$$

$$\ddot{\psi}_3=(1-2r^2)\psi_3. \tag{7.9}$$

One solution to (7.7) is

$$\psi^{(4)}(t,0)=\dot{\gamma}(t,0)=(\dot{r}(t),\ddot{r}(t),0,0).$$

This proves a solution to (7.8)

$$\psi_1(t,0)=\dot{r}(t).$$

As in §6 we set $\lambda_4=-1$ and $\psi^{(4)}(t,0)=(\dot{r}(t),\ddot{r}(t),0,0)$ and we get a second solution to (7.8): $t\to Q(t)\dot{r}(t)$ where $Q(t)=-\frac{3}{2}t-\frac{1}{2}\sinh t\cosh t+\coth t$. Since $Q(t)\dot{r}(t)$ is asymptotic to e^t at $+\infty$ and e^{-t} at $-\infty$ we set $\lambda_2=1$ and

$$\psi^{(2)}(t,0)=\left(Q(t)\dot{r}(t),(Q\dot{r})^{\cdot}(t),0,0\right).$$

We get another solution to (7.7) from

$$\psi(t)=\frac{\partial\gamma}{\partial s}(t,0)=(0,0,r(t),\dot{r}(t)).$$

This gives the solution $\psi_3(t)=r(t)$ to (7.9). We get another solution to (7.9) of the form $P(t)r(t)$ by the use of the method of variation of parameters. We find $P(t)=\frac{1}{2}t+\frac{1}{2}\sinh t\cosh t$, where the arbitrary constant has been chosen so that

$$\begin{vmatrix} r(t) & (Pr)(t) \\ \dot{r}(t) & (Pr)^{\cdot}(t)\end{vmatrix}=1.$$

Since $r(t)$ and $\dot{r}(t)$ are asymptotic to e^{-t} at $+\infty$ and e^t at $-\infty$, set $\lambda_3=1$ and $\psi^{(3)}(t,0)=(0,0,r(t),\dot{r}(t))$. Similarly, $\lambda_1=-1$ and $\psi^{(1)}(t,0)=(0,0,(Pr)(t),(Pr)^{\cdot}(t))$.

We now have a fundamental solution as described in Theorem 5.1 for the variational equation. This solution also satisfies

$$\det\left(\psi^{(1)}(0),\psi^{(2)}(0),\psi^{(3)}(0),\psi^{(4)}(0)\right)=1.$$

At this point the Δ_{ij}'s can be computed (for details see [6]). When this has been done the equations $\Sigma_{j=1}^{4}\Delta_{ij}(\xi,\theta)=0$ become

$$\begin{aligned}&(4\sin\theta\cos\theta)\varepsilon_1-\left(\pi\sin\theta\cos\omega\xi\operatorname{sech}\frac{\pi\omega}{2}\right)\varepsilon_3+\left(\pi\cos\theta\cos\omega\xi\operatorname{sech}\frac{\pi\omega}{2}\right)\varepsilon_4=0,\\ &\frac{2}{3}\varepsilon_2-\left(\pi\omega\cos\theta\sin\omega\xi\operatorname{sech}\frac{\pi\omega}{2}\right)\varepsilon_3-\left(\pi\omega\sin\theta\sin\omega\xi\operatorname{sech}\frac{\pi\omega}{2}\right)\varepsilon_4=0.\end{aligned}\tag{7.10}$$

As a special case, consider $\varepsilon_1=\varepsilon_4=0$, which represents radial symmetry and forcing in the x- direction only. By taking $\theta=0$, (7.10) becomes

$$\frac{2}{3}\varepsilon_2-\left(\pi\omega\sin\omega\xi\operatorname{sech}\frac{\pi\omega}{2}\right)\varepsilon_3=0.$$

Given ε_2, ε_3 and ω we can satisfy this equation by solving for ξ:

$$\sin\omega\xi=\frac{2\varepsilon_2}{3\varepsilon_3\pi\omega}\cosh\frac{\pi\omega}{2}.$$

Thus, for a fixed ω, we have the existence of a homoclinic orbit for $\varepsilon \in S$ where

$$S=\left\{\varepsilon \,\middle|\, \varepsilon_1=\varepsilon_4=0, \varepsilon_2 \neq 0, \varepsilon_3 \neq 0, \frac{2\varepsilon_2 \cosh(\pi\omega/2)}{3\varepsilon_3 \pi\omega}<1\right\}.$$

This result was obtained in [10].

To check for transversality, we compute the matrix C as used in Theorem 5.3.

$$\begin{aligned} C(\bar{\xi},\theta,\varepsilon)=&\begin{bmatrix} 0 & 4\cos 2\theta \\ 0 & 0 \end{bmatrix}\varepsilon_1 \\ &+\begin{bmatrix} 0 & -\cos\theta\cos\omega\xi \\ -\omega^2\cos\theta\cos\omega\xi & \omega\sin\theta\sin\omega\xi \end{bmatrix}\left(\pi\operatorname{sech}\frac{\pi\omega}{2}\right)\varepsilon \\ &+\begin{bmatrix} -\omega\cos\theta\sin\omega\xi & \sin\theta\cos\omega\xi \\ -\omega^2\sin\theta\cos\omega\xi & -\omega\cos\theta\sin\omega\xi \end{bmatrix}\left(\pi\operatorname{sech}\frac{\pi\omega}{2}\right)\varepsilon_4. \end{aligned}$$

We see that when $\varepsilon_1=\varepsilon_4=0$ and $\theta=0$

$$\det\left(C(\bar{\xi},\theta,\varepsilon)\right)=-\pi\omega^2\cos^2\omega\xi\operatorname{sech}\frac{\pi\omega}{2}\varepsilon_3.$$

Hence, for each $\varepsilon \in S$ the perturbed system has a transverse homoclinic orbit by Theorem 5.4.

By the openness of the transverse intersection condition we now have the existence of an open set $U \subset R^4$ with $S \subset U$ such that the perturbed system has a transverse homoclinic orbit for each $\varepsilon \in U$.

Acknowledgments. I wish to thank Sheldon Newhouse who supervised this work in its original form as a Ph. D. thesis. Thanks also to Richard McGehee and the other editors and referees at SIAM for their help in getting this work into print.

REFERENCES

[1] S.-N. Chow, J. K. Hale and J. Mallet-Paret, *An example of bifurcation to homoclinic orbits*, J. Differential Equations, 37 (1980), pp. 351–373.

[2] Earl A. Coddington and Norman Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[3] Izrail Salomonavich Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series, and Products*, trans. Scripta Technika, Inc., 4th ed., Academic Press, New York, 1965.

[4] B. D. Greenspan, *Bifurcations in periodically forced oscillations: subharmonics and homoclinic orbits*, Ph. D. thesis, Center for Applied Mathematics, Cornell Univ., Ithaca, NY, 1981.

[5] B. Greenspan and P. Holmes, *Homoclinic orbits, subharmonics and global bifurcations in forced oscillations*, Nonlinear Dynamics and Turbulence, C. Barenblatt, G. Looss and D. D. Joseph, eds., Pitman, London, 1981.

[6] J. R. Gruendler, *A generalization of the method of Melnikov to arbitrary dimension*, Ph. D. thesis, Univ. North Carolina, Chapel Hill, 1982.

[7] Milton M. Abramowitz and Irene Stegun, eds., *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, DC, 1964.

[8] Morris W. Hirsch and Stephen Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.

[9] P. J. Holmes, *Averaging and chaotic motions in forced oscillations*, SIAM J. Appl. Math, 38 (1980), pp. 65–80.

[10] ______, *A nonlinear oscillator with a strange attractor*, Phil. Trans. Roy. Soc. A, 292 (1979), pp. 419–448.

[11] P. J. Holmes and J. E. Marsden, *Melnikov's method and Arnold diffusion for perturbations of integrable Hamiltonian systems*, to appear.

[12] ______, *A partial differential equation with infinitely many periodic orbits: chaotic oscillations of a forced beam*, Arch. Rat. Mech. Anal., in press, 1981.

[13] V. K. MELNIKOV, *On the stability of the center for time-periodic perturbations*, Trans. Moscow Math. Soc., 12 (1963), pp. 1–56.

[14] F. C. MOON, *Experimental models for strange attractor vibrations in elastic systems*, in [17].

[15] ______, *Experiments on chaotic motions of a forced attractor*, J. Appl. Mech. Trans. ASME, 47 (1980), pp. 638–644.

[16] F. C. MOON AND P. J. HOLMES, *A magnetoelastic strange attractor*, J. Sound and Vibration, 65 (1979), pp. 275–296.

[17] PHILIP J. HOLMES, ed., *New Approaches to Nonlinear Problems in Dynamics*, Society for Industrial and Applied Mathematics, Philadelphia, 1980.

[18] J. A. SANDERS, *A note on the validity of Melnikov's method*, Rep. # 139, Wishundig Seminarium, Urije Universiteit, Amsterdam, 1980.

[19] C. GEORGHIOU, *Solution to problem* 83–10, SIAM Rev., 26 (1984), pp. 282–283.

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

TRANSCENDENTAL ESTIMATES FOR THE ADIABATIC VARIATION OF LINEAR HAMILTONIAN SYSTEMS*

GILBERT STENGLE†

Abstract. We consider a slowly varying linear Hamiltonian system of dimension $2n$, $\varepsilon\dot{u}=A(t)u$, where $A(t)$ is a real valued matrix with distinct pure imaginary eigenvalues for each $t\in[-\infty,\infty]$ and $(d/dt)^N A\in L_1(-\infty,\infty)$ for each $N>0$. Leung and Meyer (J. Differential Equations, 17 (1973), pp. 32–43) have found n independent adiabatic invariants in involution for this system and have shown that the increment in each from $t=-\infty$ to $t=+\infty$ is $O(\varepsilon^N)$ for each $N>0$ as $\varepsilon\to 0^+$. We obtain an upper estimate for the rate at which these asymptotically negligible increments tend to 0 in terms of the asymptotic growth of the L_1 norms of the elements of $(d/dt)^N A$ as $N\to\infty$.

1. Introduction. The central problem of this paper is the estimation of certain asymptotically negligible quantities arising from the slow modulation of dynamical systems. These are functions depending on a small parameter which occur in a context where methods of asymptotic expansion

$$f(\varepsilon)\sim f_0+f_1\varepsilon+f_2\varepsilon^2+\cdots$$

normally yield powerful results. However, these methods can exceptionally fail because all coefficients f_k vanish, even though $f\neq 0$. In this circumstance the asymptotic power series yields only a qualitative result: f tends to 0 rapidly as $\varepsilon\to 0$ but cannot be described quantitatively by the series. Problems of this kind are not uncommon in applications and present a notable challenge to the science of asymptotics. The survey article [M_1] of R. E. Meyer and its bibliography contain many references to asymptotically negligible quantities arising in physical science. We consider one such problem, cases of which have been the object of many previous investigations, and which, although quite special, requires methods which we believe can be applied to many other problems of this kind.

Specifically, we consider the linear Hamiltonian system $u'=A(\varepsilon\tau)u$ with slowly varying matrix $A(\varepsilon\tau)$. On bounded time intervals or even intervals of length $o(1/\varepsilon)$, this problem can be regarded as a small perturbation of a system with constant coefficients. Even for longer durations the limiting behavior of this problem as $\varepsilon\to 0^+$ is simpler in many ways than the behavior of the full problem $u'=A(\tau)u$, although the simplest perturbation arguments no longer apply. In 1963, J. E. Littlewood [L] considered the problem $u''+a^2(\varepsilon\tau)u=0$ and the associated functional $\mathscr{I}=u^2a+(u')^2/a$. A simple calculation shows $\mathscr{I}'=O(\varepsilon)$, a property which is often expressed by calling $\mathscr{I}$ an approximate or *adiabatic* invariant. However, Littlewood showed much more. He proved that $\mathscr{I}(\infty)-\mathscr{I}(-\infty)$ is asymptotically negligible as a function of ε under the hypotheses that $a>0$ for $-\infty\leqq t\leqq+\infty$ and that a is *gentle* in the sense that $a^{(N)}\in L_1(-\infty,\infty)$ for all $N>0$. In 1966 Knorr and Pfirsch [KP] showed that $\mathscr{I}(+\infty)-\mathscr{I}(-\infty)=O(\exp\{-c/\varepsilon\})$ for analytic a. Simultaneously and independently in 1973, R. E. Meyer [M_2] and Wasow [W] obtained actual asymptotic formulas for this increment under more specific hypotheses on a.

*Received by the editors July 7, 1983, and in revised form July 2, 1984.

† Department of Mathematics, Lehigh University, Bethlehem, Pennsylvania 18015.

In contrast to these two highly concrete investigations, the present paper is concerned with obtaining asymptotically negligible *upper bounds* for such increments under weaker, more generic hypotheses.

In 1977, Stengle [S], using methods which are further developed in this paper, obtained estimates precisely under Littlewood's hypotheses. In 1973 Leung and K. Meyer [LM] discovered the natural generalization of Littlewood's result to linear systems. For a Hamiltonian system of order $2n$, they found n adiabatic invariants in involution, each with an increment $\Delta(\varepsilon)$ from $-\infty$ to ∞ which is $O(\varepsilon^N)$ for each $N>0$. In this paper we obtain an upper estimate for the rate at which these asymptotically negligible increments tend to 0 as $\varepsilon\to 0^+$. Our main result, Theorem 6.1 below, gives an estimate of the form

$$\Delta(\varepsilon)=O\left(\exp-\psi\left(\log\frac{1}{\varepsilon}-\alpha\right)\right),$$

where ψ is a function determined by the asymptotic growth of the L_1 norms of the elements of d^NA/dt^N as $N\to\infty$ and α is a constant depending on the spectral properties of A.

In §2 we describe the adiabatic invariants of Leung and K. Meyer. In §3 we define an associated isospectral flow which, for reasons which appear in §6, is the basic object of our investigation. In §4 we give a quantified notion of gentleness by identifying certain algebras of gentle functions. These permit efficient description of a chain of estimates leading from the matrix A to the adiabatic invariants. Section 5 establishes smoothness properties of the associated isospectral flow from which, in §6, we derive estimates for the adiabatic invariants.

2. Basic hypotheses. The adiabatic invariants of Leung and Meyer. Assume that the real $2n$-dimensional system

$$\varepsilon\dot{u}=A(t)u \tag{2.1}$$

satisfies the following conditions.

1. The system is Hamiltonian, that is, $JA^T(t)=-A(t)J$, where $J=\left(\begin{smallmatrix}0&I\\-I&0\end{smallmatrix}\right)$.

2. The entries of A are gentle functions. (This implies that limiting values $A(\pm\infty)$ exist.)

3. For each t, $-\infty\leqq t\leqq\infty$, the eigenvalues of $A(t)$ are distinct and purely imaginary.

We say that a matrix S is *symplectic with multiplier i* or, more briefly, *symplectic* if $SJS^T=iJ$. The adiabatic invariants of Leung and Meyer [LM] can be described in the following way.

THEOREM 2.1 (Leung and Meyer). *There is a diagonalizing similarity $A\to SAS^{-1}$ such that S is symplectic-valued with gentle entries. For any constant diagonal matrix D, the function $\mathscr{I}=u^TS^TJDSu$ is an adiabatic invariant of system* (2.1) *satisfying $\mathscr{I}(+\infty)-\mathscr{I}(-\infty)=O(\varepsilon^N)$ for each $N>0$.*

We remark that the quadratic form $\mathscr{I}$ depends only on the symmetric part of the matrix JD. This contains exactly n independent parameters and gives us an n-parameter family of adiabatic invariants. We find that considering these forms collectively leads to considerable algebraic simplification of our problem.

3. A related isospectral flow. Suppose that the quadratic functional u^TCu is a true invariant of system (2.1). Then it is both necessary and sufficient that the matrix C satisfy the equation

$$\varepsilon\dot{C}+CA+A^TC=0.$$

Since $JA^T = -AJ$, this is equivalent to

$$\varepsilon(J\dot{C}) + [JC, A] = 0, \tag{3.1}$$

where $[X, Y] = XY - YX$. An equation of this form defines an isospectral matrix flow; that is, the eigenvalues of the matrix $JC(t)$ are independent of t. In our analysis, this flow is the underlying primary phenomenon from which properties of the adiabatic invariants will be derived without much difficulty. The central analytic problem of this paper is to show that if A satisfies our previous hypotheses, then (3.1) possesses a solution C which is *uniformly differentiable* in the sense that $\|C^{(N)}\|_\infty$ has a bound independent of ε for each N. It is clear that most solutions of (3.1) cannot satisfy such estimates and that, because of the parameter ε multiplying $\dot{C}$, such estimates cannot be directly derived recursively from the differential equation. We require a quantified version of such a result. This, in turn, requires a quantified description of gentle data.

4. Gentle and uniformly differentiable data. If f is a gentle function, we can give a measure of its gentleness in the following way. Let ϕ: $[0, \infty) \to (-\infty, \infty)$ be a convex monotonic increasing function such that

$$\|f^{(N+1)}\|_1 \leqq N! \exp\{\phi(N) + \alpha N + \beta\} \tag{4.1}$$

for some constants α and β (depending on f). It is evident that each gentle function satisfies such an estimate, even with $\alpha = \beta = 0$. For convenience we define $\phi(x) = +\infty$ if $x < 0$. Roughly speaking, a larger ϕ corresponds to a less gentle f, while the constants α, β represent finer shades of gentleness for which we account only qualitatively. (The occurrence of $N+1$ in (4.1) can be explained by noting that under our hypotheses, $\|f^{(N+1)}\|_1$ is actually the variation of $f^{(N)}$.) We will say that a function or vector or matrix function with components satisfying estimates (4.1) is ϕ-*gentle*. We also require a notion of ϕ-*differentiability* for functions satisfying similar uniform estimates

$$\|f^{(N)}\|_\infty \leqq N! \exp\{\phi(N) + \alpha N + \beta\}. \tag{4.2}$$

Definition. Let G_ϕ and M_ϕ be the spaces of smooth real-valued functions satisfying estimates (4.1) and (4.2) respectively.

We will also write $u \in G_\phi$ or M_ϕ to indicate that a vector or matrix function has components in G_ϕ or M_ϕ.

The following proposition gives some properties of these classes.

Proposition 4.1.

i) $G_\phi \subset M_\phi$.

ii) *G_ϕ and M_ϕ are algebras.*

iii) *If $f \in G_\phi(M_\phi)$ and $f \geqq a > 0$, then $1/f$ and $\sqrt{f} \in G_\phi(M_\phi)$.*

iv) *If $\sigma \in G_\phi$ and $\dot{\sigma} \geqq a > 0$, then $f \circ \sigma^{-1} \in G_\phi(M_\phi)$ if and only if $f \in G_\phi(M_\phi)$.*

v) *If $f \in M_\phi$, $g \in G_\phi$, then $\int_{-\infty}^{t} f\dot{g}\, ds \in G_\phi$.*

Proof. i) This is obvious.

ii) It is obvious that G_ϕ and M_ϕ are real vector spaces. Suppose $f, g \in G_\phi$. Let $u_k = (1/k!)u^{(k)}$. Then $(\dot{fg})_k = (\dot{f}g + f\dot{g})_k = \Sigma_{j=0}^{k}\{\dot{f}_j g_{k-j} + f_{k-j}\dot{g}_j\}$.

$$\|(\dot{fg})_k\|_1 \leqq \Sigma\{\|\dot{f}_j\|_1 \|g_{k-j}\|_\infty + \|\dot{g}_j\|_1 \|f_{k-j}\|_\alpha\}$$

$$\leqq 2 \sum_{j=0}^{k} \exp\{\phi(j) + \phi(k-j) + (\alpha_f + \alpha_g)k + \beta_f + \beta_g\}.$$

Since ϕ is convex, $\phi(j)+\phi(k-j)\leqq\phi(k)+\phi(0)$ and

$$\|(\dot{\overline{fg}})_k\|_1\leqq 2(k+1)\exp\{\phi(k)+\phi(0)+\alpha_k+\beta\}\leqq\exp\{\phi(k)+\alpha' k+\beta'\}.$$

The proof for M_ϕ is similar.

iii) Suppose $f\in M_\phi$, $f\geqq a>0$. Then $\|f_k\|_\infty\leqq\exp\{\phi(k)+\alpha k+\beta\}$. Let $g=1/f$. Then $(fg)_k=0$ gives

$$fg_k=-\sum_{j=0}^{k-1}f_{k-j}g_j.$$

Let $\gamma_k=\|g_k\|_\infty\exp\{-\phi(k)-(\omega+1)k\}$. Then

$$\gamma_k\leqq\frac{1}{a}\sum_{j=0}^{k-1}\gamma_j\exp\{-\phi(k)-(\alpha+1)k+\phi(k-j)+\alpha(k-j)+\beta+\phi(j)+(\alpha+1)j\}$$

$$\leqq\frac{1}{a}\sum_{j=0}^{k-1}\gamma_j\exp\{-(k-j)+\phi(0)+\beta\}\leqq\max_{j<k}\gamma_j\exp\alpha'.$$

This implies $\gamma_k\leqq\exp\{\alpha' k+\beta\}$, which is equivalent to $g\in M_\phi$.

Now suppose $f\in G_\phi$. Since $G_\phi\subset M_\phi$, we know $g\in M_\phi$. Again by the chain rule, $(fg)_{k+1}=0$ gives

$$fg_{k+1}=-\sum_{j=0}^{k}f_{k+1-j}g_j$$

or

$$\frac{1}{k+1}f\dot{g}_k=-\sum_{j=0}^{k}\frac{1}{k+1-j}\dot{f}_{k-j}g_j.$$

Using L_1-norms for $\dot{f}_{k-j}$ and L_∞-norms for the g_j gives

$$\|\dot{g}_k\|_1\leqq(k+1)^2\exp\{\phi(0)+\phi(k)+k\alpha+\beta\}\leqq\exp\{\phi(k)+k\alpha'+\beta'\}.$$

Hence $g\in G_\phi$.

The proof that $\sqrt{f}\in G_\phi$ is similar.

iv) We show that $f\in M_\phi$ implies $f\circ\sigma^{-1}\in M_\phi$. The estimates for the latter inclusion can be expressed in the form

$$\frac{1}{k!}\left|\left(g\frac{d}{dt}\right)^k f\right|\leqq\exp(\phi(k)+\alpha k+\beta),\tag{4.3}$$

where $g=1/\dot{\sigma}\in G_\phi\subset M_\phi$ by iii). Increasing ϕ by a linear function if necessary, we can suppose that $|f_k|$, $|g_k|\leqq\exp\phi(k)$. It will suffice to show that (4.3) is a consequence of these estimates.

Induction shows that $(gd/dt)^k f$ is a sum of terms of the form $g^{(r_1)}\cdots g^{(r_k)}f^{(j)}$, $\sum r_l=k-j$. Each such term has an absolute bound

$$\exp\{\sum\phi(r_l)+\phi(j)\}=\exp\left\{\sum\phi\left(\left[1-\frac{r_l}{k}\right]0+\frac{r_l}{k}k\right)+\phi\left(\left[1-\frac{j}{k}\right]0+\frac{j}{k}k\right)\right\}$$

$$\leqq\exp\{k\phi(0)+\phi(k)\}.$$

Moreover, the number of such terms is the integer obtained by choosing $f=g=e^t$ and setting $t=0$. This is precisely $k!$. Hence

$$(f\circ\sigma^{-1})^{(k)} \leqq k!\exp\{\phi(k)+k\phi(0)\};$$

that is, $f\circ\sigma^{-1}\in M_\phi$.

Conversely, suppose $f\circ\sigma^{-1}\in G_\phi$. Then by (iii), $\dot\sigma\in G_\phi \Rightarrow 1/\dot\sigma\in G_\phi$. By the above argument, $(1/\dot\sigma)\circ\sigma^{-1}=(\sigma^{-1})^{\cdot}\in G_\phi$. Hence $f\circ\sigma^{-1}\circ(\sigma^{-1})^{-1}=f\in G_\phi$. The proof for M_ϕ is similar.

(v) The proof is similar to that of (ii).

We note that each element of G_0 is holomorphic in some strip containing the real axis (Cauchy's estimate for the Taylor coefficients of an analytic function). The following lemma shows that any gentle function has a certain kind of majorant in G_0.

LEMMA 4.2. *Given $f\in G_\phi$, there exists $h\in G_0$ such that $|\dot f|\leqq \dot h$.*

Proof. Let

$$\dot h(t)=\frac{2}{\pi}\int_{-\infty}^{\infty}\frac{1}{1+(t-s)^2}\left\{|\dot f(s)|+\int_{s-1}^{s+1}|\ddot f(r)|dr\right\}ds.$$

The integrability of $\dot f$ and $\ddot f$ implies that $h\in G_0$. Moreover,

$$\begin{aligned}|\dot f(t)|&=\frac{2}{\pi}\int_{t-1}^{t+1}\frac{1}{1+(t-s)^2}|\dot f(t)|ds\\&=\frac{2}{\pi}\int_{t-1}^{t+1}\frac{1}{1+(t-s)^2}|\dot f(s)|ds+\frac{2}{\pi}\int_{t-1}^{t+1}\frac{1}{1+(t-s)^2}\{|\dot f(t)|-|\dot f(s)|\}ds\\&\leqq\frac{2}{\pi}\int_{-\infty}^{\infty}\frac{1}{1+(t-s)^2}|\dot f(s)|ds+\frac{2}{\pi}\int_{t-1}^{t+1}\frac{1}{1+(t-s)^2}\int_{s-1}^{s+1}|\ddot f(r)|dr\\&\leqq\dot h(t).\end{aligned}$$

5. Existence of uniformly differentiable isospectral flows. The main technical difficulties of our analysis are contained in the following propositions.

PROPOSITION 5.1. *If the matrix A of system* (2.1) *is ϕ-gentle, then the diagonalizing matrix S of Theorem* 2.1 *is also ϕ-gentle.*

Proof. The proof of Leung and Meyer in [LM] that A has a gentle diagonalizing transformation is based on the closure of the class of gentle scalar-valued functions under rational operations with denominators bounded away from 0 and under the formation of square roots of strictly uniformly positive elements. Proposition 4.1 implies that our more restricted classes also enjoy these closure properties. It is then easy to check that under our hypotheses the explicit construction of the symplectic diagonalizing transformation S given by Leung and Meyer also takes place within this more restricted class of data.

PROPOSITION 5.2. *Let $A=i\Lambda+\varepsilon\dot B$, where $A,B\in G_\phi$ and Λ is real diagonal with diagonal elements distinct on $[-\infty,\infty]$. Then for each constant diagonal matrix D, the initial value problem $\varepsilon\dot X+[X,A]=0$, $X(-\infty)=D$ has a solution contained in $M_{\phi+N\log N}$.*

Proof. Our method will be to derive estimates from the infinite recursive set of differential equations obtained by repeated differentiation of the system. However, some care is required to select a system which leads to simple estimates from among the many equivalent systems obtainable in this way. To accomplish this we use two devices.

The first is to choose advantageously between considering $y^{(k)}$ as the kth derivative or as the first derivative of the $(k-1)$th derivative. The second is to apply powers of a differential operator $(1/g)d/dt$ rather than d/dt for a suitably chosen unit g of G_ϕ.

We determine g in the following way. It follows from Lemma 4.2 that we can choose $h\in G_0$ so that $\max_j|\dot{\lambda}_j|\leqq \dot{h}$. We can also suppose, adding a constant to h if necessary, that $h\geqq a>0$. It then follows that for M sufficiently large, $\max_{j\neq k}|(\dot{\lambda}_j-\dot{\lambda}_k)/(\lambda_j-\lambda_k)|\leqq M\dot{h}/h$. Let $g=h^M$. Then $\max_{j\neq k}|(\dot{\lambda}_j-\dot{\lambda}_k)/(\lambda_j-\lambda_k)|\leqq \dot{g}/g$.

We now assemble the off-diagonal elements of X into a vector y, the diagonal elements into a vector z. Then, using Proposition 4.1, it can be seen that the system can be expressed in the form

$$\varepsilon\dot{y}-igwy=gw(\dot{a}y+\dot{b}z),\qquad \dot{z}=g(\dot{c}y+\dot{d}z),$$

where w is real diagonal with nonzero elements bounded away from 0, and where a, b, c, d belong to G_ϕ. Let $(\ )_k$ denote the differential operator $(k!)^{-1}(g^{-1}(d/dt))^k$. Then applying $gw(\cdot)_k g^{-1}w^{-1}$ to the first equation and $k^{-1}(\cdot)_{k-1}g^{-1}$ to the second yields

$$\begin{aligned}\varepsilon(\dot{y})_k-(igw-\varepsilon kw^{-1}\dot{w})(y)_k\\ =\varepsilon gw\Bigg\{a(\dot{y})_k+\sum_{j=1}^{k-1}\big(\big[(a)_{k-j}-j(w^{-1})_{k-j-1}\big](y)_j\\ +(b)_{k-j}(z)_j\big)+(a)_k y+(b)_k z\Bigg\},\end{aligned}$$

$$(z)_k=k^{-1}\sum_{j=0}^{k-1}\big((c)_{k-1-j}(y)_j+(d)_{k-1-j}(z)_j\big).$$

A favorable property of this system for joint asymptotics in ε and k appears in the matrix $w^{-1}\dot{w}$. Each diagonal element of this has the form

$$g^{-1}(\lambda_i-\lambda_j)^{-1}\frac{d}{dt}\big(g(\lambda_i-\lambda_j)\big)=g^{-1}\dot{g}-(\lambda_i-\lambda_j)^{-1}(\dot{\lambda}_i-\dot{\lambda}_j),$$

which, because of our choice of g, is nonnegative. These equations are equivalent to a system of integral equations of the form

$$\begin{aligned}(y)_k=\int_{-\infty}^{t}\exp\left\{\frac{i}{\varepsilon}\int_s^t gw(r)\,dr-k\int_s^t w^{-1}\dot{w}\right\}\{\dot{a}(y)_k\}\\ \cdot\sum_{j=0}^{k-1}\{(a_{kj}+jb_{kj})(y)_j+c_{kj}(z)_j\}\,ds,\end{aligned}$$

$$z_k=\delta_{k0}D+\sum_{j=0}^{k-1}\{d_{kj}(y)_j+e_{kj}(z)_j\}.$$

Since $((1/g)d/dt)^k f(t)=(d/d\tau)^k f(\sigma^{-1}(\tau))$, where $\sigma(t)=\int_0^t g(s)\,ds$, by Proposition 4.1(iv) we find that elements of all matrix functions subscripted kj satisfy a common estimate

$$\|u_i k_j\|_1\leqq\exp\{\phi(k-j)+\alpha(k-j)+\beta\}.$$

We use the maximum absolute element norm $|u|$ for vectors and the subordinate maximum absolute row sum for matrices, and we introduce scalar bounds

$$\eta_k(t)=\sup_{s\leqq t}|(y)_k(s)|\exp\{-\phi(k)-\alpha k\},$$
$$\zeta_k(t)=\sup_{s\leqq t}|(z)_k(s)|\exp\{-\phi(k)-\alpha k\}.$$

Then the integral equations imply the integral inequalities

$$\eta_k(t)-\int_{-\infty}^{t}|a(s)|\eta_k(s)\,ds$$
$$\leqq C\sum_{j=0}^{k}\left[(1+j)\eta_j(t)+\zeta_j(t)\right]$$
$$\cdot\exp\{-\phi(k)-\alpha k+\phi(j)+\alpha(j)+\phi(k-j)+\alpha(k-j)+\beta\}$$

and a similar simpler inequality for $\zeta_k(t)$. Convexity of $\phi(1+u)$ implies $-\phi(k+1)+\phi(k+1-j)+\phi(j+1)\leqq\phi(1)$. This implies, for a possibly larger C, that

$$\eta_k(t)-\int_{-\infty}^{t}|a(s)|\eta_k(s)\,ds\leqq C\sum_{j=0}^{k-1}\left[(1+j)\eta_j(t)+\zeta_j(t)\right]$$

and similarly

$$\zeta_k(t)\leqq C\sum_{j=0}^{k-1}\left[\eta_j(t)+\zeta_j(t)\right].$$

Gronwall's inequality applied to the first inequality implies

$$\eta_k(t)\leqq C\sum_{j=0}^{k-1}\left((1+j)\eta_j(t)+\zeta_j(t)\right)\int_{-\infty}^{t}|a(s)|\exp\left\{\int_s^t|a(r)|ds\right\}ds.$$

Since $|a|\in L_1$, this implies, increasing C if necessary, that η_k, ζ_k are bounded by $\rho_k e^k$, where

$$\rho_k=C\sum_{j=0}^{k-1}(j+1)e^{j-k}\rho_j.$$

This easily implies $\rho_k\leqq\exp\{k\log k+\alpha_1 k+\beta_1\}$, which in turn implies

$$|(y)_k|,\ |(z)_k|\leqq\exp\{k\log k+\alpha_2 k+\beta_2\}.$$

Again let $\sigma(t)=\int_0^t g(s)\,ds$. Then these estimates imply $X\circ\sigma^{-1}\in M_{\phi+N\log N}$. Hence by Proposition 4.1(iv), $X\in M_{\phi+N\log N}$.

6. Estimates for adiabatic invariants. The following theorem, our main result, is a direct consequence of the preceding propositions.

THEOREM 6.1. *Suppose the matrix function A is Hamiltonian, ϕ-gentle, with purely imaginary eigenvalues for $t\in[-\infty,\infty]$. Let $\psi(x)=\sup_y(xy-\phi(y)-2y\log y)$. Then there is a ϕ-gentle diagonalizing transformation $A\to SAS^{-1}$ and a constant α such that for any* *constant diagonal matrix D, the function $\mathscr{I}=u^TS^TJDSu$ is an adiabatic invariant of the system $\varepsilon\dot{y}=A(t)y$ satisfying*

$$\mathscr{I}(+\infty)-\mathscr{I}(-\infty)=O\left(\exp\left\{-\psi\left(\log\frac{1}{\varepsilon}-\alpha\right)\right\}\right)$$

as $\varepsilon\to 0^+$.

Proof. By Proposition 5.1, $S\in G_\phi$. Proposition 4.1 implies that $\Lambda=i^{-1}SAS^{-1}\in G_\phi$ and $B=\int_\infty^t SS^{-1}ds$ belong to G_ϕ. By Proposition 5.2 there is an X in $M_{\phi+N\log N}$ satisfying $\varepsilon\dot{X}+[X,i\Lambda+\varepsilon\dot{B}]=0$, $X(-\infty)=D$. Since S is symplectic, $i\Lambda+\varepsilon\dot{B}$ is again Hamiltonian and the quadratic functional v^TJXv is an actual invariant of the system $\varepsilon\dot{v}=(i\Lambda+\varepsilon\dot{B})v$; that is, u^TS^TJXSu is an actual invariant of $\varepsilon\dot{u}=Au$. Thus $\mathscr{I}(-\infty)$ exists since it is the constant value of this invariant. For the same reason, $\mathscr{I}(-\infty)=\lim_{t\to+\infty}u^TS^TJXSu$. $\mathscr{I}(+\infty)$ also exists since our hypotheses amply ensure that all solutions v are bounded, and a simple calculation shows $(d/dt)v^TJDv=v^T[\dot{B},JD]v\in L_1$.

The differential equation for X implies that each off-diagonal element X_{kl} of X, $k\neq l$ satisfies

$$X_{kl}=-\frac{\varepsilon}{\lambda_k-\lambda_l}\{\dot{X}_{kl}+[\dot{B},X]_{kl}\}. \tag{6.1}$$

Hence, since X, $\dot{X}$ and $\dot{B}$ are bounded independently of ε, we have

$$X_{kl}=O(\varepsilon),\qquad -\infty\leqq t<\infty.$$

Thus for small ε, the diagonal elements X_{kk} are uniformly close to the eigenvalues of X, which are, since $X(t)$ is isospectral, just the eigenvalues of $X(-\infty)=D$. This shows that all increments of $\mathscr{I}$ are $o(1)$. But near $t=+\infty$ we can show more. N-fold self-substitution of the relations (6.1) gives

$$X_{kl}=\left\{\frac{\varepsilon}{\lambda_l(\infty)-\lambda_k(\infty)}\right\}^N X_{kl}^{(N)}+o_t(1)$$

as $t\to\infty$. This implies

$$\overline{\lim_{t\to\infty}}\;|X_{kl}|\leqq\varepsilon^N\exp\{\phi(N)+2N\log N+\alpha N+\beta\}.$$

Thus all off-diagonal elements of X satisfy a common estimate

$$\begin{aligned}\overline{\lim_{t\to\infty}}\;|X_{kl}|&\leqq\exp\left\{\phi(N)+2N\log N+\alpha N+\beta-N\log\frac{1}{\varepsilon}\right\}\\&\leqq\exp\left\{-\psi\left(\log\frac{1}{\varepsilon}-\alpha\right)+\beta\right\}.\end{aligned}$$

Now suppose for the moment that the diagonal elements d_k of D are distinct. Then the characteristic equation of X, whose roots are precisely the d_k, has the form $\pi_k(\lambda-X_{kk})=O\{\exp-\psi(\log(1/\varepsilon)-\alpha)\}$ as $t\to\infty$. The implicit function theorem, which applied directly when the d_k are distinct, then implies that

$$X_{kk}=d_k+O_\varepsilon\left\{\exp-\psi\left(\log\frac{1}{\varepsilon}-\alpha\right)\right\}+o_t(1).$$

Thus $X=D+O_\varepsilon\{\exp-\psi(\log(1/\varepsilon)-\alpha)\}+o_t(1)$ as $t\to+\infty$. Finally

$$\begin{aligned}\mathscr{I}(-\infty)&=u^TS^TJXSu(-\infty)=\lim_{t\to+\infty}u^TS^TJXSu(t)\\&=\lim_{t\to+\infty}u^TS^TJ\left(D+O_\varepsilon\left\{\exp\psi\left(\log\frac{1}{\varepsilon}-\alpha\right)\right\}+o_t(1)\right)Su\\&=\mathscr{I}(+\infty)+O\left\{\exp-\psi\left(\log\frac{1}{\varepsilon}-\alpha\right)\right\}.\end{aligned}$$

Since our conclusion depends linearly on D (although, curiously, our intermediate steps do not), and any D is the sum of two matrices with distinct diagonal entries, our conclusion follows in general.

Acknowledgment. The author is grateful to the Department of Mathematics of Wuhan University for support during the writing of this paper.

REFERENCES

[KP] G. KNORR AND D. PFIRSCH, *The variation of the adiabatic invariant of the harmonic oscillator*, Z. Naturforschg., 21 (1966), pp. 688–693.

[L] J. E. LITTLEWOOD, *Lorentz's pendulum problem*, Ann. Physics, 21 (1963), pp. 233–242.

[LM] A. LEUNG AND K. MEYER, *Adiabatic invariants for linear Hamiltonian systems*, J. Differential Equations, 17 (1973), pp. 32–43.

[M_1] R. E. MEYER, *Exponential asymptotics*, SIAM Rev., 22 (1980), pp. 213–224.

[M_2] R. E. MEYER, *Adiabatic variation*, *Part* I. *Exponential property for the simple oscillator*, Z. Angew. Math. Phys., 24 (1973), pp. 293–303.

[S] G. STENGLE, *Asymptotic estimates for the adiabatic invariance of a simple oscillator*, this Journal, 8 (1977), pp. 640–654.

[W] W. WASOW, *Adiabatic invariance of a simple oscillator*, this Journal, 4 (1973), pp. 78–88.

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

003

ANALYSIS OF A MODEL OF PERCOLATION IN A GENTLY SLOPING SAND-BANK*

CHARLES M. ELLIOTT[†] AND AVNER FRIEDMAN[‡]

Abstract. A free boundary problem associated with the percolation of sea-water in a "nearly" flat sand-bank is considered. The wet and dry zones of the beach are studied.

1. Introduction. Consider the problem of finding a function $u=u(x,y,t)$ which at any time $t\geqq 0$ is harmonic in the rectangle

$$\Omega=\{(x,y):0<x<1,\ -b<y<0\},$$

satisfies homogeneous Neumann conditions on

$$\Gamma_1=\{(x,y):x=0,\ -b<y<0\},$$
$$\Gamma_2=\{(x,y):0<x<1, y=-b\},$$
$$\Gamma_3=\{(x,y):x=1,\ -b<y<0\},$$

and the evolutionary unilateral conditions

$$u\leqq G,\quad \frac{\partial u}{\partial t}+\frac{\partial u}{\partial y}\leqq 0,\quad (u-G)\left(\frac{\partial u}{\partial t}+\frac{\partial u}{\partial y}\right)=0\quad \text{on } \Gamma, \tag{1.1}$$

where G is a given function and

$$\Gamma=\{(x,y):0<x<1, y=0\}.$$

Such a problem arises from a model of sea-water seepage in a periodic array of symmetrical sand-banks. The domain of half a typical sand-bank is

$$\{(x,y):0<x<1,\ -b<y<\bar{\varepsilon}G(x)\},$$

where $\bar{\varepsilon}\ll 1$. The model of [1] uses the fact that $\bar{\varepsilon}$ is a small parameter in order to obtain the approximations,

$$y=\bar{\varepsilon}u(x,0,t)$$

for the equation of the surface separating the dry portion of the bank from the wet portion and

$$p(x,y,t)=-y+\bar{\varepsilon}u(x,y,t)$$

for the pressure p of the sea-water in the sand-bank. Γ is an approximation to the upper surface of the sand-bank. Let $\Gamma\equiv\Gamma_d(t)\cup\Gamma_w(t)$, where

$$\Gamma_d(t)\equiv\{(x,0):u(x,0,t)<G(x)\}$$

*Received by the editors April 24, 1984. This work is partially supported by National Science Foundation under Grant MCS 7915171.

[†] Department of Mathematics, Imperial College, London SW7, England.
[‡] Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

approximates the dry portion of the upper surface of the bank and

$$\Gamma_w(t) \equiv \{(x,0): u(x,0,t) = G(x)\}$$

approximates the wet portion of the upper surface of the bank across which sea-water is seeping out of the bank. It is assumed in the model that at time $t=0$ the sand-bank has a wet region which is known, i.e., $u(x,0,0)$ is prescribed, and that as the sea-water seeps out of the bank it instantaneously drains away. We refer to [1] and [2] for details concerning the derivation and justification of the model and for applications.

It is the purpose of this paper to obtain results regarding the regularity of u and the shape of $\Gamma_d(t)$. In §2 we prove that $u \in H^1(0,T;\ H^1(\Omega))$ and also establish an asymptotic limit for u as $t \to \infty$. Under certain conditions on the data in §3 we obtain regularity results of the form $u \in L^2(0,T;\ H^2(\Omega))$ and $\partial u/\partial t \in L^\infty(\Omega \times (0,T))$. In §4 we prove, under suitable conditions on G and $u(x,0,0)$, that

i) $\Gamma_d(t)$ is monotone increasing in t.

ii) $\Gamma_d(t)$ consists of a single interval.

The method of proof is based on finite differencing in time and penalisation. Results of the type (i) and (ii) for the nonstationary dam problem were proved in [4] using a finite difference scheme.

We shall use in §2 the notation:

$$\langle v,w\rangle = \int_\Gamma vw\,dx, \qquad (v,w) = \int_\Omega vw\,dx\,dy,$$

$$|v|_\Gamma = \langle v,v\rangle^{1/2}, \quad |v|_0 = (v,v)^{1/2}, \quad |v|_1 = (\nabla v, \nabla v)^{1/2}, \quad \|v\| = \left(|v|_0^2 + |v|_1^2\right)^{1/2}.$$

The boundary values (or trace) of a function v on Γ are denoted by v_Γ for $v \in C(\bar{\Omega})$ (or $H^1(\Omega)$). The following trace inequalities hold: there exist constants C_1 and C_2 such that for all $v \in H^1(\Omega)$

$$(1.2) \qquad |v_\Gamma|_\Gamma \leqq C_1\|v\|, \qquad |v|_0 \leqq C_2\left(|v_\Gamma|_\Gamma^2 + |v|_1^2\right)^{1/2}.$$

We shall use Young's inequality: given $p>1$, a and b nonnegative numbers, then

$$(1.3) \qquad ab \leqq \frac{a^p}{p} + \frac{b^{p/(p-1)}(p-1)}{p}.$$

2. Existence, uniqueness, regularity and asymptotic limit. It is straightforward to verify that the function u introduced in §1, if sufficiently regular, solves the evolutionary unilateral inequality $(\mathscr{P})$:

$$(2.1) \qquad u \in L^2(0,T;\ H^1(\Omega)), \qquad \frac{\partial u_\Gamma}{\partial t} \in L^2(0,T;\ L^2(\Gamma)),$$

$$(2.2) \qquad u \in \mathscr{K} \equiv \left\{v \in L^2(0,T;\ H^1(\Omega)), v_\Gamma \leqq G\right\},$$

$$(2.3) \qquad u_\Gamma(0) = U_\Gamma,$$

$$(2.4) \qquad \int_0^T \left\{\left\langle \frac{\partial u_\Gamma}{\partial t}, v_\Gamma - u_\Gamma \right\rangle + (\nabla u, \nabla v - \nabla u)\right\} dt \geqq 0 \quad \forall v \in \mathscr{K},$$

where U_Γ is given. We shall therefore consider (2.1)–(2.4) as a weak formulation of the original problem for u.

We assume that

$$(2.5)\qquad \begin{aligned}&\text{(i)}\quad U_\Gamma \in H^{1/2}(\Gamma), \qquad G \in H^1(\Gamma),\\ &\text{(ii)}\quad U_\Gamma \leqq G.\end{aligned}$$

Then there exists an extension $U \in H^1(\Omega)$ of U_Γ. The natural choice for U in the harmonic function defined by

$$(2.6)\qquad (\nabla U, \nabla v) = 0 \quad \forall v \in H^1(\Omega), \quad v_\Gamma = 0.$$

By taking a finite difference in time, $(\mathscr{P})$ may be approximated by a sequence of elliptic variational inequalities $(\mathscr{P}^\delta)$ of Signorini type. Let N be a positive integer and $\delta = T/N$. Problem $(\mathscr{P}^\delta)$ is to find a sequence $\{u^n\}_{n=1}^N$ such that for all $n \in [1, N]$:

$$(2.7)\qquad u^n \in K \equiv \{v \in H^1(\Omega), v_\Gamma \leqq G\},$$

$$(2.8)\qquad u_\Gamma^0 = U_\Gamma, \quad u^0 = U,$$

$$(2.9)\qquad \frac{1}{\delta}\langle u_\Gamma^n - u_\Gamma^{n-1}, v_\Gamma - u_\Gamma^n\rangle + (\nabla u^n, \nabla v - \nabla u^n) \geqq 0 \quad \forall v \in K.$$

THEOREM 2.1. *There exists a unique solution to* $(\mathscr{P})$ *and the solution is the limit as* $\delta \to 0$ *of the unique solution to* $(\mathscr{P}^\delta)$.

The proof will exhibit the precise nature of how the solution of $(\mathscr{P})$ is approximated by the solutions of $(\mathscr{P}^\delta)$.

Proof. To prove uniqueness, suppose u^1 and u^2 are two solutions of $(\mathscr{P})$. Let $0 < \tau < T$. Taking

$$v_1 = \begin{cases} u^2 & \text{if } t < \tau, \\ u^1 & \text{if } t > \tau, \end{cases} \qquad v_2 = \begin{cases} u^1 & \text{if } t < \tau, \\ u^2 & \text{if } t > \tau \end{cases}$$

in the appropriate inequalities (2.4) and adding yields

$$\int_0^\tau \left\{ \frac{1}{2}\frac{d}{dt}\left|u_\Gamma^1 - {}_\Gamma^2\right|_\Gamma^2 + \left|u^1 - u^2\right|_1^2 \right\} dt \leqq 0.$$

Uniqueness follows immediately upon integrating and applying inequalities (1.2).

Existence and uniqueness of the sequence $\{u^n\}_{n=1}^N$ follows from the standard theory of variational inequalities [3], [5], since the bilinear form

$$B(w, v) \equiv \frac{1}{\delta}\langle w_\Gamma, v_\Gamma\rangle + (\nabla w, \nabla v)$$

is coercive on $H^1(\Omega)$ by virtue of (1.2) and since K is a closed convex set of $H^1(\Omega)$. In order to establish the convergence of the sequence $\{u^n\}$ as $\delta \to 0$ we need to derive some estimates. Taking $v = u^{n-1}$ in (2.9) yields

$$\delta\left|\frac{u_\Gamma^n - u_\Gamma^{n-1}}{\delta}\right|_\Gamma^2 + \frac{1}{2}|u^n|_1^2 + \frac{1}{2}|u^n - u^{n-1}|_1^2 \leqq \frac{1}{2}|u^{n-1}|_1^2.$$

Summing over n, the inequality

$$(2.10)\qquad 2\sum_{j=1}^n \delta\left|\frac{u_\Gamma^j - u_\Gamma^{j-1}}{\delta}\right|_\Gamma^2 + \sum_{j=1}^n |u^j - u^{j-1}|_1^2 + |u^n|_1^2 \leqq |U|_1^2$$

is obtained.

Similarly taking $v=G\in K$ in (2.9), we obtain

$$\frac{1}{2}|u^n_\Gamma-G|^2_\Gamma-\frac{1}{2}|u^{n-1}_\Gamma-G|^2_\Gamma+\delta|u^n|^2_1\leqq\delta(\nabla u^n,\nabla G)\leqq\frac{1}{2}\delta|u^n|^2_1+\frac{1}{2}\delta|G|^2_1$$

and summing over n, results in the inequality

$$|u^n_\Gamma-G|^2_\Gamma+\delta\sum_{j=1}^{n}|u^j|^2_1\leqq|U_\Gamma-G|^2_\Gamma+T|G|^2_1. \tag{2.11}$$

Let $u^\delta(t)$ and $\bar{u}^\delta(t)$ be defined by

$$\begin{aligned} &u^\delta(t)=u^n && \text{for } t\in(t^{n-1},t^n],\\ &\bar{u}^\delta(t)=u^n+\left(\frac{t^n-t}{\delta}\right)(u^{n-1}-u^n) && \text{for } t\in(t^{n-1},t^n], \end{aligned} \tag{2.12}$$

where $t^n=n\delta$. The estimates (2.10), (2.11) which hold for all $n\in[1,N]$ imply that $u^\delta(t)$ and $\bar{u}^\delta(t)$ are uniformly bounded, independently of δ, in $L^\infty(0,T;\ H^1(\Omega))$. Also $\bar{u}^\delta_\Gamma(t)$ is uniformly bounded, independently of δ, in $H^1(0,T;\ L^2(\Gamma))$. It follows by weak compactness in a Hilbert space that there exist subsequences (still denoted by δ) such that, as $\delta\to0$,

$$\begin{aligned} &u^\delta\rightharpoonup u && \text{in } L^2(0,T;\ H^1(\Omega)),\\ &\bar{u}^\delta\rightharpoonup\bar{u} && \text{in } L^2(0,T;\ H^1(\Omega)),\\ &\bar{u}^\delta_\Gamma\rightharpoonup\bar{u}_\Gamma && \text{in } H^1(0,T;\ L^2(\Gamma)), \end{aligned} \tag{2.13}$$

and by a compactness theorem [6; p. 58]

$$\bar{u}^\delta_\Gamma\to\bar{u}_\Gamma \quad \text{in } L^2(0,T;\ L^2(\Gamma)). \tag{2.14}$$

We wish to show that $u=\bar{u}$. For this it suffices to prove that for all smooth ϕ,

$$\int_0^T\{\langle u_\Gamma-\bar{u}_\Gamma,\phi\rangle+(\nabla u-\nabla\bar{u},\nabla\phi)\}\,dt=0. \tag{2.15}$$

Since

$$\begin{aligned} &\left|\int_0^T\{\langle u^\delta_\Gamma(t)-\bar{u}^\delta_\Gamma(t),\phi(t)\rangle+(\nabla u-\nabla\bar{u},\nabla\phi)\}\,dt\right|\\ &\quad=\left|\sum_{j=1}^{N}\int_{t^{j-1}}^{t^j}\frac{(t^j-t)}{\delta}\{\langle u^j_\Gamma-u^{j-1}_\Gamma,\phi(t)\rangle+(\nabla(u^j-u^{j-1}),\nabla\phi(t))\}\,dt\right|\\ &\quad\leqq C(\phi)\sum_{j=1}^{N}\left(\delta|u^j_\Gamma-u^{j-1}_\Gamma|_\Gamma+|u^j-u^{j-1}|_1\right)\\ &\quad\leqq C(\phi)\delta^{1/2}T^{1/2}\sum_{j=1}^{N}\left(|u^j_\Gamma-u^{j-1}_\Gamma|^2_\Gamma+|u^j-u^{j-1}|^2_1\right), \end{aligned}$$

the estimate (2.10) implies that as $\delta\to0$ we obtain (2.15).

Since $\mathscr{K}$ is closed in $L^2(0,T;\ H^1(\Omega))$ and $u^\delta\in\mathscr{K}$ we have, by (2.13), that $u\in\mathscr{K}$. Note next that $\bar{u}^\delta_\Gamma(0)=U_\Gamma$ implies that $u_\Gamma(0)=U_\Gamma$. Thus it remains to show that u

satisfies the variational inequality (2.4). It is sufficient to establish the inequality for all $v\in([0,T];\ H^1(\Omega))\cap\mathcal{K}$ which is dense in $\mathcal{K}$. For such a v the sequence $v^\delta(t)$ defined by

$$v^\delta(t)=v(t^n) \quad \text{for } t\in(t^{n-1},t^n] \tag{2.16}$$

converges strongly to v in $L^2(0,T;\ H^1(\Omega))$. Taking $v=v(t^n)$ in (2.9) implies the inequality

$$\int_0^T\left\{\left\langle \frac{\partial\bar u^\delta}{\partial t}(t),v_\Gamma^\delta(t)-u_\Gamma^\delta(t)\right\rangle+(\nabla u^\delta,\nabla v^\delta-\nabla u^\delta)\right\}dt\geqq 0.$$

Passing to the limit as $\delta\to 0$ and recalling (2.13), (2.14) and the lower semicontinuity of the semi-norm $\int_0^T|\cdot|_1$, (2.4) follows.

THEOREM 2.2. *If* $(\partial U/\partial y)_\Gamma\in L^2(\Gamma)$ *then* $\partial u/\partial t$ *belongs to* $L^2(0,T;\ H^1(\Omega))$; *consequently* $u\in C([0,T];\ H^1(\Omega))$ *with* $u(0)=U$.

Proof. The inequalities (2.9) imply that

$$\frac{1}{\delta}\langle u_\Gamma^n-u_\Gamma^{n-1},u_\Gamma^n-u_\Gamma^{n+1}\rangle+(\nabla u^n,\nabla u^n-\nabla u^{n+1})\leqq 0,$$

$$\frac{1}{\delta}\langle u_\Gamma^{n+1}-u_\Gamma^n,u_\Gamma^{n+1}-u_\Gamma^n\rangle+(\nabla u^{n+1},\nabla u^{n+1}-\nabla u^n)\leqq 0$$

and adding, yields

$$\langle z_\Gamma^n-z_\Gamma^{n-1},z_\Gamma^n\rangle+\delta(\nabla z^n,\nabla z^n)\leqq 0 \quad \text{where } z^k\equiv\frac{u^{k+1}-u^k}{\delta}.$$

From this we easily obtain

$$\frac{1}{2}|z_\Gamma^n|_\Gamma^2+\sum_{j=1}^n\delta|z^j|_1^2\leqq|z_\Gamma^0|_\Gamma^2. \tag{2.17}$$

Taking $n=1$ and $v=u^0$ in (2.9) gives

$$|z_\Gamma^0|_\Gamma^2+\delta|z^0|_1^2\leqq-(\nabla U,\nabla z^0)=-\langle Z_\Gamma,z_\Gamma^0\rangle\leqq\frac{1}{2}|Z_\Gamma|_\Gamma^2+\frac{1}{2}|z_\Gamma^0|_\Gamma^2,$$

where $Z_\Gamma=(\partial U/\partial y)_\Gamma$. Thus $\partial\bar u^\delta(t)/\partial t$ is uniformly bounded in $L^2(0,T;\ H^1(\Omega))$, and as $\delta\to 0$ we have that

$$\begin{aligned}\bar u^\delta(t)&\rightharpoonup u \quad \text{in } H^1(0,T;\ H^1(\Omega)),\\ \bar u^\delta(t)&\to u \quad \text{in } L^2(0,T;\ H^1(\Omega)).\end{aligned} \tag{2.18}$$

The assertion of the theorem now follows.

Since Theorems 2.1, 2.2 hold for all $T>0$, the solution exists for all $t>0$. Denote by w a solution of the Signorini problem

$$w\in\mathcal{K}\colon(\nabla w,\nabla v-\nabla w)\geqq 0, \qquad v\in\mathcal{K}. \tag{2.19}$$

Thus setting

$$G_m=\inf_x G(x), \qquad G_M=\sup_x G(x),$$

we have that w is any constant less than or equal to G_m.

THEOREM 2.3. *If* $(\partial U/\partial y)_\Gamma \in L^2(\Gamma)$ *then*

$$\int_0^\infty |u(t)|_1^2 dt < \infty, \tag{2.20}$$

$$\lim_{t\to\infty} |u(t)|_1 = 0. \tag{2.21}$$

Proof. Taking $v=w$ in (2.4), then taking $v=u(t)$ in (2.19) and integrating over $t\in(0,T)$, and adding, results in

$$\int_0^T 0\left\{\left\langle \frac{\partial u_\Gamma}{\partial t}, u_\Gamma - w_\Gamma\right\rangle + |u(t)|_1^2\right\} dt \leqq 0.$$

Since

$$\int_0^T\left\{\left\langle \frac{\partial u_\Gamma}{\partial t}, u_\Gamma - w\right\rangle\right\} dt = \frac{1}{2}|u_\Gamma(T)|_\Gamma^2 - \frac{1}{2}|u_\Gamma(0)|_\Gamma^2 - \int_0^T\left\langle \frac{\partial u_\Gamma}{\partial t}, w\right\rangle dt$$

(2.20) follows.

To prove (2.21) we suppose, that

$$|u(\sigma_n)|_1 \geqq c > 0 \quad \text{for a sequence } \sigma_n \to \infty.$$

Then, for $\sigma_n < s < \sigma_{n+1}$,

$$|u(s) - u(\sigma_n)|_1^2 = \left\{\int_{\sigma_n}^s |u_t(t)|_1 dt\right\}^2 \leqq \int_{\sigma_n}^{\sigma_{n+1}} |u_t|_1^2 dt = \varepsilon_n.$$

Since by (2.17) and (2.18),

$$\int_0^\infty |u_t|_1^2 dt < \infty,$$

it follows that $\varepsilon_n \to 0$ if $n\to\infty$ and consequently

$$|u(s)|_1 \geqq \frac{c}{2} \quad \text{if } \sigma_n < s < \sigma_{n+1}, \quad n \text{ large enough},$$

which contradicts (2.20).

Let u_∞ and U_m be defined by

$$u_\infty = \lim_{t\to\infty} u(t) \quad \text{and} \quad U_m = \inf_{x\in\Gamma} U_\Gamma.$$

By Theorem 2.3 we have that u_∞ is a constant.

COROLLARY 2.4. *Any solution of* $(\mathcal{P})$ *satisfies*

$$U_m \leqq u \leqq G_M, \quad i.e. \quad u\in L^\infty(\Omega\times(0,T)). \tag{2.22}$$

If $(\partial U/\partial y)_\Gamma \in L^2(\Gamma)$ *then the constant* u_∞ *satisfies*

$$U_m \leqq u_\infty \leqq G_m \tag{2.23}$$

and, in particular, when $U_m = G_m$ *we have that* $u_\infty = G_m$.

Proof. Taking $v = u - (u - U_m)^- \in K$ in (2.4) results in

$$\frac{1}{2}|(u_\Gamma(t) - U_m)^-|_\Gamma^2 + \int_0^t |(u(\tau) - U_m)^-|_1^2 d\tau \leqq \frac{1}{2}|(U_\Gamma - U_m)^-|_\Gamma^2 = 0$$

and this implies (2.22). Inequality (2.23) in an immediate consequence of the fact that $u_\infty \leqq G$.

3. Further regularity. It is convenient to use the penalty method for approximating the problems $(\mathscr{P})$ and $(\mathscr{P}^\delta)$. Let $\beta_\varepsilon(\cdot) \in C^\infty$ be a family of smooth functions depending on the positive parameter ε such that

$$\beta_\varepsilon \geqq 0, \quad \beta_\varepsilon' \geqq 0, \quad \beta_\varepsilon'' \geqq 0, \text{ and } \beta_\varepsilon(s)=0 \text{ if } s \leqq 0, \quad \lim_{\varepsilon\to 0} \beta_\varepsilon(s)=\infty \text{ if } s>0. \tag{3.1}$$

Consider the two problems:
$(\mathscr{P}_\varepsilon)$ Find $u_\varepsilon(t)$ such that

$$\begin{aligned} &\Delta u_\varepsilon = 0 \text{ in } \Omega, \quad \frac{\partial u_\varepsilon}{\partial \nu} = 0 \text{ on } \bigcup_i \Gamma_i \\ &\frac{\partial u_\varepsilon}{\partial t} + \frac{\partial u_\varepsilon}{\partial y} + \beta_\varepsilon(u_\varepsilon - G) = 0 \text{ on } \Gamma, \qquad u_\varepsilon(0) = U. \end{aligned} \tag{3.2}$$

$(\mathscr{P}_\varepsilon^\delta)$ Find $\{u_\varepsilon^n\}_{n=1}^N$ such that for all $n \geqq 1$

$$\begin{aligned} &\Delta u_\varepsilon^n = 0 \text{ in } \Omega, \quad \frac{\partial u_\varepsilon^n}{\partial \nu} = 0 \text{ on } \bigcup_i \Gamma_i \\ &\frac{1}{\delta} u_\varepsilon^n + \frac{\partial u_\varepsilon^n}{\partial y} + \beta_\varepsilon(u_\varepsilon^n - G) = \frac{1}{\delta} u_\varepsilon^{nl-1} \text{ on } \Gamma, \quad u_\varepsilon^0 = U. \end{aligned} \tag{3.3}$$

Standard arguments establish the existence and uniqueness of solutions to $(\mathscr{P}_\varepsilon)$ and $(\mathscr{P}_\varepsilon^\delta)$ with the following convergence properties. Defining $u_\varepsilon^\delta(t)$ and $\bar{u}_\varepsilon^\delta(t)$ by

$$\begin{aligned} &u_\varepsilon^\delta(t) = u_\varepsilon^n, \qquad t^{n-1} < t \leqq t^n, \\ &\bar{u}_\varepsilon^\delta(t) = u_\varepsilon^n + (t^n - t)(u^{n-1} - u^n)/\delta, \qquad t^{n-1} \leqq t \leqq t^n, \end{aligned} \tag{3.4}$$

where $t^n = n\delta$, we have that as $\delta \to 0$, using the arguments of §2,

$$\begin{aligned} &u_\varepsilon^\delta \rightharpoonup u_\varepsilon \quad \text{in } L^2(0,T; H^1(\Omega)), \\ &\bar{u}_\varepsilon^\delta \rightharpoonup u_\varepsilon \quad \text{in } L^2(0,T; H^1(\Omega)), \qquad (\bar{u}_\varepsilon^\delta)_\Gamma \rightharpoonup (u_\varepsilon)_\Gamma \quad \text{in } H^1(0,T; L^2(\Gamma)) \end{aligned} \tag{3.5}$$

and as $\varepsilon \to 0$, (see [3] or [6]),

$$\begin{aligned} &u_\varepsilon^n \to u^n \quad \text{in } H^1(\Omega), \\ &u_\varepsilon \rightharpoonup u \quad \text{in } L^2(0,T; H^1(\Omega)), \qquad (u_\varepsilon)_\Gamma \rightharpoonup u_\Gamma \quad \text{in } H^1(0,T; L^2(\Gamma)). \end{aligned} \tag{3.6}$$

For the remainder of this section we make the assumptions

$$\begin{aligned} &\text{a)} \quad U \in H^2(\Omega), \\ &\text{b)} \quad \left(\frac{\partial U}{\partial y}\right)_\Gamma \in L^\infty(\Gamma), \end{aligned} \tag{3.7}$$

on the data U_Γ. It follows that u_ε^n is in $H^2(\Omega)$ for each n and so is continuous.

THEOREM 3.1. *If* (3.7) *holds then* $\partial u_\Gamma/\partial t \in L^\infty(0,T;\ L^\infty(\Gamma))$.

Proof. Consider $w^n=(u_\varepsilon^n-u_\varepsilon^{n-1})/\delta$. It follows from (3.3) and that fact that $u_\varepsilon^n\in H^2(\Omega)$, that for any positive integer p, and $n\geqq 2$

$$\begin{aligned}0=\int_\Omega(-\Delta w^n)(w^n)^p dx\,dy=\int_\Omega\nabla w^n\cdot\nabla(w^n)^p dx\,dy\\ +\frac{1}{\delta}\int_\Gamma(w^n-w^{n-1}+\beta_\varepsilon^n-\beta_\varepsilon^{n-1})(w^n)^p dx,\end{aligned}\tag{3.8}$$

where $\beta_\varepsilon^n=\beta_\varepsilon(u_\varepsilon^n-G)$. For p odd,

$$p|\nabla w|^2 w^{p-1}\geqq 0$$

and, by the monotonicity of β_ε,

$$(\beta_\varepsilon^n-\beta_\varepsilon^{n-1})w^n\cdot(w^n)^{p-1}\geqq 0.$$

Hence (3.8) implies the inequality

$$\int_\Gamma(w^n-w^{n-1})(w^n)^p dx\leqq 0,$$

which can be rewritten as

$$\frac{1}{2}\int_\Gamma\{(w^n)^2-(w^{n-1})^2+(w^n-w^{n-1})^2\}(w^n)^{p-1}dx\leqq 0.\tag{3.9}$$

From (3.9) we deduce that

$$\int_\Gamma(w^n)^{p+1}dx\leqq\int_\Gamma(w^{n-1})^2(w^n)^{p-1}dx$$

and applying Young's inequality to the right-hand side, we obtain, for $n\geqq 2$,

$$\int_\Gamma(w^n)^{p+1}dx\leqq\int_\Gamma\left\{\frac{2}{p+1}(w^{n-1})^{p+1}+\frac{p-1}{p+1}(w^n)^{p+1}\right\}dx,$$

so that,

$$\int_\Gamma(w^n)^{p+1}dx\leqq\int_\Gamma(w^{n-1})^{p+1}dx\leqq\int_\Gamma(w^1)^{p+1}dx.\tag{3.10}$$

In order to bound w^1, we note that on Γ

$$w^1+\delta\frac{\partial}{\partial y}w^1+\beta_\varepsilon(u_\varepsilon^1-G)=-U_y,$$

which implies, for any odd integer p,

$$0=\int_\Omega\nabla w^1\cdot\nabla(w^1)^p dx\,dy+\frac{1}{\delta}\int_\Gamma(w^1)^p\big(w^1+U_y+\beta_\varepsilon(u_\varepsilon^1-G)\big)dx.$$

Noting that $U\leqq G$ on Γ and $\beta_\varepsilon(s)=0$ for $s\leqq 0$, we obtain

$$\int_\Gamma(w^1)^{p+1}dx+\frac{1}{\delta}\int_\Gamma(w^1)^{p-1}w^1\big(\beta_\varepsilon(u_\varepsilon^1-g)-\beta_\varepsilon(U-G)\big)dx\leqq-\int_\Gamma(w^1)^p U_y dx$$

which implies by the monotonicity of β_ε and Young's inequality that

$$(3.11)\qquad \int_\Gamma (w^1)^{p+1}dx \leqq \int_\Gamma \left\{ \frac{p}{p+1}(w^1)^{p+1} + \frac{1}{p+1}(U_y)^{p+1} \right\} dx.$$

It follows from (3.10) and (3.11) that for all $n \geqq 1$ and odd positive integers p,

$$(3.12)\qquad \int_\Gamma (w^n)^{p+1}dx \leqq \int_\Gamma (U_y)^{p+1}dx.$$

Taking the limit as $p \to \infty$ in (3.12) and recalling that $(U_y)_\Gamma \in L^\infty(\Gamma)$ we obtain

$$(3.13)\qquad \|w^n\|_{L^\infty(\Gamma)} \leqq \|U_y\|_{L^\infty(\Gamma)}.$$

The convergence properties (3.5) and (3.6) of the sequence $\{u_\varepsilon^n\}$ imply the result of the theorem.

PROPOSITION 3.2. *If* (3.7) *holds and* $G \in C^2[0,1]$ *with* $G'(0) = G'(1) = 0$ *then for any* $t > 0$

$$(3.14)\qquad \int_\Gamma \left(\frac{\partial u}{\partial x}\right)^2_\Gamma dx + 2\int_0^t \int_\Omega \left|\nabla \frac{\partial u}{\partial x}\right|^2 dx\,dy\,dt \leqq C,$$

where C depends on U and G; consequently $u \in L^2(0, T;\, H^2(\Omega))$.

Proof. Setting $w^n = \partial u_\varepsilon^n/\partial x$, it follows from $(\mathscr{P}_\varepsilon^\delta)$ that

$$0 = \int_\Omega -\Delta w^n \cdot w^n\,dx\,dy = \int_\Omega |\nabla w^n|^2 dx\,dy + \int_\Gamma w^n \left\{ \frac{w^n - w^{n-1}}{\delta} + \frac{\partial \beta_\varepsilon}{\partial x}(u_\varepsilon^n - G) \right\} dx.$$

Rearranging this equation we obtain

$$(3.15)\qquad \int_\Omega |\nabla w^n|^2 dx\,dy + \frac{1}{2\delta}\int_\Gamma \{(w^n)^2 - (w^{n-1})^2 + (w^n - w^{n-1})^2\}\,dx$$

$$+ \int_\Gamma (w^n - G')^2 \beta_\varepsilon'(u_\varepsilon^n - G)\,dx = -\int_\Gamma G' \frac{\partial}{\partial x}\beta_\varepsilon'(u_\varepsilon^n - G)\,dx.$$

Summing (3.15) over n, noting the monotonicity of β_ε and using $G'(0) = G'(1) = 0$, we obtain

$$(3.16)\qquad \int_\Gamma (w^n)^2 dx + 2\delta \sum_{j=1}^n \int_\Omega |\nabla w^j|^2 dx\,dy \leqq \int_\Gamma \left(\frac{\partial U}{\partial x}\right)^2 dx + 2\delta \sum_{j=1}^n \int_\Gamma G''\beta_\varepsilon(u_\varepsilon^n - G)\,dx.$$

By (3.3),

$$0 = \int_\Omega \Delta u_\varepsilon^n\,dx\,dy = \int_\Gamma \frac{\partial u_\varepsilon^n}{\partial y}dx = -\int_\Gamma \left\{ \beta_\varepsilon(u_\varepsilon^n - G) + \frac{u_\varepsilon^n - u_\varepsilon^{n-1}}{\delta} \right\} dx$$

and since the estimate (3.13) holds, we obtain

$$(3.17)\qquad \int_\Gamma \beta_\varepsilon(u_\varepsilon^n - G)\,dx \leqq \mathbf{C}$$

for a constant $\mathbf{C}$ independent of ε and n. Recalling (3.7a), the boundedness of G'' and the nonnegativity of β_ε, it follows from (3.17) that the left-hand side of (3.16) is bounded independently of ε, n and δ. Passing to the limit as ε and $\delta \to 0$ we obtain (3.14).

In this section we have established that if

$$\begin{aligned} &U\in H^2(\Omega), \quad (U_y)_\Gamma\in L^\infty(\Gamma), \\ &G\in C^2[0,1], \quad G'(0)=G'(1)=0, \end{aligned} \tag{3.18}$$

then

$$\begin{aligned} &u\in L^2(0,T;\ H^2(\Omega))\cap C[0,T;\ H^1(\Omega)]\cap L^\infty(0,T;\ L^\infty(\Omega)), \\ &\frac{\partial u}{\partial t}\in L^2(0,T;\ H^1(\Omega))\cap L^\infty(0,T;\ L^\infty(\Omega)), \\ &u_\Gamma\in L^\infty(0,T;\ H^1(\Gamma)), \quad \frac{\partial u_\Gamma}{\partial t}\in L^\infty(0,T;\ L^\infty(\Gamma)). \end{aligned} \tag{3.19}$$

It follows from the Sobolev imbedding theorems that

$$\begin{aligned} &u\in L^2(0,T;\ C^{0,\lambda}(\bar\Omega)) \quad \text{for all } \lambda, \quad 0\leqq\lambda<1, \\ &u_\Gamma\in L^\infty(0,T;\ C^{0,1/2}(\bar\Gamma)). \end{aligned} \tag{3.20}$$

Set $\Gamma_T=\Gamma\times(0,T)$ and

$$\Gamma_d\equiv\{(x,t):x\in\Gamma_d(t), 0<t<T\}, \qquad \Gamma_w\equiv\{(x,t):x\in\Gamma_w(t), 0<t<T\}. \tag{3.21}$$

The variational inequality (2.4) and the regularity results (3.19) imply

$$\int_0^T\int_\Gamma\left(\frac{\partial u_\Gamma}{\partial t}+\left(\frac{\partial u}{\partial y}\right)_\Gamma\right)(v_\Gamma-u_\Gamma)\,dx\,dt\geqq0 \quad \text{for all } v\in\mathscr{K}. \tag{3.22}$$

In particular, the solution of the variational inequality satisfies the linear complementarity system

$$\left.\begin{aligned} &\frac{\partial u_\Gamma}{\partial t}+\left(\frac{\partial u}{\partial y}\right)_\Gamma\leqq0, \qquad u_\Gamma\leqq G, \\ &\left(\frac{\partial u_\Gamma}{\partial t}+\left(\frac{\partial u}{\partial y}\right)_\Gamma\right)(u_\Gamma-G)=0, \end{aligned}\right\} \quad \text{a.e. in } \Gamma_T. \tag{3.23}$$

and

$$\frac{\partial u_\Gamma}{\partial t}+\left(\frac{\partial u}{\partial y}\right)_\Gamma=0 \quad \text{a.e. in } \Gamma_d. \tag{3.24}$$

4. Some properties of $\Gamma_d(t)$.

PROPOSITION 4.1. *If either $U_\Gamma=G$ or $(\partial U/\partial y)_\Gamma\geqq0$ then $(\partial u/\partial t)\leqq0$; consequently $\Gamma_d(t)\subset\Gamma_d(t')$ for $t<t'$.*

Proof. Noting the convergence properties of the sequences $\{u^n\}_{n=1}^N$ of Theorem 2.1, we see that it is sufficient to show

$$u^n\leqq u^{n-1} \ \text{in } \Omega, \qquad u_\Gamma^n\leqq u_\Gamma^{n-1}. \tag{4.1}$$

We proceed by induction. Taking $n=1$ and $v=u^1-(u^1-U)^+$ in (2.9) and noting that

$$\int_\Omega\nabla U\cdot\nabla v\,dx\,dy=\int_\Gamma\frac{\partial U}{\partial y}v\,dx,$$

we obtain

$$\frac{1}{\delta}\int_\Gamma (u^1-U)^{+2}dx+\int_\Omega |\nabla(u^1-U)^+|^2 dx\,dy \leqq \int_\Gamma -\frac{\partial U}{\partial y}(u^1-U)^+dx.$$

Hence either $U_\Gamma = G_\Gamma$ or $(\partial U/\partial y)_\Gamma \geqq 0$ results in

$$(4.2) \qquad u^1 - U \leqq 0 \quad \text{on } \Omega \qquad u^1_\Gamma \leqq U_\Gamma.$$

Taking $v = u^n + (u^{n+1}-u^n)^+$ and $v = u^{n+1}-(u^{n+1}-u^n)^+$ in the appropriate inequalities of (2.9) results in

(4.3)

$$\int_\Gamma (u^{n+1}-u^n)^{+2}dx + \int_\Omega \delta|\nabla(u^{n+1}-u^n)^+|^2 dx\,dy \leqq \int_\Gamma (u^n-u^{n-1})(u^{n+1}-u^n)^+dx.$$

Thus (4.2) and (4.3) imply by induction that (4.1) holds.

Proposition 4.1 gives a sufficient condition for the dry portion $\Gamma_d(t)$ to monotonically increase in size. The next theorem gives sufficient conditions for $\Gamma_d(t)$ to have at most one component.

THEOREM 4.2. *Assume that* $G \in C^3[0,1]$, $U_\Gamma \in C^1[0,1]$ *and*

$$(4.4) \qquad \begin{aligned} &G'(0) \leqq 0, \quad G'(1) \leqq 0, \quad G'''(x) \geqq 0 \quad \textit{and} \\ &(U_\Gamma - G)_x \geqq 0 \quad \textit{for all } x \in (0,1). \end{aligned}$$

Then $(u-G)_x \geqq 0$; *consequently, for each* t, $\Gamma_d(t)$ *is an interval* $(0,s(t))$ *for some* $s(t) \in [0,1]$.

Proof. It is sufficient to prove that

$$(4.5) \qquad (u_\varepsilon^{n-1}-G)_x \geqq 0 \quad \text{implies} \quad (u_\varepsilon^n - G)_x \geqq 0.$$

Setting $w = (u_\varepsilon^n - G)_x$, we have

$$(4.6) \qquad \begin{aligned} &-\Delta w = G''' \geqq 0 \quad \text{in } \Omega, \qquad w = -G' \quad \text{on } \Gamma_1 \cup \Gamma_3, \\ &\frac{\partial w}{\partial y} = 0 \quad \text{on } \Gamma_2, \\ &\frac{\partial w}{\partial y} + \frac{1}{\delta}w + \beta_\varepsilon(u_\varepsilon^n - G)w = \frac{1}{\delta}(u_\varepsilon^n - G)w = \frac{1}{\delta}(u_\varepsilon^{n-1}-g)_x \quad \text{on } \Gamma. \end{aligned}$$

We claim that w is continuous at the vertices of Ω. Indeed suppose that $\bar{X}$ is the vertex $(1,0)$. We define

$$z(x,y) = z(2-x,y) \quad \text{for } z = u_\varepsilon^n$$

and similarly extend $u_{\varepsilon,\Gamma}^{n-1}$ and G by reflection across $\{x=1\}$. Then the reflected u_ε^n satisfies (3.2) in

$$\Omega^* = \{(x,y): 0 < x < 2, \ -b < y < 0\}$$

with the same Neumann condition on $\{y=0\}$; however there is a jump in the first derivative of G across $\{x=1\}$. Representing u_ε^n in an Ω^*-neighborhood of $\bar{X}$ (denote it by Ω_0) in terms of the local Neumann function, we easily find that ∇u_ε^n is continuous at $\bar{X}$. Thus w is continuous in $\bar{\Omega}$ at the vertices of Ω.

If the assertion (4.5) is not true then by the maximum principle w takes a negative minimum at a point $X \in \partial\Omega$, and

$$\frac{\partial w}{\partial \nu}(X_0) < 0 \quad (\nu \text{ the outward normal}) \tag{4.7}$$

if X_0 is not a vertex of Ω. By the continuity of w at the vertices and the conditions $w = -G' \geqq 0$ on $\Gamma_1 \cup \Gamma_3$ it is clear that X_0 cannot be a vertex. Since $\partial w/\partial y = 0$ on Γ_2, X_0 must belong to Γ and then

$$\frac{\partial w}{\partial y}(X_0) = \frac{1}{\delta}(u_\varepsilon^{n-1} - G)_x - \frac{1}{\delta} w - \beta_\varepsilon'(u_\varepsilon^n - G) w > 0 \tag{4.8}$$

by the induction hypothesis and the nonnegativity of β_ε'; this contradicts (4.7). Thus (4.5) is proved.

Remark 4.1. Numerical calculations reported in [1] in the case

$$G = \cos \pi x, \qquad U(x,y) = \cos \pi x \frac{\cosh \pi(y+b)}{\cosh \pi b}$$

illustrate the results of Proposition 4.1 and Theorem 4.2.

We now show that, under certain conditions, u_Γ is Hölder continuous.

PROPOSITION 4.3. *Assume that* (3.7) *holds*, $U_\Gamma \in C^1[0,1]$, $G \in C^3[0,1]$, $G'''(x) \geqq 0$, $(U_\Gamma - G)'(x) \geqq 0$ $0 < x < 1$, $G'(0) \leqq 0$ *and* $G'(1) \leqq 0$. *Then* $(\partial u_\Gamma/\partial x) \in L^\infty(0,T;\ L^3(\Gamma))$ *and* $u_\Gamma \in C^{2/3}(\overline{\Gamma}_T)$.

Proof. Setting $w^n = \partial u_\varepsilon^n/\partial x - G'$ whih is continuous and nonnegative by the proof of Theorem 4.2 we find that

$$\begin{aligned} 0 &= -\int_\Omega \Delta w^n \cdot (w^n)^2 dx\,dy - \int_\Omega G'''(w^n)^2 dx\,dy \\ &= 2\int_\Omega w^n |\nabla w^n|^2 dx\,dy + \int_\Gamma \left\{ (w^n)^2 \left(\frac{w^n - w^{n-1}}{\delta} \right) + (w^n)^2 \frac{\partial}{\partial x} \beta_\varepsilon(u_\varepsilon^n - G) \right\} dx \\ &\quad - \int_\Omega G'''(w^n)^2 dx\,dy. \end{aligned} \tag{4.9}$$

Since $w^n \geqq 0$ and $\beta_\varepsilon' \geqq 0$,

$$\frac{\partial}{\partial x} \beta_\varepsilon(u_\varepsilon^n - G) = w^n \beta_\varepsilon'(u_\varepsilon^n - G) \geqq 0$$

and we obtain from (4.9) the inequality

$$\frac{1}{2\delta} \int_\Gamma w^n \{ (w^n)^2 - (w^{n-1})^2 + (w^n - w^{n-1})^2 \} dx \leqq \int_\Omega G'''(w^n)^2 dx\,dy,$$

which implies the inequality

$$\int_\Gamma (w^n)^3 \leqq \int_\Gamma w^n (w^{n-1})^2 dx + 2\delta \int_\Omega G'''(w^n)^2 dx\,dy. \tag{4.10}$$

Young's inequality implies

$$\int_\Gamma w^n (w^{n-1})^2 dx \leqq \int_\Gamma \left\{ \left(\frac{w^n}{3} \right)^3 + \frac{2}{3}(w^{n-1})^3 \right\} dx$$

and (4.10) yields, upon summing over n,

$$\int_\Gamma (w^n)^3 dx \leqq \int_\Gamma (U_\Gamma' - G')^3 dx + 3\delta \sum_{j=1}^{n} \int_\Omega G'''(w^j)^2 dx\, dy. \tag{4.11}$$

The convergence of $\{u_\varepsilon^n\}$ as $\varepsilon \to 0$ and $\delta \to 0$, together with the boundedness of derivatives of U_Γ and G and the nonnegativity of w^n imply that $\partial u_\Gamma/\partial x \in L^\infty(0, T;\ L^3(\Gamma))$. Since $u_\Gamma \in L^\infty(\Gamma_T)$, the Sobolev imbedding theorem implies that $u_\Gamma \in L^\infty(0, T;\ C^{2/3}(\Gamma))$. By Theorem 3.1 we also have $\partial u_\Gamma/\partial t \in L^\infty(0, T;\ L^\infty(\Gamma)) = L^\infty(\Gamma_T)$. Hence

$$\begin{aligned} |u_\Gamma(x,t) - u_\Gamma(x',t')| &\leqq |u(x,t) - u(x',t)| + |u(x',t) - u(x',t')| \\ &\leqq C_1|x - x'|^{2/3} + C_2|t - t'|, \end{aligned}$$

so that $u_\Gamma \in C^{2/3}(\bar{\Gamma}_T)$.

A related problem of interest, [2], is the steady state free boundary problem.

$$\begin{aligned} &\Delta u = 0 \quad \text{in } \Omega, \\ &u = \gamma_1 \quad \text{on } \Gamma_1, \qquad u = \gamma_3 \quad \text{on } \Gamma_3, \quad \frac{\partial u}{\partial y} = 0 \quad \text{on } \Gamma_2, \\ &u \leqq G, \quad \frac{\partial u}{\partial y} \leqq G, \quad (u - G)\frac{\partial u}{\partial y} = 0 \quad \text{on } \Gamma, \end{aligned} \tag{4.12}$$

where $\gamma_1 \leqq G(0)$ and $\gamma_3 \leqq G(1)$. The following theorem can be proved by the method of proof of Theorem 4.2.

THEOREM 4.4. *Assume that* $G''(0) \geqq 0$, $G''(1) \geqq 0$ *and* $G^{(iv)}(x) \leqq 0$ *for* $0 < x < 1$. *Then* $(u - G)_{xx} \leqq 0$ *and thus the set* $\{x = u(x,0) < G(x)\}$ *consists of at most one interval.*

Indeed, working with $v = (u_\varepsilon - G)_{xx}$, the only point which requires clarification is that v cannot take a positive maximum at a vertex. However, since

$$\begin{aligned} &\frac{\partial v}{\partial y} + \beta_\varepsilon'' \cdot (u_x - G')^2 + \beta_\varepsilon' \cdot v = 0 \quad \text{on } \Gamma, \\ &\frac{\partial v}{\partial y} = 0 \quad \text{on } \Gamma_2, \\ &v = -G'' \quad \text{on } \Gamma_1 \cup \Gamma_3, \end{aligned}$$

we have, by reflecting across $\{x = 0\}$ and $\{x = L\}$, that v is continuous at the vertices. The assumptions on G'' at the end points imply the result.

Remark 4.2. Results of numerical calculations which illustrate Theorem 4.4 are given in [2].

REFERENCES

[1] J. M. AITCHISON, C. M. ELLIOTT AND J. R. OCKENDON, *Percolation in gently sloping beaches*, IMA J. Appl. Math., 30 (1983), pp. 269–287.

[2] J. M. AITCHISON, A. G. NEWLANDS AND C. P. PLEASE, *Percolation in intertidal sand-bands—a constrained harmonic problem*, CEGB Research Rept. TPRD/L/2344/N82 (1982), CERL, Kelvin Avenue, Leatherhead, Surrey, UK.

[3] A. Friedman, *Variational Principles and Free-Boundary Problems*, John Wiley, New York, 1982.

[4] A. Friedman and A. Torelli, *A free boundary problem connected with nonsteady filtration in a porous medium*, Nonlinear Anal. Theory Methods Appl., 1 (1977), pp. 503–545; *correction*, 2 (1978), pp. 513–518.

[5] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and their Applications*, Academic Press, New York, 1980.

[6] J. L. Lions, *Quelques méthodes de résolution de problèmes aux limites non linéaires*, Dunod, Paris, 1969.

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

A NONLINEAR EVOLUTION PROBLEM ASSOCIATED WITH AN ELECTROPAINT PROCESS*

LUIS A. CAFFARELLI† AND AVNER FRIEDMAN‡

Abstract. A model for an electropaint process can be described by a time-dependent family of harmonic functions $\phi(x,t)$ satisfying on the surface which is being painted: $h\phi_n=\phi$ where h is the thickness of the paint coat and ϕ_n is the inward normal derivative of ϕ; h is a suitable nonlinear function of the history of ϕ. We prove that the time-discretized version of the problem has a unique solution $\{\phi(x,t_i)\}$ and $\lim_{t_i\to\infty}\phi(x,t_i)$ exists and coincides with the solution $W(x)$ of an appropriate Signorini problem.

1. Introduction. A common method for painting a metal surface is electropainting. The metal workpiece is immersed in a bath of electrolyte solution containing negative ions and is made the anode, as a potential difference is applied on the outside boundary of the bath. The workpiece is then being painted by disposition from the resulting transport process.

In a recent paper [1] Aitchison, Lacy and Shillor developed a model describing the above process. They further derived a simplified model which we shall now describe.

Let Γ be the surface of the workpiece and S the (outside) boundary of the bath, and denote by Ω the domain occupied by the electrolyte solution, i.e., $\partial\Omega=S\cup\Gamma$ with Γ the inner boundary of Ω and S the outside boundary of Ω; Ω is a domain in $\mathbb{R}^N$, for any $N\geq 2$.

Denote by ϕ the electric potential, by h the thickness of the paint coat on Γ, and by ε the dissolution current (a positive dimensionless constant). Then

$$\Delta\phi=0 \quad \text{in } \Omega, \tag{1.1}$$

$$\phi=1 \quad \text{on } S, \tag{1.2}$$

$$h\frac{\partial\phi}{\partial n}=\phi \quad \text{on } \Gamma, \tag{1.3}$$

$$\frac{\partial h}{\partial t}=\frac{\partial\phi}{\partial n}-\varepsilon \quad \text{if } x\in\Gamma, \quad h(x,t)>0, \tag{1.4}$$

$$\frac{\partial h}{\partial t}=\left(\frac{\partial\phi}{\partial n}-\varepsilon\right)^+ \quad \text{if } x\in\Gamma, \quad h(x,t)=0, \tag{1.5}$$

$$h(x,0)=0 \quad \text{if } x\in\Gamma \tag{1.6}$$

where $\partial/\partial n$ is the normal derivative into Ω. The condition (1.4) means that, where $h>0$,

(1.4′) if $\partial\phi/\partial n>\varepsilon$ then the thickness h increases, whereas if $\partial\phi/\partial n<\varepsilon$ then h decreases.

Notice that (1.4), (1.5) imply that

$$h\geq 0. \tag{1.7}$$

* Received by the editors June 7, 1984. This work was partially supported by the National Science Foundation under grants 7406375 and MCS 7915171.

† University of Chicago, Chicago, Illinois.

‡ Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

It is of special interest to find the thickness distribution $h(x,t)$ of the paint coat after a sufficiently long time. Assuming that the process stabilizes as $t \to \infty$ and taking $\partial h/\partial t = 0$, the problem (1.1)–(1.6) formally stabilizes to the solution $W(x)$ of the variational inequality:

$$\begin{gathered} \Delta W = 0 \quad \text{in } \Omega, \qquad W = 1 \quad \text{on } S, \\ W \geq 0, \quad \frac{\partial W}{\partial n} - \varepsilon \leq 0, \quad \left(\frac{\partial W}{\partial n} - \varepsilon\right) W = 0 \quad \text{on } \Gamma. \end{gathered} \tag{1.8}$$

This is precisely the Signorini problem (for details and references on this problem, see [4]).

No existence, uniqueness or regularity results are known for (1.1)–(1.6). It was conjectured in [1] that

(1.9) the solution of (1.8) is a global attractor of solutions (1.1)–(1.6).

In this paper we study the problem (1.1)–(1.6) and establish (1.9) for the time-discretized version of (1.1)–(1.6). We also establish the important fact that

(1.10) there is no paint dissolution in the process (1.1)–(1.6).

Thus, although (1.4) (or (1.4′)) allows for the possibility that $h(x,t)$ may decrease in some time intervals, where $h > 0$, what actually happens is that

(1.10′) $h(x,t)$ is a nondecreasing function of t, for all $t > 0$.

In §2 we establish (1.10) (or (1.10′)) and (1.9) for smooth solutions of (1.1)–(1.6). In §4 we introduce the time-discretized version of (1.1)–(1.6) and prove that it has a unique solution $\{\phi(x,t_i)\}$. To accomplish this, we introduce in §3 a "regularized" version of the discretized system, involving a small parameter $\sigma > 0$, whereby $h(x,t_i)$ is replaced by $h(x,t_i)+\sigma$. We derive (in §3) existence, uniqueness and a priori estimates for the "σ-problem". In §4 we let $\sigma \to 0$ to obtain the solution $\{\phi(x,t_i)\}$.

In §5 we prove that $\phi(x,t_i) \to W(x)$ as $t_i \to \infty$, where $W(x)$ is the solution of (1.8).

2. No paint dissolution; convergence to *W*. Denote by w the solution of

$$\Delta w = 0 \text{ in } \Omega, \quad w = 1 \text{ on } S, \quad w = 0 \text{ on } \Gamma, \tag{2.1}$$

that is, $w(x) = \phi(x,0)$ if ϕ is a solution of (1.1)–(1.6).

Denote by W the solution of the Signorini problem (1.8); more precisely, if

$$K = \{\zeta \in H^1(\Omega), \zeta = 1 \text{ on } S, \zeta \geq 0 \text{ on } \Gamma\} \tag{2.2}$$

then W is the solution of

$$W \in K, \qquad \int_\Omega \nabla W \cdot \nabla(\zeta - W) + \varepsilon \int_\Gamma (\zeta - W) \geq 0 \quad \forall \zeta \in K. \tag{2.3}$$

We assume for simplicity that $\partial\Omega \in C^{1,\beta}$ for any $0 < \beta < 1$.

It was proved by Frehse [3] that $W \in C^1(\bar{\Omega})$ and by Caffarelli [2] (under some mild conditions) that $W \in C^{1,\alpha}(\bar{\Omega})$ for some $0 < \alpha < 1$; for further details, see [4].

Remark 2.1. If

$$w_n < \varepsilon \quad \text{along } \Gamma, \tag{2.4}$$

then $\phi(x,t)\equiv w(x)$ together with $h(x,t)=0$ for a solution of (1.1)–(1.6). Here the potential difference is too weak to cause any accumulation of paint anywhere on the workpiece Γ.

Remark 2.2. Suppose, for some $T\geq 0$,

$$\phi(x,T)=W(x) \qquad (x\in\Omega). \tag{2.5}$$

Then $h_t(x,T)=0$ (by (1.4), (1.5) and (1.8)) and

$$h(x,T)W_n(x)=W(x) \quad \text{on } \Gamma.$$

Set

$$\begin{aligned} \tilde{\phi}(x,t)&=\begin{cases} \phi(x,t) & \text{if } t\geq T,\\ W(x) & \text{if } t>T,\end{cases}\\ \tilde{h}(x,t)&=\begin{cases} h(x,t) & \text{if } t\leq T,\\ h(x,T) & \text{if } t>T.\end{cases}\end{aligned} \tag{2.6}$$

Then $(\tilde{\phi},\tilde{h})$ forms a solution of (1.1)–(1.6) for all $t>0$, with h_t continuous across $\{t=T\}$.

In the next theorem we wish to exclude the trivial case (2.4); i.e., we shall assume that

$$\phi_n(x,0)>\varepsilon \quad \text{for some points } x\in\Gamma. \tag{2.7}$$

DEFINITION 2.1. Denote by t_0 the supremum of all numbers s such that

$$h(x,t) \text{ is nondecreasing for all } x\in\Gamma, \qquad 0\leq t<s.$$

DEFINITION 2.2. By a *smooth solution* of (1.1)–(1.6) we mean a solution (ϕ,h) such that ϕ, ϕ_t, $D_x\phi$ are continuous in $\bar{\Omega}\times[0,\infty)$ and h, h_t are continuous in $\Gamma\times[0,\infty)$.

THEOREM 2.1. *Let (ϕ,h) be a smooth solution of* (1.1)–(1.6) *and let* (2.7) *hold. Then either* (i) $t_0=\infty$, *or* (ii) $t_0<\infty$ *and there exists a $t_*\in(0,t_0]$ such that $\phi(x,t)<W(x)$ if $x\in\Omega$, $0\leq t<t_*$ and $\phi(x,t_*)\equiv W(x)$.*

By Remark 2.2, if (ii) occurs then $\phi(x,t)$ and $h(x,t)$ can be modified (for $t>t_*$) into another smooth solution for which h is nondecreasing for all $t>0$.

Proof. Denote by $\partial/\partial\nu$ the exterior normal derivative along Γ. Then, by (1.3),

$$h\frac{\partial\phi}{\partial\nu}+\phi=0 \quad \text{on } \Gamma. \tag{2.8}$$

Since $h\geq 0$, we can use the strong maximum principle to deduce that ϕ cannot take positive maximum or negative minimum on Γ; therefore

$$0\leq\phi\leq 1 \quad \text{in } \Omega. \tag{2.9}$$

By the strong maximum principle we also deduce that

$$\text{if } \phi(x,t)=0 \text{ for some } x\in\Gamma,\ t>0, \text{ then } \phi_n(x,t)>0. \tag{2.10}$$

Thus, by (1.3), for any $x\in\Gamma$, $t>0$,

$$\begin{aligned} &\phi(x,t)=0 \text{ implies } h(x,t)=0,\ \phi_n(x,t)>0;\\ &\phi(x,t)>0 \text{ implies } h(x,t)>0,\ \phi_n(x,t)>0.\end{aligned}$$

We conclude that

$$\phi_n(x,t)>0 \quad \text{on } \Gamma. \tag{2.11}$$

LEMMA 2.2. *If $h(x,t_1)\geq h(x,t_2)$ and $h(x,t_1)\not\equiv h(x,t_2)$ on Γ, then $\phi(x,t_1)>\phi(x,t_2)$ in Ω.*

Proof. The function $\psi(x)=\phi(x,t_1)-\phi(x,t_2)$ is harmonic in Ω, vanishes on S and satisfies

$$h(x,t_1)\frac{\partial\psi}{\partial\nu}+\psi=-(h(x,t_1)-h(x,t_2))\frac{\partial\phi(x,t_2)}{\partial\nu}\geq 0$$
$$\text{and} \quad \not\equiv 0 \quad \text{on } \Gamma$$

(by (2.11)). Applying the strong maximum principle, we deduce that $\psi>0$ in Ω.

If $t_0=\infty$, then the assertion (i) follows. We may therefore assume that $t_0<\infty$. Then there exists a sequence (x_i,t_i) with $x_i\in\Gamma$, $t_i>t_0$ such that

$$t_i\downarrow t_0, \quad x_i\to x_0 \quad \text{if } i\to\infty$$

and

$$h_t(x_i,t_i)<0; \qquad \text{consequently also } h(x_i,t_i)>0.$$

Clearly $h_t(x_0,t_0)=0$.

LEMMA 2.3. *There holds $\phi_{nt}(x_0,t_0)\leq 0$.*

Proof. If $\phi_n(x_i,t_0)\geq\varepsilon$ then, since $\phi_n(x_i,t_i)<\varepsilon$,

$$\phi_{nt}(x_i,\tilde{t}_i)<0 \quad \textit{for some } t_0<\tilde{t}_i<t_i. \tag{2.12}$$

If, on the other hand, $\phi_n(x_i,t_0)<\varepsilon$, then $h(x_i,t_0)=0$ (for $h(x_i,t_0)>0$ implies $h_t(x_i,t_0)=\phi_n(x_i,t_0)-\varepsilon<0$, a contradiction to the definition of t_0). Since further $h(x_i,t_i)>0$, we conclude that $h_t(x_i,\hat{t}_i)>0$ for some $0<\hat{t}_i<t_i$ and thus $\phi_n(x_i,\hat{t}_i)>\varepsilon$, which yields (since $\phi_n(x_i,t_i)<\varepsilon$)

$$\phi_{nt}(x_i,\tilde{\tilde{t}}_i)<0 \quad \text{for some } \hat{t}_i<\tilde{\tilde{t}}_i<t_i. \tag{2.13}$$

The lemma now follows from (2.12), (2.13) upon taking $i\to\infty$.

From Lemma 2.2 we have

$$\phi_t\geq 0 \quad \text{if } 0<t\leq t_0. \tag{2.14}$$

LEMMA 2.4. *There holds*: $h_t(x,t_0)\equiv 0$.

Proof. Suppose $h_t(x,t_0)\not\equiv 0$. Then from the proof of Lemma 2.2 we have that

$$\phi_t(x,t_0)>0 \quad \text{in } \Omega \tag{2.15}$$

(note that, for $t=t_0$, $h(\phi_t)_\nu+(\phi_t)=-h_t\phi_\nu\geq 0$ and $\not\equiv 0$ on Γ).

Differentiating (1.3) with respect to t, we have

$$h_t\phi_n+h\phi_{nt}=\phi_t. \tag{2.16}$$

Now, at (x_0,t_0) $h_t=0$ and $\phi_{nt}\leq 0$ (by Lemma 2.3). Since also $\phi_t\geq 0$, (2.16) implies that

$$\phi_t(x_0,t_0)=0. \tag{2.17}$$

Thus applying the strong maximum principle to the harmonic function $\phi_t(x,t_0)$ (using (2.15), (2.17)), we deduce that $\phi_{nt}(x_0,t_0)>0$, a contradiction to Lemma 2.3.

LEMMA 2.5. *There exists a* $\tau \in (0,t_0)$ *such that* $\phi(x,t) < W(x)$ *if* $x \in \Omega$, $t < \tau$ *and* $\phi(x,\tau) \equiv W(x)$.

Proof. Observe that $\phi(x,0) \equiv w(x) \leq W(x)$ on Γ and, by (2.7), $w_n > \varepsilon \geq W_n$ at some points of Γ. It follows that $w \not\equiv W$ and, in fact,

$$\phi(x,0) < W(x) \quad \text{in } \Omega. \tag{2.18}$$

Set $A(t) = \{x \in \Gamma, \phi(x,t) = W(x)\}$, for any $t \leq t_0$. We claim:

(2.19) if $\phi(x,t) < W(x)$ in Ω and if $\bar{x} \in A(t)$, then $h(\bar{x},t) = 0$ (and consequently $\phi(\bar{x},t) = W(\bar{x}) = 0$).

Indeed, $\phi_n(\bar{x},t) < W_n(\bar{x},t) \leq \varepsilon$ so that if $h(\bar{x},t) > 0$, then

$$h_t(\bar{x},t) = \phi_n(\bar{x},t) - \varepsilon < 0,$$

which contradicts the definition of t_0.

On $A(0)$, $\phi_n < W_n$ (by (2.18) and the strong maximum principle) and $W_n \leq \varepsilon$ by (1.8). Thus $\phi_n < \varepsilon$ on $A(0)$ and, by continuity, $\phi_n < \varepsilon$ in some $\Gamma \times [0,\infty)$-neighborhood N of $A(0)$. This implies that $h_t \leq 0$, or $h = 0$ on N, so that $\phi = 0 \leq W$ in N. Since $\phi < W$ in $\Gamma \backslash A(0)$, it follows that $\phi \leq W$ on $\Gamma \times [0,t_1]$ for some $t_1 > 0$. Hence $\phi(x,t) < W(x)$ if $0 \leq t \leq t_1$, $x \in \Omega$.

Let τ be the supremum of all $t \in (0,t_0)$ such that $\phi(x,t) < W(x)$ in Ω.

Consider first the case $\tau < t_0$. We claim that

$$\phi(x,\tau) \equiv W(x). \tag{2.20}$$

Indeed, if (2.20) is not true, then $\phi(x,\tau) < W(x)$ in Ω. We can now use (2.19) to conclude that $h = 0$, $h_t = 0$ in some $\Gamma \times (0,\infty)$-neighborhood N_τ of $A(\tau)$ and, therefore, $\phi \leq W$ in $\Gamma \times (0,\tau+\delta)$ for some $\delta > 0$ (and $\phi \not\equiv W$ if $x \in \Omega$, $\tau < t < \tau + \delta$). It follows that $\phi(x,t) < W(x)$ if $x \in \Omega$, $0 < t < \tau + \delta$, a contradiction to the definition of τ.

We have thus completed the proof of (2.20), assuming $\tau < t_0$. It remains to consider the case $\tau = t_0$ and to prove that (2.20) holds also in this case.

If $\phi(x,t_0) \not\equiv W(x)$, then $W(x) > \phi(x,t_0)$ in Ω and

$$(W-\phi)_n > 0 \quad \text{on } S \text{ and on } A(t_0).$$

On $\Gamma \backslash A(t_0)$ we have

$$W(x) > \phi(x,t_0) \geq 0 \quad \text{and consequently } W_n = \varepsilon.$$

By Lemma 2.4 we also have $\phi_n(x,t_0) \leq \varepsilon$ on Γ. Therefore

$$(W-\phi)_n \geq 0 \quad \text{on } \Gamma \backslash A(t_0).$$

We deduce that

$$0 = \int_\Omega \Delta(W-\phi) = \int_{S \cup \Gamma} (W-\phi)_n > 0,$$

a contradiction.

Remark 2.3. One can immediately check that for any $t \in (0,\infty)$,

$$\phi(x,t) \equiv W(x) \text{ in } \Omega \text{ if and only if } h_t(x,t) \equiv 0 \text{ on } \Gamma.$$

COROLLARY 2.6. *Suppose* $\phi(x,t) \not\equiv W(x)$ *in* Ω, *for any* $t > 0$. *If* $h(\bar{x},s) > 0$ *for some* $\bar{x} \in \Gamma$, $s > 0$, *then* $h_t(\bar{x},t) > 0$ *for all* $t > s$.

Thus, once $h(\bar{x},t)$ becomes strictly positive, it continues to grow at a strictly positive rate.

To prove the corollary, we proceed by contradiction and consider the number t_1 such that $h_t(\bar{x},t)>0$ if $s<t<t_1$ and $h_t(\bar{x},t_1)=0$. Applying the argument of Lemma 2.4, we deduce that $h_t(x,t_1)\equiv 0$ and $\phi(x,t_1)\equiv W(x)$.

COROLLARY 2.7. *If* $w_n(x,0)>\varepsilon$ *for all* $x\in\Gamma$, *then* $t_0=\infty$ *and* $h_t(x,t)>0$ *for all* $x\in\Gamma$, $t>0$.

Proof. Otherwise, there is a $\tau>0$ such that $h_t(x,t)>0$ if $x\in\Gamma$, $0\le t<\tau$ and $h_t(x_0,\tau)=0$ for some $x_0\in\Gamma$. By Lemma 2.4, $h_t(x,\tau)\equiv 0$. For any $0<t_1<t_2<\tau$, set $\phi^i(x)=\phi(x,t_i)$, $h^i(x)=h(x,t_i)$. Then

$$(h^2-h^1)\phi_n^2+(\phi_n^2-\phi_n^1)h^1=\phi^2-\phi^1 \quad \text{along } \Gamma. \tag{2.21}$$

By standard estimates for harmonic functions

$$\int_\Gamma (\phi^2-\phi^1)\ge c\int_S (\phi^2-\phi^1)_n \qquad (c>0).$$

Also,

$$\int_S (\phi^2-\phi^1)_n=-\int_\Gamma (\phi^2-\phi^1)_n.$$

Hence

$$\int_\Gamma (\phi^2-\phi^1)\ge -c\int_\Gamma (\phi^2-\phi^1)_n.$$

Integrating (2.21) over Γ and using the last inequality, we find that

$$C\int_\Gamma (h^2-h^1)\ge -\int_\Gamma (\phi^2-\phi^1)_n \qquad (C>0),$$

or

$$C\int_\Gamma \left(\int_{t_1}^{t_2} h_t dt\right)\ge -\int_\Gamma \left[(\phi_n^2-\varepsilon)-(\phi_n^1-\varepsilon)\right].$$

Recalling that $h_t=(\phi_n-\varepsilon)=(\phi_n-\varepsilon)^+>0$ if $t<\tau$, and setting

$$\psi(t)=\int_\Gamma h_t(x,t),$$

we arrive at the inequality

$$C\int_{t_1}^{t_2}\psi\ge -(\psi(t_2)-\psi(t_1)),$$

or $\psi+C\psi\ge 0$. Thus ψe^{Ct} is monotone nondecreasing for $0<t<\tau$. Since $\psi(0)>0$, it follows that $\psi(\tau)>0$, a contradiction to $h_t(x,\tau)\equiv 0$.

Theorem 2.1 suggests that in the model (1.1)–(1.6) we should replace (1.4), (1.5) by

$$\frac{\partial h}{\partial t}=\left(\frac{\partial \phi}{\partial n}-\varepsilon\right)^+,$$

i.e.,

$$h(x,t)=\int_0^t(\phi_n(x,s)-\varepsilon)^+ds. \tag{2.22}$$

We next prove (1.9) for smooth solutions satisfying:

$$\sup_{t>0}\int_\Omega|\nabla\phi(x,t)|^2dx<\infty, \tag{2.23}$$

$$\sup_{x\in\Gamma,\,t>0}|\phi_n(x,t)|<\infty. \tag{2.24}$$

THEOREM 2.8. *If the assertion* (i) *of Theorem* 2.1 *holds and if* (2.33), (2.24) *are satsified, then* $\phi(x,t)\uparrow W(x)$ *pointwise in* Ω *and weakly in* $H^1(\Omega)$ *and* $h(x,t)\uparrow h(x)$ *pointwise in* Γ *as* $t\uparrow\infty$, *where* $h(x)\le 1/\varepsilon$.

Proof. Suppose $h(x,t)>1/\varepsilon$ for some $x\in\Gamma$, $t>0$. Then $(\phi_n(x,s)-\varepsilon)^+>0$ and $h(x,s)>0$ for some $0<s<t$ and, by Corollary 2.6, $\phi_n(x,t)>\varepsilon$. But then

$$\varepsilon h\le h\phi_n=\phi\le 1 \quad \text{at } (x,t),$$

so that $h(x,t)\le 1/\varepsilon$, a contradiction.

By (2.23), $\phi(x,t)\uparrow\bar{\phi}(x)$ pointwise in Ω and weakly in $H^1(\Omega)$, as $t\uparrow\infty$. For any $\zeta\in K$ (K as in (2.21))

$$\begin{aligned}\int_\Omega\nabla\phi(x,t)\cdot\nabla(\zeta-\phi(x,t))+\varepsilon\int_\Gamma(\zeta-\phi(x,t))\\ =-\int_\Gamma(\phi_n(x,t)-\varepsilon)(\zeta-\phi(x,t))\ge-\int_{\Gamma_t}(\phi_n(x,t)-\varepsilon)(\zeta-\phi(x,t))\end{aligned} \tag{2.25}$$

where $\Gamma_t=\Gamma\cap\{\phi(x,t)>\varepsilon\}$; $\Gamma_t\uparrow\Gamma_\infty$ if $t\uparrow\infty$. For any small $\eta>0$ and large $M>0$ there is a τ (depending only on η; $\tau\to\infty$ if $\eta\to0$) such that

$$\begin{aligned}\int_\tau^{\tau+M}\int_{\Gamma_t}|\phi_n-\varepsilon|=\int_\tau^{\tau+M}\int_{\Gamma_t}(\phi_n-\varepsilon)^+\le\int_\tau^{\tau+M}\int_{\Gamma_\tau}(\phi_n-\varepsilon)^++\eta M\\ \le\int_{\Gamma_\tau}h(x,\tau+M)+\eta M\le\frac{C}{\varepsilon}+\eta M;\end{aligned} \tag{2.26}$$

here we used (2.24), the fact that $\operatorname{meas}(\Gamma_\infty\backslash\Gamma_\tau)\to0$ if $\tau\to\infty$, and estimate $h\le1/\varepsilon$.

Integrating (2.25) with respect to t, $\tau<t<\tau+M$ and dividing by M, then using (2.26) and letting $M\to\infty$, $\eta\to0$, we find that $\bar{\phi}$ is the solution of the variational inequality (2.3).

From Theorems 2.1, 2.8 we obtain:

COROLLARY 2.9. *Let* (ϕ,h) *be a smooth solution of* (1.1)–(1.6) *satisfying* (2.23), (2.24) *and let* (2.7) *hold. Then there exists a* $t_*\in(0,\infty]$ *such that*

$$\begin{aligned}&\phi(x,t)<W(x) && \textit{for all } x\in\Omega,\quad 0\le t<t_*,\\ &\phi(x,t)\uparrow W(x) && \textit{weakly in } H^1(\Omega)\quad \textit{as } t\uparrow t_*.\end{aligned}$$

If $w_n(x,0)>\varepsilon$ for all $x\in\Gamma$, then $t_*=\infty$ (by Corollary 2.7). We suspect that $t_*=\infty$ in all cases.

3. A discretized-regularized approximation. We wish to study a time-discretized version of the evolutionary process

$$\Delta\phi=0 \quad \text{in } \Omega, \tag{3.1}$$

$$\phi=1 \quad \text{on } S, \tag{3.2}$$

$$h\phi_n=\phi \quad \text{on } \Gamma, \tag{3.3}$$

$$h=\int_0^t (\phi_n(x,s)-\varepsilon)^+ dx \quad \text{on } \Gamma, \tag{3.4}$$

where $\phi=\phi(x,t)$, $h=h(x,t)$. In this section we study a regularized approximation of the discretized system. This approximation is defined as follows.

Let δ and σ be any small positive numbers with

$$\sigma<\frac{1}{\varepsilon}$$

and set $t_m=m\delta$ $(m=1,2,\cdots)$. If we replace $\phi(x,t_m)$ in (3.1)–(3.4) by $\phi^m(x)$, we get

$$\Delta\phi^m=0 \quad \text{in } \Omega, \tag{3.5}$$

$$\phi^m=1 \quad \text{on } S, \tag{3.6}$$

$$h^m\frac{\partial\phi^m}{\partial n}=\phi^m \quad \text{on } \Gamma \tag{3.7}$$

where h^m is defined by

$$h^m(x)=\sigma+\delta\sum_{i=1}^{m}\left(\frac{\partial\phi^i}{\partial n}-\varepsilon\right)^+ \qquad (x\in\Gamma) \tag{3.8}$$

with $\sigma=0$; in this section we take $\sigma>0$ in order to avoid degeneracy in (3.7). In §4 we shall let $\sigma\to 0$.

As in §2 we assume, for simplicity, that

$$\partial\Omega\in C^{1,\beta} \quad \text{for any } 0<\beta<1. \tag{3.9}$$

LEMMA 3.1. *There exists a unique solution of* (3.5)–(3.8) *with* $\phi^m\in C^{1,\beta}(\overline{\Omega})$ *for any* $0<\beta<1$.

Proof. Proceeding by induction on m, it suffices to show that the elliptic problem

$$\Delta\psi=0 \quad \text{in } \Omega, \tag{3.10}$$

$$\psi=1 \quad \text{on } S, \tag{3.11}$$

$$(\gamma+\delta(\psi_n-\varepsilon)^+)\psi_n=\psi \quad \text{on } \Gamma \tag{3.12}$$

has a unique solution ψ in $C^{1,\beta}(\overline{\Omega})$, provided $\gamma\geq\sigma$, $\gamma\in C^{\beta}(\overline{\Omega})$.

The function $f(t)=(\gamma+\delta(t-\varepsilon)^+)t$ satisfies

$$f'(t)=\begin{cases}\gamma & \text{if } t<\varepsilon,\\ \gamma+\delta(t-\varepsilon)+\delta t>\gamma & \text{if } t>\varepsilon.\end{cases}$$

Therefore (3.12) can be rewritten in the form

$$\psi_n=g(\gamma,\psi) \tag{3.13}$$

with $g(s,t)$ piecewise in C^1 and $g_t>0$. The solution of (3.10)–(3.12) must a priori satisfy $0\le\psi\le1$ (by the maximum principle). Hence we may truncate $g(s,t)$ for $t<0$ and $t>1$ so that, for the new function g, $g(\gamma(x),t)$ is uniformly bounded.

Let

$$X_M=\{v\in C^\beta(\bar\Omega),\ \|v\|_{C^\beta(\bar\Omega)}\le M\},\qquad M>0.$$

For any $\bar\psi\in X_M$ we solve (3.10), (3.11) with

$$\psi_n=g(\gamma,\bar\psi)\quad\text{on }\Gamma.$$

Since $|g(\gamma,\bar\psi)|\le C$, we deduce that

$$\|\psi\|_{C^\gamma(\bar\Omega)}\le M_0\quad\text{for any }\beta\le\gamma<\beta'$$

where β' is any number in $(\beta,1)$ and M_0 is a constant depending only on C and β'. Choosing $M=M_0$, we see that the mapping T, defined by $\psi=T\bar\psi$, maps X_M into itself. It is easily seen that T is continuous and its range lies in a compact subset of X_M. Hence, by Schauder's fixed point theorem, T has a fixed point ψ, which is clearly a solution of (3.10)–(3.12).

To prove uniqueness, we take the difference of two solutions and apply to it the maximum principle.

Having established Lemma 3.1, we note that since $h^i\ge h^{i-1}$, the proof of Lemma 2.2 gives

$$\phi^i\ge\phi^{i-1}\quad\text{in }\Omega. \tag{3.14}$$

Recall also that

$$0\le\phi^i\le1\quad\text{in }\Omega, \tag{3.15}$$

by the maximum principle.

For any $x_0\in\Gamma$ denote by $j_0=j(x_0)$ the first integer j_0 (if existing) such that

$$\phi_n^i(x_0)\le\varepsilon\quad\text{if } i<j_0-1,\quad \phi_n^{j_0}(x_0)>\varepsilon. \tag{3.16}$$

Notice that $h^i(x_0)=\sigma$ if $i<j_0$ and $h^{j_0}(x_0)>h^{j_0-1}(x_0)$.

LEMMA 3.2. (i) *$h^i(x_0)$ is strictly increasing with i for all $i\ge j_0$ (and consequently $\phi_n^i(x_0)>\varepsilon$ for all $i\ge j_0$)* (ii) *$\phi^i(x)$ is strictly increasing in i for all $x\in\Omega$ and for all $i\ge j_*$ where $j_*=\min_{x_0\in\Gamma}j(x_0)$.*

Proof. Suppose the first assertion is not true. Then there exists a smallest integer m, $m\ge j_0$, such that

$$h^m(x_0)>h^{m-1}(x_0), \tag{3.17}$$

$$h^{m+1}(x_0)=h^m(x_0). \tag{3.18}$$

Then

$$\phi_n^m(x_0)>\varepsilon,\qquad \phi_n^{m+1}(x_0)\le\varepsilon. \tag{3.19}$$

Since

$$(h^{m+1}-h^m)\phi_n^{m+1}+h^m(\phi_n^{m+1}-\phi_n^m)=\phi^{m+1}-\phi^m,$$

we deduce from (3.18) and (3.19) that, at x_0, $\phi^{m+1}-\phi^m<0$, which contradicts (3.14). This completes the proof of (i). The assertion (ii) follows from (i) and the proof of Lemma 2.2.

Remark 3.1. Lemma 3.2 (i) reflects the fact (established in a different setting in Theorem 2.1 and Corollary 2.6) that once h starts strictly growing in t, it continues to strictly grow for all subsequent times.

LEMMA 3.3. *There holds*:

$$\sigma\le h^m(x)\le\frac{1}{\varepsilon}\quad \textit{for all } x\in\Gamma,\quad m\ge 1. \tag{3.20}$$

Proof. Suppose $h^m(x_0)>1/\varepsilon$. Since $\sigma<1/\varepsilon$, it follows that $\phi_n^i(x_0)>\varepsilon$ for some $i\le m$ and, by Lemma 3.2, also for $i=m$. From (3.7) we thus get $h^m(x_0)\le\phi^m/\varepsilon\le 1/\varepsilon$, a contradiction.

LEMMA 3.4. *There exist* $\alpha\in(0,1)$ *and* $C>0$ *such that for any* $0<\sigma<1/\varepsilon$, $0<\delta<1$, $m\ge1$,

$$\|\phi^m\|_{C^\alpha(\bar\Omega)}\le C. \tag{3.21}$$

Proof. Let ζ be a C^∞ function with support in the interior of the domain bounded by S. Multiplying the equation $\Delta\phi^i=0$ by $\zeta^2(\phi^i-k)^+$ ($k>0$) and integrating over Ω, we get

$$\int_\Omega\nabla\phi^i\cdot\nabla\left(\zeta^2(\phi^i-k)^+\right)=\int_\Gamma\zeta^2\frac{\partial\phi^i}{\partial\nu}(\phi^i-k)^+=-\int_\Gamma\zeta^2\frac{\phi^i}{h^i}(\phi^i-k)^+\le 0.$$

Now we can proceed as in [5, pp. 196–198] to deduce the C^α estimate on ϕ^i, independently of σ, δ, i.

LEMMA 3.5. *There exists a constant C independent of* σ, m (*but possibly depending on* δ) *such that for all* $0<\sigma<1/\varepsilon$, $m\ge1$,

$$0\le\frac{\partial\phi^m}{\partial n}\le C\quad \textit{on } \Gamma. \tag{3.22}$$

Proof. It suffices to consider points where $\partial\phi^m/\partial n>2\varepsilon$. At such points $h^m\ge\delta(\phi_n^m-\varepsilon)>\delta\varepsilon$, and since $h^m\phi_n^m=\phi^m\le1$, we get $\phi_n^m\le1/(\delta\varepsilon)$.

4. The time-discretized system. In this section we consider the time-discretized version of (3.1)–(3.4), namely,

$$\Delta\phi^m=0\quad \text{in } \Omega, \tag{4.1}$$

$$\phi^m=1\quad \text{on } S, \tag{4.2}$$

$$h^m\frac{\partial\phi^m}{\partial n}=\phi^m\quad \text{on } \Gamma, \tag{4.3}$$

with

$$h^m=\delta\sum_{i=0}^m\left(\frac{\partial\phi^i}{\partial n}-\varepsilon\right)^+\quad \text{on } \Gamma. \tag{4.4}$$

This system is obtained from (3.5)–(3.8) by taking $\sigma\to0$.

DEFINITION 4.1. A sequence $\{\phi^m(x)\}$ is said to be a solution of (4.1)–(4.4) if:

(i) ϕ^m satisfies (4.1), (4.2) and

$$\|\phi^m\|_{C^\alpha(\bar{\Omega})} \le C_1 < \infty \quad \forall m \ge 0,$$

for some $0 < \alpha < 1$;

(ii) setting $\Gamma_m = \{x \in \Gamma, \phi^m(x) > 0\}$, $\Gamma_m^0 = \Gamma \backslash \Gamma_m$, there holds

$$\phi^i \le \phi^{i+1} \quad \text{in } \Omega, \quad \text{for all } i \ge 0,$$

and, consequently, $\Gamma^i \subset \Gamma^{i+1}$;

(iii) there exist $L^\infty(\Gamma)$ functions ϕ_n^i such that

$$\int_\Omega \Delta\phi \cdot \zeta = -\int_\Gamma \phi_n^i \zeta \quad \forall \zeta \in H^1(\Omega), \quad \zeta = 0 \text{ on } S$$

(thus ϕ_n^i may be interpreted as the generalized normal derivative of ϕ^i);

(iv) $(\phi_n^i - \varepsilon)^+ \in C^\alpha(\Gamma)$ and it vanishes on Γ_i^0; consequently $h^m \in C^\alpha(\Gamma)$, where h^m is defined by (4.4);

(v) $\phi_n^i \in C^\alpha(\Gamma_i)$;

(vi) there holds:

$$h^i \phi_n^i = \phi^i \quad \text{on } \Gamma;$$

(note that on $\Gamma \backslash \Gamma_i$ this relation holds in the sense that $h^i = 0$, $\phi^i = 0$ and ϕ_n^i is an L^∞ function);

(vii) $0 \le h^i(x) \le 1/\varepsilon$.

THEOREM 4.1. *There exists a unique solution of* (4.1)–(4.4).

Proof. Denote by $\phi^{m,\sigma}$ the solution of (3.5)–(3.8), where $0 < \sigma < 1/\varepsilon$. By Lemma 3.5

$$0 \le \frac{\partial}{\partial n}\phi^{m,\sigma} \le C \quad \text{on } \Gamma. \tag{4.5}$$

By elliptic estimates, $|\nabla \phi^{m,\sigma}| \le C$ on S. Hence

$$\int_\Omega \nabla\phi^{m,\sigma} \cdot \nabla\phi^{m,\sigma} = -\int_{\Gamma \cup S} \phi_n^{m,\sigma}\phi^{m,\sigma} \le C.$$

From this and from Lemma 3.4 it follows that, for a sequence $\sigma \to 0$,

(4.6) $\phi^{m,\sigma} \to \phi^m$ weakly in $H^1(\Omega)$ and strongly in $C^\alpha(\bar{\Omega})$, for any $m \ge 0$

where α is independent of m.

In view of (4.5) we may also assume that

$$\phi_n^{m,\sigma} \to \psi^m \quad L^\infty(\Gamma)\text{-weakly star.} \tag{4.7}$$

Since

$$\int_\Omega \nabla\phi^{m,\sigma} \cdot \nabla\zeta = -\int_\Gamma \phi_n^{m,\sigma}\zeta \quad \forall \zeta \in H^1(\Omega), \quad \zeta = 0 \text{ on } S,$$

the assertion (iii) in Definition 4.1 follows. Clearly ϕ^m satisfies (4.1), (4.2), and the C^α estimate in (i) follows (with C_1 independent of m, σ) as in the proof of Lemma 3.4, since the inequality

$$\int_\Omega \nabla\phi \cdot \nabla(\zeta^2(\phi-k)^+) \le 0$$

holds for $\phi=\phi^{m,\sigma}$ and, by taking $\sigma\to 0$, also for $\phi=\phi^m$. Clearly (ii) also holds.

To prove (iv)–(vii) we begin with $i=1$. We have

$$(4.8)\qquad \left(\sigma+\left(\phi_n^{1,\sigma}-\varepsilon\right)^+\right)\phi_n^{1,\sigma}=\phi^{1,\sigma} \quad \text{on } \Gamma.$$

On Γ_1 ϕ^1 is positive and hence $(\phi_n^{1,\sigma}-\varepsilon)^+$ remains positive as $\sigma\to 0$. We can therefore rewrite (4.8), in every compact subset $\tilde{\Gamma}_1$ of Γ_1, in the form

$$\left(\sigma+\left(\phi_n^{1,\sigma}-\varepsilon\right)\right)\phi_n^{1,\sigma}=\phi^{1,\sigma}$$

for σ small enough. We then solve for $\phi_n^{1,\sigma}$: $\phi_n^{1,\sigma}=f_\sigma(\phi^{1,\sigma})$ and thus deduce that $\phi_n^{1,\sigma}\to\psi_n^1$ $(\equiv\phi_n^1)$ uniformly in every compact subset $\tilde{\Gamma}_1$ of Γ_1, with ψ_n^1 in $C^\alpha(\tilde{\Gamma}_1)$, i.e.,

$$(4.9)\qquad \phi_n^1\in C^\alpha(\Gamma_1);$$

this is the assertion (v).

Let $x_j\in\Gamma_1$, $x_j\to x\in\partial\Gamma_1$. Then

$$\left(\phi_n^1-\varepsilon\right)^+(x_j)=\lim_{\sigma\to 0}\left(\phi_n^{1,\sigma}(x_j)-\varepsilon\right)^+\le\frac{1}{\varepsilon}\lim_{\sigma\to 0}\left(\phi_n^{1,\sigma}(x_j)-\varepsilon\right)\phi_n^{1,\sigma}(x_j)$$
$$=\frac{1}{\varepsilon}\lim_{\sigma\to 0}\phi^{1,\sigma}(x_j)=\frac{1}{\varepsilon}\phi^1(x_j)\le C|x_j-x|^\alpha$$

since $\phi^1\in C^\alpha(\Gamma)$ and $\phi^1(x)=0$. From this and from (4.9) it follows that

$$(4.10)\qquad \left(\phi_n^1-\varepsilon\right)^+\in C^\alpha(\bar{\Gamma}_1) \quad \text{with } \left(\phi_n^1-\varepsilon\right)^+=0 \quad \text{on } \partial\Gamma_1.$$

On $\Gamma\backslash\Gamma_1$

$$(4.11)\qquad \phi_n^{1,\sigma}\to\phi_n^1 \quad L^\infty\text{-weakly star}$$

and

$$(4.12)\qquad \left(\phi_n^{1,\sigma}-\varepsilon\right)^+\phi_n^{1,\sigma}=\phi^{1,\sigma}\to 0 \quad \text{pointwise.}$$

It follows that, pointwise, either $(\phi_n^{1,\sigma}(x)-\varepsilon)^+\to 0$ or $\phi_n^{1,\sigma}(x)\to 0$. In both cases we have that $(\phi_n^{1,\sigma}(x)-\varepsilon)^+\to 0$ pointwise in $\Gamma\backslash\Gamma_1$. Hence, for any $\eta>0$,

$$\text{meas}\{x\in\Gamma\backslash\Gamma_1, \phi_n^{1,\sigma}(x)>\varepsilon+\eta\}\to 0$$

and together with (4.11) we see that

$$\phi_n^1\le\varepsilon+\eta \quad \text{a.e. on } \Gamma\backslash\bar{\Gamma}_1.$$

It follows that $(\phi_n^1-\varepsilon)^+=0$ on $\Gamma\backslash\bar{\Gamma}_1$. Recalling (4.10) we see that

$$(4.13)\qquad \left(\phi_n^1-\varepsilon\right)^+\in C^\alpha(\Gamma) \quad \text{with } \left(\phi_n^1-\varepsilon\right)^+=0 \text{ on } \Gamma\backslash\Gamma_1,$$

which is the assertion (iv).

Next on Γ_1 ϕ_n^1 was defined by taking $\sigma\to 0$ in (4.8). This means that

$$\left(\phi_n^1-\varepsilon\right)^+\phi_n^1=\phi_n,$$

or $h\phi_n^1=\phi_n$ on Γ_1. Since $h=0$ and $\phi_n^1=0$ on $\Gamma\backslash\Gamma_1$, the same relation $h\phi_n^1=\phi_n$ holds a.e. on $\Gamma\backslash\Gamma_1$, thereby proving (vi).

Since $(\phi_n^{1,\sigma}-\varepsilon)^+\to(\phi_n^1-\varepsilon)^+$, say weakly in $L^2(\Gamma)$, the assertion (vii) follows from Lemma 3.3. We have thus proved (iv)–(vii) for $i=1$. Since $(\phi_n^1-\varepsilon)^+$ is in $C^\alpha(\Gamma)$ and $\Gamma_2\supset\Gamma_1$, the above proof extends to the case $i=2$ and similarly we can extend it for all i.

We shall now prove uniqueness (which in particular implies that the full family $\{\phi^{i,\sigma}\}$ is convergent to $\{\phi^i\}$ as $\sigma\to 0$).

Suppose $\{\psi^i\}$ is another solution and set $\Delta_i=\{x\in\Gamma,\psi^i(x)>0\}$. We proceed by induction. Assuming that $\phi^j\equiv\psi^j$ for all $1\le j<i$, we shall prove that $\phi^i\equiv\psi^i$ (the proof works also for $i=1$).

Set

$$b=\sum_{j=1}^{i-1}\left(\phi_n^i-\varepsilon\right)^+.$$

Then, by (iii),

$$-\int_\Omega\left|\nabla(\phi^i-\psi^i)\right|^2=\int_\Gamma\left(\phi_n^i-\psi_n^i\right)\left(\phi^i-\psi^i\right). \tag{4.14}$$

Set

$$H=\left(\phi^i-\psi^i\right)\left(\phi_n^i-\psi_n^i\right).$$

On Γ_i

$$\phi^i=\left(b+\left(\phi_n^i-\varepsilon\right)^+\right)\phi_n^i$$

and on Δ_i

$$\psi^i=\left(b+\left(\psi_n^i-\varepsilon\right)^+\right)\psi_n^i.$$

Hence on $\Gamma_i\cap\Delta_i$

$$H=\left[\left(b+\left(\phi_n^i-\varepsilon\right)^+\right)\phi_n^i-\left(b+\left(\psi_n^i-\varepsilon\right)^+\right)\psi_n^i\right]\left[\phi_n^i-\psi_n^i\right]\ge 0.$$

On $\Delta_i\backslash\Gamma_i$ $\phi^i=0$, $\phi_n^i\le\varepsilon$ and $\psi_n^i>\varepsilon$. Hence

$$H=\left(\phi^i-\psi^i\right)\left(\phi_n^i-\psi_n^i\right)=-\psi^i\left(\phi_n^i-\psi_n^i\right)\ge 0.$$

Similarly, on $\Gamma_i\backslash\Delta_i$

$$H=\phi^i\left(\phi_n^i-\psi_n^i\right)\ge 0.$$

Finally on $\Gamma\backslash(\Delta_i\cup\Gamma_i)$ $\phi^i=\psi^i=0$ so that $H=0$. It follows that $H\ge 0$ on Γ and (4.14) then yields

$$-\int_\Omega\left|\nabla(\phi^i-\psi^i)\right|^2\ge 0,$$

which gives $\phi^i\equiv\psi^i$.

From the above proof of existence we easily deduce:

COROLLARY 4.2. *The solution* $\{\phi^i\}$ *satisfies*: $0 \le \phi_n^i \le C_1$ *on* Γ *and*

$$\int_\Omega |\nabla\phi^i|^2 \le C$$

where C_1, C *are positive constants independent of* i.

Remark 4.1. The assertions of Lemma 3.2 are valid also for the solution of (4.1)–(4.4).

5. Convergence to the stationary solution.

THEOREM 5.1. *The solution* $\{\phi^i\}$ *of* (4.1)–(4.4) *satisfies*: $\phi^i \to W$ *as* $i \to \infty$, *uniformly in* Ω, *where* W *is the solution of the Signorini problem* (2.3).

Proof. From the properties of the solution ϕ^i and Corollary 4.2 it follows that

$$(5.1) \qquad \phi^m \nearrow \bar\phi \quad \text{in } C^\alpha(\bar\Omega) \text{ (for some } \alpha > 0\text{)},$$

$$(5.2) \qquad \phi^m \to \bar\phi \quad \text{weakly in } H^1(\Omega),$$

for some function $\bar\phi$. For any $\zeta \in K$ (K as in (2.2)) we have

$$\int_\Omega \nabla\phi^m \cdot \nabla(\zeta - \phi^m) + \varepsilon\int_\Gamma (\zeta - \phi^m)$$
$$= -\int_\Gamma (\phi_n^m - \varepsilon)(\zeta - \phi^m)$$
$$= -\int_{\Gamma_m} (\phi_n^m - \varepsilon)(\zeta - \phi^m) - \int_{\Gamma\setminus\Gamma_m} (\phi_n^m - \varepsilon)(\zeta - \phi^m).$$

On $\Gamma\setminus\Gamma_m$ we have $\phi^m = 0$ and $\phi_n^m \le \varepsilon$. Hence

$$-\int_{\Gamma\setminus\Gamma_m} (\phi_n^m - \varepsilon)(\zeta - \phi^m) = -\int_{\Gamma\setminus\Gamma_m} (\phi_n^m - \varepsilon)\zeta \ge 0.$$

It follows that

$$(5.3) \qquad \int_\Omega \nabla\phi^m \cdot (\nabla\zeta - \phi^m) + \varepsilon\int_\Gamma (\zeta - \phi^m) \ge -\int_{\Gamma_m} (\phi_n^m - \varepsilon)(\zeta - \phi^m).$$

Recall that $\Gamma_m \uparrow$ if $m \uparrow$ and set $\Gamma_0 = \lim_{m\to\infty} \Gamma_m$. For any $\eta > 0$ there exists a measurable set $E \subset \Gamma_0$ such that $\operatorname{meas}(\Gamma_0\setminus E) < \eta$ and $\Gamma_m \supset E$ if m is sufficiently large, say $m \ge m_0$. Since $h^i(x) \le 1/\varepsilon$, we have

$$(5.4) \qquad \sum_{m=k}^{\infty} (\phi_n^m(x) - \varepsilon)^+ \le \frac{1}{\delta\varepsilon} \quad \text{if } x \in E, \quad k \ge m_0.$$

Hence for any $k \ge m_0$ and for any positive integer M,

$$(5.5) \qquad \sum_{m=k}^{k+M-1} \int_{\Gamma_m} |\phi_n^m - \varepsilon| = \sum_{m=k}^{k+M-1} \int_{\Gamma_m} (\phi_n^m - \varepsilon)^+ = \sum_{m=k}^{k+M-1} \left(\int_E + \int_{\Gamma_m\setminus E} \right)$$
$$\le \frac{1}{\delta\varepsilon} + C_1 M\eta,$$

with C_1 as in Corollary 4.2.

Summing in (5.3) over m, $k \leq m \leq k+M-1$ and using (5.5), we obtain (since $|\zeta-\phi^m| \leq C$)

$$\frac{1}{M}\sum_{k=m}^{k+M-1}\left[\int_\Omega \nabla\phi^m\cdot\nabla(\zeta-\phi^m)+\varepsilon\int_\Gamma(\zeta-\phi^m)\right]+\frac{C}{\delta\varepsilon M}+C\eta\geq 0.$$

Taking $m \to \infty$ and using (5.1), (5.2), we obtain

$$\int_\Omega \nabla\bar{\phi}\cdot\nabla(\zeta-\bar{\phi})+\varepsilon\int_\Gamma(\zeta-\bar{\phi})+\frac{C}{\delta\varepsilon M}+C\eta\geq 0.$$

Since η and $1/M$ can be taken arbitrarily small, it follows that $\bar{\phi}$ is the solution W of problem (2.3).

COROLLARY 5.2. *The limiting thickness* $h(x)=\lim_{i\to\infty}h^i(x)$ *of the paint coat is determined by*

$$h(x)=\begin{cases}\dfrac{W(x)}{W_n(x)} & \text{if } W(x)>0,\\ 0 & \text{if } W(x)=0,\end{cases}$$

and $0 \leq h(x) \leq 1/\varepsilon$.

REFERENCES

[1] J. M. AITCHISON, A. A. LACY AND M. SHILLOR, *A model for an electropaint process*, IMA J. Appl. Math., to appear.

[2] L. A. CAFFARELLI, *Further regularity for the Signorini problem*, Comm. PDE, 4 (1979), pp. 1067–1076.

[3] J. FREHSE, *On the Signorini problem and variational problems with thin obstacles*, Ann. Scu. Norm. Sup. Pisa, 4 (4) (1977), pp. 343–362.

[4] A. FRIEDMAN, *Variational Principles and Free Boundary Problems*, John Wiley, New York, 1982.

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

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

A FREE-BOUNDARY PROBLEM ARISING FROM A GALVANIZING PROCESS*

THOMAS I. VOGEL[†]

Abstract. A free-boundary problem which arises from a galvanizing process is studied. The physical problem is that of an infinite cylinder $\Omega' \times \mathbb{R}$ withdrawn from a fluid bath. Formally, this is a gravity-driven unidirectional viscous fluid flow on the exterior of the cylinder $\Omega' \times \mathbb{R}$. The mathematical problem is to find a function u with compact support in the exterior of Ω' satisfying:

$$\begin{aligned} \Delta u &= \chi_{\{u>0\}} && \text{in } \mathbb{R}^n - \Omega', \\ u &= c && \text{on } \partial\Omega' \end{aligned}$$

where χ_U is the characteristic function of U. The existence of a unique classical solution is shown under certain conditions on Ω', and asymptotic results for the thickness of the coat are obtained for large and small withdrawal speeds. If Ω' is a convex set, then the region bounded by the free surface of the fluid is shown to be convex, using level curve techniques. Finally, level curve techniques are used to bound the curvature of the free boundary in terms of that of the fixed boundary.

1. Introduction. Many industrial processes involve applying a thin coat of liquid to some material. To coat an infinitely long cylinder, a common method is to pull it out of a liquid bath so that the gravity vector points parallel to the generators of the cylinder. This is typically used for galvanizing, where the cylinder (not necessarily circular) is a wire or sheet of steel, and the liquid is molten zinc. As the cylinder moves up, it carries with it a coat of liquid, which gradually solidifies. Over a substantial length of the cylinder, the flow of the liquid is steady and straight down, with the outer boundary of the region of flow a free surface. The driving forces are gravity and viscosity.

Tuck, Bentwich and van der Hoek [7] (hereafter referred to as TBH) have recently given a formulation of this problem. Let $\Omega' \subset \mathbb{R}^2$ be the cross section of the cylinder, and let Σ be its boundary. Let $\Omega \subset \mathbb{R}^2$ be the cross section of the region of flow plus the cylinder, and let Γ be the boundary of Ω (which is free). The region of flow is exterior to the given Ω'. Then they show that under certain assumptions, the upward velocity $w(x,y)$ must satisfy

$$\begin{aligned} \Delta w &= \frac{g}{\nu} && \text{in } \Omega - \bar{\Omega}', \\ w &= W_B && \text{on } \Sigma, \\ w &= \frac{W_B}{2} && \text{on } \Gamma, \\ \frac{\partial w}{\partial n} &= 0 && \text{on } \Gamma. \end{aligned}$$

Here g is the downward acceleration due to gravity, ν is the kinematic viscosity of the liquid, and W_B is the withdrawal speed of the cylinder. It is important to keep in mind that Ω' is given, but Ω is not. The free boundary Γ is determined by the condition that the above system has a solution. The fact that we impose both Dirichlet and Neumann

*Received by the editors July 19, 1983, and in revised form May 24, 1984. This research was sponsored by the U. S. Army under contract DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under grant MCS-7927062, Mod. 2.

[†] Department of Mathematics, Texas A & M University, College Station, Texas 77843.

boundary conditions on Γ will prevent a solution w from existing for a general Ω. This situation is typical of free boundary problems, where the fact that the boundary conditions are overdetermined is compensated for by the freeness of the boundary.

The model of TBH neglects surface tension and assumes that the net rate of transport

$$Q=\iint_{\Omega-\bar{\Omega}'} w(x,y)\,dx\,dy$$

is maximized. More precisely, they show that if $w(x,y;\,U)$ satisfies

$$\begin{aligned} \Delta w &= \frac{g}{\nu} && \text{in } \tilde{U}-\bar{\Omega}', \\ w &= W_B && \text{on } \Sigma, \\ \frac{\partial w}{\partial n} &= 0 && \text{on } \partial U, \end{aligned}$$

then if $Q(U)=\iint_{U-\bar{\Omega}'} w(x,y;\,U)\,dx\,dy$ is maximized over all admissible U, the maximum will be obtained when $w(x,y;\,U)\equiv W_B/2$ on $\partial\Omega$.

In this paper, we will work with the normalization

$$u=\nu\left(w-\frac{W_B}{2}\right)\Big/g \quad \text{and} \quad c=\frac{\nu W_B}{2g},$$

so that the equations become

$$\begin{aligned} \Delta u &= 1 && \text{in } \Omega-\bar{\Omega}', \\ u &= c && \text{on } \Sigma, \\ u &= 0 && \text{on } \Gamma, \\ \frac{\partial u}{\partial n} &= 0 && \text{on } \Gamma. \end{aligned} \tag{1.1}$$

This is an example of an obstacle problem (see Friedman [2, Chap. 1]). Notice that the last condition is equivalent to $|\nabla u|=0$ on Γ, since Γ is a level surface of u. If the dependence on c is to be emphasized, we will write Γ_c and u_c.

The main purpose of this paper is to determine as much about the shape of the free boundary as possible. In §2, we show the existence of a classical solution of (1.1) in n dimensions if Ω' is starshaped with respect to a ball. That section consists mainly of arguments in TBH placed on firmer theoretical grounds. In §3, we obtain asymptotic results in two dimensions for c large and c small, and also some useful comparison results. In particular, as c tends to infinity, the free surfaces Γ_c tend to circles of radius $2\sqrt{c/\log 2c}$. In §4, we prove the convexity of the set $\{u>t\}$ for $c>t>0$ if Ω' is convex, and in §5, we show that for $n=2$, if Ω' is convex, then each point of the ridge of Ω is closer to Σ than to Γ.

2. Existence and regularity. Let $\Omega'\subset\mathbb{R}^n$ be a bounded open set which is starshaped with respect to all points contained in some ball, and let c be a positive constant. Suppose that $\Sigma=\partial\Omega'$ is sufficiently smooth. We will prove in this section that then there exists a set Ω containing Ω' with $\Gamma=\partial\Omega$ analytic and a nonnegative function u

which satisfies (1.1). This will be done variationally by using the functional

$$J_R(v)=\int_{B_R}|\nabla v|^2+2v, \tag{2.1}$$

where B_R is an open ball containing $\overline{\Omega}'$ of some sufficiently large radius R centered in Ω'. J_R will be minimized over the set $K_{c,R}=\{v\in L^1(\mathbb{R}^n),\ \nabla v\in L^2(\mathbb{R}^n),\ v=c \text{ in } \Omega',\ v=0 \text{ in } \mathbb{R}^n-B_R,\ v\geqq 0 \text{ everywhere}\}$.

THEOREM 2.1. *If Ω' is a bounded set in $\mathbb{R}^n$ and Σ $(=\partial\Omega')$ is in $C^{2+\alpha}$, then there exists a unique $u\in K_{c,R}$ such that*

$$J_R(u)=\inf_{v\in K_{c,R}} J_R(v).$$

Moreover, $u\in W^{2,p}(B_R-\overline{\Omega}')\cap W^{2,\infty}_{\mathrm{loc}}(B_R-\overline{\Omega}')$ for all $p<\infty$, where $W^{2,p}(B-\overline{\Omega}')=\{v\in L^p(B_R-\overline{\Omega}'),\ \nabla v\in L^2(B_R-\overline{\Omega}')\}$. As a consequence, u is C^1 in $B_R-\overline{\Omega}'$ (see Gilbarg and Trudinger [5]). *Moreover, u is analytic in $\Omega-\overline{\Omega}'$ and $\Delta u=1$ there, where $\Omega=\{u>0\}$.*

Proof. This follows from standard theorems (Friedman [2, §§1.3 and 1.4]).

As is well known, once we have proven this much smoothness, u must satisfy (1.1), where $\Gamma=\partial\Omega\cap B_R$. In §3, we will see that for R_0 sufficiently large, $\Omega\cap\partial B_{R_0}=\varnothing$. It is clear that for any R_1, $R_2>R_0$, the minimizers u_{R_1} and u_{R_2} will be identical. We will assume from now on that R is larger than this R_0, thus eliminating the dependence of u on R.

Note. The minimizer obtained in Theorem 2.1 satisfies $0\leqq u\leqq c$.

Proof. One easily checks that

$$J_R(u\wedge c)\leqq J_R(u),$$

where $u\wedge c=\min(u(X),c)$. Moreover, $u\wedge c\in K_{c,R}$, so that the uniqueness part of Theorem 2.1 applies.

DEFINITION 2.1. A region U is *almost starlike* with respect to a point $P\in U$ if the characteristic function χ_U is nonincreasing along rays from P. The difference between an almost starlike region and a *starlike* region is that an almost starlike region may contain a portion of a ray through P in its boundary.

LEMMA 2.2. *If $\partial\Omega'$ is $C^{2+\alpha}$ and Ω' is almost starlike with respect to the origin, then $\partial u/\partial r<0$ in $\Omega-\overline{\Omega}'$ and Ω is almost starlike with respect to the origin. (Here $r=\sqrt{x_1^2+\cdots+x_n^2}$.)*

Proof. This is proven in TBH [7] for $n=2$ by showing that $r\partial u/\partial r$ is subharmonic in $\Omega-\overline{\Omega}'$ with nonpositive boundary values. The proof is the same in n dimensions. The almost starlikeness of Ω follows, since u and hence χ_Ω is nonincreasing along rays.

THEOREM 2.3. *If Ω' is starlike with respect to each point contained in a ball B_ε, then Γ is analytic, and u satisfies*

$$\begin{aligned} \Delta u&=1 && \text{in } \Omega-\Omega',\\ u&=c && \text{on } \Sigma,\\ u&=0 && \text{on } \Gamma,\\ \frac{\partial u}{\partial n}&=0 && \text{on } \Gamma. \end{aligned}$$

Σ need not be $C^{2+\alpha}$, although it is clearly Lipschitz continuous.

Proof. First we assume that $\partial\Omega'$ is in $C^{2+\alpha}$. Since Γ is almost starlike with respect to each point in B_ε, it is therefore Lipschitz continuous. This is enough to apply a theorem of Caffarelli (Friedman [2, p. 162]) to show that Γ is C^1 and hence analytic. Once we have the smoothness of Γ, the boundary conditions on u will necessarily be satisfied.

If Σ is not $C^{2+\alpha}$, then we may approximate Ω' by an increasing nested series of sets Ω_i' with $\partial\Omega_i'$ smooth and Ω_i' starlike with respect to each point in B_ε. By standard arguments, the free boundaries Γ_i increase out to Γ, the boundary of $\{u>0\}$. Since each Γ_i is starlike with respect to B_ε, so is Γ. But then we apply the same argument as before to say that Γ is analytic and u satisfies the correct boundary conditions on Γ.

THEOREM 2.4 (uniqueness). *If Ω' is starlike, then there exists at most one solution (u,Γ) to* (1.1).

Proof. Suppose there are two solutions, (u,Γ) and (u^*,Γ^*), to (1.1). We assume that Ω' is starlike with respect to the origin. Define

$$u_r=u(rX),\quad \Omega_r=\frac{1}{r}\Omega,\quad \Omega_r'=\frac{1}{r}\Omega',\quad \Gamma_r=\frac{1}{r}\Gamma,\quad \Sigma_r=\frac{1}{r}\Omega.$$

Since Ω' is starlike, $\Omega'\subset\Omega_r'$ for $r<1$. Let $s=\sup\{r|(\Omega^*)'\subset\Omega_r'\}$. We may assume, without loss of generality, that $s\leqq 1$, for we could exchange the role of u and u^* if this were not so. Then Γ_s and Γ^* are tangent at some point Y.

The boundary of $\Omega^*\subset(\Omega_s-\overline{\Omega}_s')$ consists of $\Gamma^*\cap(\Omega_s-\overline{\Omega}_s')$ and Σ_s. On both of these surfaces, $u_s\geqq u^*$, so that $u_s\geqq u^*$ everywhere in $\Omega^*\cap(\Omega_s-\overline{\Omega}_s')$. However, at Y we have $\partial u_s/\partial n=\partial u^*/\partial n=0$, so that $u_s=u^*$ on $\Omega^*\cap(\Omega_s-\overline{\Omega}_s')$ by the strong maximum principle. Hence $s=1$ and we have the desired uniqueness.

Combining Theorems 2.3 and 2.4, we see that there exists a unique solution to (1.1) if Ω' is starlike with respect to all the points in some B_ε. If Ω' is starlike with respect to only one point, then there is at most one solution to (1.1).

3. Asymptotic results. The following comparison lemma is basic to our work, and is needed in the proof of Theorem 2.3.

LEMMA 3.1. *Let Ω' and Ω_1' satisfy the hypotheses of Theorem* 2.1 *with $\Omega_1'\subset\Omega'$, and let $c_1\leqq c$. Let u and u_1 minimize J_R with boundary values c on Σ and c_1 on Σ_1 respectively. Then $u\geqq u_1$ everywhere.*

Proof. We have

$$J_R(u\wedge u_1)+J_R(u\vee u_1)=J_R(u)+J_R(u_1),$$

by a simple computation (here $u\wedge u_1=\min(u,u_1)$, and $u\vee u_1=\max(u,u_1)$). But $u\wedge u_1\in(K_1)_{c_1,R}$ and $u\vee u_1\in K_{c,R}$, so the uniqueness result of Theorem 2.1 applies.

Note. We may use the same technique of proof to give an interesting characterization of u. Let u, Ω', Ω_1' be as above, and let v_A minimize

$$J_A(v)=\int_A|\nabla v|^2+2v,$$

where $A\subset B_R$, over the set $\{v\in L^1(A),\ \nabla v\in L^2(A),\ v=c_1 \text{ on } \Sigma_1,\ v=0 \text{ on } \partial A\}$ (here we are not requiring v to be nonnegative). Then $u\geqq v_A$ everywhere. This is proven by the same argument as in the proof of Theorem 3.1.

We may therefore characterize $u(X)$ when Σ is $C^{2+\alpha}$ and Ω' is starlike with respect to a ball as

$$u(X)=\sup_{\substack{A\supset\Omega'\\ \partial A \text{ smooth}}} v_A(X),$$

where $v_A(X)$ solves the Dirichlet problem

$$\begin{aligned} v_A &= c && \text{on } \Sigma, \\ v_A &= 0 && \text{on } \partial A, \\ \Delta v_A &= 1 && \text{in } A - \overline{\Omega}'. \end{aligned}$$

$v_A(X)$ will not in general be nonnegative.

We now deal exclusively with the case $n=2$.

To determine the asymptotic behavior of Γ as $c \to \infty$ and $c \to 0$, it is necessary to look at radial solutions. That is, given ρ, the radius of the circular fixed boundary, and $c>0$, we seek a $v_{\rho,c}(r)$ to solve

$$\begin{aligned} &v''_{\rho,c}(r) + \frac{1}{r} v'_{\rho,c}(r) = 1, \\ &v_{\rho,c}(\rho) = c, \\ &v_{\rho,c}(\gamma) = 0, \\ &v'_{\rho,c}(\gamma) = 0, \end{aligned} \tag{3.1}$$

where γ, to be determined, is the radius of the free boundary. One can calculate that γ is given by the implicit relation

$$c = \frac{\rho^2 - \gamma^2}{4} + \frac{\gamma^2}{2} \log\left(\frac{\gamma}{\rho}\right), \tag{3.2}$$

where $c>0$, $\rho>0$, $\gamma>\rho$, and log is $\log_e$.

LEMMA 3.2. *For $R_0 = R_0(c, \Omega')$ sufficiently large, $\Omega \cap \partial B_{R_0} = \varnothing$, where c is fixed and B_{R_0} is centered in Ω'.*

Proof. Since Ω is contained in some ball B_ρ, the result will follow if it is proven for symmetric solutions. But γ in (3.2) will be bounded if ρ and c are bounded, since the highest order term in γ on the right-hand side, $\gamma^2 \log \gamma$, must be bounded.

This lemma was already used extensively in §2.

From the radial solutions we may investigate the behavior as $c \to \infty$ for a larger class of Ω''s.

THEOREM 3.3. *Let Ω' be a bounded set containing a ball $B_\varepsilon(0)$ around the origin. Then both $d(\Gamma_c, 0) = \inf_{X \in \Gamma_c} |X|$ and $d_1(\Gamma_c, 0) = \sup_{X \in \Gamma_c} |X|$ are equal to*

$$2\sqrt{\frac{c}{\log 2c}} + o\left(\sqrt{\frac{c}{\log 2c}}\right)$$

as $c \to \infty$. (Here Γ is subscripted to emphasize the dependence on c.) Less formally, Γ_c is asymptotic to a circle of radius $2\sqrt{c/\log 2c}$ as c tends to infinity.

Proof. From Lemma 3.1, we know that Γ_c is contained in the annulus centered at the origin with inner radius $\gamma(\varepsilon, c)$ and outer radius $\gamma(p, c)$, where B_ρ contains Ω'. Therefore, we must show that if $\gamma(p, c)$ satisfies (3.2), then as $c \to \infty$,

$$\gamma = 2\sqrt{\frac{c}{\log 2c}} + o\left(\sqrt{\frac{c}{\log 2c}}\right),$$

where dependence on ρ will only appear in the second term.

First, it is clear from (3.1) that γ cannot stay bounded as c tends to infinity, and that $\partial\gamma/\partial c>0$. Dividing by c we obtain

$$1=\frac{\rho^2}{4c}-\frac{\gamma^2}{4c}+\frac{\gamma^2\log\gamma}{2c}-\frac{\gamma^2\log\rho}{2c}.$$

For the largest order term on the right-hand side, we must have

$$\lim_{c\to\infty}\frac{\gamma^2\log\gamma}{2c}=1,$$

and the other terms must go to zero.

If we write

$$\gamma=2f(c)\sqrt{\frac{c}{\log 2c}},$$

then

$$\lim_{c\to\infty}2f^2(c)\left[\frac{\log 2+(1/2)\log c-(1/2)\log(\log 2c)+\log f(c)}{\log 2+\log c}\right]=1. \tag{3.3}$$

We can observe from the above expression that $f(c)$ is bounded. The largest order term on the left of (3.3) is

$$f^2(c)\frac{\log c}{\log c+\log 2}$$

which must approach 1 as $c\to\infty$. The other terms in (3.2) will go to zero. We then conclude that

$$\lim_{c\to\infty}\frac{\gamma(\rho,c)}{2\sqrt{c/\log 2c}}=1$$

so that $\gamma(\rho,c)=2\sqrt{c/\log 2c}+o(\sqrt{c/\log 2c})$, as desired.

We now estimate the thickness of the coat $\Omega-\bar{\Omega}'$ as c tends to zero if $\Sigma\in C^{2+\alpha}$. Fix a point P on Σ, and let ρ and ρ_1 be two radii so that a ball of radius ρ is contained in Ω' and tangent to Σ at P, and a ball of radius ρ_1 is exterior to Ω' and tangent to Σ at P. One choice for ρ is $1/\kappa(\Sigma)$, where $\kappa(\Sigma)$ is the maximum curvature of Σ. If Ω' is convex, then ρ_1 can be chosen to be infinity. From Lemma 3.1, we have

$$d(\Gamma,P)\geqq\gamma(\rho,c)-\rho.$$

For an upper bound on $d(\Gamma,P)$, we must look at interior radially symmetric solutions. That is, for the fixed radius ρ_1, we seek a function $v_{\rho_1,c}$ solving (3.1) for a value $\gamma_1<\rho_1$. The same calculation as before yields that γ_1 solves the implicit relation (3.2). Here, now, we seek a root γ_1 less than ρ_1. We conclude

$$\rho_1-\gamma_1(\rho_1,c)\geqq d(\Gamma,P)\geqq\gamma(\rho,c)-\rho. \tag{3.4}$$

A straightforward calculation using (3.2) yields that

$$\lim_{c\to 0}\frac{c}{(\rho-\gamma)^2}=\lim_{c\to 0}\frac{c}{(\rho_1-\gamma_1)^2}=\frac{1}{2},$$

for $\rho_1\neq+\infty$. If $\rho_1=+\infty$, then the upper bound for the thickness of the coat is $\sqrt{2c}$.

To sharpen the asymptotics in (3.4), we must analyze $\rho-\gamma$ and $\rho_1-\gamma_1$ more closely for small c. One can calculate that

$$\lim_{c\to 0}\frac{1}{\rho-\gamma}\left(\frac{c}{(\rho-\gamma)^2}-\frac{1}{2}\right)=-\frac{1}{6\rho}, \tag{3.5}$$

by substituting (3.2) in for c, and taking the limit as γ approaches ρ.

Now, letting $\rho-\gamma=\sqrt{2c}f(c)$, where $\lim_{c\to 0}f(c)=-1$, and substituting into (3.5) one obtains

$$\lim_{c\to 0}\frac{1+f(c)}{\sqrt{c}}=\frac{1}{3\rho\sqrt{2}},$$

after some manipulation. Therefore,

$$\rho-\gamma=-\sqrt{2c}+\frac{c}{3\rho}+o(c).$$

Similarly, for the interior radially symmetric solution,

$$\rho_1-\gamma_1=\sqrt{2c}+\frac{c}{3\rho_1}+o(c).$$

We have proven the following theorem:

THEOREM 3.4. *Let Ω' be a set with $C^{2+\alpha}$ boundary Σ satisfying the hypothesis of Theorem 2.3. Let P be a point of Σ and let ρ and ρ_1 be the radii of disks tangent to Σ at P which are interior and exterior to Ω', respectively. Then $\sqrt{2c}-c/3\rho+o(c)\leqq d(P,\Gamma)\leqq\sqrt{2c}+c/3\rho_1+o(c)$. If Ω' is convex, then the right-hand side is simply $\sqrt{2c}$.*

Note. This is similar to a result obtained by Friedman and Phillips [3] for an interior free boundary problem for a more general equation.

Remark. If Σ has an angle at P, then the free boundaries for the scaled functions $u_c=(1/c)u(\sqrt{c}X)$ will approach the free boundary corresponding to a wedge as fixed boundary. Thus, to investigate the asymptotic thickness for small c when Σ has corners, one must look at wedge solutions (see TBH [7] for some numerical results).

4. Convexity. In this section we investigate what happens if Ω' is assumed to be convex.

THEOREM 4.1. *If Ω' is convex, then the sets $\{u>r\}$, $r\geqq 0$ are convex, including $\Omega=\{u>0\}$. (This is true in n dimensions.)*

Proof. (This approach was suggested by Daniel Phillips.) Assume first that $\partial\Omega'$ is smooth. From Caffarelli and Spruck [1], we know that if u_p satisfies the free boundary problem

$$\begin{aligned}
\Delta u_p&=u_p^p &&\text{on } \Omega_p-\bar{\Omega}',\\
u_p&=1 &&\text{on } \Sigma,\\
u_p&=0 &&\text{on } \Gamma_p,\\
|\nabla u_p|&=0 &&\text{on } \Gamma_p,
\end{aligned}$$

then Ω_p and all the sets $\{u_p > r\}$, $r > 0$ are convex. We deal with the particular u_p which minimizes

$$J_p(v) = \int_{B_R - \overline{\Omega}} \frac{|\nabla v|^2}{2} + \frac{1}{p+1} v^{p+1}.$$

These functions have been studied by Phillips [6]. It is not difficult to show that the functions u_p are uniformly bounded in $W^{1,2}(B_R)$ as p tends to zero. Therefore a subsequence converges weakly in $W^{1,2}$ to some function u, which must be the unique minimizer to our original functional (2.1). Using the Rellich lemma, $u_p(x) \to u(x)$ pointwise almost everywhere in B_R, by going to another subsequence. (This is a standard technique: see Friedman [2].) Let $A \subset B_R$ be the set on which u_p converges pointwise to u. If the level sets of u are not convex, then there are three colinear points X, Y, Z in B_R with $u(Y) < \min(u(X), u(Z))$, and Y between X and Z. Since $\mu(A) = \mu(B_R)$, where μ is Lebesgue measure, we may assume that X, Y and Z are contained in A. But this, combined with the pointwise convergence of $\{u_p\}$ contradicts the convexity of the level sets of u_p. I now present an independent proof of the convexity of the free surface in 2 dimensions which is more elementary.

LEMMA 4.2. *Let $(x(s), y(s))$, $0 \leqq s \leqq l$ be a parametrization of the free boundary curve Γ. Suppose at $X_0 = (x(s_0), y(s_0))$, $x(s)$ has a local extremum. Then there is a level curve $\{u_y = 0\}$ extending into Ω from X_0.*

Proof. Let $\varepsilon_k \to 0$ be a decreasing sequence such that $\{u = \varepsilon_k\}$ is a C^∞ curve. We have that $\{u = \varepsilon_k\}$ will contain a point X_k near X_0 with a locally extreme x value for small enough ε_k, with $\lim_{k\to\infty} X_k = X_0$. Since $\{u = \varepsilon_k\}$ has a vertical tangent at X_k, $u_y(X_k) = 0$. But u_y is harmonic in $\Omega - \overline{\Omega}'$, so that the properties of its level curves are well known. In particular, the set $\{u_y = 0\}$ must consist of piecewise analytic curves with a finite number of branch points. Therefore, some analytic curve along which $u_y = 0$ must start at X_0 and extend into Ω.

Alternate proof of convexity of Ω for $n = 2$. Suppose that Ω is not convex. Assume first that $\partial\Omega'$ is smooth. We can then rotate the coordinate system so that the x coordinate has a local extremum no Γ for at least four points X_1, X_2, X_3 and x_4. At each point X_i there is a level curve γ_i, on which $u_y = 0$, extending into Ω. Since u_y is identically zero on Γ and u_y is harmonic, it follows that no γ_i can both start and end on Γ, nor can any two γ_i's meet at a branch point or a point of Σ. Since γ_i cannot terminate in the region $\Omega - \overline{\Omega}'$, it follows that these curves must terminate at four distinct points $Y_i \in \Sigma$. However, the normal derivative of u is nonzero on Σ, so that u_y can equal zero only at the two points of Σ where the normal is horizontal, since Σ is smooth. This contradicts the fact that the Y_i are distinct. If Σ is not smooth, we can approximate it from within by smooth sets Ω'_k.

Note. The method of the alternate proof generalizes to elliptic operators with constant coefficients

$$a^{ij}u_{ij} + b^i k u_i + cu = f(u), \qquad i,j = 1,2$$

with $c \leqq 0$ and $f'(u) \geqq 0$, and with the same boundary conditions on Σ and Γ as before.

5. The ridge of Ω. In this section we prove that each point of the ridge (defined later) of Ω must be closer to Ω' than Γ if Ω' is convex. This shows that our intuition is correct: going from Ω' to Ω smooths out corners. This result is then used to bound the curvature of Γ. We first need another level curve lemma.

LEMMA 5.1. *Let $\gamma_b \subset \Omega - \bar{\Omega}'$ be a smooth curve along which $\psi(r,\theta) \equiv r^2/2 - ru_r$ is a constant b. Then u_θ is strictly monotone along γ_b. Specifically, if γ_b is traversed so that $\{\psi > b\}$ lies to the right and $\{\psi < b\}$ lies to the left, then u_θ is strictly increasing. This does not depend on where the origin for polar coordinates is placed.*

Proof. The functions ψ and u_θ are harmonic conjugates, so that this follows from a well-known result. See Friedman and Vogel [4] for a proof.

LEMMA 5.2. *At every point $P \in \Gamma$, there begins at least one level curve of ψ. If P is a local extremum of ψ restricted to Γ (we will write this as $\psi|_\Gamma$), then there are at least two level curves of ψ beginning at P and going into Ω.*

Proof. Since u_θ and ψ_θ are harmonic conjugates, the normal derivative of ψ at P equals the tangential derivative of u_θ at P which is zero (since $u_\theta = 0$ on Γ). By the boundary point lemma, ψ cannot attain a local extremum on Γ, hence every point of Γ is the start of a level curve of ψ.

To prove the second assertion of the lemma, assume that $\psi(P)$ is a strict local minimum of $\psi|_\Gamma$. Since $\psi(P)$ cannot be a local minimum of ψ in any $B_R(P) \cap \Omega$, it follows that there is a region $Q = \{\psi < \psi(P)\}$ which contains P in its closure. But ∂Q contains no points of Γ except for P in some neighborhood of P; so we conclude that there are at least two curves $\{\psi = \psi(P)\}$ beginning at P and going into Ω, as desired.

Now suppose that the origin 0 for polar coordinates is placed outside of Ω'. Introduce the following notation:

$$\Sigma_1 = \{X \in \Sigma | u_r(X) > 0\},$$
$$\Sigma_2 = \{X \in \Sigma | u_r(X) < 0\},$$
$$\Sigma^+ = \{X \in \Sigma | u_\theta(X) > 0\},$$
$$\Sigma^- = \{X \in \Sigma | u_\theta(X) < 0\}.$$

In addition, Σ_1^+, Σ_1^-, etc., are intersections of the appropriate sets above.

LEMMA 5.3. *There is precisely one point on Σ_1 at which $u_\theta = 0$, and this is the closest point of Σ to* 0.

Proof. Suppose $Y \in \Sigma$ satisfies $u_\theta(Y) = 0$, $u_r(Y) > 0$. Then the tangent l to Σ at Y is perpendicular to the line OY, and Ω' lies to one side of l. Since $u_r(Y) > 0$, Ω' lies on the far side of l from 0. It is clear then, that Y is the unique closest point of Σ to 0.

Hence we know that Σ_1 is divided into two segments, Σ_1^+ and Σ_1^-, and a point Σ_1^0, where $u_\theta = 0$.

DEFINITION 5.1. The *ridge* R of Ω is the set of all points $X_0 \in \Omega$ such that $d(X) \equiv \operatorname{dist}(X, \partial\Omega)$ is not in $C^{1,1}(V)$ for any neighborhood V of X.

Let $R_0 = \{X_0 \in \Omega | d(X_0) = |X_0 - Y| = |X_0 - Z|$ for two distinct $Y, Z \in \Gamma\}$, and $R_1 = \{X_0 \in \Omega|$ there exists precisely one point $Y \in \Gamma$ with $d(X_0) = |X_0 - Y|$ and X_0 is the center of the osculating circle at $Y\}$. Then $R = R_0 \cup R_1$ and, since Ω is convex, $R = \bar{R}_0$ (Friedman [2, Chap. 2, §7]).

THEOREM 5.4. *If $X_0 \in R_0$ and Ω' is convex, then* $\operatorname{dist}(X_0, \Gamma) > \operatorname{dist}(X_0, \Sigma)$. *In consequence, if $X_0 \in R$, then* $\operatorname{dist}(X_0, \Gamma) \geq \operatorname{dist}(X_0, \Sigma)$.

Proof. Suppose this is not the case, and let X_0 be the polar origin. Let P_1 and P_2 be points of Γ such that $d(X_0) = |X_0 - P_1| = |X_0 - P_2| = t$. Since $u_r \equiv 0$ on Γ, $\psi|_\Gamma$ has a local minimum at P_1 and P_2. Therefore, from Lemma 5.2, there are level curves γ_1^+, γ_1^- starting at P_1, and γ_2^+, γ_2^- starting at P_2, on which $\psi \equiv t^2/2$. As γ_1^+ and γ_2^+ are traversed in the direction away from Γ, u_θ increases, and as γ_1^- and γ_2^- are traversed in this direction, u_θ decreases. Since by assumption, the distance from X_0 to each point of Σ is greater than t, all of the level curves $\gamma_1^{\pm}$, $\gamma_2^{\pm}$ must terminate on Σ_1. Indeed, γ_1^+ and γ_2^+ must end on Σ_1^+, and γ_1^- and γ_2^- must end on Σ_1^-.

Let $U \subset \Omega - \bar{\Omega}'$ be the open set whose boundary consists of γ_1^+, the segment of Σ_1^+ between the endpoints of γ_1^+ and γ_2^+, γ_2^+, and Γ from P_1 to P_2. Here the condition that $U \subset \Omega - \bar{\Omega}'$ forces the direction that Γ is traversed from P_1 to P_2 for ∂U. Let $U^+ = \{X \in U | \psi(X) > t^2/2\}$ and $U^- = \{X \in U | \psi(X) < t^2/2\}$. Then neither U^+ nor U^- is empty, and either $\gamma_1^+ \subset \partial U^+$ and $\gamma_2^+ \subset \partial U^-$ or vice versa. For if this were not the case, then Lemma 5.1 would be violated, since γ_1 and γ_2 both have a region lying to their right as they are traversed from Γ to Σ, where $\psi > t^2/2$.

We are now led to a contradiction, since there must then be a curve $\gamma^* \subset U$ on which $\psi \equiv t^2/2$ which goes from Σ^+ to Γ to separate U^+ and U^-. Then u_θ will be increasing along γ^* from Σ^+ to Γ, violating the free boundary condition.

As a corollary, we get a rough bound on the curvature of Γ.

COROLLARY 5.5. *Assume that Ω' is convex, and Σ is $C^{2+\alpha}$, and let $\kappa(\Gamma_c)$ be the maximum of the curvature of Γ_c, and $\kappa(\Sigma)$ be the maximum of the curvature of Σ. Then*

$$\frac{2}{\gamma(c, 1/\kappa(\Sigma)) - 1/\kappa(\Sigma)} \geqq \kappa(\Gamma_c).$$

Proof. At each point X of Σ we may place a circle of radius $1/\kappa(\Sigma)$ contained in Ω' and tangent to Σ at X. From the proof of Theorem 3.6, we know that at each point $X \in \Sigma$ the distance from X to Γ is at least $\gamma(c, 1/\kappa(\Sigma)) - 1/\kappa(\Sigma)$. Now consider the ball $B_{1/\kappa(\Gamma)}$ of radius $1/\kappa(\Gamma)$ osculating at the point of greatest curvature of Γ. From Theorem 5.4, $B_{1/\kappa(\Gamma)}$ must contain a point of Σ, hence

$$\frac{2}{\kappa(\Gamma)} \geqq \gamma\left(c, \frac{1}{\kappa(\Sigma)}\right) - \frac{1}{\kappa(\Sigma)},$$

yielding the desired result.

Acknowledgments. I would like to thank Avner Friedman, Ernie Tuck and Daniel Phillips for valuable conversations.

REFERENCES

[1] L. A. CAFFARELLI AND J. SPRUCK, *Convexity properties of solutions to some classical variational problems*, Comm. PDE, 11 (1982), pp. 1337–1379.

[2] A. FRIEDMAN, *Variational Principles and Free Boundary Problems*, John Wiley, New York, 1982.

[3] A. FRIEDMAN AND D. PHILLIPS, *The free boundary of a semilinear elliptic equation*, Trans. Amer. Math. Soc., 282 (1984), pp. 153–182.

[4] A. FRIEDMAN AND T. I. VOGEL, *Cavitational flow in a channel with oscillatory wall*, Nonlinear Anal., 8 (1984), pp. 115–132.

[5] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.

[6] D. PHILLIPS, *A minimization problem and the regularity of solutions in the presence of a free boundary*, Indiana Univ. Math. J., 13 (1983), pp. 1–17.

[7] E. O. TUCK, M. BENTWICH AND J. VAN DER HOEK, *The free-boundary problem for gravity-driven unidirectional flow*, IMA J. Appl. Math., 30 (1983), pp. 191–208.

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

006

A NONLINEAR PSEUDOPARABOLIC DIFFUSION EQUATION*

MICHAEL BÖHM†‡ AND R. E. SHOWALTER†§

Abstract. Diffusion in a fissured medium with absorption or partial saturation effects leads to a pseudoparabolic equation nonlinear in both the enthalpy and the permeability. The corresponding initial-boundary value problem is shown to have a solution in various Sobolev–Besov spaces, and sufficient conditions are given for the problem to be well-posed.

Introduction. This is the second of two papers dealing with a certain pseudoparabolic diffusion equation. It was shown in [6] (see also [2]) that diffusion processes in fissured media lead to the following problem:

Let $G \subset \mathbb{R}^N$ be a bounded domain (the place where the diffusion process takes place), and denote by $S := [0,T]$ a finite time interval. We are looking for functions $u = u(x,t)$ (concentration) and $v = v(x,t)$ (a flow potential), such that

$$\begin{aligned} &u' + \frac{1}{\varepsilon}(\alpha(u) - v) = f_1, \\ &-\operatorname{div}(k(u)\nabla v) + \frac{1}{\varepsilon}(v - \alpha(u)) = f_2, \\ &u(x,0) = u_0(x), \qquad v_{\partial G} = 0. \end{aligned} \tag{0.1}$$

Here $u' := \partial u/\partial t$, "div" denotes the usual divergence operator, "∇" stands for the gradient with respect to $x = (x_1, \cdots, x_N) \in \mathbb{R}^N$, and $t \in S$. The functions $f_i = f_i(x,t)$, $i = 1,2$, and u_0 are given and $k = k(u)$, $\alpha = \alpha(u)$ are specified by properties of the field or medium. For each $u \in L^1(G)$ define $A_u := -\operatorname{div}(k(u)\nabla)$ and consider this elliptic operator subject to Dirichlet boundary conditions. By eliminating v in (0.1) we obtain the following equivalent ordinary differential equation involving only the single variable u

$$\begin{aligned} &u'(t) + \frac{1}{\varepsilon}\left(I - (I + \varepsilon A_u)^{-1}\right)\alpha(u) = f_1 + (I + \varepsilon A_u)^{-1} f_2, \\ &u(0) = u_0. \end{aligned} \tag{0.2}$$

Applying $I + A_{u(t)}$ to both sides of the equation in (0.2), one formally obtains the pseudo-parabolic problem

$$\begin{aligned} &u' + \varepsilon A_u(u') + A_u(\alpha(u)) = (I + \varepsilon A_u)(f_1) + f_2, \\ &u(0) = u_0, \\ &\alpha(u)|_{\partial G} = 0. \end{aligned} \tag{0.3}$$

*Received by the editors July 21, 1983, and in revised form May 1, 1984.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

‡ On leave from Sektion Mathematik der Humboldt-Universität zu Berlin, 1086 Berlin, PSF 1297, German Democratic Republic.

§ On leave from Department of Mathematics, University of Texas, Austin, Texas 78712.

We note that problems of the type (0.3) also arise in the quite different contexts of heat conduction modelled by two-temperature systems [8], certain weak formulations of two-phase Stefan problems [6], [15] and in the description of some non-Newtonian fluids [9], [17]. Further references can be found in [7].

In [6] we considered the case $k=k(x,t)$, $\alpha=\alpha(u)$ monotone, Lipschitz. There we showed existence and uniqueness of solutions under fairly weak assumptions on the data. Furthermore, for this case comparison and maximum principles were shown. Here we are concerned with the additional nonlinearity $k=k(u)$ which arises in the diffusion model [6] due to saturation or absorption effects on the permeability. Specific properties of k and α, as they are used later are listed below under (H1)–(H6).

We prove the existence of solutions to (0.2) or (0.3) in various spaces. Theorems 2.1, 2.3 and Corollary 2.2 contain existence results for solutions $u=u(t)$ taking their values in $BV(G)$, $W_0^{s,p}(G)$ and $H_0^1(G)$, respectively. As the formulation of Theorem 2.1 shows, there remains a "gap". If $p\in[1,N/2)$, $s\in(0,1]$, then there are solutions in $W_0^{s,p}(G)$ (provided, u_0, $f(t)\in W_0^{s,p}(G)$). If $p>N/2$, we only have some (sufficiently small) $s>0$, such that $u(t)\in W_0^{s,p}(G)$. Theorem 2.6 deals with $W^{2,p}$-existence of the solutions in the two-dimensional case. By interpolation methods we obtain results for $W_0^{1+\tau,p}$-existence for $\tau\in(0,1)$ (Corollary 2.4). As a consequence we get some sufficient conditions on the data which imply that $u(t)\in W_0^{1,p}(G)$ for $N=2$ and for certain $p>2$ (Corollary 2.5). These seem not to be optimal since the assumed regularity of the data is higher than that of the solutions obtained. We continue by proving a uniqueness- and continuous-dependence result, the assumptions of which can be met at least in the one- and two-dimensional cases. The final theorem states some useful pointwise estimates, which in particular imply a weak maximum principle for (0.2).

The paper is organized as follows. Section 1 contains notations and lists some function spaces which we use. Section 2 contains the precise formulation of the results. Section 3 is concerned with the proofs. We conclude with a short appendix which presents some facts on interpolation.

1. Notation and spaces. Let $G\subset\mathbb{R}^N$ be a smooth and bounded domain, $\Gamma:=\partial G$ the boundary,

$S:=[0,T]$—a finite (time) interval, $S_t:=[0,t]$ for $t\in S$,
$Q_t:=(0,t)\times G$, $Q:=(0,T)\times G$
$D^\alpha:=D^{\alpha_1}D^{\alpha_2}\cdots D^{\alpha_N}$ for a multiindex $\alpha=(\alpha_1,\cdots,\alpha_N)$, $D^{\alpha_i}:=\partial/\partial x_i$,
$x=(x_1,\cdots,x_N)\in G$,
$s\in[0,2]$, $p\in[1,\infty]$, $r\in[1,\infty]$, $\sigma:=Ns+p$ if $s\in(0,1)$,
$\lambda\in[0,1]$, $k\in\mathbb{N}$,

$$W^{s,p}(G):=\begin{cases} W^{s,p}(G)\text{—the usual Sobolev space, if } s \text{ is an integer,}^1\\ B^s_{p,p}(G)\text{—the usual Besov space, if } s \text{ is not an integer,}\end{cases}$$

the norm in $W^{s,p}(G)$ is denoted by $\|\cdot\|_{s,p}$,

$L^p(G):=$ the usual space of p-integrable real-valued functions, $L^p(G)$ normed by the usual L^p-norm $|\cdot|_p$. *Special case.* If $p=2$ and $s=1$, then $|v|:=|v|_2$, $\|v\|:=\|v\|_{1,2}$.

"$(\cdot,\cdot)$" denotes the usual scalar product in $L^2(G)$ and the dual pairing between $H^{-1}(G)$ and $H_0^1(G)$, "$((\cdot,\cdot))$" stands for the scalar product in $H_0^1(G)$.

By $W_0^{s,p}(G)$ we denote the Sobolev—Besov space of those functions in $W^{s,p}(G)$ having zero-trace on ∂G. $W_0^{1,p}(G)$ is assumed to be normed by $\|v\|_{1,p}:=\sum_{|\alpha|=1}|D^\alpha v|_p$.

[1] A complete reference for these spaces may be found in Adams [1].

This norm is on $W_0^{1,p}(G)$ equivalent to the usual $W^{1,p}(G)$-norm, so that we do not introduce extra notation.

Furthermore:

$C^k(\bar{G}) := \{v\colon \bar{G} \to \mathbb{R}, k$ times continuously differentiable$\}$, equipped with the usual max-norm $\|\cdot\|_{C^k(\bar{G})}$,

$C^{0,\mu}(\bar{G})$ denotes the space of all Hölder continuous functions with Hölder exponent $\mu \in (0,1]$, $\|\|_{C^{0,\mu}(\bar{G})}$-norm of this space.

$BV(G) := \{v \in L^1(G)\colon \|v\|_{BV(G)} < \infty\}$-the set of all L^1-functions with finite total variation,

$$\|v\|_{BV(G)} := |v|_{L^1(G)} + [v]_{BV(G)},$$

$$[v]_{BV(G)} := \sup\left\{\int_G v \operatorname{div} \vec{w}\, dx, \vec{w} \in C_0^1(\bar{G})^N, |\vec{w}(x)| \leqq 1 \text{ for all } x \in \bar{G}\right\}.$$

$C_0^1(G) := \{v \in C^1(G)\colon \operatorname{supp} v \subset\subset G\}$.

Generally, we do not distinguish in our notation between the norms in the space $\{V, |\cdot|\}$ and the norm in $V \times V \times \cdots \times V$, e.g., "$|\nabla g|_p$" means $|\nabla g|_{L^p(G)^N}$. The same applies for scalar products.

If $\{V, |\cdot|\}$ is a normed space, then:

$L^r(S, V) := L^r(0, T; V)$—the usual space of V-valued, to the power r Bochner-integrable functions on S and equipped with the norm $\|\|_{L^r(S,V)}$ (sometimes we also write $||_{L^r(S,V)}$ for the same standard norm).

$W^{1,r}(S, V) := \{v \in L^r(S, V)\colon\ v' \in L^r(S, V)\}$.The norm in $W^{1,r}(S, V)$ is $\|v\|_{W^{1,r}(S,V)} := |v|_{L^r(S,V)} + |v'|_{L^r(S,V)}$.

$C(S, V) := C(0, T; V)$—the space of all continuous functions mapping S in V. $\|v\|_{C(S,V)} := \max\{|v(s)|\colon s \in S\}$,

$C^1(S, V) := \{v \in C(S, V)\colon v' \in C(S, V)\}$, $\|v\|_{C^1(S,V)} := \|v\|_{C(S,V)} + \|v'\|_{C(S,V)}$.

$C^{0,\mu}(S, V) := \{v \in C(S, V)\colon \|v\|_{C^{0,\mu}(S,V)} < \infty\}$.

$$\|v\|_{C^{0,\mu}(S,V)} := \|v\|_{C(S,V)} + \sup\left\{\frac{|v(s) - v(t)|}{|t-s|^{\mu}}, t \neq s, t, s \in S\right\}.$$

By "$\subset$" we denote (beside set theoretic inclusion) continuous imbeddings, and "$\subset\subset$" denotes compact imbeddings.

"c" always stands for a nonnegative constant. Sometimes, we indicate on what quantities c might depend.

"$\to$" denotes strong convergence, "$\rightharpoonup$"-weak convergence, "$\overset{*}{\rightharpoonup}$"-weak-star convergence.

2. Results. We will use the following hypotheses on the coefficients k and α, respectively.

(H1) $k\colon \mathbb{R} \to \mathbb{R}$-continuous, and there are constants k_0, k_1 such that $0 < k_0 \leqq k(u) \leqq k_1$ for $u \in \mathbb{R}$,

(H2) $\alpha\colon \mathbb{R} \to \mathbb{R}$-Lipschitz continuous with Lipschitz constant L_α,

(H3) α is monotone and $\alpha(0) = 0$,

(H4) $k \in W^{1,\infty}(\mathbb{R})$, $|k'|_\infty =: L_k$,

(H5) $\alpha \in W^{2,\infty}(\mathbb{R})$, $\|\alpha\|_{2,\infty} =: \alpha_1$,

(H6) $\gamma := k \cdot \alpha'$-Lipschitz continuous with Lipschitz constant L_γ.

Now we are going to formulate what we understand by a solution of (0.2) and (0.3), respectively. Assume k and α are sufficiently regular so that all the appearing terms make sense. For the sake of (purely technical) simplicity we set $f_1=f$ and $f_2=0$ hereafter.

Formulation of (0.2) *as an ordinary differential equation*. Let us assume for a moment, we are given a sufficiently regular solution u, v of (0.1). If $u(t)\in L^1(G)$, then $\hat{k}(t):=k(u(t))\in L^\infty(G)$ (k as in (H1)). Set $A(t)=-\operatorname{div}(\hat{k}(t)\nabla(\cdot))$. By the existence theory for elliptic operators which are subject to Dirichlet boundary conditions, we can define

$$B(t):=(I+\varepsilon A(t))^{-1}\quad (I=\text{identity})$$

and we have a continuous linear operator

$$B(t)\colon L^p(G)\to L^p(G)\quad \text{for all } p\in[1,\infty]$$

(see Lemma 3.1 below). Now, set $A_v:=-\operatorname{div}(k(v)\nabla(\cdot))$ for $v\in L^1(G)$. By the preceding remarks,

$$B_v:=(I+\varepsilon A_v)^{-1}\colon L^p(G)\to L^p(G)\quad \text{for all } p\in[1,\infty].$$

Set for abbreviation

$$A_v^\varepsilon:=\frac{1}{\varepsilon}(I-B_v).$$

Thus, (0.2) is equivalent to

$$u'(t)+A_{u(t)}^\varepsilon(\alpha(u(t)))=f(t)\quad \text{for } t\in S. \tag{2.1}$$

This leads to the formulation of (0.2) as an ordinary differential equation in $L^p(G)$: Let

$$r\in[1,\infty],\quad p\in[1,\infty],\quad f\in L^r(S,L^p(G)),\quad u_0\in L^p(G).$$

We call $u\in W^{1,r}(S,L^p(G))$ a solution, if (2.1) holds for *a.a.* $t\in S$ as an equation in $L^p(G)$ and

$$u(0)=u_0\quad \text{in } L^p(G). \tag{2.2}$$

Formulation in variational form. Formally applying $I+\varepsilon A_{u(t)}$ on both sides of (2.1), we obtain (0.3). Given $r\in[1,\infty]$, $p\in[1,\infty]$, $1/p+1/p'=1$, $u_0\in W_0^{1,p}(G)$, $f\in L^r(S, W_0^{1,p}(G))$, we call $u\in W^{1,r}(S, W_0^{1,p}(G))$ a solution of (0.3) if

$$\begin{aligned} (2.3)\quad &(u'(t),v)+\varepsilon(k(u(t))\nabla u'(t),\nabla v)+(k(u(t))\nabla\alpha(u(t)),\nabla v)\\ &\qquad=(f(t),v)+\varepsilon(k(u)\nabla f,\nabla v)\quad \text{for all } v\in W_0^{1,p'}(G),\quad \text{a.a. } t\in S,\\ (2.4)\quad &u(0)=u_0. \end{aligned}$$

Notice that under appropriate regularity assumptions on u, the fact that u satisfies (2.1), (2.2) implies that it also satisfies (2.3), (2.4) and vice versa.

Our results are as follows.

THEOREM 2.1. *Assume* (H1), (H2),

$$u_0\in W_0^{s,p}(G),\qquad f\in C(S,W_0^{s,p}(G))$$

and one of the following conditions is satisfied:

(a) $N\geqq 2$, $p\in[1,\min\{2N/(N+2),N/2\}]$, $s\in(0,\ 1]$,

(b) $2 \leqq N \leqq 6, p \in [2N/(N+2), N/2), s \in (0, (2-N)/2 + N/p)$,

(c) $p > \max\{1, N/2\}, s \in (0,1]$ *sufficiently small*,

(d) $N = 1, p = 1, s \in (0,1]$.

Then (2.1), (2.2) *has a solution* $u \in C^1(S, W_0^{s,p}(G))$.

Remark 2.1. Let k and α be as in Theorem 2.1. If $u_0 \in BV(G)$, $f \in C(S, BV(G))$, then (2.1), (2.2) is solvable by a $u \in C^1(S, BV(G))$. For merely integrable right-hand sides f we have

COROLLARY 2.2. *Let* k, α, p *and* s *be as in Theorem* 2.1, $r \in [1, \infty]$. *If* $f \in L^r(S, W_0^{s,p}(G))$, $u_0 \in W_0^{s,p}(G)$, *then there is a solution* $u \in W^{1,r}(S, W_0^{s,p}(G))$ *satisfying* (2.1), (2.2). *The corresponding remark holds for* $f \in L^r(S, BV(G))$, $u_0 \in BV(G)$.

THEOREM 2.3. *Let* $N \geqq 1$, $r \in [1, \infty]$.

(a) *If* $f \in L^r(S, H_0^1(G))$, $u_0 \in H_0^1(G)$, *then* (2.3), (2.4) *has a solution* $u \in W^{1,r}(S, H_0^1(G))$.

(b) *If* $f \in C(S, H_0^1(G))$, *then* $u \in C^1(S, H_0^1(G))$.

We remark that the proofs also yield several estimates for norms of u in terms of the data.

Looking at (2.1), (2.2), one should expect u to be exactly as regular as u_0 and f, since $B_{u(t)}$ is for many function spaces at least a regularity-preserving operator. But the nonlinearity of the problem causes some problems. With respect to a higher than square integrability of the first derivatives we get only a partial result which will be a consequence of Theorem 2.6 formulated below and the following corollary, the formulation of which seems to be rather technical.

COROLLARY 2.4. *Let* $N = 2$, *and assume* (H1), (H2), (H4), (H5) *and* (H6). *Furthermore, let* θ, $\tau \in (0,1)$, $p^* \geqq 2$, a, $a' > 1$, $p > 2$ *be such that*

$$\frac{1}{a} + \frac{1}{a'} = 1, \quad \theta + \tau \leqq 1, \quad \theta = \frac{a'p-2}{2a'(p-1)}, \quad \frac{1}{p^*} = \frac{(1-\tau)}{2} + \frac{\tau}{p}, \quad \tau \geqq \frac{2}{p^*} - \frac{2}{ap}.$$

If $u_0 \in W_0^{1+\tau, p^*}(G)$, $f \in C(S, W_0^{1+\tau, p^*}(G))$, *then* (2.3) (2.4) *has a solution* $u \in C^1(S, W_0^{1+\tau, p^*}(G))$. *Furthermore*, $u \in C^1(S, W_0^{1, ap}(G))$.

To illustrate the assumptions under which this corollary is valid, we formulate

COROLLARY 2.5.

(a) *Let* $a > 1, p > 2$ *and set*

$$\tau := \frac{ap-2}{2a(p-1)}, \quad p^* := \frac{2ap(p-1)}{p(a+1)-2}, \quad \theta := \frac{a(p-2)+2}{2a(p-2)}.$$

These numbers satisfy the conditions imposed in Corollary 2.4.

(b) *Vice versa—let* $N = 2, k$ *and* α *as in Corollary* 2.4, $q > 2$ *a given number. We have for a solution* $u \in C^1(S, W_0^{1,q}(G))$, *provided the data satisfy*

$$u_0 \in W_0^{1+\tau, p^*}(G), \qquad f \in C\big(S, W_0^{1+\tau, p^*}(G)\big),$$

where for a $p \in [2, q)$

$$p^* = \frac{2q(p-1)}{q+p-2}, \quad \theta = \frac{q(1-1/p)+2}{2q(1-2/p)}, \quad \tau = \frac{q-2}{2q(1-1/p)}.$$

Concerning $W^{2,p}$-regularity, we have

THEOREM 2.6. *Let $N=2$, k and α as in Corollary* 2.4, $\tau\in[1,\infty]$, *and* $p\in[2,\infty]$.

(a) *If* $f\in L^r(S, W^{2,p}(G)\cap W_0^{1,p}(G))$, $u_0\in W^{2,p}(G)\cap W_0^{1,p}(G)$, *then we have for a solution of* (2.1), (2.2),

$$u\in W^{1,r}\big(S, W^{2,p}(G)\cap W_0^{1,p}(G)\big).$$

(b) *If* $f\in C(S, W^{2,p}(G)\cap W_0^{1,p}(G))$, *then* $u\in C^1(S, W^{2,p}(G)\cap W_0^{1,p}(G))$.

The next theorem reflects some sufficient conditions, which ensure unique solvability and continuous dependence of the solution on the data for problem (2.3), (2.4). In particular, these conditions are automatically fulfilled if $N=1$. For $N=2$, Corollary 2.5 provides information about such properties of u_0 and f giving at least one sufficiently regular solution which meets these conditions.

THEOREM 2.7. *Let $N\geqq 1$, and k satisfying* (H1), (H4), (H6). *Set $p>2$ if $N=2$, otherwise, $p:=N$.*

(a) *If* $u_0\in H_0^1(G)$, $f\in L^1(S, W_0^{1,p}(G))$ *and if there is at least one solution of* (2.3), (2.4) *with* $u\in W^{1,1}(S, W_0^{1,p}(G))$, *then* (2.1), (2.2) *is uniquely solvable.*

(b) *The map*

$$\{u_0,f\}\in H_0^1(G)\times L^r\big(S, W^{1,p}(G)\big)\mapsto u\in W^{1,r}\big(S,H_0^1(G)\big)$$

is locally Lipschitz for all $r\in[1,\infty]$.

Finally, we obtain some pointwise estimates on solutions and briefly indicate their usefulness.

THEOREM 2.8. *Let k and α satisfy* (H1)–(H3) *and let* $u\in W^{1,1}(S,L^1(G))$ *be a solution of* (2.1), (2.2). *Then, for a.a.* $t\in S$ *we have*

$$|u^+(t)|_\infty\leqq|u_0^+|_\infty+\int_0^t|f^+(s)|_\infty\,ds,$$

$$|u^-(t)|_\infty\leqq|u_0^-|_\infty+\int_0^t|f^-(s)|_\infty\,ds.$$

In particular, if $u_0(x)\geqq 0$ a.e. $f(s,x)\geqq 0$ a.e., then $u(t,x)\geqq 0$ a.e. If, in addition, there is a number $c_0>0$ with $\alpha(c_0)=0$ and $u_0(x)\geqq c_0$ a.e. in G, then $u(t,x)\geqq c_0$ a.e. in 0.

The preceding is particularly relevant in the diffusion model of [6] where there is some interval $[0,L]$ on which α is identically zero. (This occurs because of partial saturation or absorption in the model.) To illustrate the usefulness of Theorem 2.8, suppose in this situation we know only that $k(u)$ is defined and continuous at each $u>0$. With u_0 and f as given in Theorem 2.8, we choose k_0 to be the minimum and k_1 the maximum of $k(\cdot)$ on the interval $[c_0,\|u_0\|_{L^\infty}+\int_0^1\|f(s)\|_{l^\infty}\,ds]$. Then extend k outside this interval so as to satisfy (H1). By Theorem 2.8 it follows the solution is independent of the extension, so we may assume without loss of generality that the original k satisfies (H1). These remarks are useful in the diffusion model [6] where possibly $k(u)\to+\infty$ as $u\to 0^+$ or $k(u)\to 0$ as $u\to+\infty$.

3. Proofs. The formulation of (0.2) as an operator equation (2.1) involves the resolvent $B_v:=(I+\varepsilon A_v)^{-1}$. We have to justify that B_v exists. The point which has to be observed is that for fixed v the operator A_v has a coefficient $k=k(v)$, which is due to the lack of regularity of k not too smooth. The following lemma lists some properties of the resolvent B of a related operator A.

LEMMA 3.1. *Let $\hat{k}$: $\mathbb{R}^n \to \mathbb{R}$ be measureable, $0 < k_0 \leqq \hat{k}(x) \leqq k_1$ for a.a. $x \in \mathbb{R}^n$ (k_0, k_1 as in the definition of $k(\cdot)$). Set $A := -\operatorname{div}(\hat{k}(x)\nabla(\cdot))$, $B := (I + \varepsilon A)^{-1}$, and consider the elliptic operator A as subject to Dirichlet boundary conditions. In each of the following cases B is defined and we have*

(i) *If $1 < p \leqq 2N/(N+2)$, $1/p^* = 1/p - 1/N$, then B: $L^p(G) \to W_0^{1,p^*}(G)$.*

(ii) *If $p > N/2$, then there is a $\lambda \in (0,1)$ such that B: $L^p(G) \to H_0^1(G) \cap C^{0,\lambda}(\overline{G})$.*

(iii) *If $p = 1$, $q \in [1, N(N-1))$, then B: $L^1(G) \to W_0^{1,q}(G)$.*

(iv) *If $2N/(N+2) < p \leqq \min\{N/2, 2\}$, then B: $L^p(G) \to W_0^{1,2}(G)$.*

(v) *If $2 \leqq p < N$, $1/p^* = 1/p - 1/N$, then B: $L^p(G) \to W_0^{1,2}(G) \cap L^{p^*}(G)$.*

In each of these cases, B is a linear and continuous map and its norm depends at most on $G, p^*, p, N, \lambda, \varepsilon k_0, k_1$. λ depends at most on $G, p, N, \varepsilon k_0, k_1$.

Proof. (i), (iii) are part of [16, Thm. 4.5], for (ii) see [12, Chap. III], (iv) and (v) follows from (i). □

The next lemma list some known imbedding properties.

LEMMA 3.2. *Consider the situation in Lemma* 3.1 *and let p, p^*, λ, N be as in* (i)–(v). *Take in* (i)–(iv) $s \in (0,1)$, *in* (v) $s \in (0(2-N)/2 + N/p)$. *Then*

(i)[2] $W^{1,p^*}(G) \subset\subset W^{s,p^*}(G) \subset W^{s,p} \subset\subset L^p(G)$

(ii) $W^{s,p}(G) \subset\subset L^p(G)$, $C^{0,\lambda}(\overline{G}) \subset L^\infty(G)$

(iii) $W^{1,p^*}(G) \subset\subset W^{s,p^*}(G) \subset W^{s,1}(G) \subset\subset L^1(G)$

(iv) $W^{1,p^*}(G) \subset\subset W^{s,p}(G) \subset\subset L^p(G)$

(v) $W^{1,2}(G) \subset\subset W^{s,p}(G) \subset\subset L^p(G)$, *if* $N \leqq 6$.

Proof. See [1], [3]. The compactness results from $W^{1,p} \subset\subset L^p$ for any $p \geqq 1$, and the fact, that[3] $W^{s,p} = [L^p, W^{1,p}]_s$ and general interpolation theory. □

COROLLARY 3.3. *Let $v \in L^1(G)$, set, as before, $A_v := -\operatorname{div}(k(v(x))\nabla(\cdot))$, $B_v := (I + \varepsilon A_v)^{-1}$. B_v has exactly the same properties as B in Lemma* 3.1. *Furthermore, $A_v^\varepsilon := (1/\varepsilon)(I - B_v)$: $L^p(G) \to L^p(G)$ is Lipschitz. Thus, $A_v^\varepsilon \circ \alpha$: $L^p(G) \to L^p(G)$ is Lipschitz.*

It follows that (2.1) can be considered as an ordinary differential equation in $L^p(G)$. Moreover, we have the following.

LEMMA 3.4. *Let $v \in L^1(S, L^1(G))$ be given. Then*

(i) *For each $w \in L^p(G)$*

$$t \in S \to A_{v(t)}^\varepsilon(w) \in L^p(G) \tag{3.1}$$

is measureable.

(ii) *If $v \in C(S, L^1(G))$, then the map in* (3.1) *is continuous.*

Proof. (i) see [14].

(ii) Let $t_n \to t$ in S, set $g_n := (I + \varepsilon A_{v(t_n)})^{-1}(w)$. By definition g_n satisfies

$$\left(I + \varepsilon A_{v(t_n)}\right)(g_n) = w. \tag{3.2}$$

Consider case (i) in Lemma 3.1. We have $\|g_n\|_{1,p^*} \leqq c|w|_p$. Therefore, for a subsequence $g_{n_j} \to \tilde{g}$ in $L^{p^*}(G)$ and $g_{n_j} \rightharpoonup \tilde{g}$ in $W_0^{1,p^*}(G)$. By the continuity of $k, \tilde{g}$ satisfies

$$\left(I + \varepsilon A_{v(t)}\right)\tilde{g} = w. \tag{3.2$'$}$$

[2] "$\subset$" denotes algebraic and topological imbedding, "$\subset\subset$" the compact imbedding.

[3] "$[\cdot, \cdot]$" is the complex interpolation space generator (see appendix).

But, (3.2′) is uniquely solvable, so that $g_n \to \tilde g$ in $L^p(G)$, which proves the assertion. The other cases in Lemma 3.1 can be dealt with in a similar manner. □

Proof of Theorem 2.1. Fix $\bar u \in C(S, L^p(G))$. By Corollary 3.3 and Lemma 3.4, there is exactly one solution

$$u \in C^1(S, L^p(G)) \tag{3.3}$$

satisfying

$$u'(t) + A\frac{\varepsilon}{\bar u(t)}\alpha(u(t)) = f(t), \qquad u(0) = u_0. \tag{3.4}$$

We employ Schauder's theorem to show that the map

$$\mathscr{T}\colon C(S, L^p(G)) \to C(S, L^p(G))$$

defined by (3.3), (3.4) has a fixed point. Obviously, a fixed point of $\mathscr{T}$ solves (2.3), (2.4). The next lemma summarizes some properties of $\mathscr{T}$. In particular, the proof yields several estimates of the solutions of (2.1), (2.2).

LEMMA 3.5. (i) *$\mathscr{T}$ maps $C(S, L^1(G))$ into a bounded subset of $C^1(S, L^p(G))$.*

(ii) *Let s, N be as in Lemma* 3.2, *$t \in S$, $\bar u \in C(S, L^1(G))$. Then $(\mathscr{T}\bar u)(t)$ is in a bounded subset of $W^{s,p}(G)$.*

(iii) *$\mathscr{T}\colon C(S, L^1(G)) \to C(S, L^p(G))$ is continuous.*

Proof. (i). Integrate (3.4) over $(0,t)$, and take the $L^p(G)$-norm on both sides. Then, by Lemmas 3.1, 3.2 and α's Lipschitz continuity

$$\begin{aligned} |u(t)|_p &\leqq |u_0|_p + \int_0^t \left|A^{\varepsilon}_{\bar u(s)}\alpha(u(s))\right|_p ds + \int_0^t |f(s)|_p\, ds \\ &\leqq |u_0|_p + c\int_0^t |\alpha(u(s))|_p\, ds + \int_0^t |f(s)|_p\, ds \\ &\leqq |u_0|_p + c\int_0^t \left\{|\alpha(0)|\cdot|G| + L_\alpha |u(s)|_p\right\} ds + \int_0^t |f(s)|_p\, ds; \end{aligned}$$

hence, by Gronwall's inequality

$$|u|_{C(0,t;L^p(G))} \leqq c\left\{|u_0|_p + 1 + |f|_{L^1(S,L^p(G))}\right\}, \tag{3.5}$$

where $c = c(\varepsilon k_0, k_1, |G|, N, p, L_\alpha, \lambda, s, T)$. By (3.4) and (3.5)

$$|u|_{C^1(0,t;L^p(G))} \leqq c\left\{|u_0|_p + 1 + |f|_{C(S,L^p(G))}\right\}. \tag{3.6}$$

To obtain further estimates, we notice that u as a solution of an ordinary differential equation is the $C(S, L^p(G))$-limit of the sequence $\{u_n\}$ defined by

$$u_1 := u_0, \tag{3.7}$$

$$u_{n+1}(t) = u_0 + \int_0^t f(s)\, ds - \int_0^t A^{\varepsilon}_{\bar u(s)}\alpha(u_n(s))\, ds,\ t \in S. \tag{3.8}$$

We have

LEMMA 3.6. *Let $u_n \in C(S, W^{s,p}(G))$, where s, p are taken as in Lemma* 3.1, 3.2, *$s \in (0, \lambda)$ if $p > N/2$ (λ arises in Lemma* 3.1, (ii)). *Then*

a) *$u_{n+1} \in C^1(S, W^{s,p}(G))$ and*

b)

$$\|u_n\|_{C(S,W^{s,p}(G))} \leqq c\left\{\|u_0\|_{s,p} + 1 + \|f\|_{L^1(S,W^{s,p}(G))}\right\}, \tag{3.9}$$

with c as in (3.5).

Proof. By assumption $u_n(t) \in W^{s,p}(G)$; hence $\alpha(u_n) \in W^{s,p}(G) \subset L^p(G)$. By Lemma 3.1, 3.2 $A^{\varepsilon}_{\bar{u}(t)}(\alpha(u_n(t))) \in W^{s,p}(G)$ if $p \leqq N/2$, and

$$\text{(3.10)} \qquad \left\|A^{\varepsilon}_{\bar{u}(t)}\left(\alpha\left(u_n(t)\right)\right)\right\|_{s,p} \leqq \frac{1}{\varepsilon}\left\|\alpha\left(u_n(t)\right)\right\|_{s,p} + \frac{c}{\varepsilon}\left|\alpha\left(u_n(t)\right)\right|_p$$

$$\leqq \frac{c}{\varepsilon}\left\|u_n(t)\right\|_{s,p} \qquad (c \text{ as in } (3.5)).$$

Therefore, by (3.8), $u_{n+1}(t) \in W^{s,p}(G)$. The t-continuity properties follow as in Lemma 3.4. Equations (3.8) and (3.10) yield (3.9) by an iteration argument (cf. [4]). To show (3.9) for $p > N/2$, take the $W^{s,p}$-norm on both sides of (3.8). Thus

$$\text{(3.11)} \qquad \|u_{n+1}(t)\|_{s,p} \leqq \|u_0\|_{s,p} + \int_0^t \|f(s)\|_{s,p} ds + \frac{1}{\varepsilon}\int_0^t \|\alpha(u_n(s))\|_{s,p} ds$$

$$+ \frac{1}{\varepsilon}\int_0^t \left\|\left(I + \varepsilon A_{\bar{u}(s)}\right)^{-1} \alpha\left(u_n(s)\right)\right\|_{s,p} ds.$$

We have $\|\alpha(u_n)\|_{s,p} \leqq L_\alpha \|u_n\|_{s,p}$ and for

$$g(t,x) := \left(I + \varepsilon A_{\bar{u}(t)}\right)^{-1}\left(\alpha\left(u_n(t)\right)\right)$$

by Lemma 3.1, 3.2.

(ii)

$$\|g(t)\|^p_{s,p} = \iint_{GG} \frac{|g(t,x) - g(t,y)|^p}{|x-y|^\mu}\, dx\, dy, \qquad \mu = N + sp$$

$$\leqq \iint_{GG} c \cdot |x-y|^{\lambda p - \mu}\, dx\, dy \leqq \text{const.}$$

Therefore (3.11) and Gronwall's inequality imply (3.9). □

To complete the proof of Lemma 3.5(ii), we note that (3.9) and (3.6) imply

$$\text{(3.12)} \qquad \|u_n\|_{C^1(S, W^{s,p}(G))} \leqq c\left\{\|u_0\|_{s,p} + \|f\|_{C(S, W^{s,p}(G))} + 1\right\},$$

where c is as in (3.5). By weak-star compactness, $u \in W^{1,\infty}(S, W^{s,p}(G))$ and by (3.4), $u \in C^1(S, W^{s,p}(G))$, where (3.12) implies

$$\text{(3.13)} \qquad \|u\|_{C^1(S, W^{s,p}(G))} \leqq c\left\{\|u_0\|_{s,p} + \|f\|_{C(S, W^{s,p}(G))} + 1\right\}.$$

This proves (ii) of Lemma 3.5. To see the continuity of $\mathscr{T}$ let $\bar{u}_k \to \bar{u}$ in $C(S, L^1(G))$, set $g_k := B_{\bar{u}_k}(\alpha(u_k)), u_k := \mathscr{T}(\bar{u}_k)$, so that g_k and u_k, resp. satisfy

$$\text{(3.14)} \qquad \varepsilon A_{\bar{u}_k}(g_k) = -g_k + \alpha(u_k),$$

$$\text{(3.15)} \qquad u_k'(t) + \frac{1}{\varepsilon}\alpha(u_k(t)) = \frac{1}{\varepsilon} g_k(t) + f(t), \qquad u_k(0) = u_0.$$

By estimates (3.6), (3.13) and α's Lipschitz continuity there is a subsequence

$$\text{(3.16)} \qquad \begin{aligned} &u_{k_j} \overset{*}{\rightharpoonup} u \quad \text{in } W^{1,\infty}(S, W^{s,p}(G)), \\ &\alpha(u_{k_j}) \overset{*}{\rightharpoonup} \alpha(u) \quad \text{in } L^\infty(S, W^{s,p}(G)). \end{aligned}$$

Using the estimates for u_k from (3.15) (cf. estimates (3.6), (3.13)) to estimate g_k in (3.14) ((3.14) is an elliptic problem for g_k, various norms of g_k can be estimated in terms of $\alpha(u_k)$ by Lemma 3.1, and $\alpha(u_k)$ can be estimated by (3.13), (3.6)), we arrive for $p \leqq (N/2)$ at

$$|g_k|_{L^\infty(S,W^{s,p}(G))} \leqq c|\alpha(u_k)|_{L^\infty(S,L^p(G))} \leqq c\left\{1+|u_k|_{L^\infty(S,L^p(G))}\right\}$$

$$\leqq c\left\{1+|u_0|_p+|f|_{L^1(S,L^p(G))}\right\},$$

and for $p > N/2$, at

$$|g_k|_{L^\infty(S,H_0^1(G))} \leqq c\left\{1+|u_0|_p+|f|_{L^1(S,L^p(G))}\right\}.$$

Therefore, we have for a subsequence

$$g_{k_j} \overset{*}{\rightharpoonup} g \text{ in } L^\infty\left(S, W_0^{1,p}(G)\right). \tag{3.17}$$

Equations (3.16), (3.17) and relation (3.14) imply that g satisfies

$$\varepsilon A_{\bar{u}}(g) = -g+\alpha(u). \tag{3.18}$$

(To pass to the limit $j \to \infty$ in (3.14), use the strong convergence properties of $k(\bar{u}_{k_j})$—remember, $\bar{u}_k \to \bar{u}$ in $C(S, L^1(G))$ and k is continuous.) Since (3.18) is uniquely solvable, we have $u_k \overset{*}{\rightharpoonup} u$ in $W^{1,\infty}(S, W^{s,p}(G))$, i.e., the whole sequence converges, in particular $u_k = \mathscr{T}(\bar{u}_k) \to u = \mathscr{T}(\bar{u})$ in $C(S, L^p(G))$, i.e., $\mathscr{T}$is continuous, which finishes the proof of Lemma 3.5. □

COROLLARY 3.7. *Lemma* 3.5 *implies that $\mathscr{T}$ has a fixed point u. u solves problem* (2.3), (2.4) *and satisfies the estimates given by* (3.6), (3.9). *Moreover, u is—according to* (3.7), (3.8)—*the limit of the sequence $\{u_n\}$ defined by*

$$\begin{aligned} &u_1 := u_0, \\ &u_{n+1}(t) = u_0 + \int_0^t f(s)\,ds - \int_0^t A^{\varepsilon}_{u_n(s)}(\alpha(u_n(s)))\,ds, \end{aligned} \tag{3.19}$$

and

$$u_n \to u \quad \text{in } C(S,L^p(G)), \qquad u_n \overset{*}{\rightharpoonup} u \quad \textit{in } W^{1,\infty}(S,L^p(G)). \tag{3.20}$$

This finishes the proof of Theorem 2.1. □

Proof of Remark 2.1. Construct as in (3.3), (3.4) an operator

$$\mathscr{T}\colon C(S,L^1(G)) \to C^1(S,L^1(G)).$$

One has already estimate (3.6), so that Lemma 3.5(i) follows. The third statement of this lemma has already been proved and the second has to be changed to

LEMMA 3.5′. (ii′) *If $\bar{u} \in C(S, L^1(G))$, then $(\mathscr{T}u)(t)$ is in a bounded subset of $BV(G)$, which does not depend either on t or on $\bar{u}$.*

To see this, look at the iteration procedure (3.7), (3.8) which yields u as the $C(S,L^1(G))$—limit of $\{u_n\}$. By Lemma 3.1(iii) we have $(I+\varepsilon A_{\bar{u}(s)})^{-1}\alpha(u_n(s))\in W^{1,p^*}(G)\subset BV(G)$ so that by the same lemma

$$\|A^{\varepsilon}_{\bar{u}(s)}\alpha(u_n(s))\|_{BV(G)}\leqq c\cdot\frac{1}{\varepsilon}\left\{\|\alpha(u_n)\|_{BV(G)}+\|\alpha(u_n)\|_{L^1(G)}\right\}$$

$$\leqq c\cdot\frac{1}{\varepsilon}\left\{1+\|u_n\|_{BV(G)}\right\}.$$

Taking the $BV(G)$-norm on both sides of (3.8) yields

$$\|u_n\|_{C^1(S,BV(G))}\leqq c\left\{\|u_0\|_{BV(G)}+\|f\|_{C(S,BV(G))}+1\right\}.$$

Since $BV(G)\subset\subset L^{p^*}(G)\subset L^1(G)$ for $p^*\in(1,N/(N-1))$, Arzela–Ascoli's theorem yields for a subsequence

$$u_{n_j}\to u \quad\text{in } C(S,L^{p^*}(G)).$$

Because of the reflexivity of $BV(G)$ we have by weak-star compactness

$$u_{n_j}\overset{*}{\rightharpoonup}u \quad\text{in } W^{1,\infty}(S,BV(g)) \quad\text{and}$$

$$\|u\|_{W^{1,\infty}(S,BV(G))}\leqq c\left\{\|u_0\|_{BV(G)}+1+\|f\|_{C(S,BV(G))}\right\}.$$

Since u satisfies (3.4), we obtain after a short calculation

$$\begin{aligned}&u'\in C(S,BV(G)) \quad\text{and}\\ &\|u\|_{C^1(S,BV(G))}\leqq c\left\{\|u_0\|_{BV(G)}+1+\|f\|_{C(S,Bv(G))}\right\}.\end{aligned}\tag{3.21}$$

Therefore $\mathscr{T}$ has a fixed point $u\in C^1(S,BV(G))$ which solves (2.3), (2.4). □

Proof of Corollary 2.2. We modify estimates (3.6), (3.13). One has

$$|u'(t)|_p\leqq|A^{\varepsilon}_{u(t)}(\alpha(u(t)))|_p+|f(t)|_p$$

$$\leqq c\left\{1+|u(t)|_p\right\}+|f(t)|_p$$

from arguments which led to estimate (3.5). Taking (3.5) into account, one gets for $r\in[1,\infty]$

$$|u'|_{L^r(S,L^p(G))}\leqq c\left\{1+|u_0|_p+|f|_{L^2(S,L^p(G))}\right\}.\tag{3.22}$$

Similarly, (3.11) implies

$$\|u\|_{C(S,W^{s,p}(G))}\leqq c\left\{1+\|u_0\|_{s,p}+|f|_{L^1(S,W^{s,p}(G))}\right\},$$

and (2.3) yields

$$\|u'(t)\|_{s,p}\leqq\|A^{\varepsilon}_{u(t)}(\alpha(u(t)))\|_{s,p}+\|f(t)\|_{s,p}.$$

Therefore,

$$|u'|_{L^r(S,W^{s,p}(G))}\leqq c\left\{1+\|u_0\|_{s,p}+|f|_{L^r(S,W^{s,p}(G))}\right\}.\tag{3.23}$$

□

Proof of Theorem 2.3. We employ the Galerkin method. Let $\{w_i\} \subset H_0^1(G) \cap H^2(G)$ be an orthonormal eigenvalue basis of the Laplacian subject to Dirichlet boundary conditions, i.e,.

$$-\Delta w_i = \lambda_i w_i,\ (w_i, w_j) = \delta_{ij},\ w_i \perp w_j \quad \text{in } H_0^1(G). \tag{3.24}$$

Set $V_n := \operatorname{span}\{w_1, \cdots, w_n\}$, denote by P_n the orthogonal projection in $H_0^1(G)$ onto V_n. We are looking for an absolutely continuous

$$u_n(t) := \sum_{j=1}^{n} h_{nj}(t) w_j, \tag{3.25$'$}$$

which satisfies

$$\begin{aligned}(u_n'(t), v) + \varepsilon(k(u_n)\nabla u_n', \nabla v) + (k(u_n)\nabla\alpha(u_n), \nabla v)\\ = \varepsilon(k(u_n)\nabla f, \nabla v) + (f, v) \quad \forall v \in V_n, \text{ a.a. } t \in S,\end{aligned} \tag{3.25}$$

$$u_n(0) = u_{0n} := P_n u_0. \tag{3.26}$$

By using a fixed point argument and applying the results of [6] or a reduction of (3.25), (3.26) to an ordinary differential equation (cf. [5, Lemma 1]) to obtain the standard form for an application of Caratheodory's theorem, one shows, that (3.25′)–(3.26) is (at least) locally solvable. The following lemma implies that these solutions are globally defined. One has

LEMMA 3.8. *There is a constant* $c = c(\varepsilon k_0, k_1, L_\alpha, T, |G|, \varepsilon)$ *such that for* $r \in [1, \infty]$

$$\|u_n\|_{W^{1,r}(S, H_0^1(G))} \leqq c\left\{\|u_0\| + \|f\|_{L^r(S, H_0^1(G))}\right\}. \tag{3.27}$$

Proof of Lemma 3.8. Choose in (3.25) $v := u_n'(t)$, use the boundedness properties of $k(\cdot)$ and $\alpha(\cdot)$ and apply Hölder's and Young's inequalities

$$\begin{aligned}&|u_n'(t)|^2 + \varepsilon k_0 |\nabla u_n'(t)|^2\\ &\quad k_1 L_\alpha |\nabla u_n||\nabla u_n'| + \varepsilon k_1 |\nabla f||\nabla u_n'| + |f| \cdot |u_n'|\\ &\quad \leqq \frac{\varepsilon k_0}{2}|\nabla u_n'|^2 + c\left\{|\nabla u_n|^2 + |\nabla f|^2 + |f|\right\} + \frac{1}{2}|u_n'|^2,\end{aligned} \tag{3.28}$$

$$|\nabla u_n'(t)|^{\hat r} \leqq c\left\{|f|^{\hat r} + |\nabla f|^{\hat r} + |\nabla u_{0n}|^{\hat r} + \int_0^t |\nabla u_n'(s)|^{\hat r}\, ds\right\}. \tag{3.29}$$

If $r < \infty$, then $\hat r := r$; if $r = \infty$, then $\hat r := 1$.

Gronwall's inequality in connection with (3.26) yields

$$\|u_n'\|_{L^r(S, H_0^1(G))} \leqq c\left\{\|f\|_{L^r(S, H_0^1(G))} + \|u_0\|\right\}$$

and therefore (3.27). The usual compactness argument completes the proof and shows that (2.1), (2.2) has a solution, which satisfies

$$\|u\|_{W^{1,r}(S, H_0^1(G))} \leqq c\left\{\|u_0\| + \|f\|_{L^r(S, H_0^1(G))}\right\}. \tag{3.30}$$

□

Before proving Corollaries 2.4, 2.5, we prove Theorem 2.6.

Proof of Theorem 2.6(a). We use the Galerkin method and continue at (3.25), (3.26). We have already estimate (3.27) and will show that

$$\|u_n\|_{W^{1,r}(S, W^{2,p}(G) \cap W_0^{1,p}(G))} \leqq \text{const}. \tag{3.31}$$

To see this, note that due to (3.24)

$$v := -\Delta u_n'(T) \in V_n$$

is an admissible function in (3.25). Integration by parts and reordering yield

$$\begin{aligned}
&\left|\nabla u_n'(t)\right|^2 + \varepsilon\big(k(u_n)\Delta u_n', \Delta u_n'\big)\\
&\qquad = \varepsilon\big(k'(u_n)\nabla u_n', \nabla u_n v\big) + (f, v) - \varepsilon\big(k'(u_n)\nabla u_n, \nabla f v\big)\\
&\qquad\quad - \varepsilon\big(k(u_n)\Delta, f, v\big) + \big(k'(u_n)\Delta u_n \cdot \nabla u_n, \alpha'(u_n) v\big)\\
&\qquad\quad + \Big(k(u_n)\alpha''(u_n)|\nabla u_n|^2, v\Big) + \big(k(u_n)\alpha'(u_n)\Delta u_n, v\big).
\end{aligned}$$

By (H1), (H2), (H4), (H5), Hölder's, Young's and one of Sobolev's inequalities we obtain

$$\begin{aligned}
&\left|\nabla u_n'(t)\right|^2 + \varepsilon k_0 \left|\Delta u_n'(t)\right|^2\\
&\quad \leqq \varepsilon L_k |\nabla u_n'|_4 |\nabla u_n|_4 |v|_2 + |f|_2 |v|_2 + \varepsilon L_k |\nabla u_n|_4 |\nabla f|_4 |v|_2\\
&\qquad + \varepsilon k_1 |\Delta f|_2 |v|_2 + L_k \alpha_1 |\nabla u_n|_4^2 |v|_2 + k_1 \alpha_1 |\nabla u_n|_4^2 |v|_2 + k_1 \alpha_1 |\Delta u_n|_2 |v|_2\\
&\quad \leqq \frac{\varepsilon k_0}{4} |v|_2^2 + c\Big\{|\nabla u_n'|_2 |\Delta u_n'|_2 |\nabla u_n|_2 + |f|_2^2\\
&\qquad\qquad + |\nabla u_n||\Delta u_n||\nabla f|_4^2 + |\Delta f|_2^2 + |\nabla u_n|^2 |\Delta u_n|^2 + |\Delta u_n|^2\Big\}.
\end{aligned}$$

(This is the only place where the restriction $N=2$ is essential. The case $N=3$ allows only estimates like $|\nabla u_n'|_4 \leqq c|\nabla u_n'|_2^{1/3} |\Delta u_n'|_2^{2/3}$, which finally would require some restrictions on $|\nabla f_4$ and $\|u_0\|_{2,2}$, to obtain (3.31).)

From now on we use again our convention concerning the notation of L^2-norms. Using again Young's inequality, we arrive at

$$\begin{aligned}
\left|\Delta u_n'(t)\right|^2 &\leqq c\Big\{|\nabla u_n'|^2 |\nabla u_n|^2 |\Delta u_n|^2 + |f|^2 + |\Delta f|^2\\
&\qquad + |\nabla u_n||\Delta_n||\nabla f|_4^2 + |\nabla u_n|^2 |\Delta u_n|^2 + |\Delta u_n|^2\Big\}\\
|\Delta u_n'(t)| &\leqq c\Big\{|\nabla u_n'||\nabla u_n||\Delta u_n| + |f| + |\nabla u_n| + |\Delta u_n||\nabla f|_4^2\\
&\qquad + |\Delta f| + |\nabla u_n||\Delta u_n| + |\nabla u_n||\nabla f|_4^2 + |\Delta u_n|\Big\}.
\end{aligned}$$

By (3.27),

$$|\nabla u_n|_{C(S, L^2(G)^N)} \leqq c\left\{\|u_0\| + \|f\|_{L^1(S, H_0^1(G))}\right\} =: \bar{c},$$

so that

$$(3.32)\qquad |\Delta u_n'(t)| \leqq c\left\{ \bar{c}|\nabla u_n'|\left(|\Delta u_{0n}| + \int_0^t |\Delta u_n'(s)|\,ds\right) + \|f\|_{2,2} + \bar{c} + \left(|\Delta u_{0n}| + \int_0^t |\Delta u_n'(s)|\,ds\right)|\nabla f|_4^2 + \cdots \right\}.$$

If $r=\infty$, then estimate (3.30) and Gronwall's inequality yield $|\Delta u_n'|_{C(0,t;\,L^2(G))} \leqq \text{const.}$, where the const. depends on

$$(3.33)\qquad \varepsilon k_0,\ k_1,\ L_k,\ T, N=2,\ |G|,\ |f|_{L^r(S,W^{1,4}(G))},\ |\Delta u_0|_2.$$

If $r\in[1,\infty)$, take on both sides of (3.32) the rth power and integrate over $(0,t)\subset[0,T]$, which yields by Gronwall's inequality

$$|\Delta u_n'|_{L^r(0,t;\,L^2(G))} \leqq \text{const.}$$

Because of $u_n/\partial G = \Delta u_n'/\partial G = 0$, this implies

$$(3.34')\qquad \|u_n\|_{W^{1,r}(S,W^{2,2}(G)\cap H_0^1(G))} \leqq \text{const} =: \hat{c},$$

$\hat{c}$ as in (3.33). The usual compactness argument finishes the proof and yields an estimate

$$(3.34)\qquad \|u\|_{W^{1,r}(S,W^{2,2}(G)\cap H_0^1(G))} \leqq \hat{c}$$

for the solution u of (2.1), (2.2).

Proof of part (b). (Sketch, a similar argument is used in the proof of Corollary 2.4.) Under the conditions of part (b) we set that (2.1), (2.2) holds in $L^1(G)$ a.e. Therefore, we can multiply equation (2.2) on both sides by $v := -|\Delta u'(t)|^{p-2}\Delta u'(t)$ and arrive, using similar arguments as in (a) and estimate (3.34), at

$$(3.35)\qquad \|u\|_{W^{1,r}(S,W^{2,p}(G))} \leqq c\left\{1 + |\Delta f|_{L^r(S,L^p(G))} + |\Delta u_0|_p\right\}$$

which is sufficient to complete the proof. c depends via (3.34) on the data. □

Proof of Corollary 2.4. We approximate f and u_0 respectively by regular f_δ, $u_{0\delta}$, and obtain by Theorem 2.6 regular solutions u_δ for which we show the estimate (3.39). The basic tool to obtain (3.39) are the estimates (3.42), (3.43) for a related linear problem (3.40). Let

$$(3.36)\qquad f_\delta \in C\left(S, W^{2,\infty}(G)\cap W_0^{1,\infty}(G)\right),\ u_{0\delta}\in W^{2,\infty}(G)\cap W_0^{1,\infty}(G)$$

such that

$$(3.37)\qquad f_\delta \to f \quad \text{in } C\left(S, W_0^{1+\tau,p^*}(G)\right), \qquad u_{0\delta}\to u_0 \quad \text{n } W_0^{1+\tau,p^*}(G).$$

By Theorem 2.6, there are solutions

$$(3.38)\qquad u_\delta \in C^1\left(S, W^{2,\infty}(G)\cap W_0^{1,\infty}(G)\right),$$

satisfying (2.1), (2.2) with $u_{0\delta}$, f_δ as data. Set $\hat{k}(t,x) := k(u_\delta(t,x))$, $\hat{\gamma}(t,x) := \hat{k}(t,x)\cdot \alpha'(u_\delta(t,x))$. By our hypotheses, $0<k_0\leqq\hat{k}(t,x)\leqq k_1$ $\forall t\in S$, $x\in G$ and $|\hat{\gamma}|_\infty<\infty$. It is our goal to show the following estimate. There is a constant c merely depending on the bounds of the coefficients k, α and their derivatives as appearing in (H1)–(H6) and on

$$\|u_0\|_{1+\tau,p^*},\ \|f\|_{C(S,W^{1+\tau,p^*}(G))},$$

such that

$$\|u_\delta\|_{C^1(S,W_0^{1+\tau,p^*}(G))} \leqq c. \tag{3.39}$$

To this end consider for given g, v_0 the linear problem

$$\begin{aligned} &v' - \varepsilon \operatorname{div}(\hat{k}\nabla v') - \operatorname{div}(\hat{\gamma}\nabla v) = g - \varepsilon \operatorname{div}(\hat{k}\nabla\ ), \\ &v(0) = v_0. \end{aligned} \tag{3.40}$$

We note that choosing $g := f_\delta$, $v_0 := u_{0\delta}$, $v := u_\delta$ satisfies this equation.

Denote by $S_t := [0,t]$, $t \in S$, a subinterval of S. As in Theorems 2.3 and 2.5 one shows the existence of solution operators $P_{i,t}$, $i = 0, 1$,

$$\begin{aligned} &P_{0,t}\colon H_0^1(G) \times C(S_t, H_0^1(G)) \to C(S_t, H_0^1(G)), \\ &P_{1,t}\colon (W^{2,p}(G) \cap W_0^{1,p}(G)) \times C(S_t, W^{2,p}(G)) \to C(S_t, W^{2,p}(G) \cap W_0^{1,p}(G)), \\ &P_{i,t}\colon \{v_0, g\} \mapsto v',\ v_0, g, v, v' \quad \text{as in (3.40)}. \end{aligned}$$

(The numbers p and p^* are related by the hypotheses of this corollary.) These operators are linear and bounded in the given spaces and we denote by $M_{i,t}$ their respective norms. By interpolation it follows that $P_{1,t}$ can be restricted to $W_0^{1+\tau,p^*}(G) \times C(S_t, W^{1+\tau,p^*}(G))$ (see Lemma A5, appendix) and, denoting the restriction by $P_{\tau,t}$,

$$P_{\tau,t}\colon W_0^{1+\tau,p^*}(G) \times (G) \times C(S_t, W_0^{1+\tau,p^*}(G)) \to C(S_t, W_0^{1+\tau,p^*}(G)), \qquad \tau \in (0,1).$$

$P_{\tau,t}$ is linear and bounded and its norm $M_{\varepsilon,t}$ can be estimated by

$$M_{\tau,t} \leqq c M_{0,t}^{1-\tau} M_{1,t}^{\tau} \tag{3.41}$$

with some numerical constant depending on τ, but independent of t. We prove the following

LEMMA 3.9. *There exists a constant c, depending at most on τ, G, T, p^* the several bounds of the coefficients as appearing in the hypotheses of this corollary, such that*

$$\text{(a)} \quad M_{0,t} \leqq c, \tag{3.42}$$

$$\text{(b)} \quad M_{1,t} \leqq c\left\{1 + \|u_\delta\|_{C(S_t,W^{1+\tau,p^*}(G))}^{1/(1-\theta)}\right\}. \tag{3.43}$$

Proof of Lemma 3.9. First we notice that due to the rather technical assumptions of this corollary we have

$$W_0^{1+\tau,p^*}(G) \subset W_0^{1,\alpha p}(G), \tag{3.44}$$

$$[W_0^{1,2}(G), W^{2,p}(G) \cap W_0^{1,p}(G)]_\tau = W_0^{1+\tau,p^*}(G) \tag{3.45}$$

(see the appendix, Lemmas A4, A6). The proof of (a) is at the beginning almost identical to that of estimate (3.27) in Lemma 3.8. To show (3.43) we begin with the case $p = 2$. Computing the div-terms in (3.40), multiplying by $w := -\Delta v'(t)$ and integrating over G by parts, one obtains for $s \in S_t$

$$|\nabla v'(s)|^2 + (\varepsilon\hat{k}\Delta v', \Delta v')$$
$$= (\varepsilon\hat{k}\Delta g + \hat{\gamma}\Delta v, \Delta v') + (g, -\Delta v') + (\varepsilon\nabla\hat{k} \cdot \nabla v' + \nabla\hat{\gamma} \cdot \nabla v, -\Delta v') + (\varepsilon\nabla\hat{k} \cdot \nabla g, \Delta v').$$

To obtain (3.43), it is sufficient to assume that

$$\|v_0\|_{2,p} + \|g\|_{C(S_t, W^{2,p}(G))} \leqq 1. \tag{3.46}$$

Using the boundedness properties of $\hat{k}$ and $\hat{\alpha}$ and Hölder's inequality, we obtain with $c = c(\varepsilon k_0, k_1, L_\alpha)$

$$|\nabla v'(s)|^2 + \varepsilon k_0 |\Delta v'(s)|^2$$
$$\leqq c\{|\Delta g| + |\Delta v| + |g| + |\nabla\hat{k}|_4 |\nabla v'|_4 + |\nabla\hat{\gamma}|_4 |\nabla v|_4 + |\nabla\hat{k}|_4 |\nabla g|_4\} |\Delta v'|.$$

By $v'/\partial G = 0$, there is a constant $c = c(G)$,

$$|\Delta v'(s)|_p + |v'(s)|_p \geqq c\|v'(s)\|_{2,p}; \tag{3.47}$$

by interpolation, $|\nabla v'|_4^2 \leqq c(G)|\Delta v'|_2 \|v'\|_{2,2}$. This and (3.46) imply by using Young's inequality and (3.42),

$$\|v'(s)\|_{2,2} \leqq c\Big\{\|g\|_{2,2} + \|v_0\|_{2,2} + \int_0^s \|v'(r)\|_{2,2}\,dr$$
$$+ |\nabla\hat{k}|_4^2 |\nabla v'|_2 + |\nabla\hat{\gamma}|_4^2 |\nabla v|_2 + \|v\|_{2,2} + |\nabla\hat{\gamma}|_4 \|g\|_{2,2}\Big\}$$
$$\leqq c\Big\{1 + |\nabla\hat{k}|_4^2 + |\nabla\hat{\gamma}|_4^2 + \int_0^s |v'(r)|_{2,2}\,dr\Big\}.$$

Because of

$$\begin{aligned} |\nabla\hat{k}|_4 + |\nabla\hat{\gamma}|_4 &\leqq c\|u_\delta\|_{1,4} && \text{(cf. def. of } \hat{k}, \hat{\gamma}) \\ &\leqq \|u_\delta\|_{3/2,2} && \text{(cf. (3.44))}, \end{aligned}$$

this and Gronwall's inequality imply (3.43) for $p = 2$, $\tau = \frac{1}{2}$. To deal with the general case $p \geqq 2$, $\tau \in (0,1)$ as in the assumptions of this corollary, we compute in (3.40) the div-terms, multiply by $w := -|\Delta v'(s)|^{p-2}\Delta v'(s)$, integrate over G parts, use the boundedness of $\hat{k}$ and $\hat{\gamma}$ as implied by (H1), (H2) and apply Hölder's inequality to arrive at

$$|\Delta v'(s)|_p \leqq c\Big\{|v'(s)|_p + |g|_p + |\nabla\hat{k} \cdot \nabla g|_p + |\hat{k}_{\Delta g}|_p$$
$$+ |\hat{\gamma}\Delta v|_p + |\nabla\hat{\gamma} \cdot \nabla v|_p + |\nabla\hat{k} \cdot \nabla v'|_p\Big\} |\Delta v'(s)|_p. \tag{3.48'}$$

By (3.42),

$$|v'(s)|_p \leqq c(G)|v'|_{C(S_t, H_0^1(G))} \leqq c\{\|v_0\| + \|g\|_{C(S_t, H_0^1(G))}\},$$

so that by (3.46) and the embedding theorems

$$|v'(s)|_p \leqq c.$$

Furthermore, $|\nabla g|_{a'p} \leqq c\|g\|_{2,p}$ (we have $N=2$, $p \geqq 2!$), so that by (3.46) $|\nabla g|_{a'p} \leqq c$ (indep. of g). By Sobolev's inequalities $|\nabla v'|_{a'p} \leqq c(G,a',p,p)|\nabla v'|_2^{1-\theta}\|v'\|_{2,p}^{\theta}$ (θ defined in the assumptions) $|v|_{a'p} \leqq c(G,a',p,p)|\nabla v'|_2^{1-\theta}\|v'\|_{2,p}^{\theta}$,

$$\begin{aligned}|\nabla v'(s)|_2 + |\nabla v(s)|_2 &\leqq c\left\{\|v_0\| + \|g\|_{C(S_t,H_0^1(G))}\right\} \quad \text{by (3.42)}\\ &\leqq c \quad \text{by (3.46)}.\end{aligned}$$

Finally, by (3.44), $|\nabla \hat{k}\|_{ap} + |\nabla \hat{\gamma}|_{ap} \leqq c\|u_\delta\|_{1,ap} \leqq c\|u_\delta\|_{1+\tau,p^*}$. Therefore, (3.47), (3.48′) and Young's inequality imply

$$\begin{aligned}\|v'(s)\|_{2,p} &\leqq c\left\{1 + \|u_\delta\|_{1+\tau,p^*}^{1/(1-\theta)} + \|v(s)\|_{2,p}\right\}\\ &\leqq c\left\{1 + \|u_\delta\|_{1+\tau,p^*}^{1/(1-\theta)} + \|v_0\|_{2,p} + \int_0^s |v'(r)|_{2,p}\,dr\right\}.\end{aligned}$$

Set for abbreviation $V := W_0^{1+\tau,\tau^*}(G)$. Gronwall's inequality and (3.46) show that

$$\begin{aligned}&|v'|_{C(S_t,W^{2,p}(G))} \leqq c\left\{1 + \|u_\delta\|_{C(S_t,V)}^{1/(1-\theta)}\right\}\cdot\exp(cT),\\ &\text{where } c = c\left(\varepsilon k_0, k_1, L_\alpha, |G|, T, p, \theta, \tau, p^*, |k'|_\infty, |\alpha''|_\infty\right),\end{aligned} \tag{3.48}$$

which implies (3.43). Lemma 3.9 is proved. □

To show (3.9), we first notice, that by (3.37), (3.41)–(3.43) and by $\theta + \tau \leqq 1$,[4]

$$\begin{aligned}\|u'_\delta\|_{C(S_t,V)} &\leqq c\left\{1 + \|u_\delta\|_{C(S_t,V)}^{1/(1-\theta)}\right\}^{\tau}\left\{\|u_{0\delta}\|_V + \|f_\delta\|_{C(S_t,V)}\right\}\\ &\leqq c\left\{1 + \|u_\delta\|_{C(S_t,V)}\right\}\left\{\|u_0\|_V + \|f\|_{C(S,V)}\right\}\\ &\leqq c\left\{1 + \|u_\delta\|_{C(S_t,V)}\right\}\\ &\leqq c\left\{1 + \|u_{0\delta}\|_V + \int_0^t \|u'_\delta(s)\|_V\,ds\right\}.\end{aligned}$$

Gronwall's inequality and (3.37) yield (3.39). Therefore, problem (3.40) with $u_{0\delta}$, f_δ as data has u_δ as a solution, which by (3.39) satisfies

$$\|u_\delta\|_{W^{1,\infty}(S,V)} \leqq \text{const}.$$

The usual compactness argument yields a (sub-) sequence

$$u_\delta \overset{*}{\rightharpoonup} u \quad \text{in } W^{1,\infty}(S,V).$$

By means of the approximate equation (3.40) one easily shows that u satisfies (2.1), (2.2). This completes the proof of Corollary 2.4. □

Proof of Corollary 2.5. (a) follows directly from Corollary 2.4.

(b) set $a := q/p$ and apply (a). Then, $u \in C^1(S, W_0^{1+\tau,p^*}(G))$ and by (3.44), $u \in C^1(S, W_0^{1,q}(G))$. □

Proof of Theorem 2.7. Let u_i be a solution of (2.1), (2.2) with respect to the initial value u_{0i} and the right-hand side f_i, $i = 1, 2$. Set for abbreviation $f := f_1 - f_2$, $w := u_1 - u_2$,

[4] $V := W_0^{1+\tau,p^*}(G)$, $\|\cdot\|_V := \|\cdot\|_{1+\tau,p^*}$.

$w_0 := u_{01} - u_{02}$ and assume

$$u_2 \in W^{1,1}(S, W_0^{1,p}(G)), \quad u_1 \in W^{1,1}(S, H_0^1(G)),$$
$$f_2 \in L^1(S, W^{1,p}(G)), \quad f_1 \in L^1(S, H_0^1(G)).$$

Subtract (2.2) for u_2 from that for u_1, choose w' as a test-function in the variational formulation, integrate over G and make some rearrangements. Thus, for $t \in S$

$$\begin{aligned}
&|w'(t)|^2 + \varepsilon k_0 |\nabla w'(t)|^2 \\
&\quad \leqq |w'(t)|^2 + \varepsilon(k(u_1)\nabla w', \nabla w') \\
&\quad = (f, w') + \varepsilon(k(u_1)\nabla f, \nabla w') + \varepsilon((k(u_1) - k(u_2))\nabla f_2, \nabla w') \\
&\qquad + \varepsilon((k(u_1) - k(u_2))\nabla u_2', \nabla w') + (\gamma(u_1)\nabla w, \nabla w') \\
&\qquad + ((\gamma(u_1) - \gamma(u_2))\nabla u_2, \nabla w').
\end{aligned}$$

By using the Lipschitz continuity of k and γ and applying Hölder's inequality, we obtain

$$\begin{aligned}
&|w'(t)|^2 + \varepsilon k_0 |\nabla w'(t)|^2 \\
&\leqq |f||w'| + \left\{ \varepsilon k_1 \|f\| + \varepsilon L_k |w|_{p'} \|f_2\|_{1,p} + \varepsilon L_k |w|_{p'} \|u_2'\|_{1,p} + |\gamma|_\infty \|w\| + L_\gamma \|u_2\|_{1,p} |w|_{p'} \right\} \|w'(t)\|.
\end{aligned}$$

By the imbedding theorems, $|w|_{p'} \leqq c(G)\|w\|$. Also, $\|w(t)\|^r \leqq \|w_0\|^r + c\int_0^t \|w(r)\|^r dr$, if $r \in [1, \infty)$, with $c = c(r, T)$. This and Young's inequality imply for $\bar{r} := r$ if $r < \infty$, $\bar{r} = 1$ if $r = \infty$,

$$\psi := \|f_2\|_{1,p}^{\bar{r}} + \|u_r'\|_{1,p}^{\bar{r}} + \|u_2\|_{1,p}^{\bar{r}},$$
$$\|w'(t)\|^r \leqq \varepsilon k_1 \|f(t)\|^r = c\psi(t) \left\{ \|w_0\|^r + \int_0^t \|w'(s)\|^r ds \right\}.$$

Integration and Gronwall's inequality yield

$$\|w'\|_{L^r(S, H_0^1(G))} \leqq c\left\{ \|f\|_{L^r(S, H_0^1(G)} + \|w_0\| \cdot |\psi|_{L^1(S)} \right\} \cdot \exp\left\{ c|\psi|_{L^1(S)} \right\}.$$

For $r := \infty$ we obtain

$$\|w'\|_{L^\infty(S, H_0^1(G))} \leqq c\left\{ \|f\|_{L^\infty(S, H_0^1(G))} + \|w_0\| \cdot |\psi|_{L^\infty(S)} \right\} \cdot \exp\left\{ c|\psi|_{L^\infty(S)} \right\}.$$

This proves (b) of the theorem, and (a) follows by setting $u_{01} = u_{02}, f_1 = f_2$. □

Proof of Theorem 2.8. First of all, notice that $\hat{k}(t) := (u(t, \cdot)) \in L^\infty(G)$, $\hat{k}(t, x) \geqq k_0 > 0$ for a.a. $t \in S$, $x \in G$. Thus, $A(t) := -\operatorname{div}(\hat{k}(t)\nabla(\cdot))$ is an m-accretive operator on $L^1(G)$. Now, [6, Thm. 1] yields the desired estimates. If $u_0 \geqq 0, f(T) \geqq 0$, then obviously, $|u(t)|_{L^\infty(G)} = 0$, i.e., $u(t, s) \geqq 0$ a.e. □

Appendix. We list some facts on complex interpolation of B-spaces and, in particular, Sobolev spaces. Basic references are Bergy–Löfstrom [3], Lions–Magenes [13], and Triebel [18].

Let $\tau \in [0,1]$ be a parameter, $\bar{A} := \{A_0, A_1\}$, $\bar{B} := \{B_0, B_1\}$—two couples of compatible B-spaces "$[\cdot,\cdot]_\tau$" denotes the interpolation functor for the complex interpolation method (cf. [3], [13], [18]), $\bar{A}_\tau := [A_0, A_1]_\tau$. We have

LEMMA A1. $[A_0 \times B_0, A_1 \times B_1]_\tau = [A_0, A_1]_\tau \times [B_0, B_1]_\tau$ *(algebraically and topologically).*

LEMMA A2. *Let* $\mathscr{P} \in \mathscr{L}(A_i, B_i)$ *with norm* M_i, $i=0,1$. *Then* $\mathscr{P} \in \mathscr{L}(\bar{A}_\tau, \bar{B}_\tau)$ *and* $|\mathscr{P}|_{\mathscr{L}(\bar{A}_\tau, \bar{B}_\tau)} \leqq M_0^{1-\tau} M_1^\tau$.

LEMMA A3. $[C(S, A_0), C(S, A_1)]_\tau = C(S, \bar{A}_\tau)$ *(algebraically and topologically).*

LEMMA A4. (a) *Let* $s_i \in \mathbb{R}$, $p_i \in (1,\infty)$, $\tau \in (0,1)$, $s^* := (1-\tau)s_0 + \tau s_2$,

$$\frac{1}{p^*} = \frac{1-\tau}{p_0} + \frac{\tau}{p_1},$$

$s_0 \neq s_1$, $G \subset \mathbb{R}^N$ *a (bounded) domain. Then* $[W^{s_0,p_0}(G), W^{s_1,p_1}(G)]_\tau = W^{s^*,p^*}(G)$.

(b) *Let*

$$p \in (1,\infty), \qquad \tau \in (0,1), \quad \frac{1}{p} = \frac{1-\tau}{2} + \frac{\tau}{p}.$$

Then $[W_0^{1,2}(G), W^{2,p}(G) \cap W_0^{1,p}(G)]_\tau = W_0^{1+\tau,\bar{p}}(G)$.

LEMMA A5. *Let* τ, p *and* $\bar{p}$ *be as in Lemma* A4(b), $S := [0,T]$ *an interval of any finite length,*

$\mathscr{P}_1\colon (W^{2,p}(G) \cap W_0^{1,p}(G)) \times C(S, W^{2,p}(G)) \to C(S, W^{2,p}(G) \cap W_0^{1,p}(G))$
a linear and bounded operator with norm M_1,

$\mathscr{P}_0$ *a linear and bounded extension of* $\mathscr{P}_1$, *such that*
$\mathscr{P}_0\colon W_0^{1,2}(G) \times C(S, W^{1,2}(G)) \to C(S, W_0^{1,2}(G))$.

Then

$\mathscr{P}_\tau :=$ *restriction of* $\mathscr{P}_0$ *to* $W_0^{1+\tau,\bar{p}}(G) \times C(S, W^{1+\tau,\bar{p}}(G))$ *maps continuously into* $C(S, W_0^{1+\tau,\bar{p}}(G))$.

Let M_0 *denote the norm of* $\mathscr{P}_0$, M_τ *that of* $\mathscr{P}_\tau$. *Then*

$$M_\tau \leqq M_0^{1-\tau} M_1^\tau.$$

Note that this estimate does not depend on S.

References/proofs. Lemma A1 follows from [3, Thm. 4.1.2]. Lemma A2 is a special case of a theorem in [11]. Lemma A3 can be found in [13], if A_1, A_0 are Hilbert spaces, otherwise cf. [10]. Lemma A4(a)—cf. [3, Thm. 6.4.5], (b)—cf. [18, Thm. 4.3.3]. Lemma A5 follows from Lemmas A1–A4.

Lemma A6. [3, Thm. 6.5.1]. Let $s_i \in \mathbb{R}$, $p_i > 1$, $i=1,2$. If $s_1 - N/p_1 \geqq s_2 - N/p_2$, then $W^{s_1,p_1}(G) \subset W^{s_2,p_2}(G)$.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] G. I. BARENBLATT, I. P. ZHELTOV AND I. N. KOCHINA, *Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks*, J. Appl. Math. Mech., 24 (1960), pp. 1286–1303.

[3] J. BERGH AND J. LÖFSTROM, *Interpolation Spaces—an Introduction*, Grundlehren 223, Springer, Berlin, 1976.

[4] M. BÖHM, *On some nonlinear evolution equations*, Seminar–Bericht, Sektion Mathematik an der Humboldt Universität zu Berlin, Berlin.

[5] ———, *On a non-homogeneous non-Newtonian fluid*, LCDS Report #83-8, Brown University, Providence, RI.

[6] M. BÖHM AND R. E. SHOWALTER, *Diffusion in fissured media*, to appear.

[7] R. W. CARROLL AND R. E. SHOWALTER, *Singular and Degenerate Cauchy Problems*, Academic Press, New York, 1976.

[8] P. J. CHEN AND M. E. GURTIN, *On a theory of heat conduction involving two temperatures*, Z. Angew. Math. Phys., 19 (1968), pp. 614–627.

[9] R. HUIGOL, *A second order fluid of the differential type*, J. Non-linear Mech., 3 (1968), pp. 471–482.

[10] H. KOMATSU, *Fractional powers of operators*, VI, *Interpolation of non-negative operators*, J. Fac. Sci. Univ. Tokyo, 19 (1972), pp. 1–63.

[11] S. G. KREIN AND JU. I. PETUNIN, *Scales of Banach Spaces*, Uspekhi Mat. Nauk, 21 (1966), pp. 89–168. (In Russian.)

[12] O. A. LADYZHENSKAYA AND N. N. URALT'SEVA, *Linear and Quasilinear Elliptic Equations*, Nauka, Moscow (1964); English transl., Academic Press, New York, 1968.

[13] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, *Vol.* I, Springer, Berlin, 1976.

[14] R. E. SHOWALTER, *Existence and representation theorems for a semi-linear Sobolev equation in Banach Space*, this Journal, 3 (1972), pp. 527–543.

[15] R. E. SHOWALTER, *Two Stefan problems for a degenerate parabolic system*, in Proc. International Conference on Dynamical Systems, L. Cesari and A. R. Bednarek, eds., Academic Press, New York, 1982.

[16] G. STAMPACCHIA, *Le problème de Dirichlet pour les èquations elliptiques du second ordre*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.

[17] T. W. TING, *Certain non-steady flows of second order fluids*, Arch. Rat. Mech. Anal., 14 (1963), pp. 1–26.

[18] H. TRIEBEL, *Interpolation Theory*, *Function Spaces*, *Differential Operators*, North-Holland Math. Library, vol. 18, North-Holland, Amsterdam, 1978.

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

ON THE SCHRÖDINGER SINGULAR PERTURBATION PROBLEM*

H. O. FATTORINI[†]

Abstract. The Schrödinger singular perturbation problem arises in quantum mechanics and consists of approximating certain solutions of the Klein–Gordon equations by solutions of the Schrödinger equation in functions of a small parameter. We consider here an abstract version for operator equations in Banach spaces.

Key words. singular perturbation, operator equations, Schrödinger equation

1. Introduction. Let A be a densely defined linear operator in a Banach space E. We consider the abstract Cauchy problems

$$\varepsilon^2 u''(t;\,\varepsilon) - iu'(t;\,\varepsilon) = Au(t;\,\varepsilon) + f(t;\,\varepsilon) \qquad (-\infty < t < \infty), \tag{1.1}$$

$$u(0;\,\varepsilon) = u_0(\varepsilon), \qquad u'(0;\,\varepsilon) = u_1(\varepsilon), \tag{1.2}$$

and

$$u'(t) = iAu(t) + if(t) \qquad (-\infty < t < \infty), \tag{1.3}$$

$$u(0) = u_0. \tag{1.4}$$

The *Schrödinger singular perturbation problem* is that of showing that

$$u(t;\,\varepsilon) \to u(t) \tag{1.5}$$

as $\varepsilon \to 0$, where convergence in (1.5) can be understood in various senses. This type of problem arises in relativistic quantum mechanics (see [11] for details), where one considers functions of the form

$$u(x,t) = v(x,t)\exp(i\mathrm{m}c^2 t/h),$$

where v is a solution of the Klein–Gordon equation for a free particle

$$-h^2 v_{tt} = -h^2 c^2 \Delta v + \mathrm{m}^2 c^4 v$$

(here m is the mass of the particle, h Planck's constant and c the speed of light). It follows that $u(x,t)$ satisfies the equation

$$\frac{h}{2\mathrm{m}c^2} u_{tt} - iu_t = \frac{h}{2\mathrm{m}} \Delta u,$$

and thus an abstract differential equation of the form (1.1) (with $f(t;\,\varepsilon) = 0$) is obtained; since $h/2\mathrm{m}c^2 \ll 1$ we expect u to be an approximation of the solution of the Schrödinger equation

$$-ihu_t = \frac{h^2}{2\mathrm{m}} \Delta u.$$

The Schrödinger singular perturbation problem for the abstract initial value problem (1.1)–(1.2) was considered by Veselić (see [16], [17] and references therein) and by

*Received by the editors September 27, 1983, and in revised form July 5, 1984. This work was supported in part by the National Science Foundation under grant MCS 82-00645.

[†] Department of Mathematics, University of California, Los Angeles, California 90024.

Schoene [11]. Schoene considers the case where A is a self-adjoint negative definite operator in a Hilbert space H. In this situation the solutions of (1.1) and (1.3) can be explicitly computed by means of the functional calculus for self-adjoint operators, and the proof of (1.5) can thus be essentially reduced to that for the scalar case $H=\mathbb{C}$, $A=\lambda$ ($\lambda<0$). Using the author's results in [4], Schoene's theory can be extended to the case where A generates a strongly continuous cosine function $\mathscr{C}(t)$ uniformly bounded in $-\infty<t<\infty$ (see §3). However, no results for general A exist. We develop here a theory that is sufficiently inclusive to handle the second order partial differential operator

$$Au=\sum_{j=1}^{m}\sum_{k=1}^{m}a_{jk}(x)D^jD^ku+\sum_{j=1}^{m}b_j(x)D^ju+c(x)u. \tag{1.6}$$

To gain insight into the Schrödinger singular perturbation problem it is useful to compare it with the *parabolic singular perturbation problem*

$$\varepsilon^2u''(t;\varepsilon)+u'(t;\varepsilon)=Au(t;\varepsilon)+f(t;\varepsilon)\quad (t\geqq 0), \tag{1.7}$$

$$u(0;\varepsilon)=u_0(\varepsilon),\qquad u'(t;\varepsilon)=u_1(\varepsilon), \tag{1.8}$$

where the limiting initial value problem is

$$u'(t)=Au(t)\quad (t\geqq 0), \tag{1.9}$$

$$u(0)=u_0. \tag{1.10}$$

A *solution* of (1.1) or (1.7) is a twice continuously differentiable E-valued function such that $u(t)\in D(A)$ and the equation is satisfied; solutions of (1.3) and (1.9) are accordingly defined. The natural assumption on both initial value problems (1.1)–(1.2) and (1.7)–(1.8) is that solutions should exist for arbitrary $u_0(\varepsilon)$, $u_1(\varepsilon)$ in a dense subspace $D\subseteq E$. Moreover, arbitrary solutions should depend continuously on their initial values in the sense that

$$\|u(t;\varepsilon)\|\leqq C(T;\varepsilon)\big(\|u(0;\varepsilon)\|+\|u'(0;\varepsilon)\|\big)\quad (0\leqq t\leqq T). \tag{1.11}$$

It is easy to see by means of a change of variable that this assumption, for any of the two initial value problems, holds if and only if it holds for the second order initial value problem

$$u''(t)=Au(t), \tag{1.12}$$

$$u(0)=u_0,\qquad u'(0)=u_1. \tag{1.13}$$

The *propagators* or *solution operators* of (1.12), (1.13) are defined in the subspace D of initial data by $\mathscr{C}(t)u_0=u(t)$, $\mathscr{S}(t)u_1=v(t)$, where $u(\cdot)$ (resp. $v(\cdot)$) is the solution of (1.12), (1.13) with $u(0)=u_0$, $u'(0)=0$ (resp. $v(0)=0$, $v'(0)=u_1$), and extended by continuity to all of E; both are strongly continuous in $-\infty<t<\infty$ with

$$\mathscr{S}(t)u=\int_0^t\mathscr{C}(s)u\,ds. \tag{1.14}$$

Moreover, there exist constants C_0, ω such that

$$\|\mathscr{C}(t)\|\leqq C_0e^{\omega|t|}\quad (-\infty<t<\infty). \tag{1.15}$$

For additional details see [2, p. 86]; we note that all the properties asked of the second order initial value problem (1.12)–(1.13) are equivalent to the requirement that A should generate a *strongly continuous cosine function* $\mathscr{C}(t)$ (see [2, p. 91]).

The change of variable $v(t;\,\varepsilon)=e^{t/2\varepsilon}u(\varepsilon t;\,\varepsilon)$ (see [6, p. 531]) produces the (unique) solution of (1.7)–(1.8) in terms of the propagator $\mathscr{C}(t)$ of (1.12)–(1.13):

$$\begin{aligned}
u(t;\,\varepsilon)&=e^{-t/2\varepsilon^2}\mathscr{C}\left(\frac{t}{\varepsilon}\right)u_0(\varepsilon)\\
&\quad+\frac{te^{-t/2\varepsilon^2}}{\varepsilon^2}\int_0^{t/\varepsilon}\frac{I_1\Big(\big((t/\varepsilon^2)-s^2\big)^{1/2}/2\varepsilon\Big)}{\big((t/\varepsilon)^2-s^2\big)^{1/2}}\mathscr{C}(s)\left(\frac{1}{2}u_0(\varepsilon)\right)ds\\
&\quad+\frac{e^{-t/2\varepsilon^2}}{\varepsilon}\int_0^{t/\varepsilon}I_0\left(\frac{\big((t/\varepsilon)^2-s^2\big)^{1/2}}{2\varepsilon}\right)\mathscr{C}(s)\left(\frac{1}{2}u_0(\varepsilon)+\varepsilon^2u_1(\varepsilon)\right)ds\\
&\quad+e^{-t/2\varepsilon^2}\int_0^{t/\varepsilon}\left(\int_0^{t/\varepsilon-s}I_0\left(\frac{\big((t/\varepsilon-s)^2-\sigma^2\big)^{1/2}}{2\varepsilon}\right)\mathscr{C}(\sigma)\,d\sigma\right)e^{s/2\varepsilon}f(\varepsilon s;\,\varepsilon)\,ds\\
&=e^{-t/2\varepsilon^2}\mathscr{C}\left(\frac{t}{\varepsilon}\right)u_0(\varepsilon)+\mathfrak{K}(t;\,\varepsilon)\left(\frac{1}{2}u_0(\varepsilon)\right)\\
&\quad+\mathfrak{S}(t;\,\varepsilon)\left(\frac{1}{2}u_0(\varepsilon)+\varepsilon^2u_1(\varepsilon)\right)+\int_0^t\mathfrak{S}(t-s;\,\varepsilon)f(s;\,\varepsilon)\,ds\\
&=\mathfrak{C}(t;\,\varepsilon)u_0(\varepsilon)+\mathfrak{S}(t;\,\varepsilon)\big(\varepsilon^2u_1(\varepsilon)\big)+\int_0^t\mathfrak{S}(t-s;\,\varepsilon)f(s;\,\varepsilon)\,ds
\end{aligned}\tag{1.16}$$

(see [18, p. 77] for the definition of the Bessel functions I_0, I_1). Formula (1.16), which *defines* the operator functions $\mathfrak{K}$, $\mathfrak{C}$ and $\mathfrak{S}$ provides a solution of (1.7)–(1.8) if $u_0(\varepsilon)$, $u_1(\varepsilon)\in D(A)$; in general, $u(t;\,\varepsilon)$ will only be a weak or generalized solution (see [6, p. 533]). A solution (with the same reservations) of (1.1)–(1.2) can be obtained taking $u(it;\,i\varepsilon)$, where $u(t;\,\varepsilon)$ is the function given by (1.16). Noting that $I_0(-ix)=I_0(ix)=J_0(x)$ and that $I_1(-ix)=-I_1(ix)=-iJ_1(x)$ (see [18, pp. 15, 77]), we obtain the following expression for the solution of (1.1)–(1.2), where the operators $\mathfrak{K}_i$, $\mathfrak{C}_i$ and $\mathfrak{S}_i$ are defined:

$$\begin{aligned}
u(t;\,\varepsilon)&=e^{it/2\varepsilon^2}\mathscr{C}\left(\frac{t}{\varepsilon}\right)u_0(\varepsilon)\\
&\quad-\frac{te^{it/2\varepsilon^2}}{\varepsilon}\int_0^{t/\varepsilon}\frac{J_1\Big(\big((t/\varepsilon)^2-s^2\big)^{1/2}/2\varepsilon\Big)}{\big((t/\varepsilon)^2-s^2\big)^{1/2}}\mathscr{C}(s)\left(\frac{1}{2}u_0(\varepsilon)\right)ds\\
&\quad-\frac{ie^{it/2\varepsilon^2}}{\varepsilon}\int_0^{t/\varepsilon}J_0\left(\frac{\big((t/\varepsilon)^2-s^2\big)^{1/2}}{2\varepsilon}\right)\mathscr{C}(s)\left(\frac{1}{2}u_0(\varepsilon)+i\varepsilon^2u_1(\varepsilon)\right)ds\\
&\quad+e^{it/2\varepsilon^2}\int_0^{t/\varepsilon}\left(\int_0^{t/\varepsilon-s}J_0\left(\frac{\big((t/\varepsilon-s)^2-\sigma^2\big)^{1/2}}{2\varepsilon}\right)\mathscr{C}(\sigma)\,d\sigma\right)e^{is/2\varepsilon}f(\varepsilon s;\,\varepsilon)\,ds\\
&=e^{it/2\varepsilon^2}\mathscr{C}\left(\frac{t}{\varepsilon}\right)u_0(\varepsilon)+\mathfrak{K}_i(t;\,\varepsilon)\left(\frac{1}{2}u_0(\varepsilon)\right)-i\mathfrak{S}_i(t;\,\varepsilon)\left(\frac{1}{2}u_0(\varepsilon)+i\varepsilon^2u_1(\varepsilon)\right)\\
&\quad+\int_0^t\mathfrak{S}_i(t-s;\,\varepsilon)f(s;\,\varepsilon)\,ds\\
&=\mathfrak{S}_i(t;\,\varepsilon)u_0(\varepsilon)+\mathfrak{S}_i(t;\,\varepsilon)\big(\varepsilon^2u_1(\varepsilon)\big)+\int_0^t\mathfrak{S}_i(t-s;\,\varepsilon)f(s;\,\varepsilon)\,ds.
\end{aligned}\tag{1.17}$$

The obvious difference between the representation (1.16) for (generalized) solutions of the initial value problem (1.7)–(1.8) and the analogous representation (1.17) for the initial value problem (1.1)–(1.2) lies in the different asymptotic behavior of the integrands. In formula (1.16) we can combine the asymptotic estimate $|I_\nu(x)|=O(x^{-1/2}e^x)$ as $x\to\infty$ ([18, p. 203]) with the decreasing factor $e^{-t/2\varepsilon^2}$ and the bound (1.15) for $\|\mathscr{C}(\cdot)\|$ to show that we can take limits directly as $\varepsilon\to 0$ using the dominated convergence theorem: the results are uniform for bounded $\|u_0(\varepsilon)\|$, $\|\varepsilon^2 u_1(\varepsilon)\|$ and can be expressed as follows:

$$(1.18)\qquad \mathfrak{C}(t;\varepsilon)\to S(t),\qquad \mathfrak{S}(t;\varepsilon)\to S(t),$$

where $S(t)$ is the semigroup generated by A. This semigroup is given by the *abstract Weierstrass formula*

$$(1.19)\qquad S(t)u=\frac{1}{(\pi t)^{1/2}}\int_0^\infty e^{-s^2/4t}\mathscr{C}(s)u\,ds,$$

which shows, among other things, that $S(t)$ can be extended to a semigroup $S(\zeta)$ analytic in $\operatorname{Re}\zeta>0$. Convergence in (1.18) is in the uniform topology of operators, uniformly on compacts of $t\geq 0$ outside of an "initial layer" of order $t(\varepsilon)$ with $t(\varepsilon)/\varepsilon^2\to\infty$ (see [6, p. 539]). In contrast, the asymptotic estimate $|J_\nu(x)|=O(x^{-1/2})$ combined with the indifferent factor e^{it/ε^2} and (1.15) do not allow direct passage to the limit; in fact, doing so formally, one can only hope for a relation of the type of (1.18) for $\mathfrak{C}_i$, $\mathfrak{S}_i$ with limit

$$(1.20)\qquad S_i(t)u=S(it)u=\frac{1}{(\pi it)^{1/2}}\int_0^\infty e^{is^2/4t}\mathscr{C}(s)u\,ds,$$

which does not make sense even in the most favorable case where $\|\mathscr{C}(t)\|$ is uniformly bounded in $-\infty<t<\infty$. Hence, computation of the limit (1.5) will have to be carried out by indirect means. In particular, and in contrast with the parabolic singular perturbation problem, the existence of the group $S_i(t)=\exp(iAt)=S(it)$ is not assured by the existence-uniqueness assumption for (1.1)–(1.2) above but will follow from the stronger assumptions in §2. (Note that $S_i(t)$ is the "boundary value" of the analytic semigroup $S(\zeta)$.) As may be expected, the convergence results for the operators $\mathfrak{C}_i(t;\varepsilon)$ and $\mathfrak{S}_i(t;\varepsilon)$, which will be called in what follows the *propagators* or *solution operators* of (1.1)–(1.2), are different in character from those for $\mathfrak{C}(t;\varepsilon)$, $\mathfrak{S}(t;\varepsilon)$; in particular, there is no analogue of (1.18) and $\mathfrak{S}_i(t;\varepsilon)$ is not even strongly convergent. Differences will be pointed out below as they arise.

2. Assumptions on the initial value problem. We shall require that (1.1)–(1.2) (or, equivalently, the initial value problem (1.12)–(1.13)) satisfy the existence—uniqueness assumptions of §1, or, equivalently, that A generate a strongly continuous cosine function. An additional assumption will be needed:

Assumption 2.1. Let $\mathfrak{C}_i(t;\varepsilon)$, $\mathfrak{S}_i(t;\varepsilon)$ be the propagators of (1.1), (1.2) (see §1). Then there exist constants C_0, C_1, ω independent of t and ε such that

$$(2.1)\qquad \|\mathfrak{C}_i(t;\varepsilon)\|\leqq C_0e^{\omega|t|},\qquad \|\mathfrak{S}_i(t;\varepsilon)\|\leqq C_1e^{\omega|t|}\qquad(-\infty<t<\infty,\,0\leqq\varepsilon\leqq\varepsilon_0).$$

Assumption 2.1 can be given a simpler form as follows. Let $u(t)$ be a solution of (1.1)–(1.2) with $f(t;\varepsilon)=0$. Set $v(t)=e^{-it/2\varepsilon}u(\varepsilon t)$ or, equivalently, $u(t)=e^{it/2\varepsilon^2}v(t/\varepsilon)$.

Then $v(t)$ satisfies

$$v''(t)=\left(A-\frac{1}{4\varepsilon^2}I\right)v(t) \qquad (-\infty<t<\infty), \tag{2.2}$$

$$v(0)=u_0(\varepsilon), \qquad v'(0)=-\frac{i}{2\varepsilon}u_0(\varepsilon)+\varepsilon u_1(\varepsilon). \tag{2.3}$$

Accordingly, $u(t)$ can be written in terms of the propagators $\mathscr{C}(t;\ A-(2\varepsilon)^{-2}I)$, $\mathscr{S}(t;\ A-(2\varepsilon)^{-2}I)$ of the initial value problem (2.2)–(2.3):

$$\begin{aligned} u(t)&=e^{it/2\varepsilon^2}\mathscr{C}\left(\frac{t}{\varepsilon};\ A-(2\varepsilon)^{-2}I\right)u_0(\varepsilon) \\ &\quad+e^{it/2\varepsilon^2}\mathscr{S}\left(\frac{t}{\varepsilon};\ A-(2\varepsilon)^{-2}I\right)\left(-\frac{i}{2\varepsilon}u_0(\varepsilon)+\varepsilon u_1(\varepsilon)\right). \end{aligned} \tag{2.4}$$

(We note, incidentally, that formula (1.17) for solving (1.1)–(1.2) in terms of the propagators of (1.12)–(1.13) can be obtained from (2.4) using the "numerical perturbation" formulas in [14], although this will play no role in what follows.) Using (2.4), we deduce that

$$\mathfrak{C}_i(t;\ \varepsilon)=e^{it/2\varepsilon^2}\mathscr{C}\left(\frac{t}{\varepsilon};\ A-(2\varepsilon)^{-2}I\right)-\frac{i}{2\varepsilon}e^{it/2\varepsilon^2}\mathscr{S}\left(\frac{t}{\varepsilon};\ A-(2\varepsilon)^{-2}I\right), \tag{2.5}$$

$$\mathfrak{S}_i(t;\ \varepsilon)=\varepsilon^{-1}e^{it/2\varepsilon^2}\mathscr{S}\left(\frac{t}{\varepsilon};\ A-(2\varepsilon)^{-2}I\right), \tag{2.6}$$

or

$$\mathscr{C}\left(t;\ A-(2\varepsilon)^{-2}I\right)=e^{-it/2\varepsilon}\mathfrak{C}_i(\varepsilon t;\ \varepsilon)-\frac{i}{2}e^{-it/2\varepsilon}\mathfrak{S}_i(\varepsilon t;\ \varepsilon), \tag{2.7}$$

$$\mathscr{S}\left(t;\ A-(2\varepsilon)^{-2}I\right)=\varepsilon e^{-it/2\varepsilon}\mathfrak{S}_i(\varepsilon t;\ \varepsilon). \tag{2.8}$$

Accordingly, Assumption 2.1 will hold if and only if

$$\left\|\mathscr{C}\left(t;\ A-(2\varepsilon)^{-2}I\right)\right\|\leqq C_0'e^{\omega\varepsilon|t|}, \quad \left\|\mathscr{S}\left(t;\ A-(2\varepsilon)^{-2}I\right)\right\|\leqq C_1\varepsilon e^{\omega\varepsilon|t|} \qquad (-\infty<t<\infty,\ 0<\varepsilon\leqq\varepsilon_0). \tag{2.9}$$

We shall examine in §4 a class of operators satisfying Assumption 2.1; as seen in §5, these operators include the partial differential operator (1.6) under adequate assumptions on the coefficients.

We point out below a consequence of Assumption 2.1.

THEOREM 2.2. *Let the operator A satisfy Assumption* 2.1. *Then iA generates a strongly continuous group $S_i(\cdot)$ such that*

$$\|S_i(t)\|\leqq C_0e^{\omega|t|} \qquad (-\infty<t<\infty), \tag{2.10}$$

with C_0 and ω the constants in (2.1).

Proof. If $u\in D(A^2)$ then it follows from formula (1.17) (or from the considerations at the beginning of this section) that $\mathfrak{C}_i(t;\ \varepsilon)u$ is four times continuously differentiable, $\mathfrak{C}_i''(t;\ \varepsilon)u\in D(A)$ and

$$\varepsilon^2\mathfrak{C}_i''''(t;\ \varepsilon)u-i\mathfrak{C}_i''(t;\ \varepsilon)u=A\mathfrak{C}_i''(t;\ \varepsilon)u \qquad (-\infty<t<\infty). \tag{2.11}$$

On the other hand, we have

$$\mathfrak{C}_i''(0;\,\varepsilon)u=\varepsilon^{-2}\big(A\mathfrak{C}_i(0;\,\varepsilon)u+i\mathfrak{C}_i'(0;\,\varepsilon)u\big)=\varepsilon^{-2}Au,$$
$$\mathfrak{C}_i'''(0;\,\varepsilon)u=\varepsilon^{-2}\big(A\mathfrak{C}_i'(0;\,\varepsilon)u+i\mathfrak{C}_i''(0;\,\varepsilon)u\big)=i\varepsilon^{-4}Au,$$

thus it follows from (1.17) that

$$\mathfrak{C}_i''(t;\,\varepsilon)u=\varepsilon^{-2}\mathfrak{C}_i(t;\,\varepsilon)Au+\varepsilon^{-2}\mathfrak{S}_i(t;\,\varepsilon)Au.$$

Accordingly, it follows from Assumption 2.2 that

$$\|\mathfrak{C}_i''(t;\,\varepsilon)u\|\leqq C_0\varepsilon^{-2}e^{\omega|t|}\|Au\| \qquad (-\infty<t<\infty,\, 0<\varepsilon\leqq\varepsilon_0). \tag{2.12}$$

We take now a sequence $\{\varepsilon_n\}$, $1>\varepsilon_1>\varepsilon_2>\cdots>0$ to be specified later. For $u\in D(A)$ we have

$$\varepsilon_n^2\big(\mathfrak{C}_i''(t;\,\varepsilon_n)u-\mathfrak{C}_i''(t;\,\varepsilon_{n+1})u\big)-i\big(\mathfrak{C}_i'(t;\,\varepsilon_n)u-\mathfrak{C}_i'(t;\,\varepsilon_{n+1})u\big)$$
$$=A\big(\mathfrak{C}_i(t;\,\varepsilon_n)u-\mathfrak{C}_i(t;\,\varepsilon_{n+1})u\big)+\big(\varepsilon_n^2-\varepsilon_{n+1}^2\big)\mathfrak{C}_i''(t;\,\varepsilon_{n+1})u.$$

Using (1.17) again,

$$\mathfrak{C}_i(t;\,\varepsilon_n)u-\mathfrak{C}_i(t;\,\varepsilon_{n+1})u=\big(\varepsilon_n^2-\varepsilon_{n+1}^2\big)\int_0^t\mathfrak{S}_i(t-s;\,\varepsilon_n)\mathfrak{C}_i''(s;\,\varepsilon_n)u\,ds.$$

In view of (2.12) and Assumption 2.1, if $u\in D(A^2)$ we have

$$\|\mathfrak{C}_i(t;\,\varepsilon_n)u-\mathfrak{C}_i(t;\,\varepsilon_{n+1})u\|\leqq C_0^2|t|e^{\omega|t|}\left(1-\frac{\varepsilon_{n+1}^2}{\varepsilon_n^2}\right)\|Au\|. \tag{2.13}$$

Selecting now a sequence $\{\varepsilon_n\}$ that tends to zero sufficiently slowly (say, $\varepsilon_n=n^{-1/2}$), we show that $\{\mathfrak{C}_i(t;\,\varepsilon_n)u\}$ is a Cauchy sequence, uniformly with respect to t on compact subsets of $-\infty<t<\infty$. Using the uniform bound (2.1) and the denseness of $D(A^2)$, we deduce that $\{\mathfrak{C}_i(t;\,\varepsilon_n)\}$ converges strongly, uniformly on compacts of $-\infty<t<\infty$ to a strongly continuous operator valued function $S_i(s)$ satisfying the estimate (2.10).

Denote by $\hat{\mathfrak{C}}_i(\lambda;\,\varepsilon)u$ the Laplace transform of $\mathfrak{C}_i(t;\,\varepsilon)u$. Writing (1.1) for $\mathfrak{C}_i(t;\,u)$ $(u\in D(A))$ and taking into account that $\mathfrak{C}_i(0;\,\varepsilon)u=u$, $\mathfrak{C}_i'(0;\,\varepsilon)u=0$, we obtain, after integrating from 0 to t two times,

$$\varepsilon^2\big(\mathfrak{C}_i(t;\,\varepsilon)u-u\big)-i\int_0^t\big(\mathfrak{C}_i(s;\,\varepsilon)u-u\big)\,ds=\int_0^t(t-s)A\mathfrak{C}_i(s;\,\varepsilon)u\,ds.$$

Taking Laplace transforms, we obtain

$$\varepsilon^2\big(\lambda^2\hat{\mathfrak{C}}_i(\lambda;\,\varepsilon)u-\lambda u\big)-i\big(\lambda\hat{\mathfrak{C}}_i(\lambda;\,\varepsilon)u-u\big)=A\hat{\mathfrak{C}}_i(\lambda;\,\varepsilon)u,$$

where the introduction of A under the integral sign is easily justified. It follows that

$$\hat{\mathfrak{C}}_i(\lambda;\,\varepsilon)u=(\varepsilon^2\lambda-i)R(\varepsilon^2\lambda^2-i\lambda;\,A)u$$

for $\lambda>\omega$. Putting $\varepsilon=\varepsilon_n$ with $\{\varepsilon_n\}$ a sequence as above and taking limits, we obtain

$$\hat{S}_i(\lambda)u=-iR(-i\lambda;\,A)u=R(\lambda;\,iA)u$$

for $u \in D(A)$, and thus for all $u \in E$. It follows then that $S_i(t)$ is a strongly continuous group with infinitesimal generator iA, which ends the proof of Theorem 2.2.

We note that our proof of Theorem 2.2 includes a result on the convergence of $\mathfrak{C}_i(t;\ \varepsilon)$ to $S_i(t)$. However, this will be considerably improved in §3.

3. Convergence results. Let $u \in D(A^2)$. The function $u(t;\ \varepsilon) = \mathfrak{C}_i'(t;\ \varepsilon)u$ is a solution of the homogeneous equation (1.1) with $u(0;\ \varepsilon) = 0$, $u'(0;\ \varepsilon) = \mathfrak{C}_i''(0;\ \varepsilon)u = \varepsilon^{-2}A\mathfrak{C}_i(0;\ \varepsilon)u + i\varepsilon^{-2}\mathfrak{C}_i'(0;\ \varepsilon)u = \varepsilon^{-2}Au$. It follows that

$$\mathfrak{C}_i'(t;\ \varepsilon)u = \mathfrak{S}_i(t;\ \varepsilon)Au. \tag{3.1}$$

On the other hand, $v(t;\ \varepsilon)) = \mathfrak{S}_i'(t;\ \varepsilon)u$ is also a solution of the homogeneous equation (1.1) with $v(0;\ \varepsilon) = \mathfrak{S}_i'(0;\ \varepsilon)u = \varepsilon^{-2}u$, $v'(0;\ \varepsilon) = \mathfrak{S}_i''(0;\ \varepsilon)u = \varepsilon^{-2}A\mathfrak{S}_i(0;\ \varepsilon)u + i\varepsilon^{-2}\mathfrak{S}_i'(0;\ \varepsilon)u = i\varepsilon^{-4}u$, so that

$$\mathfrak{S}_i'(t;\ \varepsilon)u = \varepsilon^{-2}\mathfrak{C}_i(t;\ \varepsilon)u + i\varepsilon^{-2}\mathfrak{S}_i(t;\ \varepsilon)u. \tag{3.2}$$

Since all operators in (3.2) are bounded, the equality can be extended to all $u \in E$. We combine (3.1) and (3.2), the latter inequality written for an element of the form Au. The result is

$$\mathfrak{C}_i'(t;\ \varepsilon)u = iA\mathfrak{C}_i(t;\ \varepsilon)u - i\varepsilon^2\mathfrak{S}_i'(t;\ \varepsilon)Au. \tag{3.3}$$

Accordingly,

$$\mathfrak{C}_i(t;\ \varepsilon)u - S_i(t)u = -i\varepsilon^2\int_0^t S_i(t-s)\mathfrak{S}_i'(s;\ \varepsilon)Au\,ds = -i\varepsilon^2\int_0^t \mathfrak{S}_i'(t-s;\ \varepsilon)S_i(s)Au\,ds.$$

If $u \in D(A^2)$, we can integrate by parts. The result is

$$\mathfrak{C}_i(t;\ \varepsilon)u - S_i(t)u = -i\varepsilon^2\mathfrak{S}_i(t;\ \varepsilon)Au - i\varepsilon^2\int_0^t \mathfrak{S}_i(t-s;\ \varepsilon)S_i(s)A^2u\,ds. \tag{3.4}$$

A similar expression was used by Kisyński [8] for the parabolic singular perturbation problem. As an immediate consequence of (3.4) we obtain:

THEOREM 3.1. *Let A be an operator satisfying Assumption* 2.1 *and let $u(t;\ \varepsilon)$ be a solution of the homogeneous problem* (1.1)–(1.2), *$u(t)$ a solution of the homogeneous problem* (1.3)–(1.4) *with $u_0 \in D(A^2)$. Then we have*

$$\begin{aligned} \|u(t;\ \varepsilon) - u(t)\| \leqq {} & C_1\varepsilon^2 e^{\omega|t|}\left(\|Au_0\| + C_1|t|\,\|A^2u_0\|\right) \\ & + C_0 e^{\omega|t|}\|u_0(\varepsilon) - u_0\| + C_1 e^{\omega|t|}\varepsilon^2\|u_1(\varepsilon)\| \qquad (-\infty < t < \infty). \end{aligned} \tag{3.5}$$

Theorem 3.1 implies that when $u \in D(A^2)$ we have $\|u(t;\ \varepsilon) - u(t)\| = O(\varepsilon^2)$ uniformly on compacts of $-\infty < t < \infty$ if $\|u_0(\varepsilon) - u_0\| = O(\varepsilon^2)$, $\|u_1(\varepsilon)\| = O(1)$ as $\varepsilon \to 0$.

Estimates of the same sort can be easily obtained for the derivative $u'(t;\ \varepsilon)$ if $u_0 \in D(A^3)$ and $u_0(\varepsilon) \in D(A)$. In fact, $v(t;\ \varepsilon) = u'(t;\ \varepsilon)$ is the generalized solution of the homogeneous equation (1.1) satisfying $v(0;\varepsilon) = u'(0;\varepsilon) = u_1(\varepsilon)$, $v'(0;\varepsilon) = u''(0;\varepsilon) = \varepsilon^{-2}(Au_0(\varepsilon) + iu_1(\varepsilon))$. On the other hand, $v(t) = u'(t)$ is the solution of the homogeneous equation (1.3) with $v(0) = u'(0) = iAu_0$. Applying Theorem 3.1 to $v(t;\ \varepsilon)$, $v(t)$, we obtain:

THEOREM 3.2. *Let A be as in Theorem* 3.1 *and let $u(t;\varepsilon)$ be a solution of the homogeneous problem* (1.1)–(1.2) *with $u_0(\varepsilon) \in D(A)$, $u(t)$ a solution of the homogeneous* *problem* (1.3)–(1.4) *with* $u_0 \in D(A^3)$. *Then we have*

$$\|u'(t;\,\varepsilon)-u'(t)\| \leqq C_1\varepsilon^2 e^{\omega|t|}\big(\|A^2u_0\|+C_1|t|\,\|A^3u_0\|\big) \tag{3.6}$$
$$+e^{\omega|t|}\big(C_0\|u_1(\varepsilon)-iAu_0\|+C_1\|u_1(\varepsilon)-iAu_0(\varepsilon)\|\big).$$

It follows from this result that if $u_0 \in D(A^3)$ and $u_0(\varepsilon) \in D(A)$ then $\|u'(t;\,\varepsilon)-u'(t)\|=O(\varepsilon^2)$ uniformly on compact subsets of $-\infty<t<\infty$ if $\|u_1(\varepsilon)-iAu_0\|=O(\varepsilon^2)$ and $\|u_1(\varepsilon)-iAu_0(\varepsilon)\|=O(\varepsilon^2)$ or, equivalently, if $\|u_1(\varepsilon)-iAu_0\|=O(\varepsilon^2)$ and $\|Au_0(\varepsilon)-Au_0\|=O(\varepsilon^2)$.

Theorems 3.1 and 3.2 allow us to deduce convergence results for arbitrary initial conditions.

THEOREM 3.3. *Let* $u(t;\,\varepsilon)$ *be a generalized solution of the homogeneous system* (1.1)–(1.2) *with* $u_0(\varepsilon)$, $u_1(\varepsilon) \in E$, $u(t)$ *a generalized solution of the homogeneous system* (1.3)–(1.4) *with* $u_0 \in E$. *Assume that*

$$u_0(\varepsilon)\to u_0,\ \varepsilon^2u_1(\varepsilon)\to 0 \quad \textit{as } \varepsilon\to 0. \tag{3.7}$$

Then

$$u(t;\,\varepsilon)\to u(t) \quad \textit{as } \varepsilon\to 0 \tag{3.8}$$

uniformly on compacts of $-\infty<t<\infty$.

Proof. Pick $\delta>0$ and choose $\bar{u} \in D(A^2)$ such that $\|\bar{u}-u_0\| \leqq \delta$. Let $\bar{u}(t)$ be the solution of the initial value problem (1.3)–(1.4) with $\bar{u}(0)=\bar{u}$. Applying Theorem 3.1, we deduce that

$$\|u(t;\,\varepsilon)-u(t)\| \leqq \|u(t;\,\varepsilon)-\bar{u}(t)\|+\|\bar{u}(t)-u(t)\|+C_1\varepsilon^2e^{\omega|t|}\big(\|A\bar{u}\|+C_1|t|\,\|A^2\bar{u}\|\big) \tag{3.9}$$
$$+C_0e^{\omega|t|}\|u_0(\varepsilon)-\bar{u}\|+C_1e^{\omega|t|}\varepsilon^2\|u_1(\varepsilon)\|+C_1\delta e^{\omega|t|}$$
$$\leqq C_1\varepsilon^2e^{\omega|t|}\big(\|A\bar{u}\|+C_1|t|\,\|A^2\bar{u}\|\big)+C_0e^{\omega|t|}\|u_0(\varepsilon)-u_0\|$$
$$+C_1e^{\omega|t|}\varepsilon^2\|u_1(\varepsilon)\|+2C_0\delta e^{\omega|t|}.$$

Taking $\varepsilon>0$ sufficiently small, we can obviously make the right-hand side of (3.9) $\leqq 3C_0\delta e^{\omega a}$ in $|t| \leqq a>0$. This ends the proof.

THEOREM 3.4. *Let* $u(t;\,\varepsilon)$, $u(t)$ *be as in Theorem* 3.3. *Assume that* $u_0(\varepsilon)$, $u_0 \in D(A)$ *and that*

$$Au_0(\varepsilon)\to Au,\ u_1(\varepsilon)\to iAu_0 \quad \textit{as } \varepsilon\to 0. \tag{3.10}$$

Then

$$u'(t;\,\varepsilon)\to u'(t) \quad \textit{as } \varepsilon\to 0,$$

uniformly on compacts of $-\infty<t<\infty$.

Proof. Given $\delta>0$ choose $\bar{u} \in D(A^3)$ such that $\|A\bar{u}-A\bar{u}_0\| \leqq \delta$ so that $\|\bar{u}'(t)-u'(t)\|=\|S_i'(t)(\bar{u}-u_0)\|=\|S_i(t)(A\bar{u}-Au_0)\| \leqq C_0\delta e^{\omega|t|}$. This time we use (3.6) with $\bar{u}$ instead of u_0; the details are omitted.

It is natural to ask whether a result of the type of Theorem 3.3 exists where the convergence of $u(t;\,\varepsilon)$ to $u(t)$ is uniform with respect to u_0 for bounded $\|u_0\|$. To decide this and other related questions we examine the case where E is the Hilbert space l^2 of all complex valued sequences $\{u_n;\ 1 \leqq n < \infty\}$ such that $\|u\| = (\Sigma |u_n|^2)^{1/2} < \infty$ and A is the (self-adjoint) operator

$$A\{u_n\} = \{\mu_n u_n\}, \tag{3.11}$$

where $\{\mu_n\}$ is a sequence of real numbers bounded above (the domain of A consists of all $\{u_n\}$ such that the right-hand side of (3.11) belongs to l^2). We shall also use the space $E = \mathbb{C}^m$ with its ordinary Euclidean norm and an operator A of the form (3.11). We check easily that the solution operators of (1.1)–(1.2) corresponding to the operator A are

$$\mathfrak{C}_i(t;\ \varepsilon)\{u_n\} = \left\{ \frac{\lambda_n^+(\varepsilon) e^{\lambda_n^-(\varepsilon)t} - \lambda_n^-(\varepsilon) e^{\lambda_n^+(\varepsilon)t}}{\lambda_n^+(\varepsilon) - \lambda_n^-(\varepsilon)} u_n \right\}, \tag{3.12}$$

$$\mathfrak{S}_i(t;\ \varepsilon)\{u_n\} = \left\{ \frac{e^{\lambda_n^+(\varepsilon)t} - e^{\lambda_n^-(\varepsilon)t}}{\varepsilon^2(\lambda_n^+(\varepsilon) - \lambda_n^-(\varepsilon))} u_n \right\}, \tag{3.13}$$

where $\lambda_n^+(\varepsilon)$, $\lambda_n^-(\varepsilon)$ are the roots of the characteristic polynomial

$$\varepsilon^2\lambda^2 - i\lambda - \mu_n = 0, \tag{3.14}$$

$$\lambda_n^+(\varepsilon) = \frac{i}{2\varepsilon^2}\left(1 + \left(1 - 4\varepsilon^2\mu_n\right)^{1/2}\right), \qquad \lambda_n^-(\varepsilon) = \frac{i}{2\varepsilon^2}\left(1 - \left(1 - 4\varepsilon^2\mu_n\right)^{1/2}\right) \tag{3.15}$$

(note that, since the sequence $\{\mu_n\}$ is bounded above, the roots $\lambda_n^+(\varepsilon)$, $\lambda_n^-(\varepsilon)$ will be different for ε sufficiently small).

Example 3.5. *Convergence in Theorem* 3.3 *is not uniform with respect to* u_0 *even if* $\|u_0\|$ *is bounded*. We take $E = l^2$, $\mu_n = -n^2$, $\varepsilon_n^2 = 3/4n^2$. Then

$$\lambda_n^+(\varepsilon_n) = 2n^2 i, \qquad \lambda_n^-(\varepsilon_n) = \frac{-2n^2 i}{3}$$

and, since $S_i(t)\{u_n\} = \{e^{-in^2t}u_n\}$,

$$\|\mathfrak{C}_i(t;\ \varepsilon_n)\{u_n\} - S_i(t)\{u_n\}\| \geqq \left| \frac{3}{4} e^{-i(2n^2/3)t} + \frac{1}{4} e^{-i(2n^2)t} - e^{-in^2t} \right| |u_n|.$$

In contrast, in the abstract parabolic case (see §1), $\mathfrak{C}(t;\ \varepsilon) \to S(t)$ in the uniform topology of operators uniformly on compacts of $t \geqq 0$ except in an "initial layer" $0 \leqq t \leqq t(\varepsilon)$ where $t(\varepsilon)/\varepsilon^2 \to \infty$ (see [6, Thm. 3.2]).

Example 3.6. *The condition* $\varepsilon^2 u_1(\varepsilon) \to 0$ *in Theorem* 3.3 *for convergence of* $u(t;\ \varepsilon)$ *cannot be weakened*. We take $E = \mathbb{C}^1$ and rewrite formula (3.13) as follows.

$$\mathfrak{S}_i(t;\ \varepsilon)\varepsilon^2 u_1(\varepsilon) = 2ie^{it/2\varepsilon^2} \frac{\sin\left(\left(1 + 4\varepsilon^2\mu\right)^{1/2}/2\varepsilon^2\right)t}{\left(1 + 4\varepsilon^2\mu\right)^{1/2}} \varepsilon^2 u_1(\varepsilon), \tag{3.16}$$

which does not have a limit as $\varepsilon \to 0$ unless $\varepsilon^2 u_1(\varepsilon) \to 0$. In contrast, in the abstract parabolic case, $\mathfrak{S}(t;\ \varepsilon) \to S(t)$ in the uniform operator topology under the same conditions as $\mathfrak{C}(t;\ \varepsilon) \to S(t)$ (see Example 3.5).

Example 3.7. *The rate of convergence in Theorem* 3.1 *is best possible*. We use the space $E = l^2$ and the operator A in (3.11). Write

$$\mathfrak{C}_i(t;\ \varepsilon)\{u_n\} = \{\rho_n(t;\ \varepsilon)u_n\}.$$

After some computation with Taylor series, we see that

$$(3.17)\quad \rho_n(t;\,\varepsilon)=e^{i\mu_n t}+i\varepsilon^2\mu_n^2 e^{i\mu_n t}-\varepsilon^2\mu_n e^{i\mu_n t}-\varepsilon^2\mu_n e^{ir_n(\varepsilon)t}+O(\varepsilon^4)\qquad(\varepsilon\to 0),$$

where $r_n(\varepsilon)$ is a real number. Assume that $\{u_n\}\in l^2$ is such that

$$\|\mathfrak{C}_i(t;\,\varepsilon)\{u_n\}-S_i(t)\{u_n\}\|<C\varepsilon^2$$

as $\varepsilon\to 0$. Rewrite this inequality as

$$\sum\varepsilon^{-4}\left|\rho_n(t;\,\varepsilon)-e^{i\mu_n t}\right|^2|u_n|^2\leqq C^2.$$

Taking (3.17) into account and keeping in mind that $\mu_n\to\infty$, we obtain that

$$\sum\mu_n^4|u_n|^2<\infty,$$

so that $u\in D(A^2)$. We recall that in the abstract parabolic case, convergence of order ε^2 can be obtained under the weaker assumption that $u\in D(A)$ (see [6, Thm. 5.2]).

Example 3.8. *Convergence in Theorem* 3.3 *is not uniform in* $t\geqq 0$, *even if* $\omega=0$ *in Assumption* 2.1. We take here $E=\mathbf{C}^1$, $Au=\mu u$ with $\mu<0$, $u_0(\varepsilon)=u$, $u_1(\varepsilon)=0$; the fact that $\mathfrak{C}_i(t;\,\varepsilon)u$ does not converge uniformly to $S_i(t)=e^{i\mu t}$ is an obvious consequence of the fact that $e^{\lambda^+(\varepsilon)t}$ does not converge uniformly to $e^{i\mu t}$ in $t\geqq 0$.

4. Verification of Assumption 2.1. We examine in this section operators that satisfy Assumption 2.1, beginning with the case where $E=H$ is a Hilbert space and A a normal operator. It follows from the general theory ([14], [2], [3]) or directly that A generates a strongly continuous cosine function $\mathscr{C}(t)$ satisfying

$$(4.1)\qquad \mathscr{C}(t)\leqq Ce^{\omega|t|}\qquad(-\infty<t<\infty)$$

if and only if $\sigma(A)$, the spectrum of A, is contained in the region

$$(4.2)\qquad \pi(\omega)=\left\{\lambda:\operatorname{Re}\lambda\leqq\omega^2-\frac{(\operatorname{Im}\lambda)^2}{4\omega^2}\right\}.$$

($\pi(\omega)$ is the region to the left of the parabola passing through ω^2, $\pm 2i\omega$.) We note in passing that $\mathscr{C}(t)$ can be computed using the functional calculus for normal operators:

$$\mathscr{C}.(t)u=c(t;\,A)u=\int_{\sigma(A)}c(t;\,\mu)P(d\mu)u,$$

where $P(d\mu)$ is the resolution of the identity for A and $c(t;\,\mu)=\cosh t\mu^{1/2}=1+t^2\mu/2!+t^4\mu^2/4!+\cdots$. Moreover, the constant C in (4.1) can be taken equal to 1. (In fact the estimate can be improved to $\|\mathscr{C}(t)\|\leqq\cosh\omega t$.)

THEOREM 4.1. *The operator A satisfies Assumption* 2.1 *if and only if $\sigma(A)$ is contained in a half-strip.*

$$(4.3)\qquad \operatorname{Re}\mu\leqq a,\qquad |\operatorname{Im}\mu|\leqq b.$$

Proof. We have seen (Theorem 2.2) that Assumption 2.1 implies that iA generates a strongly continuous group. Since, on the other hand, A generates a cosine function it follows that $\sigma(A)$ is contained in the intersection of a horizontal strip with a region $\pi(\omega)$. This intersection is itself contained in a half-strip of the form (4.3).

Conversely, assume that $\sigma(A)$ is contained in a region of the form (4.3). Then $\sigma(A-(2\varepsilon)^{-2}I)\subseteq\pi(\omega(\varepsilon))$ if and only if $b\leqq\omega(\varepsilon)$, where

$$a-\frac{1}{4\varepsilon^2}=\omega(\varepsilon)^2-\frac{b^2}{4\omega(\varepsilon)^2}$$

so that, if $\varepsilon\leqq\frac{1}{2}a^{1/2}$ we have

$$\omega(\varepsilon)=\frac{1}{2^{1/2}}\left\{\left(\left(\frac{1}{4\varepsilon^2}-a\right)^2+b^2\right)^{1/2}-\frac{1}{4\varepsilon^2}+a\right\}^{1/2}\leqq(1-4a\varepsilon^2)^{-1/2}b\varepsilon$$

and the first inequality (2.1) is verified. We note next that

$$\mathscr{S}\left(t;A-(2\varepsilon)^{-2}I\right)=s\left(t;A-(2\varepsilon)^{-2}I\right),$$

where $s(t;\ \mu)=\mu^{-1/2}\sinh t\mu^{1/2}=t-t^3\mu^2/3!+t^5\mu^3/5!-\cdots$. Accordingly, the norm $\|\mathscr{S}(t;\ A-(2\varepsilon)^{-2}I)\|$ does not surpass the supremum of $|\mu^{-1/2}\sinh t\mu^{1/2}|$ in the half strip defined by

$$\text{(4.4)}\qquad \operatorname{Re}\mu\leqq a-\frac{1}{4\varepsilon^2},\qquad |\operatorname{Im}\mu|\leqq b.$$

If μ belongs to the region defined by (4.3), then

$$|\mu^{1/2}|>\left(\frac{1}{4\varepsilon^2}-a\right)^{1/2}=\frac{(1-4a\varepsilon^2)^{1/2}}{2\varepsilon}.$$

Hence

$$\text{(4.5)}\qquad \left\|\mathscr{S}\left(t;A-(2\varepsilon)^{-2}I\right)\right\|\leqq2(1-4a\varepsilon^2)^{-1/2}\varepsilon e^{\omega\varepsilon|t|}$$
$$\leqq2\varepsilon(1-4a\varepsilon^2)^{-1/2}\exp\left((1-4a\varepsilon^2)^{-1/2}b\varepsilon|t|\right),$$

which is the second inequality (2.9). This ends the proof of Theorem 4.1.

We note the important particular case where A is *self-adjoint* with $-A\geqq0$, in which case we can take $\omega=0$ in (2.9). Another case that can be reduced to this is covered by the following result:

THEOREM 4.2. *Let A generate a uniformly bounded cosine function $\mathscr{C}(t)$, i.e.,*

$$\text{(4.6)}\qquad \|\mathscr{C}(t)\|\leqq C\qquad(-\infty<t<\infty),$$

in a Hilbert space H. Then A satisfies Assumption 2.1 *with $\omega=0$.*

Proof. It was proved in [4] that if A is an operator that satisfies the assumptions in Theorem 4.2 then there exists a (self-adjoint) bounded, invertible operator Q and a self-adjoint operator B with $B\leqq0$ such that

$$\text{(4.7)}\qquad A=Q^{-1}BQ.$$

Using an obvious notation, $\mathscr{C}(t;\ A)=Q^{-1}\mathscr{C}(t;\ B)Q$ and $\mathscr{S}(t;\ A)=Q^{-1}\mathscr{S}(t;\ B)Q$, thus it follows from Theorem 4.1 that A satisfies Assumption 2.1 as claimed. To compute explicitly the constants C_0, C_1 in (2.1) or the constants C_0', C_1 in (2.9), explicit estimates for the norms $\|Q\|$, $\|Q^{-1}\|$ are needed. These are given in [4], but it is not clear they are best possible.

Another class of operators satisfying Assumption 2.1 is described in the next theorem.

THEOREM 4.3. *Let E be an arbitrary Banach space, A_0 an operator satisfying Assumption* 2.1, *B a bounded operator. Then $A=A_0+B$ also satisfies Assumption* 2.1.

Proof. If A_0 generates a strongly continuous cosine function $\mathscr{C}(t;\, A)$ and B is a bounded operator, then ([5, Thm. 2.1]) $A=A_0+B$ also generates a strongly continuous cosine function $\mathscr{C}(t;\, A_0+B)$. This cosine function can be expressed by the formula

$$(4.8)\qquad \mathscr{C}(t;A)u=\mathscr{C}(t;A_0+B)u=\mathscr{C}(t;A_0)u+\mathscr{C}(t;A_0)*B\mathscr{S}(t;A_0)u \\ +\mathscr{C}(t;A_0)*B\mathscr{S}(t;A_0)*B\mathscr{S}(t;A_0)u+\cdots,$$

where, for each $u\in E$ the series is uniformly convergent on compact subsets of $-\infty<t<\infty$. The corresponding formula for $\mathscr{S}$ is

$$(4.9)\qquad \mathscr{S}(t;A)=\mathscr{S}(t;A_0+B)=\mathscr{S}(t;A_0)+\mathscr{S}(t;A_0)*B\mathscr{S}(t;A_0) \\ +\mathscr{S}(t;A_0)*B\mathscr{S}(t;A_0)*B\mathscr{S}(t;A_0)\cdots$$

the series being uniformly convergent in the uniform topology of operators, for $-\infty<t<\infty$. In both formulas, the convolution $(F*G)(t)$ of two strongly continuous operator-valued functions $F(t)$ and $G(t)$ is defined by

$$(F*G)(t)u=\int_0^t F(t-s)G(s)u\,ds.$$

In fact, [5, Thm. 2.1] is considerably more general; for results in the same vein see [12] and [15].

Let A_0 satisfy Assumption 2.1. Using (4.8) to express $\mathscr{C}(t;\, A_0+B-(2\varepsilon)^{-2}I)=\mathscr{C}(t;\, A_0-(2\varepsilon)^{-2}I+B)$ and estimating the convolutions in an obvious way, we obtain

(4.10)

$$\left\|\mathscr{C}\left(t;A_0+B-(2\varepsilon)^{-2}I\right)\right\|\leqq C_0'e^{\omega\varepsilon|t|}+C_0'C_1\|B\|\,|t|\varepsilon e^{\omega\varepsilon|t|}+C_0'C_1^2\|B\|^2\frac{|t|^2}{2!}\varepsilon^2e^{\omega\varepsilon|t|}+\cdots \\ =C_0'e^{(\omega+C_1\|B\|)\varepsilon|t|}\quad(-\infty<t<\infty,\,0\leqq\varepsilon\leqq\varepsilon_0).$$

A similar estimation of $\mathscr{S}(t;\, A_0+B-(2\varepsilon)^{-2}I)$ written using (4.9) yields

(4.11)

$$\left\|\mathscr{S}\left(t;A_0+B-(2\varepsilon)^{-2}I\right)\right\|\leqq C_1\varepsilon e^{\omega\varepsilon|t|}+C_1^2\|B\|\,|t|\varepsilon^2e^{\omega\varepsilon|t|}+C_0^3\|B\|^2\frac{|t|^2}{2!}\varepsilon^3e^{\omega\varepsilon|t|}+\cdots \\ =C_1\varepsilon e^{(\omega+C_1\|B\|)\varepsilon|t|}\qquad(-\infty<t<\infty,\,0\leqq\varepsilon\leqq\varepsilon_0).$$

This ends the proof of Theorem 4.3.

In the rest of this section we examine an extension of Theorems 3.1 and 3.2 obtained by means of interpolation theory. We assume $A=A_0+B$ where B is bounded and A_0 is self-adjoint and bounded above (that A satisfies Assumption 2.1 has been proved in Theorem 4.3). To simplify, we also assume that $\sigma(A_0)\subseteq(-\infty,0)$, which can always be achieved by an obvious decomposition of A_0. Finally, we assume that B maps the domain of A_0 into itself, that is,

$$(4.12)\qquad BD(A_0)\subseteq D(A_0).$$

We recall (see [10, Chap. IX]) that if $\zeta = \sigma + i\tau$ is an arbitrary complex number the fractional powers $(-A_0)^\zeta$ can be defined using the functional calculus for A_0,

$$(-A_0)^\zeta u = \int_0^\infty \mu^\zeta P(-d\mu)u,$$

where $P(d\mu)$ is the resolution of the identity for A_0. We have

$$\|(-A_0)^\zeta u\|^2 = \|(-A_0)^{\sigma+i\tau} u\|^2 = \int_0^\infty \mu^{2\sigma} \|P(-d\mu)u\|^2. \tag{4.13}$$

Let $Q: H \to H$ be a linear operator such that

$$\|Qu\| \leqq K_0\|u\| \quad (u \in H), \qquad \|Qu\| \leqq K_2\|A_0^2 u\| \qquad \big(u \in D(A_0^2)\big) \tag{4.14}$$

and consider the H-valued holomorphic function

$$\varphi(\sigma + i\tau) = (-A_0)^{-\sigma - i\tau} Qu$$

for $u \in H$ fixed. Making use of (4.14) and applying Hadamard's three-lines theorem [9] to φ, we deduce that

$$\|\varphi(\sigma + i\tau)\| \leqq K_0^{(2-\sigma)/2} K_2^{\sigma/2} \qquad (0 \leqq \sigma \leqq 2); \tag{4.15}$$

hence

$$\|Qu\| \leqq K_0^{(2-\sigma)/2} K_2^{\sigma/2} \|(-A_0)^\sigma u\| \qquad \big(u \in D((-A_0)^\sigma u), \quad 0 \leqq \sigma \leqq 2\big). \tag{4.16}$$

We apply this argument to the operator $\mathfrak{C}_i(t;\, \varepsilon) - S_i(t)$. Using (2.1), we obtain

$$\|\mathfrak{C}_i(t;\, \varepsilon)u - S_i(t)u\| \leqq 2C_1 e^{\omega|t|}\|u\|. \tag{4.17}$$

The second estimate is less trivial, since (3.4) provides bounds in terms of $\|Au\|$ and of $\|A^2u\|$ rather than in terms of $\|A_0^2\|$ as needed. To perform the conversion, we note that

$$\begin{aligned} A^2 A_0^{-2} &= (A_0 + B)^2 A_0^{-2} = \big(A_0 + A_0 B + B A_0 + B^2\big) A_0^2 \\ &= I + A_0 B A_0^{-2} + B A_0^{-1} + B^2 A_0^{-2}, \end{aligned} \tag{4.18}$$

where the first, third and fourth operators are trivially bounded (in fact, even $A_0 B A_0^{-1}$ is bounded because of (4.12) and of the closed graph theorem). On the other hand, we have

$$A A_0^{-2} = (A_0 + B) A_0^{-2} = A_0^{-1} + B A_0^{-2}. \tag{4.19}$$

It follows from (4.18) and (4.19) that

$$\|\mathfrak{C}_i(t;\, \varepsilon)u - S_i(t)u\| \leqq C_1' \varepsilon^2 (1 + |t|) e^{\omega|t|} \|A_0^2 u\| \qquad \big(u \in D(A_0^2)\big). \tag{4.20}$$

Combining (4.17) and (4.18) with the preceding remarks, we deduce that if $u \in D((-A_0)^\sigma)$, $0 < \sigma \leqq 2$, we have

$$\begin{aligned} \|\mathfrak{C}_i(t;\, \varepsilon)u - S_i(t)u\| &\leqq C(\sigma)\varepsilon^\sigma (1 + |t|)^{\sigma/2} e^{\omega|t|} \|(-A_0)^\sigma u\| \\ &\qquad \big(u \in D((-A)^\sigma), \ -\infty < t < \infty\big). \end{aligned} \tag{4.21}$$

THEOREM 4.4. *Let $E=H$ be a Hilbert space, $A=A_0+B$ with B bounded and A self-adjoint and bounded above, and let $u(t;\ \varepsilon)$ be a solution of the homogeneous problem* (1.1)–(1.2), *$u(t)$ a solution of the homogeneous problem* (1.3)–(1.4) *with $u\in D((-A_0)^\sigma)$, $0<\sigma\leqq 2$. Then, if* (4.12) *holds, there exists a constant $C(\sigma)$ such that*

$$\|u(t;\ \varepsilon)-u(t)\|\leqq C(\sigma)\varepsilon^\sigma(1+|t|)^{\sigma/2}e^{\omega|t|}\|(-A)^\sigma u_0\| \\ +C_0e^{\omega|t|}\|u_0(\varepsilon)-u_0\|+C_1e^{\omega|t|}\varepsilon^2\|u_1(\varepsilon)\| \qquad (-\infty<t<\infty). \tag{4.22}$$

The proof follows that of Theorem 3.1. Equation (4.17) implies that when $u\in D((-A_0)^\sigma)$, $1<\sigma\leqq 2$, we have $\|u(t;\ \varepsilon)-u(t)\|=O(\varepsilon^\sigma)$ uniformly on compacts of $t\geqq 0$ if $\|u_0(\varepsilon)-u_0\|=O(\varepsilon^\sigma)$, $\|u_1(\varepsilon)\|=O(\varepsilon^{\sigma-2})$ as $\varepsilon\to 0$.

THEOREM 4.5. *Let E, A be as in Theorem* 4.3, *and let $u(t;\ \varepsilon)$ be a solution of the homogeneous problem* (1.1)–(1.2) *with $u_0(\varepsilon)\in D(A)$, $u(t)$ a solution of the homogeneous problem* (1.3)–(1.4) *with $u_0\in D((-A_0)^{1+\sigma})$. Then there exists a constant $C(\sigma)$ such that*

$$\|u'(t;\ \varepsilon)-u'(t)\|\leqq C(\sigma)\varepsilon^\sigma(1+|t|)^{\sigma/2}e^{\omega|t|}\|(-A)^{1+\sigma}u_0\| \\ +e^{\omega|t|}\big(C_0\|u_1(\varepsilon)-iAu_0\|+C_1\|u_1(\varepsilon)-iAu_0(\varepsilon)\|\big). \tag{4.23}$$

As a consequence, we deduce that $\|u'(t;\ \varepsilon)-u'(t)\|=O(\varepsilon^\sigma)$ if $u_0\in D(A^3)$, $u_0(\varepsilon)\in D(A)$ and $\|u_1(\varepsilon)-iAu_0\|=O(\varepsilon^\sigma)$, $\|u_1(\varepsilon)-iAu_0(\varepsilon)\|=O(\varepsilon^\sigma)$ or, equivalently, if $\|u_1(\varepsilon)-iAu_0\|=O(\varepsilon^\sigma)$ and $\|Au_0(\varepsilon)-Au_0\|=O(\varepsilon^\sigma)$.

5. Elliptic differential operators. We examine the operator (1.6) written in *divergence* or *variational* form,

$$A=\sum_{j=1}^{m}\sum_{k=1}^{m}D^j\big(a_{jk}(x)D^ku\big)+\sum_{j=1}^{m}\hat{b}_j(x)D^ju+c(x)u \tag{5.1}$$

in an arbitrary domain Ω of m-dimensional Euclidean space $\mathbb{R}^m$; here $D^j=\partial/\partial x_j$ and $x=(x_1,\cdots,x_m)$. (Note that (5.1) and (1.6) are equivalent entities only if the coefficients $a_{ij}(x)$ are differentiable.) The symbol $A(\beta)$ will denote the restriction of (5.1) obtained by imposition of a boundary condition β, either the *Dirichlet boundary condition*

$$u(x)=0 \qquad (x\in\Gamma) \tag{5.2}$$

or the *variational boundary condition*

$$D^\nu u(x)=\sum\sum a_{jk}(x)\nu_jD^ku(x)=\gamma(x)u(x) \qquad (x\in\Gamma), \tag{5.3}$$

where Γ is the boundary of Ω and $\nu=(\nu_1,\cdots,\nu_m)$ is the exterior normal (unit) vector at x; $D^\nu u$ is called the *conormal derivative* of u at x. The hypotheses on the coefficients of A are as follows. We assume that the a_{jk} and c are *measurable* and *bounded* in Ω. Complex values for c are allowed; the a_{jk} are real and satisfy the *uniform ellipticity condition*

$$\sum\sum a_{jk}(x)\xi_j\xi_k\geqq\kappa|\xi|^2 \qquad (\xi\in\mathbb{R}^m) \tag{5.4}$$

for some $\kappa>0$. Finally, we assume that the first order coefficients $\hat{b}_j(x)$ are *imaginary*, that is

$$\hat{b}_j(x)=ib_j(x) \tag{5.5}$$

with b_j real (we shall see later that this requirement cannot be eliminated) and that *each* b_j *belongs to* $W^{1,\infty}(\Omega)$ (i.e., it has first order partials in $L^{\infty}(\Omega)$).

We begin with the case of the Dirichlet boundary condition (5.2). Let $H_0^1(\Omega)=W_0^{1,2}(\Omega)$ be the closure of $\mathscr{D}(\Omega)$, the space of all Schwartz test functions with support contained in Ω, in $H^1(\Omega)=W^{1,2}(\Omega)$. For $u,v\in H_0^1(\Omega)$ define

$$(5.6)\qquad (u,v)=\int_{\Omega}\left\{\sum\sum a_{jk}D^j\bar{u}D^kv-\frac{i}{2}\sum b_j\left(D^j\bar{u}v-\bar{u}D^jv\right)+\alpha\bar{u}v\right\}dx,$$

where $\alpha>0$ is a parameter to be fixed below. Obviously, $(u,v)_\alpha$ is linear in v, conjugate linear in u, and we check easily that $(v,u)_\alpha=\overline{(u,v)}_\alpha$. Using the uniform ellipticity assumption (5.4), the inequality $|(D^ju)v|\leqq(\varepsilon/2)|D^j\bar{u}|^2+(1/2\varepsilon)|v|^2$ and its counterpart for $|\bar{u}D^jv|$ we easily show that if α is large enough, the first inequality

$$(5.7)\qquad c^2(u,u)\leqq(u,u)_\alpha\leqq C^2(u,u)\qquad\left(u\in H_0^1(\Omega)\right)$$

holds for some $c>0$, where (u,v) is the original scalar product of $H_0^1(\Omega)$; the second inequality (5.7) is a consequence of the assumptions on the coefficients. From now on we shall assume $H_0^1(\Omega)$ is endowed with the scalar product (5.6) and its associated norm $\|u\|_\alpha=(u,u)_\alpha^{1/2}$.

We define an operator $A_0(\beta)$ as follows: $u\in D(A_0(\beta))$ if and only if $u\in H_0^1(\Omega)$ and the functional $w\to(u,w)_\alpha$ is continuous in the norm of $L^2(\Omega)$; if v is the element of $L^2(\Omega)$ such that $(u,w)_\alpha=(v,w)$ then we set $A_0(\beta)u=\alpha u-v$. This definition can be abbreviated as follows:

$$(5.8)\qquad \left((\alpha I-A_0(\beta))u,w\right)=(u,w)_\alpha\qquad\left(w\in H_0^1(\Omega)\right).$$

It is routine to check $A_0(\beta)$ is symmetric and densely defined, and that its construction does not depend on α.

Let v be an arbitrary element of $L^2(\Omega)$. The functional $w\to(v,w)$ is continuous in $L^2(\Omega)$, thus in $H_0^1(\Omega)$, hence there exists $u\in H_0^1(\Omega)$ with $(u,w)_\alpha=(v,w)$; this means that $u\in D(A_0(\beta))$ and $\alpha u-A_0(\beta)u=v$. Since the same argument works for any $\lambda\geqq\alpha$, we have shown that

$$(5.9)\qquad (\lambda I-A_0(\beta))D(A_0(\beta))=E\qquad(\lambda\geqq\alpha).$$

Moreover, $((\lambda I-A_0(\beta))u,u)=(u,u)_\lambda>0$ so that $\lambda I-A_0(\beta)$ is one-to-one for $\lambda\geqq\alpha$. We also obtain as a byproduct of (5.9) that $(\lambda I-A_0(\beta))^{-1}$ is *bounded*, so that $\lambda\in\rho(A_0(\beta))$ if $\lambda\geqq\alpha$. This implies that $A_0(\beta)$ is self-adjoint (see [10, p. 322]).

The full operator $A(\beta)$ is constructed by perturbation. Let

$$(5.10)\qquad Bu=\frac{i}{2}\sum\left(b_jD^ju-D^j(b_ju)\right)+cu=-\frac{i}{2}\sum\left(D^jb_j\right)u+cu.$$

Obviously, B is a bounded operator. We define

$$(5.11)\qquad A(\beta)=A_0(\beta)+B$$

and it follows from Theorem (4.3) that $A(\beta)$ satisfies Assumption 2.1.

The case of boundary conditions (5.3) is slightly different. Here the basic space is $H^1(\Omega)$ instead of $H_0^1(\Omega)$ and we assume Ω to be bounded and of class $C^{(1)}$ (see [1]) so that the following particular case of Sobolev's imbedding theorem holds:

THEOREM 5.1. *There exists a constant C such that*

$$\|u\|_{L^q(\Gamma)} = \left(\int_\Gamma |u|^q d\sigma\right)^{1/q} \leqq C\|u\|_{W^{1,p}(\Omega)}, \tag{5.12}$$

where

$$1 \leqq p < m, \qquad 1 \leqq q \leqq \frac{(m-1)p}{m-p}, \tag{5.13}$$

Γ is the boundary of Ω and $d\sigma$ is the hyperarea differential on Γ.

For proofs of considerably more general facts see [1].

Theorem 5.1 will be used as follows. Let γ be measurable and bounded on Γ and let $\varphi, \psi \in \mathscr{D}$. Then, due to (5.12) for $p = q = 1$ we have

$$\left|\int_\Gamma \gamma\bar{\varphi}\psi\, d\sigma\right| \leqq C'\int_\Gamma |\bar{\varphi}\psi|\, d\sigma \leqq C\|\bar{\varphi}\psi\|_{W^{1,1}(\Omega)}.$$

But

$$\begin{aligned} \|\bar{\varphi}\psi\|_{W^{1,1}(\Omega)} &= \|\bar{\varphi}\psi\|_{L^1(\Omega)} + \sum \|D^j(\bar{\varphi}\psi)\|_{L^1(\Omega)} \\ &\leqq \|\bar{\varphi}\psi\|_{L^1(\Omega)} + \sum \|(D^j\bar{\varphi})\psi\|_{L^1(\Omega)} + \sum \|\bar{\varphi}D^j\psi\|_{L^1(\Omega)} \\ &\leqq \|\bar{\varphi}\|_{L^2(\Omega)}\|\psi\|_{L^2(\Omega)} + \left(\sum \|D^j\bar{\varphi}\|_{L^2(\Omega)}\right)\|\psi\|_{L^2(\Omega)} + \|\bar{\varphi}\|_{L^2(\Omega)}\sum \|D^j\psi\|_{L^2(\Omega)} \\ &\leqq \|\bar{\varphi}\|_{L^2(\Omega)}\|\psi\|_{L^2(\Omega)} + m^{1/2}\left(\sum \|D^j\bar{\varphi}\|^2_{L^2(\Omega)}\right)^{1/2}\|\psi\|_{L^2(\Omega)} \\ &\qquad + m^{1/2}\|\varphi\|_{L^2(\Omega)}\left(\sum \|D^j\psi\|^2_{L^2(\Omega)}\right)^{1/2} \\ &\leqq C(\alpha)\|\varphi\|_\alpha\|\psi\|_\alpha, \end{aligned} \tag{5.14}$$

where we check easily that $C(\alpha) \to 0$ as $\alpha \to \infty$.

We introduce a functional in $\mathscr{D} \subseteq H^1(\Omega)$ as follows:

$$(u,v)'_\alpha = (u,v)_\alpha - \int_\Gamma \left(\gamma + \frac{i}{2}\sum b_j\right)\bar{u}v\, d\sigma.$$

Obviously, $(u,v)'_\alpha$ has all the properties of a scalar product, both inequalities (5.7) being valid for $(u,u)'_\alpha$ for α sufficiently large in view of (5.12) and subsequent comments. Since by Theorem 5.1, $\mathscr{D}$ is dense in $H^1(\Omega)$ we can extend $(u,v)'_\alpha$ to $H^1(\Omega)$ preserving in particular (5.7). From then on construction of the operator $A_0(\beta)$ proceeds in the same way as for the Dirichlet boundary conditions: we can condense the diverse steps in the equation

$$((\alpha I - A_0(\beta))u, w) = (u,w)'_\alpha \qquad (w \in H_0^1(\Omega)). \tag{5.15}$$

The operator $A_0(\beta)$ is again self-adjoint: the full operator $A(\beta)$ is obtained by formula (5.11), where B is the bounded operator defined by (5.10). It follows again from

Theorem 4.2 that $A(\beta)$ satisfies Assumption 2.1. Summarizing:

THEOREM 5.2. *Let* Ω *be a domain in* $\mathbb{R}^m$, A *the operator* (5.1) *with* a_{jk}, $c \in L^\infty(\Omega)$, $\hat{b}_j \in W^{1,\infty}(\Omega)$. *Assume, moreover that the* a_{jk} *are real and satisfy the uniform ellipticity assumption* (5.4) *and that the* $\hat{b}_j$ *are purely imaginary. If* β *is the Dirichlet boundary condition* (5.2), *the operator* $A(\beta)$ *defined by* (5.8) *and* (5.11) *satisfies Assumption* 2.1. *If* Ω *is bounded and of class* $C^{(1)}$ *and* β *is the boundary condition* (5.3) *with* γ *measurable and bounded in* Γ, *then the operator* $A(\beta)$ *defined by* (5.15) *and* (5.11) *satisfies Assumption* 2.1.

Obvious generalizations of this result are possible: for instance, we may only assume that Γ (but not necessarily Ω) is bounded, or we may relax (5.4) to

$$\sum \sum a_{jk}(x)\xi^j\xi^k \geqq 0 \quad \text{for } \xi \in \mathbb{R}^m$$

and only require $\operatorname{Re} c$ to be bounded above, at least when no first order terms are present. On the other hand, the requirement that the a_{jk} be real and the b_{jk} be imaginary cannot be omitted, as the following example shows. Let $m=1$, $\Omega=\mathbb{R}$, A the constant coefficient operator

$$Au = au'' + bu' + cu.$$

Using the Fourier–Plancherel transform, we show that

$$\sigma(A) = \{-a\sigma^2 - ib\sigma + c : -\infty < \sigma < \infty\}$$

so that: a) $\sigma(A)$ will not be contained in a region of the form (4.2) if a is not real; b) if b is not imaginary, $\sigma(A)$ will not be contained in a half-strip of the form (4.3).

Remark 5.3. Theorem 4.4 has an interesting application here. Although $D((-A_0(\beta))^\sigma)$ is not easily identifiable even for $\sigma=1$, one can show, using the arguments in [7], that

$$D\big((-A_0(\beta))^{1/2}\big) = H_0^1(\Omega),$$

when β is the Dirichlet boundary condition (5.2), or

$$D\big((-A_0(\beta))^{1/2}\big) = H^1(\Omega),$$

when β is the variational boundary condition (5.3). We shall use this for $\Omega=\mathbb{R}^m$, in which case the boundary condition is irrelevant and $H_0^1(\Omega)=H^1(\Omega)$. Condition (4.12) will hold if $\sum D^j b_j$, $c \in W^{2,\infty}(\Omega)$ so that $\|u(t;\ \varepsilon)-u(t)\| = O(\varepsilon^{1/2})$ if $u_0 \in H_0^1(\Omega)$ and $\|u_0(\varepsilon)-u_0\| = O(\varepsilon^{1/2})$, $\|u_1(\varepsilon)\| = O(\varepsilon^{-3/2})$.

6. The inhomogeneous equation.

As pointed out in §3, the explicit solution of (1.1) with null initial conditions $u_0(\varepsilon)$, $u_1(\varepsilon)$ is

$$u(t;\ \varepsilon) = \int_0^t \mathfrak{S}_i(t-s;\ \varepsilon) f(s;\ \varepsilon)\, ds. \tag{6.1}$$

We have already noted (in Example 3.6) that $\mathfrak{S}_i(t;\ \varepsilon)$ is not even strongly convergent as $\varepsilon \to 0$. However (and somewhat surprisingly) (6.1) turns out to translate convergence of $f(t;\ \varepsilon)$ into convergence of $u(t;\ \varepsilon)$ at least for a class of operators containing the differential operators in §5.

THEOREM 6.1. *Let $E=H$ be a Hilbert space, $A=A_0+B$, where A_0 is a self-adjoint operator bounded above, B a bounded operator, $T>0$ and $\{f(s;\,\varepsilon);\ 0<\varepsilon\leqq\varepsilon_0\}$ a family of functions in $L^1(-T,T;\,H)$ such that*

$$f(s;\,\varepsilon)\to f(s)\quad \text{as } \varepsilon\to 0 \tag{6.2}$$

in $L^1(-T,T;\,H)$. Finally, let $u(t;\,\varepsilon)$ be the (weak) solution of the initial value problem

$$\varepsilon^2u''(t;\,\varepsilon)-iu'(t;\,\varepsilon)=Au(t;\,\varepsilon)+f(t;\,\varepsilon)\qquad (|t|\leqq T), \tag{6.3}$$

$$u(0;\,\varepsilon)=0,\qquad u'(0;\,\varepsilon)=0. \tag{6.4}$$

Then

$$u(t;\,\varepsilon)\to u(t) \tag{6.5}$$

uniformly in $|t|<T$, where $u(t;\,\varepsilon)$ is the weak solution of

$$u'(t)=iAu(t)+if(t), \tag{6.6}$$

$$u(0)=0. \tag{6.7}$$

Proof. We can obviously assume that $\sigma(A_0)\subseteq(0,\infty)$ (if not we incorporate into B the "part" of A_0 with spectrum in $\mu\geq 0$). We shall first show Theorem 6.1 for A_0 and then add the "perturbation" B, considering first the case $f(t;\,\varepsilon)=f(t)$ independent of ε. Let $P(d\mu)$ be the resolution of the identity for A_0 and $\mathfrak{S}_i(t;\,\varepsilon;\,A_0)$ the (second) propagator of (6.3) with $B=0$. The same argument used in §3 shows that

$$\mathfrak{S}_i(t;\,\varepsilon;\,A_0)u=\int_{-\infty}^{0}\mathfrak{s}(t;\,\varepsilon;\,\mu)P(d\mu)u \tag{6.8}$$

for $u\in E$, where

$$\mathfrak{s}(t;\,\varepsilon;\,\mu)=\frac{e^{\lambda^+(\mu;\,\varepsilon)t}-e^{\lambda^-(\mu;\,\varepsilon)t}}{\varepsilon^2(\lambda^+(\mu;\,\varepsilon)-\lambda^-(\mu;\,\varepsilon))} \tag{6.9}$$

and $\lambda^+(\mu;\,\varepsilon)=i(2\varepsilon^2)^{-1}(1+(1-4\varepsilon^2\mu)^{1/2})$, $\lambda^-(\mu;\,\varepsilon)=i(2\varepsilon^2)^{-1}(1-(1-4\varepsilon^2\mu)^{1/2})$ are the roots of the characteristic polynomial $\varepsilon^2\lambda^2-i\lambda-\mu=0$ $(-\infty<\mu\leqq 0)$. Let $0\leqq t\leqq T$. We can write

$$u(t;\,\varepsilon)=\int_0^t\mathfrak{S}_i(t-s;\,\varepsilon;\,A_0)f(s)\,ds=\int_{-\infty}^{0}P(d\mu)\int_0^t\mathfrak{s}(t-s;\,\varepsilon;\,\mu)f(s)\,ds \tag{6.10}$$

after an easily justified interchange in the order of integration. We note next that

$$\left\|\int_0^t\mathfrak{s}(t-s;\,\varepsilon;\,\mu)f(s)\,ds\right\|\leqq 2\|f\|_{L^1(-T,T;\,E)}. \tag{6.11}$$

On the other hand,

$$\begin{aligned}\int_0^t\mathfrak{s}(t-s;\,\varepsilon;\,\mu)f(s)\,ds&=\frac{e^{\lambda^+(\mu;\,\varepsilon)t}}{i(1-4\varepsilon^2\mu)^{1/2}}\int_0^t e^{-\lambda^+(\mu;\,\varepsilon)s}f(s)\,ds\\&\quad-\frac{e^{\lambda^-(\mu;\,\varepsilon)t}}{i(1-4\varepsilon^2\mu)^{1/2}}\int_0^t e^{-\lambda^-(\mu;\,\varepsilon)s}f(s)\,ds\\&=I_1(t;\,\mu;\,\varepsilon)+I_2(t;\,\mu;\,\varepsilon)=I(t;\,\mu;\,\varepsilon).\end{aligned} \tag{6.12}$$

Since $\lambda^-(\mu;\,\varepsilon)\to i\mu$ as $\varepsilon\to 0$ we deduce that, for μ fixed,

$$I_2(t;\,\mu;\,\varepsilon)\to ie^{i\mu t}\int_0^t e^{-i\mu s}f(s)\,ds \tag{6.13}$$

uniformly in $0\leqq t\leqq T$. To handle the first integral we note that $\lambda^+(\mu;\,\varepsilon)\to i\infty$ and use the following uniform version of the Riemann–Lebesgue lemma: if $g(t)$ is a (scalar or vector-valued) function in $L^1(0,T)$ then

$$\lim_{\alpha\to\infty}\int_0^t e^{i\alpha s}g(s)\,ds=0 \tag{6.14}$$

uniformly in $0\leqq t\leqq T$; the proof is achieved by approximating g in the L^1 norm by smooth functions. Applying (6.14) to the first integral in (6.11), we obtain

$$I_1(t;\,\mu;\,\varepsilon)\to 0 \quad \text{as } \varepsilon\to 0 \tag{6.15}$$

uniformly in $0\leqq t\leqq T$.

Assume that $u(t;\,\varepsilon)\not\to u(t)$ uniformly in $0\leqq t\leqq T$. Then there exists a sequence $\{t_n\}$, $0\leqq t_n\leqq T$ and a sequence $\{\varepsilon_n\}$, $\varepsilon_n\to 0$ such that $\|u(t_n;\,\varepsilon_n)-u(t_n)\|\geqq\delta>0$ for all n. However, using (6.15) and convergence of $I_2(t;\,\mu;\,\varepsilon)$ we obtain, using a variant of Lebesgue's dominated convergence theorem, that $\|u(t_n;\,\varepsilon_n)-u(t_n)\|\to 0$, a contradiction. A similar argument takes care of the range $-T\leqq t\leqq 0$. The case where f depends on ε is handled by writing

$$\begin{aligned} u(t;\,\varepsilon)&=\int_0^t\mathfrak{S}_i(t-s;\,\varepsilon;\,A_0)f(s;\,\varepsilon)\,ds\\ &=\int_0^t\mathfrak{S}_i(t-s;\,\varepsilon;\,A_0)(f(s;\,\varepsilon)-f(s))\,ds+\int_0^t\mathfrak{S}_i(t-s;\,\varepsilon;\,A)f(s)\,ds\end{aligned} \tag{6.16}$$

and making use of the uniform bound (2.1).

The general case is disposed of as follows. From (2.6) and the perturbation formula (4.9), we have

$$\begin{aligned}\mathfrak{S}_i(t;\,\varepsilon;\,A)u&=\mathfrak{S}_i(t;\,\varepsilon;\,A_0+B)u\\ &=\mathfrak{S}_i(t;\,\varepsilon;\,A_0)u+\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*B\mathfrak{S}_i(t;\,\varepsilon;\,A_0)u\\ &\quad+\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*B\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*B\mathfrak{S}_i(t;\,\varepsilon;\,A_0)u+\cdots;\end{aligned} \tag{6.17}$$

hence

$$\begin{aligned}u(t;\,\varepsilon)=\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*f(t;\,\varepsilon)+\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*B\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*f(t;\,\varepsilon)\\ +\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*B\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*B\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*f(t;\,\varepsilon)+\cdots.\end{aligned} \tag{6.18}$$

Now, using (2.1) we show that the nth term of the series (6.18) is bounded in norm by

$$C_1^n\|B\|^{n-1}\frac{|t|^{n-1}}{(n-1)!}e^{\omega t}\|f(t;\,\varepsilon)\|_{L^1(-T,T;H)}.$$

On the other hand, repeatedly using the previously proved result on convergence of $\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*f(t;\,\varepsilon)$ in each term of (6.18), we deduce that $\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*B\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*f(t;\,\varepsilon)$, $\mathfrak{S}(t;\,\varepsilon;\,A_0)*B\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*B\mathfrak{S}_i(t;\,\varepsilon;\,A_0)*f(t;\,\varepsilon),\cdots$ all converge uniformly in $|t|\leqq T$; the limit of the nth term of (6.17) is

$$S_i(t;\,A_0)*BS_i(t;\,A_0)*\cdots*BS_i(t;\,A_0)*f(t),$$

thus the sum of the series converges uniformly, as $\varepsilon \to 0$, to

$$S_i(t; A_0) * f(t) + S_i(t; A_0) * BS_i(t; A_0) * f(t)$$
$$+ S_i(t; A_0) * BS_i(t; A_0) * BS_i(t; A_0) * f(t) + \cdots = S_i(t; A_0 + B) * f(t),$$

where $S_i(t; A_0)$ (resp. $S_i(t; A_0 + B)$) is the group generated by iA_0 (resp. by $i(A_0 + B)$). This completes the proof of Theorem 6.1.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] H. O. FATTORINI, *Ordinary differential equations in linear topological spaces* I, J. Differential Equations, 5 (1969), pp. 72–105.

[3] ______, *Ordinary differential equations in linear topological spaces* II, J. Differential Equations, 6 (1969), pp. 50–70.

[4] ______, *Uniformly bounded cosine functions in Hilbert space*, Indiana Univ. Math. J., 20 (1970), pp. 411–425.

[5] ______, *Un teorema de perturbación para generadores de funciones coseno*, Rev. Unión Matemática Argentina, 25 (1971), pp. 200–211.

[6] ______, *Singular perturbation and boundary layer for an abstract Cauchy problem*, J. Math. Anal. Appl., 97 (1983), pp. 529–571.

[7] J. GOLDSTEIN, *Semigroups and second-order differential equations*, J. Functional Analysis, 4 (1969), pp. 50–70.

[8] J. KISYŃSKI, *On second order Cauchy's problem in a Banach space*, Bull. Acad. Polonaise des Sciences, 43 (1970), pp. 371–374.

[9] S. G. KREIN, JU. I. PETUNIN AND E. M. SEMENOV, *Interpolation of Linear Operators*, Nauka, Moscow, 1978, (Russian); English translation, American Mathematical Society, Providence, RI, 1982.

[10] F. RIESZ AND B. SZ.-NAGY, *Functional Analysis*, Ungar, New York, 1955.

[11] A. Y. SCHOENE, *Semi-groups and a class of singular perturbation problems*, Indiana Univ. Math. J., 20 (1970), pp. 247–263.

[12] M. SHIMIZU AND I. MIYADERA, *Perturbation theory for cosine families in Banach spaces*, Tokyo J. Math., 1 (1978), pp. 333–343.

[13] M. SOVA, *Cosine operator functions*, Rozprawy Mat., 49 (1966), pp. 47.

[14] ______, *Perturbations numériques des evolutions paraboliques et hyperboliques*, Časopis Pešt. Mat., 96 (1971), pp. 406–407.

[15] C. C. TRAVIS AND G. F. WEBB, *Perturbation of strongly continuous cosine family generators*, Colloq. Math., 45 (1981), pp. 277–285.

[16] K. VESELIĆ, *The nonrelativistic limit of the Dirac equation and the spectral concentration*, Glas. Mat., (24) (1969), pp. 231–241.

[17] ______, *On the nonrelativistic limit of the bound states of the Klein–Gordon equation*, J. Math. Anal. Appl., 96 (1983), pp. 63–84.

[18] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Cambridge Univ. Press, Cambridge, 1948.

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

008

ORDER STARS, APPROXIMATIONS AND FINITE DIFFERENCES III. FINITE DIFFERENCES FOR $u_t = \omega u_{xx}$*

A. ISERLES[†]

Abstract. Given the finite difference discretization

$$\sum_{j=-r}^{r} \alpha_j(\mu) U_{m+j}^{n+1} = \sum_{j=-r}^{r} \beta_j(\mu) U_{m+j}^{n}, \qquad \mu = \frac{\omega \Delta t}{(\Delta x)^2}.$$

of the differential equation $u_t = \omega u_{xx}$, $\omega \in \mathbb{C}$, $\operatorname{Re} \omega \geqq 0$, we prove that a unique choice of coefficients gives order $4r+1$ and that no higher order method exists. Furthermore, we show that this highest order method is stable for every $r \geqq 1$. Our analysis uses order stars of first and second kind, in conjunction with Padé theory and computational complex analysis.

1. Introduction. An extensive effort has been devoted in the last few decades to the determination of the highest possible order of a "stable" scheme. The results to date are of interest in several branches of numerical analysis of differential equations:

(a) The order of a k-step one-derivative zero-stable method for ODE's may not exceed $2[(k+2)/2]$ [4], [11];

(b) The order of a multistep n-derivative A-stable method for ODE's may not exceed $2n$ [5], [16];

(c) The order of a stable semi-discretization or a stable full discretization of $u_t = u_x$ may not exceed $\min\{2r+2, 2s, r+s\}$ (for an explicit scheme) or $\min\{4r+2, 4s, 2r+2s\}$ (for an implicit scheme), where r and s denote the number of points to the left and to the right, respectively, along the x-axis [8], [12], [13];

(d) The order of a monotone method for $u_t = u_x$ may not exceed 1 [6] (monotonicity is important here to ensure that no spurious oscillations occur in the numerical solution [15]);

(e) The order of a monotone method for $u_t = u_{xx}$ may not exceed 1 [2] (the importance of monotonicity in this case is in the conservation of positive initial data).

In the present paper we investigate the relationship between order and stability of full discretizations of the equation

$$u_t = \omega u_{xx}, \quad \omega \in \mathbb{C}, \quad \operatorname{Re} \omega \geqq 0. \tag{1}$$

Note that $\omega \in \mathbb{R}$, $\omega > 0$, gives the parabolic *heat-conduction equation*, whilst $\omega \in i\mathbb{R}$ results in the hyperbolic *linearized Schrödinger equation*. It turns out that, unlike in examples (a)–(e), the order does not compete with stability, that is to say the highest-order method is stable. Perhaps surprisingly, it turns out that the order of that optimal method exceeds the number of available degrees of freedom.

Let

$$\sum_{j=-r}^{r} \alpha_j(\mu) U_{m+j}^{n+1} = \sum_{j=-r}^{r} \beta_j(\mu) U_{m+j}^{n}, \qquad \mu := \frac{\omega \Delta t}{(\Delta x)^2}, \tag{2}$$

*Received by the editors June 14, 1983, and in revised form October 24, 1983.

[†] King's College, University of Cambridge, Cambridge CB2 1ST, England.

be a full discretization of the equation (1), where U_m^n denotes an approximation to $u(n\Delta t, m\Delta x)$. We set

$$R(z;\mu) := \frac{\sum_{j=0}^{2r}\beta_{j-r}(\mu)z^j}{\sum_{j=0}^{2r}\alpha_{j-r}(\mu)z^j}.$$

It can be easily shown (cf. [9, Chaps. 7 and 9]) that the scheme (2) is of *order p* in Δx (Δt being linked to Δx via the Courant number μ) if and only if

$$(3) \qquad R(z;\mu) = e^{\mu(\ln z)^2} + c(z-1)^{p+1} + O\left(|z-1|^{p+2}\right), \qquad c \neq 0,$$

and it is *stable* if and only if both

$$(4) \qquad \left|R(e^{i\theta};\mu)\right| \leqq 1, \qquad 0 \leqq \theta \leqq 2\pi$$

for a range of values $0 < \mu \leqq \mu_0$, say, and

(5) R has r poles inside and r poles outside the complex unit circle.

The inequality (4) is the familiar *von Neumann condition* for stability and it is sufficient in the case of a Cauchy problem, i.e. when the initial data is given along the whole x-axis. It is augmented by (5), the *Wiener–Hopf condition*, to cater for equations (1) with initial conditions along a semi-finite or compact interval (strictly speaking, (4) and (5) are together equivalent to stability only if either zero boundary conditions are given or if $r=1$. If $r \geqq 2$ and boundary conditions are nonzero, then extra conditions must be imposed on the special schemes which need be applied in the vicinity of the boundary [7]).

The simplest case, $r=1$, is already surprising enough: the popular Crank–Nicolson scheme with

$$R(z;\mu) = \frac{z+\mu(z-1)^2/2}{z-\mu(z-1)^2/2}$$

is of order 3 and it is stable. However, a less known method, the Crandall scheme [3] with

$$R(z;\mu) = \frac{z+(1/12+\mu/2)(z-1)^2}{z+(1/12-\mu/2)(z-1)^2},$$

is stable and of order 5. The main result of this paper is that for every $r \geqq 1$ there exists a unique method of order $4r+1$, that this method is stable and that no other method with the same r can exceed this order.

It follows from (3) that the order analysis is equivalent to the determination of the block structure along the diagonals of the Padé tableau of $f(x) = \exp(\mu(\ln z)^2)$ at $z_0 = 1$. Note that the Taylor expansion of this function is unknown. Luckily, the desired information can be obtained by indirect means. This will be done in two stages: first, in §2, we prove that the order of the Padé approximation is at least $4r+1$ by using the classical Padé theory. In fact, we prove a more general statement, pertaining to Padé approximations at $z_0=1$ of arbitrary analytic functions F such that $F(z)=F(1/z)$. In §3 we use the order star theory to deduce that the order is *at most* $4r+1$. Finally, in §4 we apply standard complex analysis in conjunction with order stars to prove that the method of order $4r+1$ is stable for every $\omega\in\mathbb{C}, \operatorname{Re}\omega\geqq 0$. Note that $\operatorname{Re}\omega\geqq 0$ is also the condition for the asymptotic boundedness of the analytic solution of (1).

This is the place to mention that methods for (1) can be generalized for the equation

$$u_t=\nabla^2 u, \quad u=u(t,\underline{x}), \quad \underline{x}\in\mathbb{R}^N.$$

The simplest approach is by using product methods. In this case Crandall's scheme for $N=2$ yields

$$\begin{aligned}
&\left(\frac{25}{36}+\frac{5}{3}\mu+\mu^2\right)U^{n+1}_{m,k}\\
&\quad+\left(\frac{5}{72}-\frac{1}{3}\mu-\frac{1}{2}\mu^2\right)\left(U^{n+1}_{m+1,k}+U^{n+1}_{m-1,k}+U^{n+1}_{m,k+1}+U^{n+1}_{m,k-1}\right)\\
&\quad+\left(\frac{1}{144}-\frac{1}{12}\mu+\frac{1}{4}\mu^2\right)\left(U^{n+1}_{m+1,k+1}+U^{n+1}_{m-1,k+1}+U^{n+1}_{m+1,k-1}+U^{n+1}_{m-1,k-1}\right)\\
&\quad=\left(\frac{25}{36}-\frac{5}{3}\mu+\mu^2\right)U^{n}_{m,k}\\
&\quad+\left(\frac{5}{72}+\frac{1}{3}\mu-\frac{1}{2}\mu^2\right)\left(U^{n}_{m+1,k}+U^{n}_{m-1,k}+U^{n}_{m,k+1}+U^{n}_{m,k-1}\right)\\
&\quad+\left(\frac{1}{144}+\frac{1}{12}\mu+\frac{1}{4}\mu^2\right)\left(U^{n}_{m+1,k+1}+U^{n}_{m-1,k+1}+U^{n}_{m+1,k-1}+U^{n}_{m-1,k-1}\right),
\end{aligned}$$

$$U^n_{m,k}\approx u(n\Delta t; m\Delta x, k\Delta x), \qquad \mu:=\frac{\omega\Delta t}{(\Delta x)^2},$$

a stable fifth order method. However, a much neater and more useful scheme can be obtained with small effort, namely

$$\begin{aligned}
&\left(\frac{2}{3}+2\mu\right)U^{n+1}_{m,k}+\left(\frac{1}{12}-\frac{1}{2}\mu\right)\left(U^{n+1}_{m+1,k}+U^{n+1}_{m-1,k}+U^{n+1}_{m,k+1}+U^{n+1}_{m,k-1}\right)\\
&\quad=\left(\frac{2}{3}-\frac{4}{3}\mu\right)U^{n}_{m,k}+\left(\frac{1}{12}+\frac{1}{6}\mu\right)\left(U^{n}_{m+1,k}+U^{n}_{m-1,k}+U^{n}_{m,k+1}+U^{n}_{m,k-1}\right)\\
&\quad+\frac{1}{6}\mu\left(U^{n}_{m+1,k+1}+U^{n}_{m-1,k+1}+U^{n}_{m+1,k-1}+U^{n}_{m-1,k-1}\right).
\end{aligned}$$

Also this method is stable and of order five. However, it leads to a more sparse matrix which can be solved more economically.

2. On diagonal Padé approximations to F, $F(z)=F(1/z)$. Let F be a complex function, analytic about $z_0=1$ and satisfying the functional equation

$$F(z)=F\left(\frac{1}{z}\right) \tag{6}$$

for every z in an open neighbourhood of that point. Further, let $R=P/Q$, $\deg P = \deg Q = n$, be a rational approximation to F at $z_0=1$.

$$R(z)=F(z)+c(z-1)^{p+1}+O\left(|z-1|^{p+2}\right), \qquad c\neq 0. \tag{7}$$

Given that

$$P^*(z):=z^nP\left(\frac{1}{z}\right), \qquad Q^*(z):=z^nQ\left(\frac{1}{z}\right),$$

it follows from (6) that

$$R\left(\frac{1}{z}\right)=\frac{P^*(z)}{Q^*(z)}=F(z)+c(1-z)^{p+1}+O\left(|z-1|^{p+2}\right). \tag{8}$$

Thus

$$P(z)Q^*(z)-P^*(z)Q(z)=O\left(|z-1|^{p+1}\right). \tag{9}$$

Let R be the $[n/n]$ diagonal Padé approximation to F. Then $p \geqq 2n$ [1] and (9) implies that

$$R(z)=R\left(\frac{1}{z}\right).$$

It now follows from (7) and (8) that p must be odd.

THEOREM 1. *The diagonal of the Padé tableau of F is composed out of nontrivial blocks* (*i.e. q-by-q blocks with* $q>1$).

Proof. Follows at once from our analysis by the standard Padé theory [1]: given that the $[n/n]$ Padé approximation exists $p>2n$ implies that it must coincide with the $[n/n+1]$ and $[n+1/n]$ approximations. Hence it necessarily belongs to a nontriival block. □

The importance of Theorem 1 is clear; if we can prove that the order of the $[2n/2n]$ Padé approximation to F cannot exceed $4n+1$ then it must be *exactly* $4n+1$ and the approximation must belong to 2-by-2 block. This will be done in the next section for the function

$$F(z)=e^{\mu(\ln z)^2},$$

which is central to our analysis of numerical methods for $u_t=\omega u_{xx}$.

As an aside we note that the set of all functions F that satisfy the functional equation (6) and are sufficiently smooth can be readily characterized. Let F be analytic in $\mathbb{C}/(-\infty,0)$ and entire in the covering Riemann surface, such that

$$F(z)=F\left(\frac{1}{z}\right)$$

holds for every $z\in\mathbb{C}/(-\infty,0)$. We set

$$\tilde{F}(z):=F(e^z), \qquad z\in\mathbb{C}.$$

Since F is entire in the covering Riemann surface, $\tilde{F}$ is entire in $\mathbb{C}$. Furthermore, by (6). $\tilde{F}$ is even and so its Taylor expansion is of the form

$$\tilde{F}=\sum_{k=0}^{\infty}\tilde{F}_kz^{2k}.$$

Let

$$g(z)=\sum_{k=0}^{\infty}\tilde{F}_k z^k.$$

Then g is an entire function and

$$F(z)=g\big((\ln z)^2\big).$$

Therefore every function F is merely an entire function acting on $(\ln z)^2$. Moreover, unless g is constant, f has a branch cut along $(-\infty,0)$. Thus, if f is entire then it must be constant.

The results of the present section can be generalized to cater for functions F that satisfy the functional equation

$$F(z)=z^L F\left(\frac{1}{z}\right),$$

where L is an arbitrary integer. Such functions are obtained when noncentred schemes are used to solve the differential equation (1). However, as there is no apparent advantage to be gained from such schemes, we do not explore this topic further.

3. The maximal order of approximations to $f(z)=e^{\mu(\ln z)^2}$. In the present section we use the theory of order stars [10] to bound the order of rational approximation to

$$f(z)=e^{\mu(\ln z)^2}, \qquad \mu\in\mathbb{C}$$

at $z_0=1$. Indeed, one of the purposes of the present paper is to demonstrate the usefulness of order stars in solving problems in numerical analysis.

Firstly we will show, by applying order stars of the second kind that the order of a rational $[2n/2n]$ approximation to $(\ln z)^2$ at $z_0=1$ may not exceed $4n+1$. This will lead, by a simple argument, to the proof that $h>0$ exists such that, subject to $|\mu|\leqq h$, rational $[2n/2n]$ approximations to f have at most order $4n+1$. Secondly we will use order stars of the first kind to prove that no $[2n/2n]$ approximation to f with pure imaginary μ may exceed order $4n+1$. This result will be subsequently applied in §4 to stability analysis.

The theory of order stars was extensively explained in [10]. In the present paper we follow the notation and the terminology of [10]. In particular, the phrases "Proposition 1" etc. refer to that paper.

Let $\tilde{R}$ be a rational $[n/n]$ function with real cofficients that approximates $(\ln z)^2$ at $z_0=1$.

$$\tilde{R}(z)=(\ln z)^2+d(z-1)^{p+1}+O\big(|z-1|^{p+2}\big), \qquad d\neq 0.$$

We form the order star of the second kind with respect to

$$\sigma(z):=\tilde{R}(e^z)-z^2 \tag{10}$$

(cf. Fig. 1). The following facts, that can be eaisly derived from (10), are important to our analysis:

(a) If $\operatorname{Im} z\geqq -\pi$, then

$$\operatorname{Re}\sigma(z+2\pi i)=\operatorname{Re}\sigma(z)+4\pi^2+4\pi\operatorname{Im} z\geqq\operatorname{Re}\sigma(z).$$

Therefore, if $\operatorname{Im} z\geqq 0$ and $z\in A$, then also $z+2\pi i\in A$.

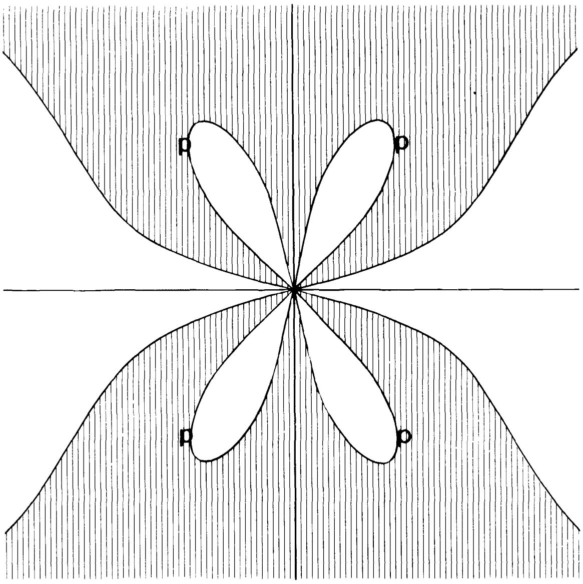


FIG. 1. *Order star of the second kind for*

$$\tilde{R}(z)=\frac{(z-1)^2}{1/12+5z/6+z^2/12}=(\ln z)^2+O\left(|z-1|^6\right)$$

in the rectangle $0<|\mathrm{Re}\, z|$, $|\mathrm{Im}\, z|\leqq 2\pi$. *Poles of* $\tilde{R}$ *are denoted by* P.

(b) The order star is symmetric with respect to the real axis. Therefore it follows from (a) that $\mathrm{Im}\, z\leqq\pi$ and $z\in A$ imply that $z-2\pi i\in A$.

LEMMA 2. *Let* $x+(\pi-\delta)i\in A$, *where* x, $\delta\in\mathbb{R}$ *and* $0\leqq\delta\leqq\pi$. *Then necessarily* $x+(\pi+\delta)i\in A$.

Proof. We set $z=x+(\pi-\delta)i$. It follows from (b) that $\bar{z}\in A$. Moreover, $\mathrm{Im}\,\bar{z}\geqq-\pi$ and (a) imply that $\bar{z}+2\pi i\in A$. This completes the proof, since $\bar{z}+2\pi i=x+(\pi+\delta)i$. □

(c) Let us suppose that $\tilde{R}$ is of the form

$$\tilde{R}(z)=\frac{\sum_{k=M_1}^{M_2}\tilde{a}_k z^k}{\sum_{k=N_1}^{N_2}\tilde{b}_k z^k},\qquad \tilde{a}_{M_1},\tilde{a}_{M_2},\tilde{b}_{N_1},\tilde{b}_{N_2}\neq 0,$$

where $\max\{M_2,N_2\}=n$, $\min\{M_1,N_1\}=0$. Therefore $\tilde{R}(z)$ has $(N_1-M_1)_+$ poles at ∞ and $(M_2-N_2)_+$ poles at the origin. We denote

$$I_0:=\{z\in\mathbb{C}\colon |\mathrm{Im} z|\leqq\pi\},\quad I_+:=\{z\in\mathbb{C}\colon \mathrm{Im}\, z>\pi\},\quad I_-:=\{z\in\mathbb{C}\colon \mathrm{Im}\, z<-\pi\}.$$

It follows from [10] that σ has at most $(M_2-N_2)_++1$ sectors of D in I_0 that approach $+\infty$ and at most $(N_1-M_1)_++1$ sectors of D there that approach $-\infty$.

(d) If a sector of D that approaches $\pm\infty$ in $I_+\cup I_-$ belongs to a D-region that adjoins the origin then, by Lemma 2, that D-region must also contain a sector of D that approaches $\pm\infty$ in I_0.

(e) All the finite essential singularities of the order star (which, of course. correspond to transformed poles of $\tilde{R}$), are repeated with a period of $2\pi i$. By Proposition 8 every bounded D-region and every bounded A-region must contain an essential singularity on its boundary. Let us assume that a D-region that adjoins the origin has an essential singularity from I_+ at $x_0+(\pi+\delta)i$, $\delta>0$ say, on its boundary. By Lemma 2 also the essential singularity at

$$x_0+\left(\left(\left[\frac{\delta}{\pi}\right]+1\right)\pi-\delta\right)i\in I_0$$

must belong to the boundary of the region.

LEMMA 3. *The order p of $\tilde{R}(z)$ as an approximation to* $(\ln z)^2$ *at* $z_0=1$ *may not exceed* $2n+1$.

Proof. We count the sectors of D that adjoin the origin in the order star of σ. Let

$$S:=(N_1-M_1)_+ + (M_2-N_2)_+$$

denote the number of poles of $\tilde{R}(z)$ at 0 and ∞. It follows by (c) that at most $S+2$ D-regions approach $\pm\infty$ in I_0 and, by (d), only these sectors need be considered in our count of sectors of D that approach the origin. In other words, at most $S+2$ sectors of D at the origin belong to unbounded D-regions.

Since all the remaining sectors of D necessarily belong to bounded D regions, their number is restricted, according to (e), by the number of finite essential singularities in I_0. Since $\tilde{R}(z)$ has at most $n-S$ poles in $\mathbb{C}/\{0\}$, there are in I_0 at most $2(n-S)$ finite essential singularities (because every essential singularity z_0 with $\operatorname{Im} z_0=\pm\pi$ is counted twice).

Thus the number of sectors of D at the origin is at most

$$\{S+2\}+\{2(n-S)\}=2n-S+2\leqq 2n+2.$$

According to Proposition 7 this number equals $p+1$, yielding the desired result. □

It follows at once from our method of proof that if $p=2n+1$, then $N_1=M_1=0$, $N_2=M_2=n$, $S=0$ and all the poles of $\tilde{R}(z)$ are negative. This is an interesting point, since the negative ray is a branch cut of $(\ln z)^2$.

We note in passing an interesting consequence of Lemma 3. Let

$$\sum_{j=-r}^{r} \nu_j U'_{m+j}(t)=\frac{\omega}{(\Delta x)^2}\sum_{j=-r}^{r}\delta_j U_{m+j}(t) \tag{11}$$

be an implicit semi-discretization of the differential equation (1), $U_m(t)\approx u(t,m(\Delta x))$, $m\in\mathbb{Z}$. Then, if

$$\tilde{R}(z):=\frac{\sum_{j=-r}^{r}\delta_j z^j}{\sum_{j=-r}^{r}\gamma_j z^j},$$

it follows easily from [9, Chap. 7] that the method is of order p if and only if $\tilde{R}$ is an order p approximation to $(\ln z)^2$ at $z_0=1$. Since $f(z)=(\ln z)^2$ satisfies $f(z)=f(1/z)$, we can use the theory of §2 in conjunction with the last theorem to prove that for every r there exists a unique implicit semi-discretization of (1) that attains order $4r+1$ and that no other such method may exceed this order. This, however, is of only marginal interest, since the solution of the linear ordinary differential system (11) by a numerical method of order $4r+1$ will necessarily increase the bandwidth, offsetting the benefits of high order. Apparently, it pays to discretize both space and time variables in (1) simultaneously.

Let $R(z;\mu)=P(z;\mu)/Q(z;\mu)$ be an approximation to $f(z)=e^{\mu(\ln z)^2}$ at $z_0=1$. We assume that $\deg P$, $\deg Q \leqq n$ (as polynomials in z) and that both functions are analytic in μ in the neighbourhood of $\mu=0$.

THEOREM 4. *There exists $h_n>0$ such that, subject to $0<|\mu|<h_n$, the order p of the approximation $R(z;\mu)$ may not exceed $2n+1$.*

Proof. We differentiate the expression

$$R(z;\mu)=e^{\mu(\ln z)^2}+c(\mu)(z-1)^{p+1}+O\left(|z-1|^{p+2}\right) \tag{12}$$

with respect to μ and set $\mu=0$. Since $R(z;0)=1$, we obtain

$$\tilde{R}(z)=(\ln z)^2+c'(0)(z-1)^{p+1}+O\left(|z-1|^{p+2}\right),\quad \tilde{R}=\frac{\tilde{P}}{\tilde{Q}},\quad \deg\tilde{P},\deg\tilde{Q}\leqq n.$$

Therefore $\tilde{R}$ is an approximation to $(\ln z)^2$ at least of order p. The statement of the theorem follows at once by Lemma 3. □

COROLLARY 5. *The order of the $[2n/2n]$ Padé approximation to* $\exp(\mu(\ln z)^2)$ *at $z_0=1$ is exactly $4n+1$ for every $0<|\mu|<h_n$.*

Proof. The coefficients of the approximation are analytic in μ, as can be seen at once from Lemma 7 and the Padé theory [1]. It has been proved in §2 that the order is at least $4n+1$. This, together with the last theorem, furnishes the desired result. □

Corollary 5 does not prevent an existence of a complex μ that gives higher order—it merely shows that this may not happen in a punctured neighbourhood of the origin. However, we will need in §4 a stronger result for the special case $\mu=it$, $t\in\mathbb{R}$, namely that the order is $4n+1$ regardless of the size of t. This is proved by order stars of the first kind:

We consider the order star with respect to

$$\sigma(z):=e^{-itz^2}R(e^z;it),$$

where $R(z;it)$ is a $[2n/2n]$ approximation to $\exp(it(\ln z)^2)$, $t\in\mathbb{R}/\{0\}$, of order $p\geqq 4n+1$. Let

$$R(z;it)=\frac{P(z;it)}{Q(z;it)}.$$

The order condition (12) implies that

$$|P(e^{i\theta};it)|^2-|Q(e^{i\theta};it)|^2=O(\theta^{p+1})=O\left((1-\cos\theta)^{[p/2]+1}\right).$$

Therefore, since both $|P(e^{i\theta};it)|^2$ and $|Q(e^{i\theta};it)|^2$ are polynomials in $1-\cos\theta$ of degree $2n-1<[p/2]+1$ it follows that

$$|P(e^{i\theta};it)|^2-|Q(e^{i\theta};it)|^2\equiv 0$$

and

$$|R(e^{i\theta};it)|\equiv 1,\qquad \theta\in\mathbb{R}.$$

Similarly, given $x\in\mathbb{R}$, (12) implies that

$$|P(e^x;it)|^2-|Q(e^x;it)|^2=O(x^{p+1})=O\left((e^x-1)^{p+1}\right)$$

and, since both $|P(e^x;it)|^2$ and $|Q(e^x;it)|^2$ are polynomials in e^x-1 of degree $4n<p+1$, it is true that

$$|R(e^x;it)|\equiv 1, \qquad x\in\mathbb{R}.$$

In other words, both the real and the pure imaginary axes in the order star belong to ∂ (cf. Fig. 2).

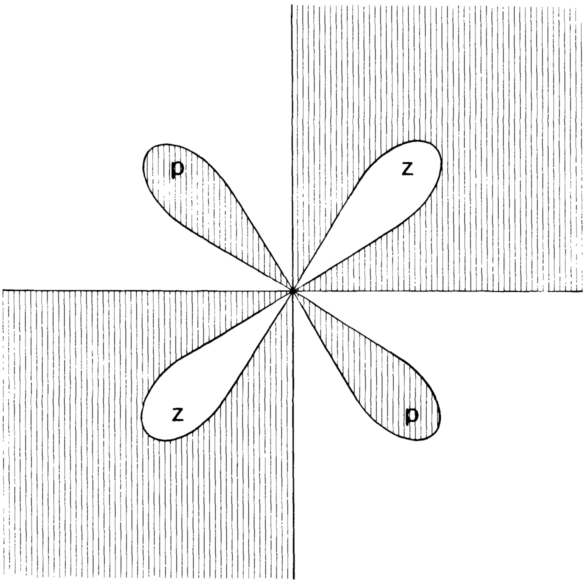


FIG. 2. *Order star of the first kind for*

$$R(z,2i)=\frac{z+(1/12+i)(z-1)^2}{z+(1/12-i)(z-1)^2}=e^{2i(\ln z)^2}+O\left(|z-1|^6\right),$$

in the rectangle $0\leq|\mathrm{Re}\, z|$, $|\mathrm{Im}\, z|\leq\pi/2$. *P and Z denote poles and zeros of R respectively.*

Proceeding as before, we show that

$$|\sigma(x+2\pi i)|=|\sigma(x)|e^{4\pi x}, \qquad x\in\mathbb{R}.$$

Hence, since $|\sigma(x)|\equiv 1$, $x+2\pi i\in A$ if $x>0$ and $x+2\pi i\in D$ if $x<0$.

It follows from Proposition 5 that $\mathrm{ind}(\infty)=2$ and ∞ is a regular point of ∂. Let us assume that $t>0$. Then it is obvious from

$$|\sigma(z)|=e^{2t(\mathrm{Re}\, z)(\mathrm{Im}\, z)}(1+o(1)) \qquad (|z|\gg 0),$$

that A tends to infinity in $\Omega_+ := \{z\in\mathbb{C}\colon (\mathrm{Re}\, z)(\mathrm{Im}\, z)>0\}$, D approaches infinity in $\Omega_- := \{z\in\mathbb{C}\colon (\mathrm{Re}\, z)(\mathrm{Im}\, z)<0\}$ and $\mathbb{R}$, $i\mathbb{R}\subset\partial$. Thus, all the D-regions in Ω_+ are bounded and, since σ is meromorphic, it follows from Proposition 2 that the sum of the multiplicities of these regions cannot exceed $2n$, the number of zeros of σ in I_0 (since $x+2\pi i$ belongs to A for $x>0$ and to D if $x<0$). It follows that at most $2n$ sectors of D reach the origin in Ω_+. Thus, since $\mathbb{R}$, $i\mathbb{R}\subset\partial$ and the origin is a regular point of ∂, at most $2n+2$ sectors of D reach the origin in Ω_-. It now follows by Proposition 1 that

$$p\leq(2n)+(2n+2)-1=4n+1. \tag{13}$$

Given $t<0$ we can show that (13) is true by exchanging the roles of A and D, i.e. counting sectors of A at the origin.

THEOREM 6. *The order of the* $[2n/2n]$ *Padé approximation to* $\exp(it(\ln z)^2)$ *at* $z_0=1$ *is exactly* $4n+1$ *for every* $t\in\mathbb{R}$, $t\neq 0$.

Proof. Since, by the analysis of §2, the order is at least $4n+1$, the desired result is a consequence of the inequality (13). □

4. Stability analysis. In the present section we aim to prove that the highest order method of Corollary 5 satisfies the von Neumann condition (4) and the Wiener–Hopf condition (5) and, consequently, is stable.

Let

$$R(z;\mu)=\frac{P(z;\mu)}{Q(z;\mu)}=e^{\mu(\ln z)^2}+O\left(|z-1|^{2n+2}\right),\qquad \deg P,\deg Q=n=2r+1,$$

(as polynomials in z). It follows that

$$R(e^{i\theta};\mu)=e^{-\mu\theta^2}+O(\theta^{2n+2})$$

and

$$\frac{1}{R(e^{i\theta};-\mu)}=e^{-\mu\theta^2}+O(\theta^{2n+2}).$$

Therefore

$$(14)\quad P(e^{i\theta};\mu)P(e^{i\theta};-\mu)-Q(e^{i\theta};\mu)Q(e^{i\theta};-\mu)=O(\theta^{2n+2})=O\left((1-e^{i\theta})^{2n+2}\right).$$

The function

$$P(e^{i\theta};\mu)P(e^{i\theta};-\mu)-Q(e^{i\theta};\mu)Q(e^{i\theta};-\mu)$$

is a polynomial of degree $2n$ in $1-e^{i\theta}$. Hence, it is a consequence of (14) that this polynomial identically vanishes and so

$$(15)\qquad P(e^{i\theta};\mu)=Q(e^{i\theta};-\mu).$$

We now proceed in a manner diametrically different from the standard stability analysis—instead of fixing μ and showing that the von Neumann condition, say, is satisfied for every $0\leqq\theta\leqq 2\pi$, we fix $0\leqq\theta\leqq 2\pi$ and intend to prove that

$$|R(e^{i\theta};\mu)|\leqq 1$$

for a range of values of μ.

To emphasize the dependence of μ, we write

$$G_\theta(\mu):=R(e^{i\theta};\mu).$$

Because of (15) it is true that

$$G_\theta(\mu)=\frac{g_\theta(\mu)}{g_\theta(-\mu)},$$

where $g_\theta(\cdot)$ is a polynomial in $e^{i\theta}$.

LEMMA 7. *The kth coefficient of the Taylor expansion of* $f(z)=\exp(\mu(\ln z)^2)$ *about* $z_0=1$ *is* $[k/2]$*th degree polynomial in* μ *for every* $k\geqq 0$.

Proof. Let

$$f(z)=\sum_{k=0}^{\infty}\frac{1}{k!}C_k(\mu)(z-1)^k.$$

It is easily proved by induction that

$$\frac{d^k}{dz^k}f(z)=\left(\sum_{j=0}^{k}a_{k,j}(\ln z)^j\right)z^{-k}f(z),\qquad k\geqq 0, \tag{16}$$

where the coefficients $a_{k,j}$, $0\leqq j\leqq k$, $k\geqq 0$, satisfy the partial difference equation

$$\begin{aligned} &a_{k+1,0}=-ka_{k,0}+a_{k,1};\\ &a_{k+1,j}=-2\mu a_{k,j-1}-ka_{k,j}+(j+1)a_{k,j+1},\qquad 1\leqq j\leqq k;\\ &a_{k+1,k+1}=2\mu a_{k,k} \end{aligned} \tag{17}$$

for every $k\geqq 0$. Since $a_{0,0}=1$, it follows at once by induction that each $a_{k,j}$ is a polynomial in μ of degree $[(k+j)/2]$.

The formula (16) implies that

$$C_k(\mu)=a_{k,0},$$

a polynomial of degree $[k/2]$. □

It is straightforward to prove from the last lemma, by using the Padé theory, that g_θ is a polynomial in μ. However, we need a more elaborate piece of information, namely the exact degree of that polynomial.

Let us denote by $\alpha_{k,j}$ the coefficient of $\mu^{[(k+j)/2]}$ in $a_{k,j}$. It follows from (17) that the $\alpha_{k,j}$'s satisfy the difference equation

$$\begin{aligned} &k+j\text{ even:}\qquad \alpha_{k+1,j}=2\alpha_{k,j-1}-k\alpha_{k,j}+(j+1)\alpha_{k,j+1};\\ &k+j\text{ odd;}\qquad \alpha_{k+1,j}=2\alpha_{k,j-1}+(j+1)\alpha_{k,j+1}, \end{aligned} \tag{18}$$

where we set $\alpha_{k,j}=0$ for $j\leqq 1$ or $j\geqq k+1$.

LEMMA 8. *$(-1)^{k+j}\alpha_{k,j}>0$ for every $0\leqq j\leqq k$.*

Proof. By induction on $k+j$. It follows from (18) that

$$\begin{aligned} \alpha_{k+1,k-2s}&=2\alpha_{k,k-1-2s}-k\alpha_{k,k-2s}+(k-2s+1)\alpha_{k,k+1-2s}\\ &=-\left\{2|\alpha_{k,k-1-2s}|+k|\alpha_{k,k-2s}|+(k-2s+1)|\alpha_{k,k+1-2s}|\right\}<0,\qquad 0\leqq s\leqq\left[\frac{k}{2}\right] \end{aligned}$$

and

$$\begin{aligned} \alpha_{k+1,k+1-2s}&=2\alpha_{k,k-2s}+(k-2s+2)\alpha_{k,k-2s+2}\\ &=2|\alpha_{k,k-2s}|+(k-2s+2)|\alpha_{k,k-2s+2}|>0,\qquad 0\leqq s\leqq\left[\frac{k+1}{2}\right]. \end{aligned}$$

Hence the lemma is true. □

It follows that each $C_k(\mu)$ is a polynomial of exact degree $[k/2]$. The C_k's for $0\leqq k\leqq 9$ are given in Table 1.

TABLE 1

The Taylor coefficients of $f(z)=\exp(\mu(\ln z)^2)$ *about* $z_0=1$.

$$C_0(\mu)\equiv 1;$$
$$C_1(\mu)\equiv 0;$$
$$C_2(\mu)=\mu;$$
$$C_3(\mu)=-\mu;$$
$$C_4(\mu)=\frac{11}{12}\mu+\frac{1}{2}\mu^2;$$
$$C_5(\mu)=-\frac{5}{6}\mu-\mu^2;$$
$$C_6(\mu)=\frac{137}{180}\mu+\frac{17}{12}\mu^2+\frac{1}{6}\mu^3;$$
$$C_7(\mu)=-\frac{7}{10}\mu-\frac{7}{4}\mu^2-\frac{1}{2}\mu^3;$$
$$C_8(\mu)=\frac{363}{560}\mu+\frac{967}{480}\mu^2+\frac{23}{24}\mu^3+\frac{1}{24}\mu^4;$$
$$C_9(\mu)=-\frac{761}{1260}\mu-\frac{89}{40}\mu^2-\frac{3}{2}\mu^3-\frac{1}{6}\mu^4.$$

LEMMA 9. *Let* $Q(e^{i\theta};\mu)=g_\theta(-\mu)=\sum_{k=0}^{n}q_k e^{ik\theta}$. *Then each* q_k, $1\leqq k\leqq n$, *is a polynomial in* μ *of exact degree* $\frac{1}{2}n(n-1)+k$, *whereas* q_0 *is of exact degree* $\frac{1}{2}n^2$.

Proof. It follows from the Padé theory [1] that, subject to the normalization

$$q_0=-\left[\begin{array}{cccc} C_n & C_{n-1} & \cdots & C_1 \\ C_{n+1} & C_n & \cdots & C_2 \\ \hdashline C_{2n-1} & C_{2n-2} & \cdots & C_n \end{array}\right], \tag{19}$$

each q_k, $1\leqq k\leqq n$, is of the form

$$q_k=\left[\begin{array}{ccccccc} C_{n+1} & C_n & \cdots & C_{n+2-k} & C_{n-k} & \cdots & C_1 \\ C_{n+2} & C_{n+1} & \cdots & C_{n+1-k} & C_{n-1-k} & \cdots & C_2 \\ \hdashline C_{2n} & C_{2n-1} & \cdots & C_{2n+1-k} & C_{2n-1-k} & \cdots & C_n \end{array}\right].$$

Since each C_j is a polynomial in μ, so is q_k and it follows from Lemma 8 that, because n is even, the coefficient of the highest power of μ in q_k is

$$(-1)^n C_{2n}C_{2(n-1)}\cdots C_{2(n-k+1)}C_{2(n-k)-1}C_{2(n-k)-3}\cdots C_1\neq 0.$$

Therefore, by that lemma, q_k is a polynomial of the exact degree $\frac{1}{2}n(n-1)+k$, $1\leq k\leq n$.

Also q_0 is a polynomial in μ. The highest power of μ in q_0 is obtained, by Lemma 8, from (19) by following the main diagonal. This gives degree $\frac{1}{2}n^2$. □

It is a consequence of the last lemma that the degree of g_θ as a polynomial in μ is $\frac{1}{2}n(n+1)$, regardless of the value of $\theta\neq 0 \bmod 2\pi$.

LEMMA 10. *The lowest power of* μ *in each* q_k, $1\leq k\leq n$, *is* μ^n.

Proof. Follows in exactly the same manner as Lemmas 7–9, by considering the lowest powers of C_j instead of the highest powers. □

Since the factor of μ^n is repeated in both numerator and denominator of G_θ it can be removed. We obtain a rational $[m/m]$ function in μ, where $m=\frac{1}{2}n(n-1)$. In other words, for every value of $0<\theta<2\pi$ G_θ has m poles and m zeros in $\mathbb{C}$ and neither zero nor pole can travel to ∞.

Let $R^*(z)=P^*(z)/P^*(-z)$ denote the $[m/m]$ Padé approximation to the exponential [1]. Since

$$G_\theta(\mu)=e^{-\mu\theta^2}=O(\theta^{2n+2}),$$

it follows that

$$G_\theta(\mu)=R^*(-\mu\theta^2)+O(\theta^{2\min\{m,n\}+1}). \tag{20}$$

Let ξ_0 be a zero of P^* and $\mu_0:=-\xi_0/\theta^2$. Then it follows from (20) that

$$g_\theta(\mu_0)=O(\theta^{2\min\{m,n\}+1}).$$

Hence, if $|\theta|\to 0$, then zeros of $g_\theta(\mu)$ and $P^*(-\mu\theta^2)$ are arbitrarily close. Since P^* and g_θ are of the same degree in μ, this is true regarding all their zeros.

According to [16] all the zeros of P^* are in $\mathbb{C}^-=\{\mu\in\mathbb{C}\colon \operatorname{Re}\mu<0\}$. Therefore, for $0<|\theta|\ll 2\pi$ all the zeros of G_θ are in $\mathbb{C}^+=\{\mu\in\mathbb{C}\colon \operatorname{Re}\mu>0\}$ and all the poles are in $\mathbb{C}^-$.

LEMMA 11. *G_θ is analytic in* $\operatorname{cl}\mathbb{C}^+=\mathbb{C}^+\cup\mathbb{R}$ *for every value of* $0\leqq\theta\leqq 2\pi$.

Proof. We already know that G_θ is analytic there for $0\leqq\theta\ll 2\pi$. Let us assume that there exists θ_0 in $(0,2\pi)$ such that G_{θ_0} has a pole in $\operatorname{cl}\mathbb{C}^+$. The poles of G_θ are a continuous function of θ, since no pole can jump through ∞. Therefore there must exist θ_1 in $(0,2\pi)$ such that G_{θ_1} has a pole on $i\mathbb{R}$. However, since $|G_\theta(it)|\equiv 1$ for every real t, it follows that this pole coalesces there with a zero. Hence $m_1\leqq m-1$ exists such that G_θ is a $[m_1/m_1]$ rational function. Moreover, this reduction must also lower the degree of G_θ as a rational function in $e^{i\theta}$, since the locus of this zero-pole pair is a nontrivial function of θ. This is a contradiction of Theorem 6, since $G_\theta(it)\equiv R(e^{i\theta};it)$ is an approximation of order $2n+1$ to $e^{-it\theta^2}$. Consequently, no such θ_0 exists and G_θ is analytic in $\operatorname{cl}\mathbb{C}^+$. □

We can now formulate and prove the main theorem of this section.

THEOREM 12. *The approximation $R(z;\mu)$ of order $2n+1$ corresponds to a stable method for every $\mu\in\operatorname{cl}\mathbb{C}^+$.*

Proof. The function G_θ is analytic for every $\mu\in\operatorname{cl}\mathbb{C}^+$ and $|G_\theta(\mu)|\equiv 1$ for every $\mu\in\mathbb{R}$. Therefore, by the maximal modulus principle,

$$|R(e^{i\theta};\mu)|=|G_\theta(\mu)|<1, \qquad \mu\in\mathbb{C}^+. \tag{21}$$

Since this is true for every $0\leqq\theta\leqq 2\pi$. It follows by (4) that the underlying full discretization (2) of the differential equation (1) satisfies the von Neumann condition for every $\mu\in\operatorname{cl}\mathbb{C}^+$.

The Wiener–Hopf condition is satisfied for $\mu\in i\mathbb{R}$, since then (cf. the proof of Theorem 6) $Q(e^z;\mu)$ has exactly $n/2$ zeros in Ω_+ and $n/2$ zeros in Ω_-.

Let us assume that this condition is violated by some $\mu_1\in\mathbb{C}^+$. Since, by Lemma 9, the zeros of Q are a continuous function of μ and the Wiener-Hopf condition holds for $\mu\in\mathbb{R}$ it follows that there exists $\mu_2\in\mathbb{C}^+$ such that $Q(\cdot,\mu_2)$ has a zero on the perimeter of the unit disk, at $e^{i\theta_2}$ say. This, however, is impossible, since G_{θ_2} is analytic in $\mathbb{C}^+$. Consequently no such μ_1 exists and the Wiener-Hopf condition holds for every $\mu\in\operatorname{cl}\mathbb{C}^+$. This completes the proof of stability. □

Note that the highest-order method of the last theorem is, by (21), dissipative for $\mu\in\mathbb{C}^+$ and conservative for $\mu\in\mathbb{R}$. This mimics the behaviour of the analytic solution of the differential equation (1).

REFERENCES

[1] G. A. BAKER, *Essentials of Padé Approximants*, Academic Press, New York, 1975.

[2] C. BOLLEY AND M. CROUZEIX, *Conservation de la positivité lors de la discretization des problèmes d'évolution paraboliques*, RAIRO Analyse Numér., 12 (1978), pp. 237–245.

[3] S. H. CRANDALL, *An optimum implicit recurrence formula for the heat conduction equation*, J. Assoc. Comput. Mach., 13 (1955), pp. 318.

[4] G. DAHLQUIST, *Convergence and stability in numerical integration of ordinary differential equations*, Math. Scand., 4 (1956), pp. 33–53.

[5] ———, *A special stability problem for linear multistep methods*, BIT, 3 (1963), pp. 27–43.

[6] S. K. GODUNOV, *A difference method for the numerical calculation of discontinuous solutions of hydrodynamic equations*, Mat. Sbornik, 47 (1959), pp. 271–306.

[7] B. GUSTAFSSON, H.-O. KREISS AND A. SUNDSTROM, *Stability theory of difference approximations for mixed initial boundary value problems* II, Math. Comp., 26 (1972), pp. 649–686.

[8] A. ISERLES, *Order stars and a saturation theorem for first-order hyperbolics*, IMA J. Numer. Anal., 2 (1982), pp. 49–61.

[9] ———, *Numerical Analysis of Differential Equations*, to appear.

[10] ———, *Order stars, approximations and finite differences. I. The general theory of order stars*, this Journal 16 (1985), pp. 559–576.

[11] A. ISERLES AND S. P. NØRSETT, *A proof of the first Dahlquist barrier by order stars*, BIT, 25 (1985), to appear.

[12] A. ISERLES AND G. STRANG, *The optimal accuracy of difference schemes*, Trans. Amer. Math. Soc., 227 (1983), pp. 779–803.

[13] A. ISERLES AND R. A. WILLIAMSON, *Order and stability of implicit difference schemes*, IMA J. Numer. Anal., 4 (1984), pp. 289–307.

[14] E. D. RAINVILLE, *Special Functions*, Macmillan, New York, 1967.

[15] P. L. ROE, *Numerical algorithms for the linear wave equation*, Tech. Report, Royal Aircraft Establishment, Bedford, England, 1980.

[16] G. WANNER, E. HAIRER, AND S. P. NØRSETT, *Order stars and stability theorems*, BIT, 18 (1978), pp. 475–489.

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

009

EXTENDED INITIAL AND FORCING FUNCTION SEMIGROUPS GENERATED BY A FUNCTIONAL EQUATION*

OLOF J. STAFFANS[†]

Abstract. We present two types of semigroups generated by functional equations of the form

$$x(t)+\mu * x(t)=f(t), \qquad t\geqq 0,$$
$$x(t)=\varphi(t), \qquad t\leqq 0.$$

One of them is an extended initial function type semigroup, and the other an extended forcing function type semigroup. These two types are adjoints of each other in the sense that the adjoint of a semigroup of one of the two types is of the other type. They are also equivalent in the sense that there is a one-to-one, bicontinuous mapping of the state space to itself, which maps a semigroup of the second type into a semigroup of the first type. In particular, it suffices to study the asymptotic behavior of one of the two types of semigroups, because the results can easily be transferred to the other type of semigroups.

1. Introduction. We discuss two types of semigroups generated by functional equations of the form

$$x(t)+\mu * x(t)=f(t), \qquad t\in \mathbf{R}^+, \tag{1.1}$$

with initial condition

$$x(t)=\varphi(t), \qquad t\in \mathbf{R}^-. \tag{1.2}$$

Here $\mathbf{R}^+=[0,\infty)$, $\mathbf{R}^-=(-\infty,0]$, the values of x, f and φ lie in $\mathbf{R}^n$, and μ is an n by n matrix valued measure on $\mathbf{R}^+$, which is not allowed to have a point mass at zero. The convolution $\mu * x$ is defined a.e. by

$$\mu * x=\int_{[0,\infty)} [d\mu(s)]\,x(t-s), \qquad t\in \mathbf{R}^+. \tag{1.3}$$

Equation (1.1) is fairly closely related to the retarded equation

$$\frac{d}{dt}x(t)+\nu * x(t)=f(t), \qquad t\in \mathbf{R}^+, \tag{1.4}$$

and the more general neutral equation

$$\frac{d}{dt}(x(t)+\mu * x(t))+\nu * x(t)=f(t), \qquad t\in \mathbf{R}^+. \tag{1.5}$$

Here ν is another n by n matrix valued measure on $\mathbf{R}^+$. For a long time (1.4) was the most studied equation out of (1.1), (1.4) and (1.5), but lately also (1.1) and (1.5) have received considerable attention. In this paper we shall discuss only (1.1), but the results can be modified so that they apply also to (1.4) and (1.5) (see [33]). We have chosen to first study (1.1) rather than (1.4) or (1.5) for the simple reason that technically (1.1) is a simpler equation than (1.4) and especially (1.5). Once one understands the behavior of (1.1) it is easier to understand the corresponding results for (1.4) and (1.5) given in [33].

* Received by the editors May 24, 1983 and in revised form July 15, 1984.

[†] Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo 15, Finland.

The idea behind the classical semigroup approach of Hale [10], applied to (1.1), is the following: Take f in (1.1) to be zero, and solve (1.1) with initial condition (1.2). Fix $t>0$, and define

$$\begin{aligned} x_t(s) &= x(s+t), \qquad s \in \mathbf{R}, \\ \varphi_t(s) &= x(s+t), \qquad s \in \mathbf{R}^-. \end{aligned}$$

Then x_t is again a solution of (1.1) with $f=0$, and with initial function φ_t. The mapping which takes φ into φ_t turns out to be a semigroup. We shall call this semigroup for the initial function semigroup generated by (1.1).

Miller and Sell [22], and later also Diekmann and van Gils [7], [8], [9] use a different approach. Roughly, what they do is to take φ to be zero in (1.2) instead of taking f to be zero in (1.1), and argue as above. The exact procedure is slightly more complicated to describe, but conceptually it is quite similar to the procedure leading to the initial function semigroup. We shall call Miller's and Sell's semigroup for the forcing function semigroup generated by (1.1). In a sense it describes the evolution of the forcing function in (1.1) rather than the evolution of the initial function in (1.2).

When the forcing function semigroup was discovered, it was not immediately related to the initial function semigroup. In 1976 Burns and Herdman [3] proved that for a certain equation of the type (1.4), the initial function and the forcing function types of semigroup are adjoints of each other. Since then the same result has been extended in the finite delay case to equation (1.1) [7] and to equation (1.5) [27]. From Theorems 3.1–3.3 below one can conclude that the same result is true for equation (1.1) in the infinite delay setting which we use here (the corresponding infinite delay result for (1.5) is given in [33]).

In the finite delay case the initial function and the forcing function semigroup generated by (1.1) are equivalent to each other, i.e. they can be mapped continuously and one-to-one onto each other (see [9], [19], [27], or the summary paper [32]). In the infinite delay case they are not in general equivalent (see [32]).

In 1982, the author made the trivial observation that one can allow both φ and f to be nonzero in the argument sketched above which leads to the initial function semigroups for (1.1), (1.4) and (1.5) [31]. Doing so one gets certain combined initial-forcing function semigroups. As the method used to construct these semigroups is a trivial modification of the method used to construct the usual initial functions semigroups, we shall call these semigroups for extended initial function semigroups.

The two main reasons for this work were that we wanted to find the adjoints of the extended initial function semigroups, and to investigate what the exact relationship is between the extended initial function semigroups and the standard forcing function semigroups. As we already mentioned above, here we only treat the simplest case (1.1), and return to (1.5) in [33].

For the benefit of the reader (who is not likely to be familiar with the rather abstract paper [31]) we first describe our extended initial function semigroup for (1.1) in §2. The state space which we use is of the same type as the state space in [28], apart from the fact that here the state space contains both initial functions and forcing functions. For more details the reader is referred to [28] (or to [30], where the same type of spaces appear). It has an infinite delay, and the assumption on the kernel is minimal. We already mentioned above that in earlier comparable works f is taken to be zero (or φ to be zero in the forcing function semigroups). In addition, in these works either the delay is finite, or the kernel is required to have more smoothness than we require, so even in the case $f=0$ we get certain technical improvements of earlier results.

In §3 we get to the heart of the matter, and compute the adjoint of the extended initial function semigroup. This straightforward computation has been inspired mainly by [3], [7] and [34] (later the author found out that the same type of computations have been made in [2], [6] and [19] for the retarded case, and in [27] for the neutral case). The final result of the computation has a very simple interpretation: The adjoint of the extended initial function semigroup is an "extended forcing function semigroup", which one obtains from the standard forcing function semigroup by adding an initial function component. This initial function component has no influence whatsoever on the forcing function part of the semigroup. It merely records the old values of the solution x, which are normally lost in a semigroup of forcing function type.

As a result of the computation in §3 we now have two combined initial-forcing function type semigroups, an extended initial function semigroup, and an extended forcing function semigroup. As we show in §4, it is an almost trivial task to prove that the two initial-forcing function semigroups are equivalent to each other. The equivalence operator is an extremely simple one: To go in one direction one just adds an initial function correction to the forcing function, and to go in the other direction one subtracts the same correction term. Recall that without the extension, the initial and the forcing function semigroups are not equivalent to each other in general in our infinite delay setting.

The fact that the two initial-forcing function semigroups are equivalent provides us with an answer to the second of the two questions which motivated this work. The extended initial function semigroup contains not only the ordinary initial function semigroup. It also contains the ordinary forcing function semigroup in the sense that the extended initial function semigroup can be mapped continuously onto the ordinary forcing function semigroup. This mapping is not in general one-to-one, due to the fact that the mapping which deletes the initial function component of the extended forcing function semigroup is not one-to-one.

In §§5 and 6 we show how the equivalence relation between the two different extended semigroups can be used to transfer some known results from one of the two semigroups to the other. More specifically, in §5 we describe the generators of the two semigroups, and in §6 we show how one under appropriate assumptions can decompose the two semigroups into parts with different exponential growth rate. The discussion in §6 makes fairly heavy use of the asymptotic results for the neutral equation given in [28] and [30].

2. The state space and the extended initial function semigroup. We shall throughout use a state space of the type $\mathscr{B}\times\mathscr{F}$, where $\mathscr{B}$ is a space of initial functions, defined on $\mathbf{R}^-$, and $\mathscr{F}$ is a space of forcing functions, defined on $\mathbf{R}^+$. We let both $\mathscr{B}$ and $\mathscr{F}$ be of the "fading memory type" described in [28]. For the convenience of the reader, let us here recall the most important results concerning these spaces from [28].

Let η: $\mathbf{R}\to(0,\infty)$ be a continuous function, normalized so that $\eta(0)=1$, and define

$$\rho_\eta(t)=\sup_{s\in\mathbf{R}}\frac{\eta(s+t)}{\eta(s)}\qquad(t\in\mathbf{R}).\tag{2.1}$$

Suppose that $\rho_\eta(t)$ is finite for each t, and continuous at zero. Observe that ρ_η is submultiplicative, i.e.

$$\rho_\eta(s+t)\leqq\rho_\eta(s)\rho_\eta(t)\qquad(s,t\in\mathbf{R}),\tag{2.2}$$

and that

$$\eta(s+t) \leqq \rho_\eta(s)\eta(t) \qquad (s,t \in \mathbf{R}). \tag{2.3}$$

The continuity of ρ_η at zero together with (2.2) implies that ρ_η is continuous. We call a function η of this type an *influence function*, and call ρ_η the *dominating function* induced by η.

Influence functions can be used to define certain *memory spaces*. We let $L^p(\mathbf{R};\mathbf{R}^n;\eta)$ $(1 \leqq p \leqq \infty)$ be the Banach space of measurable functions y: $\mathbf{R} \to \mathbf{R}^n$, with norm

$$\|y\| = \begin{cases} \left[\int_{\mathbf{R}} \left[\eta(t)\|y(t)\|\right]^p dt\right]^{1/p} & (1 \leqq p < \infty), \\ \operatorname*{ess\,sup}_{t \in \mathbf{R}} \eta(t)\|y(t)\| & (p = \infty) \end{cases}$$

The translation operator τ_t, defined by

$$\tau_t y(s) = y(s+t) \qquad (s,t \in \mathbf{R})$$

(for almost all s) is a continuous linear operator in $L^p(\mathbf{R};\mathbf{R}^n;\eta)$ (for fixed t), with norm

$$\|\tau_t\| = \rho_\eta(-t), \tag{2.4}$$

as is easily seen (cf. [28, Lemma 2.2.]). It is strongly continuous in t for $1 \leqq p < \infty$, but not for $p = \infty$. Therefore, we shall also consider the Banach space $BUC(\mathbf{R};\mathbf{R}^n;\eta)$ of continuous functions y: $\mathbf{R} \to \mathbf{R}^n$ such that ηy is uniformly continuous, with the norm of $L^\infty(\mathbf{R};\mathbf{R}^n;\eta)$. In this space τ_t is strongly continuous (cf. [28, Lemmas 2.2 and 2.5]). We let $BC_0(\mathbf{R};\mathbf{R}^n;\eta)$ consist of those functions y in $BUC(\mathbf{R};\mathbf{R}^n;\eta)$ which satisfy $\eta(t)y(t) \to 0$ $(t \to \pm\infty)$. Clearly $BC_0(\mathbf{R};\mathbf{R}^n;\eta)$ is a closed subspace of $BUC(\mathbf{R};\mathbf{R}^n;\eta)$, which in turn is a closed subspace of $L^\infty(\mathbf{R},\mathbf{R}^n;\eta)$.

In the preceding memory function space notation, when we replace $\mathbf{R}$ by $\mathbf{R}^-$ or $\mathbf{R}^+$ or some other interval I, then we mean the space which one gets by restricting each function in the memory space in question to $\mathbf{R}^+$ or $\mathbf{R}^-$ or I.

We denote the space of all real n by n matrices by $\mathbf{R}^{n\times n}$, and let $M(\mathbf{R};\mathbf{R}^{n\times n};\rho_\eta)$ (where ρ_η is defined as in (2.1)) be the set of $\mathbf{R}^{n\times n}$-valued measures μ on $\mathbf{R}$, satisfying

$$\|\mu\| = \int_{\mathbf{R}} \rho_\eta(t)\,d|\mu|(t) < \infty.$$

If μ is supported on $\mathbf{R}^+$, then we write $\mu \in M(\mathbf{R}^+;\mathbf{R}^{n\times n};\rho_\rho)$.

It was shown in [28] that if $\mu \in M(\mathbf{R};\mathbf{R}^{n\times n};\rho_\eta)$ and x belongs to either $L^p(\mathbf{R};\mathbf{R}^n;\eta)$, $1 \leqq p \leqq \infty$, or to $BUC(\mathbf{R};\mathbf{R}^n;\eta)$, or to $BC_0(\mathbf{R};\mathbf{R}^n;\eta)$, then $\mu * x$, defined a.e. by

$$(\mu * x)(t) = \int_{\mathbf{R}} [d\mu(s)]\,x(t-s),$$

belongs to the same space as x, and that $\|\mu * x\| \leqq \|\mu\|\|x\|$. If $\mu \in M(\mathbf{R}^+;\mathbf{R}^{n\times n};\rho_\eta)$, then we can also define the convolution $\mu * x$ of μ with a function x whose restriction to $(-\infty,T)$ belongs to $L^p((-\infty,T);\mathbf{R}^n;\eta)$ for some p, $1 \leqq p \leqq \infty$, and every T, $-\infty < T < \infty$. In this case also the restriction of $\mu * x$ to $(-\infty,T)$ belongs to $L^p((-\infty,T);\mathbf{R}^n;\eta)$ for all T. The same statement is true with L^p replaced by BUC and by BC_0.

Two measures μ and ν can also be convolved with each other, and $\|\mu * \nu\| \leqq \|\mu\|\,\|\nu\|$ (see [28]).

If $\mu \in M(\mathbf{R}^+; \mathbf{R}^{n\times n}; \rho_\eta)$, and μ has no point mass at zero, then equation (1.1) has a fundamental solution χ, which is a measure on $M(\mathbf{R}^+; \mathbf{R}^{n\times n}; e^{-dt})$ for some sufficiently large number d. This measure satisfies

$$\chi + \mu * \chi = \chi + \chi * \mu = \delta, \tag{2.5}$$

where δ is the identity point mass at zero. The solution x of (1.1) with initial condition (1.2) is

$$x = \varphi + \chi * (f + F\varphi), \tag{2.6}$$

where we have defined f and $F\varphi$ to be zero on $(-\infty, 0)$, φ to be zero on $\mathbf{R}^+$, and $F\varphi$ is the initial function correction

$$(f\varphi)(t) = -\int_{(t,\infty)} [d_\mu(s)]\varphi(t-s), \qquad t \geqq 0 \tag{2.7}$$

to the forcing function. For details, see [28], [29] and [31].

LEMMA 2.1. *Let Y be one of the spaces L^p, $1 \leqq p \leqq \infty$, or BUC or BC_0, let η be an influence function with associated dominating function ${}_\eta$, and define $\mathscr{B} = Y(\mathbf{R}^-; \mathbf{R}^n; \eta)$, $\mathscr{F} = Y(\mathbf{R}^+; \mathbf{R}^n; \eta)$. Suppose that $\mu \in M(\mathbf{R}^+; \mathbf{R}^{n\times n}; \rho_\eta)$ has no point mass at zero. For each $\varphi \in \mathscr{B}$ and each $f \in \mathscr{F}$, let $x(\varphi, f)$ be the solution of* (1.1) *with initial condition* (1.2), *and define $T(t)(\varphi, f) = (x_t(\varphi, f), f_t)$, where $x_t(\varphi, f)$ is the restriction of $\tau_t x(\varphi, f)$ to $\mathbf{R}^-$, and f_t is the restriction of $\tau_t f$ to $\mathbf{R}^+$. Then in the L^p-case, $1 \leqq p < \infty$, $T(t)$ is a strongly continuous semigroup in $\mathscr{B} \times \mathscr{F}$, and in the continuous case, $T(t)$ is a strongly continuous semigroup in the space*

$$\{(\varphi, f) \in \mathscr{B} \times \mathscr{F} \mid M(\varphi, f) = 0\},$$

where

$$M(\varphi, f) = f(0) - \varphi(0) - \mu * \varphi(0). \tag{2.8}$$

Again, for details, see [28] and [31]. As we mentioned above, we shall call the semigroup $T(t)$ in Lemma 2.1 for the extended initial function semigroup. Observe that $M(\varphi, f)$ also can be expressed in terms of F, namely

$$M(\varphi, f) = (f + F\varphi)(0) - \varphi(0). \tag{2.9}$$

3. The adjoint of the extended initial function semigroup. The semigroup $T(t)$ has an adjoint semigroup $T^+(t)$, which we want to compute. In the reflexive case when $\mathscr{B} \times \mathscr{F} = L^p(\mathbf{R}^-; \mathbf{R}^n; \eta) \times L^p(\mathbf{R}^+; \mathbf{R}^n; \eta)$ for some p, $1 < p < \infty$, one has $T^+(t) = T^*(t)$, where for each t, $T^*(t)$ is the adjoint of $T(t)$. In the nonreflexive cases $T^+(t)$ is a restriction of $T^*(t)$ to a subspace of the dual space of $\mathscr{B} \times \mathscr{F}$. Therefore, let us first compute $T^*(t)$.

Before we can find $T^*(t)$ we have to fix a representation of the dual space of $\mathscr{B} \times \mathscr{F}$. We first consider the case when $\mathscr{B} \times \mathscr{F} = L^p(\mathbf{R}^-; \mathbf{R}^n; \eta) \times L^p(\mathbf{R}^+; \mathbf{R}^n; \eta)$, with $1 \leqq p \leqq \infty$ (the case BC_0 will be treated later, but we shall not discuss the dual of BUC). Let q be the conjugate index p. We can identify the dual of $L^p(\mathbf{R}; \mathbf{R}^n; \eta)$ with e.g. $L^q(\mathbf{R}; \mathbf{R}^n; \tilde{\eta})$, where

$$\tilde{\eta}(t) = [\eta(-t)]^{-1}, \qquad t \in \mathbf{R},$$

through the duality mapping $\langle\ ,\ \rangle: L^q(\mathbf{R};\mathbf{R}^n;\tilde{\eta})\times L^p(\mathbf{R};\mathbf{R}^n;\eta)\to\mathbf{R}$ defined by

$$\langle x^*,x\rangle=\int_{\mathbf{R}} x^*(-s)x(s)\,ds. \tag{3.1}$$

Here we think of elements in $L^q(\mathbf{R};\mathbf{R}^n;\tilde{\eta})$ as row vector functions, and those in $L^p(\mathbf{R};\mathbf{R}^n;\eta)$ as column vector functions, so that $x^*(-s)x(s)$ is the inner product in $\mathbf{R}^n$ of $x^*(-s)$ and $x(s)$. Defining

$$x^**x(t)=\int_{\mathbf{R}} x^*(t-s)x(s)\,ds,\qquad t\in\mathbf{R},$$

we have

$$\langle x^*,x\rangle=x^**x(0),$$

and this is one reason for the introduction of the extra minus sign in the right-hand side of (3.1).

The preceding duality mapping carries over directly to $\mathscr{B}\times\mathscr{F}$, and the dual of $\mathscr{B}\times\mathscr{F}$ becomes $\mathscr{F}^*\times\mathscr{B}^*$, where $\mathscr{F}^*=L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$ and $\mathscr{B}^*=L^q(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$. Here the duality mapping takes the form

$$\begin{aligned}\langle(f^*,\varphi^*),(\varphi,f)\rangle&=\langle\varphi^*,\varphi\rangle+\langle f^*,f\rangle\\&=\int_{\mathbf{R}^-}\varphi^*(-s)\varphi(s)\,ds+\int_{\mathbf{R}^+}f^*(-s)f(s)\,ds\\&=\varphi^**\varphi(0)+f^**f(0),\end{aligned}$$

where the convolution formula is valid provided the functions are extended to all of $\mathbf{R}$ in such a way that either φ^* vanishes on $(-\infty,0)$ or φ vanishes on $(0,\infty)$, and either f^* vanishes on $(0,\infty)$ or f vanishes on $(-\infty,0)$. In particular, if all the functions are extended by zero outside of the original domain of definition, then the convolution formula is valid. Observe that $\mathscr{F}^*$ is a space of functions defined on $\mathbf{R}^-$, and that $\mathscr{B}^*$ is a space of functions defined on $\mathbf{R}^+$ (this is due to the extra minus sign in (3.1)).

Let $(f^*,\varphi^*)\in\mathscr{F}^*\times\mathscr{B}^*$, and denote $T^*(t)(f^*,\varphi^*)$ by $(\tilde{f},\tilde{\varphi})$. Then by the definition of an adjoint operator, and by Lemma 2.1, for all $(\varphi,f)\in\mathscr{B}\times\mathscr{F}$,

$$\begin{aligned}\langle(\tilde{f},\tilde{\varphi}),(\varphi,f)\rangle&=\langle(f^*,\varphi^*),T(t)(\varphi,f)\rangle\\&=\langle(f^*,\varphi^*),(x_t,f_t)\rangle,\end{aligned}\tag{3.2}$$

where x_t and f_t are defined as in Lemma 2.1.

To get any further we have to replace x in (3.2) by $\varphi+\chi*(f+F\varphi)$, as in (2.6). In the L^p-case, $1<p<\infty$, there is nothing wrong with formula (2.6), but in the continuous case and in the L^1-case, it is convenient to make a very small change in (2.6). We replace (2.6) by

$$x=\varphi+\chi*(f+G\varphi),\tag{3.3}$$

where we define f and $G\varphi$ to be zero on $\mathbf{R}^-$ (and not just on $(-\infty,0)$), φ to be zero on $(0,\infty)$ (instead of on $\mathbf{R}^+$), and

$$(G\varphi)(t)=-\int_{[t,\infty)}[d\mu(s)]\varphi(t-s),\qquad t>0\tag{3.4}$$

(the only difference between F and G is that in the continuous case, F is right continuous, whereas G is left continuous, so $F\varphi(t)=G\varphi(t)$ in all points of continuity). We also define $y(t)=0$, $t\leqq 0$, $y(t)=x(t)$, $t>0$. Then we have $x=\varphi+y$, and

$$y=\chi*f+\chi*G.$$

Substitute this together with the definition of $T(t)$ back into (3.2), define $\varphi^*(t)=0$, $t<0$, $f^*(t)=0$, $t\geqq 0$, and use the fact that convolution commutes with translation to get

$$\begin{aligned}\tilde{\varphi}*\varphi(0)+\tilde{f}*f(0)&=\varphi^**\tau_t(\varphi+y)(0)+f^**\tau_tf(0)\\&=\tau_t(\varphi^**(\varphi+y))(0)+\tau_t(f^**f)(0)\\&=\varphi^**(\varphi+y)(t)+f^**f(t)\\&=\varphi^**\varphi(t)+\varphi^**(\chi*f+\chi*G\varphi)(t)+f^**f(t)\\&=\varphi^**\varphi(t)+\varphi^**\chi*G\varphi(t)\\&\quad+((\varphi^**\chi+f^*)*f)(t)\\&=(\tau_t\varphi^*)*\varphi(0)+[\tau_t(\varphi^**\chi)]*G\varphi(0)\\&\quad+[\tau_t(\varphi^**\chi+f^*)]*f(0).\end{aligned}$$

As φ and f can be varied independently, we must have

$$\tilde{\varphi}*\varphi(0)=(\tau_t\varphi^*)*\varphi(0)+[\tau_t(\varphi^**\chi)]*G\varphi(0) \tag{3.5}$$

and

$$\tilde{f}*f(0)=[\tau_t(\varphi^**\chi+f^*)]*f(0) \tag{3.6}$$

for all $\varphi\in\mathscr{B}$ and all $f\in\mathscr{F}$. It follows from (3.6) that

$$\tilde{f}=\tau_t(\varphi^**\chi+f^*),$$

or rather, $\tilde{f}$ is the restriction of this function to $\mathbf{R}^-$. This equation has the following simple interpretation: Define

$$x^*(t)=f^*(t), \qquad t<0, \tag{3.7}$$

and for positive t, let $x^*(t)$ be the solution of

$$x^*(t)+\int_{[0,t]}x^*(t-s)\,d\mu(s)=\varphi^*(t), \qquad t\geqq 0. \tag{3.8}$$

Then $x^*=f^*+\varphi^**\chi$, and $\tilde{f}$ is the restriction to $\mathbf{R}^-$ of τ_tx^*. Observe that although f^* is an "initial function", it is ignored in the process of solving (3.8), and for positive t, $x^*(t)$ is independent of f^*. Also observe that φ^* is a forcing function in (3.8), not an initial function.

An equivalent way of writing (3.8) is

$$x^*(t)+x^**\mu(t)=\varphi^*(t)+f^**\mu(t), \qquad t\geqq 0, \tag{3.9}$$

where we again interpret f as zero on $\mathbf{R}^+$. If we define $y^* = x^* - f^*$, then

$$(3.10) \qquad y^*(t) = 0, \qquad t < 0,$$

$$(3.11) \qquad y^*(t) + y^* * \mu(t) = \varphi^*, \qquad t \geqq 0,$$

y is given by $y = \varphi^* * \chi$, and $x^*(t) = y^*(t)$ for $t \geqq 0$.

Now let us go back to the eqμation (3.5) for $\tilde{\varphi}$. The first term on the right-hand side does not cause any problems. In the second term, replace $\varphi^* * \chi$ by y^*, where y^* is the solution of (3.10) and (3.11), to get

$$[\tau_t(\varphi^* * \chi)] * G\varphi(0) = (\tau_t y^*) * G\varphi(0).$$

Clearly

$$(\tau_t y^*) * G\varphi(0) = (G^* \tau_t y^*) * \varphi(0),$$

where G^* is the adjoint of G, mapping $\mathscr{F}^*$ into $\mathscr{B}^*$.

Let us compute G^*. Take an arbitrary $f^* \in \mathscr{F}^*$, and define $f^*(t) = 0$, $t \geqq 0$. Then for all $\varphi \in \mathscr{B}$, vanishing on $(0, \infty)$,

$$G^* f^* * \varphi(0) = f^* * G\varphi(0).$$

By the definition (3.4) of G,

$$\begin{aligned} f^* * G\varphi(0) &= -\int_{(0,\infty)} f^*(-s)(\mu * \varphi)(s)\,ds \\ &= -f^* * \mu * \varphi(0). \end{aligned}$$

As this is true for all $\varphi \in \mathscr{B}$, we have $G^* f^* = -f^* * \mu$, i.e.

$$(3.12) \qquad G^* f^*(t) = -\int_{(t,\infty)} f^*(t-s)\,d\mu(s), \qquad t \geqq 0.$$

Having found G^*, we once more go back to $\tilde{\varphi}$. For all $\varphi \in \mathscr{B}$,

$$\tilde{\varphi} * \varphi(0) = (\tau_t \varphi^* + G^* \tau_t y^*) * \varphi(0),$$

so clearly

$$\tilde{\varphi} = \tau_t \varphi^* + G^* \tau_t y^*,$$

or rather, $\tilde{\varphi}$ is the restriction of this function to $\mathbf{R}^+$. The interpretation of this equation is the following. Solve the equations (3.10) and (3.11) to get y^*. Let $\tilde{\varphi}$ be the left-translate of φ^*, plus the term $G^* \tau_t y^*$. This term is the left-translate of the correction which has to be added to the forcing function φ^* if we want to replace y^* by zero on the initial interval $[0, t)$, and still let (3.11) be valid on $[t, \infty)$.

In the reflexive case $1 < p < \infty$, the adjoint semigroup $T^+(t)$ of $T(t)$ is equal to the adjoint $T^*(t)$ of $T(t)$ [26, p. 277], and we have the following result:

THEOREM 3.1. *Let $1 < p < \infty$, let $1/p + 1/q = 1$, and take $\mathscr{B} = L^p(\mathbf{R}^-; \mathbf{R}^n; \eta)$, $\mathscr{F} = L^p(\mathbf{R}^+; \mathbf{R}^n; \eta)$, $\mathscr{F}^* = L^q(\mathbf{R}^-; \mathbf{R}^n; \tilde{\eta})$, and $\mathscr{B}^* = L^q(\mathbf{R}^+; \mathbf{R}^n; \tilde{\eta})$. Suppose that $\mu \in M(\mathbf{R}^+; \mathbf{R}^{n \times n}; \rho_\eta)$ has no point at mass zero. Then the adjoint semigroup $T^+(t)$ of $T(t)$ is strongly continuous on $\mathscr{F}^* \times \mathscr{B}^*$, and*

$$T^+(t)(f^*, \varphi^*) = (x_t^*, \varphi_t^* + G^*(x_t^* - f_t^*)),$$

where x^* is the solution of (3.7) *and* (3.9), *x_t^* is the restriction to $\mathbf{R}^-$ of $\tau_t x^*$, φ_t^* is the restriction to $\mathbf{R}^+$ of $\tau_t\varphi^*$, and*

$$G^*(x_t^*-f_t^*)(s)=-\int_{(s,s+t]}x^*(t+s-v)\,d\mu(v),\qquad s\geqq 0.$$

The conclusion of Theorem 3.1 is true also for $p=1$ with $q=\infty$, except for the strong continuity. The requirement that $T^*(t)(f^*,\varphi^*)$ has to be continuous in t puts some additional conditions on $(f^*,\varphi^*)\in L^\infty(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times L^\infty(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$. The mapping which maps t into the restriction of $\tau_t x^*$ to $\mathbf{R}^-$ is continuous from $\mathbf{R}^+$ into $L^\infty(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$ if and only if $f^*\in BUC(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$, and x^* is continuous. As x^* and f^* are continuous, we can require all the preceding equations (3.7)–(3.12) to hold pointwise rather than almost everywhere. By (3.7) and (3.8), x^* is continuous at zero if and only if

$$N^*(f^*,\varphi^*)=0, \tag{3.13}$$

where

$$N^*(f^*,\varphi^*)=\varphi^*(0)-f^*(0) \tag{3.14}$$

(recall that we do not allow μ to have a point mass at zero). Suppose that $f^*\in BUC(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$, and that (3.13) holds. Then $x_t^*\to f^*$ in $L^\infty(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$ as $t\to 0+$, so by the continuity of G^*, we have $G^*x_t^*\to G^*f^*$ in $L^\infty(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$ as $t\to 0+$. This means that the second component $\varphi_t^*+G^*(x_t^*-f_t^*)$ of $T^*(t)(f^*,\varphi^*)$ tends to φ^* in $L^\infty(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$ as $t\to 0+$ if and only if $\varphi_t^*-G^*f_t^*$ tends to $\varphi^*-G^*f^*$ in $L^\infty(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$. Now, by (3.12), we have

$$\varphi_t^*-G^*f_t^*=\tau_t\big(\varphi^*+f^**\mu\big)=\tau_t\big(\varphi^*-G^*f^*\big),$$

where we have defined $f^*(t)=0$ for $t\geqq 0$. Thus, $\varphi_t^*-G^*f_t^*\to\varphi^*-G^*f^*$ if and only if $\varphi+f^**\mu\in BUC(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$. In other words, we have the following result:

THEOREM 3.2. *Let $\mathscr{B}=L^1(\mathbf{R}^-;\mathbf{R}^n;\eta)$, $\mathscr{F}=L^1(\mathbf{R}^+;\mathbf{R}^n;\eta)$, $\mathscr{F}^+=BUC(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$, $\mathscr{B}^*=L^\infty(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$, $\mathscr{B}^+=BUC(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$. Suppose that $\mu\in M(\mathbf{R}^+;\mathbf{R}^{n\times n};\rho_\eta)$ has no point mass at zero. Then the adjoint semigroup $T^+(t)$ is strongly continuous in the space*

$$\big\{(f^+,\varphi^+)\in\mathscr{F}^+\times\mathscr{B}^*\,|\,\varphi^+-G^*f^+\in\mathscr{B}^+,\ and\ N^*(f^+,\varphi^+)=0\big\},$$

and $T^+(t)(f^+,\varphi^+)$ is defined in the same way as in Theorem 3.1.

The dual of the space $BC_0(\mathbf{R};\mathbf{R}^n;\eta)$ can be identified with the space $M(\mathbf{R};\mathbf{R}^n;\tilde{\eta})$ of measures on $\mathbf{R}$, with weight $\tilde{\eta}$. One can compute $T^*(t)$ in this space, too, and get the same the same formulas as above, but in $M((-\infty,0);\mathbf{R}^n;\tilde{\eta})\times M(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$ rather than in $L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times L^q(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$. This time the requirement that $T^+(t)$ has to be strongly continuous cuts down the measure spaces to the corresponding L^1-spaces. In other words, we have the following result:

THEOREM 3.3. *Let $\mathscr{B}=BC_0(\mathbf{R}^-;\mathbf{R}^n;\eta)$, $\mathscr{F}=BC_0(\mathbf{R}^+;\mathbf{R}^n;\eta)$, $\mathscr{F}^+=L^1(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$, $\mathscr{B}^+=L^1(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$, and let μ have no point mass at zero. Then the adjoint semigroup $T^+(t)$ is strongly continuous in $\mathscr{F}^+\times\mathscr{B}^+$, and it is defined in the same way as in Theorem* 3.1.

In the sequel, rather than discussing $T^+(t)$ and equations (3.7)–(3.11), we shall discuss the corresponding transposed equations. In addition we throughout replace all notations which refer to adjoint spaces by the same type of notations which are used in (1.1) and (1.2). More specifically, we replace $\tilde{\eta}$ by η, the transpose $\tilde{\mu}$ of μ by μ, x^* by x, f^* by φ, φ^* by f, y^* by y, G^* by F, the operator N^* by an operator N, and substitute

$R(t)$ for $T^+(t)$. Then the equations (3.7)–(3.14) become

$$x(t)=\varphi(t), \qquad t<0, \tag{3.15}$$

$$x(t)+\int_{[0,t]}[d\mu(s)]x(t-s)=f(t), \qquad t\geqq 0, \tag{3.16}$$

$$x(t)+\mu * x(t)=f(t)-F\varphi(t), \qquad t\geqq 0, \tag{3.17}$$

$$y(t)=0, \qquad t<0, \tag{3.18}$$

$$y(t)+\mu * y(t)=f, \qquad t\geqq 0, \tag{3.19}$$

$$N(\varphi,f)=0, \tag{3.20}$$

$$N(\varphi,f)=f(0)-\varphi(0). \tag{3.21}$$

With the new notations, the semigroup which we found above can be described as follows:

LEMMA 3.4. *Let Y be one of the spaces L^p, $1\leqq p\leqq\infty$, or BUC, or BC_0, let η be an influence function with associated dominating function ρ_η, and define $\mathscr{B}=Y(\mathbf{R}^-;\mathbf{R}^n;\eta)$, $\mathscr{F}=Y(\mathbf{R}^+;\mathbf{R}^n;\eta)$. Suppose that $\mu\in M(\mathbf{R}^+;\mathbf{R}^{n\times n};\rho_\eta)$ has no point mass at zero. In the L^p-case, $1\leqq p<\infty$, define $\mathscr{S}$ by $\mathscr{S}=\mathscr{B}\times\mathscr{F}$, and in the continuous case, define $\mathscr{S}$ by*

$$\mathscr{S}=\{(\varphi,f)\in\mathscr{B}\times L^\infty(\mathbf{R}^+;\mathbf{R}^n;\eta)\,|\,f-F\varphi\in\mathscr{F},\ \text{and}\ N(\varphi,f)=0\}.$$

For each $(\varphi,f)\in\mathscr{S}$, let $x(\varphi,f)$ be the solution of (3.15) *and* 3.16, *and define*

$$R(t)(\varphi,f)=(x_t(\varphi,f),f_t+F(x_t-\varphi_t)),$$

where $x_t(\varphi,f)$ is the restriction of $\tau_t x(\varphi,f)$ to $\mathbf{R}^-$, f_t is the restriction of $\tau_t f$ to $\mathbf{R}^+$, and

$$F(x_t-\varphi_t)(s)=-\int_{(s,s+t]}[d\mu(v)]x(t+s-v), \qquad s\geqq 0.$$

Then $R(t)$ is a strongly continuous semigroup in $\mathscr{S}$.

Here the case $Y=BC_0$ is new in the sense that it has not been mentioned before, but it can be obtained by restricting the corresponding semigroup defined in BUC to BC_0. If μ is absolutely continuous, i.e. $d\mu(s)=a(s)ds$ for some function $a\in L^1(\mathbf{R};\mathbf{R}^{n\times n};\rho_\eta)$, then in the continuous case the condition $f-F\varphi\in\mathscr{F}$ is equivalent to $f\in\mathscr{F}$. In this case the state space simply becomes $\mathscr{B}\times\mathscr{F}$ (this is the situation discussed in [7] and [8]). The condition $f-F\varphi\in\mathscr{F}$ can be replaced by another condition, which makes no reference to φ. Namely, if we define $\varphi(t)=0$ for $t\geqq 0$, $y(t)=\varphi(0)(1-t)=f(0)(1-t)$ for $0\leqq t<1$, and $y(t)=0$ otherwise, then $F\varphi(t)=-\mu * \varphi(t)$ for $t\geqq 0$, and $\varphi+y\in BC_0(\mathbf{R};\mathbf{R}^n;\eta)$. This implies that also $\mu *(\varphi+y)\in BC_0(\mathbf{R};\mathbf{R}^n;\eta)$, and the condition $f-F\varphi\in\mathscr{F}$ is satisfied if and only if $f-\mu * y$, restricted to $\mathbf{R}^+$, belongs to $\mathscr{F}$. This means among others that f has exactly the same discontinuities as $\mu * y$ has. This function again has exactly the same discontinuities as the function $\mu([0,t])\varphi(0)$. Thus, the function $f-F\varphi$ is continuous if and only if $f(t)-\mu([0,t])f(0)$ is continuous.

We leave it to the reader to prove that the adjoint semigroup of a semigroup of type $R(t)$ in the cases $Y=L^1$ and $Y=BC_0$ is a semigroup of type $T(t)$ in BUC and L^1, respectively.

4. The extended initial and forcing function semigroups are equivalent. In the preceding section we saw that the adjoint of a semigroup of type $T(t)$ in L^p, $1\leqq p<\infty$, and in BC_0, is a semigroup of type $R(t)$, and vice versa. Here we shall show that there is a one-to-one, bicontinuous mapping D of the state space onto itself (or of the state space of $R(t)$ onto the state space of $T(t)$ in the continuous case) such that $R(t)=D^{-1}T(t)D$ and $T(t)=DR(t)D^{-1}$.

Let $1\leqq p\leqq\infty$, and $\mathscr{B}=L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)$, $\mathscr{F}=L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$. We define the continuous operator Q from $\mathscr{B}\times\mathscr{F}$ into itself by

$$Q(\varphi,f)=(0,-F\varphi), \tag{4.1}$$

where F is the operator defined in (2.7). Clearly $Q^2=0$, so if we let I be the identity operator, and define D by

$$D=I+Q, \tag{4.2}$$

or equivalently,

$$D(\varphi,f)=(\varphi,f-F\varphi),$$

then D is invertible, and

$$D^{-1}=I-Q. \tag{4.3}$$

If one wants to give a verbal description of the preceding definitions one can say that D is the operator which subtracts the initial function correction from the forcing function, and D^{-1} is the operator which adds the same correction to the forcing function.

LEMMA 4.1. *Define the semigroups $T(t)$ and $R(t)$ as in Lemma* 2.1 *and Lemma* 3.4. *Then in all the different cases $Y=L^p$, $Y=BUC$ and $Y=BC_0$, we have $R(t)=D^{-1}T(t)D$, and $T(t)=DR(t)D^{-1}$, with D defined as in* (4.2).

Proof. It suffices to consider the case $Y=L^p$, $1\leqq p\leqq\infty$, because the continuous cases are special cases of $Y=L^\infty$.

Let $(\varphi,f)\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$, and observe that the second component of $D(\varphi,f)$, which is $f-F\varphi$, is exactly the forcing function used in (3.17). Let x be the solution of (3.17) with initial condition (3.15). Then by Lemma 3.4,

$$R(t)(\varphi,f)=(x_t,f_t+Fx_t-F\varphi_t),$$

where we have defined $\varphi(t)=0$ for $t\geqq0$, and let φ_t be the restriction to R^- of $\tau_t\varphi$. On the other hand, we can also interpret (3.17) and (3.15) as a special case of (1.1) and (1.2), (with f replaced by $f-F\varphi$), and with this interpretation we have (cf. Lemma 2.1)

$$\begin{aligned}T(t)D(\varphi,f)&=(x_t,f_t-F\varphi_t)\\&=(x_t,f(t)-F\varphi_t+Fx_t)-(0,Fx_t)\\&=R(t)(\varphi,f)+QR(t)(\varphi,f)=DR(t)(\varphi,f).\end{aligned}$$

As D is invertible, this gives the conclusion of Lemma 4.1.

5. The generators of the extended semigroups. As a first application of Lemma 4.1, we shall compute the generator of the semigroup $R(t)$ with the aid of the generator of the semigroup $T(t)$. The generator A of $T(t)$ is very easy to compute. By the definition of the generator of a semigroup, (φ,f) belongs to the domain $\mathscr{D}(A)$ of A, if $\lim_{t\to0+}t^{-1}[T(t)(\varphi,f)-(\varphi,f)]$ exists, and for $(\varphi,f)\in\mathscr{D}(A)$,

$$A(\varphi,f)=\lim_{t\to0+}t^{-1}[T(t)(\varphi,f)-(\varphi,f)].$$

It is clear from the definition of $T(t)$ that $A(\varphi,f)$ must equal (φ',f'), where we use the prime to denote differentiation (e.g. in the distribution sense, or in the L^p-sense).

Moreover, $(\varphi,f)\in\mathscr{D}(A)$ if and only if $(\varphi,f)\in\mathscr{B}\times\mathscr{F}$, f is differentiable with $f'\in\mathscr{F}$, and the solution x of (1.1) and (1.2) has a derivative x', whose translates $\tau_t x'$, restricted to $\mathbf{R}^-$, belongs to $\mathscr{B}$. In other words, the generator A can be described as follows:

THEOREM 5.1. *The domain $\mathscr{D}(A)$ of the generator A of $T(t)$ is given by*

$$\mathscr{D}(A)=\left\{(\varphi,f)\in\mathscr{B}\times\mathscr{F}\middle|(\varphi',f')\in\mathscr{B}\times\mathscr{F},\ M(\varphi,f)=0\right\}$$

in the L^p-case, $1\leqq p<\infty$, where M is the operator defined in (2.8), *and by*

$$\mathscr{D}(A)=\{(\varphi,f)\in\mathscr{B}\times\mathscr{F}|(\varphi',f')\in\mathscr{B}\times\mathscr{F},$$

$$\text{and } M(\varphi,f)=M(\varphi',f')=0\}$$

in the continuous case. In both cases, $A(\varphi,f)=(\varphi',f')$.

To get the generator B of $R(t)$ we use Lemma 4.1 and Theorem 5.1. Clearly,

$$\lim_{t\to0+}t^{-1}[R(t)(\varphi,f)-(\varphi,f)]=D^{-1}\lim_{t\to0+}[T(t)D(\varphi,f)-D(\varphi,f)],$$

provided one of these two limits exists. In other words, $B=D^{-1}AD$, with domain $\mathscr{D}(B)=D^{-1}\mathscr{D}(A)$.

THEOREM 5.2. *Both in the L^p-case and in the continuous case, the generator B of $R(t)$ is given by*

$$B(\varphi,f)=\left(\varphi',(f(t)-\mu([0,t])\varphi(0))'\right) \tag{5.1}$$

for all (φ,f) in the domain $\mathscr{D}(B)$ of B. In the L^p-case,

$$\mathscr{D}(B)=\{(\varphi,f)\in\mathscr{B}\times\mathscr{F}|B(\varphi,f)\in\mathscr{B}\times\mathscr{F},\text{ and }N(\varphi,f)=0\},$$

where N is the operator defined in (3.21), *and in the continuous case,*

$$\mathscr{D}(B)=\{(\varphi,f)\in\mathscr{B}\times L^\infty(\mathbf{R}^+;\mathbf{R}^n;\eta)|\varphi'\in\mathscr{B},$$

$$(f-F\varphi),(f-F\varphi)'\in\mathscr{F},\ \text{and } N(\varphi,f)=NB(\varphi,f)=0\}.$$

If we replace $\varphi(0)$ in (5.1) by $f(0)$, the description of $\mathscr{D}(B)$ above in the L^p-case has the property that the conditions on f are independent of the conditions on φ. One could also in the continuous case give a description of $\mathscr{D}(B)$ of the same type (cf. the discussion following Lemma 3.4).

Proof. Let $(\varphi,f)\in\mathscr{D}(B)$, and define $g=f-F\varphi$. Then $D(\varphi,f)=(\varphi,g)$, so $(\varphi,g)\in\mathscr{D}(A)$, i.e. $(\varphi,g)\in\mathscr{B}\times\mathscr{F}$, $(\varphi',g')\in\mathscr{B}\times\mathscr{F}$, $M(\varphi,g)=0$, and in the continuous case, also $M(\varphi',g')=0$. Moreover, $A(\varphi,g)=(\varphi',g')$. Thus,

$$B(\varphi,f)=D^{-1}AD(\varphi,f)=D^{-1}A(\varphi,g)=D^{-1}(\varphi',g')$$
$$=(I-Q)(\varphi',g')=(\varphi',g'+F\varphi')=\left(\varphi',(f-F\varphi)'+F\varphi'\right),$$

which gives us the preliminary formula

$$B(\varphi,f)=\left(\varphi',(f-F\varphi)'+F\varphi'\right)$$

for B. Define $y(t)=0$ for $t<0$, and $y(t)=\varphi(0)$, $\varphi(t)=0$ for $t\geqq0$. Then $F\varphi=-\mu*\varphi$, and

$$F\varphi'=-\mu*(\varphi+y)'=-(\mu*(\varphi+y))'.$$

Thus

$$(5.2)\qquad (f-F\varphi)'+F\varphi'=(f+\mu*\varphi)'-(\mu*(\varphi+y))'=(f-\mu*y)'.$$

As

$$\mu*y(t)=\mu([0,t])\varphi(0),\qquad t\geqq 0,$$

we get the formula (5.1).

Let us turn to the two descriptions of $\mathscr{D}(B)$. In both the L^p-case and the continuous case, $N(\varphi,f)=0$ if and only if $M(\varphi,g)=0$ (see (2.9) and (3.21)). In the continuous case we can evaluate (5.2) at zero to get

$$g'(0)+F\varphi'(0)=(f-\mu*y)'(0),$$

which means that $M(\varphi',g')=0$ if and only if $NB(\varphi,f)=0$. In the continuous case, the conditions φ, $\varphi'\in\mathscr{B}$, $(f-F\varphi)$, $(f-F\varphi)'\in\mathscr{F}$ are clearly equivalent to the requirement that (φ,g), $(\varphi',g')\in\mathscr{B}\times\mathscr{F}$. In the L^p-case, $F\varphi'\in\mathscr{F}$, and $(f-F\varphi)'=(f-\mu*y)'-F\varphi'$, so we can replace $(f-F\varphi)'\in\mathscr{F}$ by $(f-\mu*y)'\in\mathscr{F}$, and get $B(\varphi,f)\in\mathscr{B}\times\mathscr{F}$. The proof of theorem 5.2 is thereby complete.

6. The asymptotic behavior of the extended semigroups. Below we shall look at the asymptotic behavior of $T(t)$ and $R(t)$. The discussion of the asymptotic behavior of $T(t)$ is based on [28] and [30] (although the equation discussed there was a differentiated one rather than the undifferentiated equation (1.1)). Once the behavior of $T(t)$ is known, the results can be transferred to $R(t)$ with the aid of Lemma 4.1.

In the sequel, we suppose throughout that

$$(6.1)\qquad \text{(equation (1.1) is noncritical with respect to } \eta)$$

i.e. there exists a measure $\nu\in M(\mathbf{R};\mathbf{R}^{n\times n};\rho_\eta)$ satsifying (2.5) with χ replaced by ν. In general, ν does not vanish on $\mathbf{R}^-$. If it does, then the fundamental solution χ in (2.5) belongs to $M(\mathbf{R}^+;\mathbf{R}^n;\rho_\eta)$, and $\nu=\chi$. See e.g. [16] for conditions which imply that ν exists.

With the aid of ν we can split the solution x of (1.1), (1.2) into two components $x=x_S+x_U$, where

$$(6.2)\qquad x_S=\varphi+\nu*(f+F\varphi)$$

and

$$(6.3)\qquad x_U=(\chi-\nu)*(f+F\varphi)$$

(here we have used the same conventions as in (2.6)). Neither x_S nor x_U satisfies in general (1.2). Instead they satisfy

$$(6.4)\qquad x_S(t)+\mu*x_S(t)=\begin{cases}\mu*\varphi(t), & t<0,\\ f(t), & t\geqq 0,\end{cases}$$

and

$$(6.5)\qquad x_U(t)+\mu*x_U(t)=0,\qquad -\infty<t<\infty.$$

We define the *stable part* $T_S(t)$ and the *unstable part* $T_U(t)$ of $T(t)$ by

$$(6.6)\qquad T_S(t)(\varphi,f)=((x_S)_t,f_t),$$

$$(6.7)\qquad T_U(t)(\varphi,f)=((x_U)_t,0),$$

where $(x_S)_t$ and $(x_U)_t$ are defined analogously to x_t in Lemma 2.1. Then $T(t)=T_S(t)+T_U(t)$. The operators $T_S(0)$ and $T_U(0)$ are projection operators, which split the state space of $T(t)$ into a *stable subspace* S and an *unstable subspace* U. Both S and U are invariant under $T(t)$, the restriction of $T(t)$ to S equals $T_S(t)$ (restricted to S), and the restriction of $T(t)$ to U equals $T_U(t)$ (restricted to U). Because of (2.4), $\|T_S(t)\|=O(\rho_\eta(-t))$ as $t\to\infty$. The growth rate of $T_U(t)$ is bigger at infinity than the growth rate of $T_S(t)$, except when $\nu=\chi$, and $U=\{0\}$. The restriction of the semigroup $T_U(t)$ to U can be extended to a group. If the singular part of μ is small enough (e.g. zero), and if the determinant $\det(I+\hat{\mu}(z))$ of the Laplace transform of $\delta+\mu$ is bounded away from zero in the half plane $\operatorname{Re} z\geqq\alpha$, where

$$\alpha=-\lim_{t\to\infty} t^{-1}\log\left(\rho_\eta(t)\right),$$

with the exception of finitely many points in $\operatorname{Re} z>\alpha$ where it may be zero, then (6.1) is satisfied, and U is finite dimensional (its dimension is the same as the sum of the dimensions of the singularities of $(I+\hat{\mu})$). In this case, all "initial functions" in U are exponential polynomials related to the singular points of $(I+\hat{\mu})$.

We can use Lemma 4.2 to get a similar decomposition of $R(t)$ into a stable part $R_S(t)$ and an unstable part $R_U(t)$. One simply defines $R_S(t)$ and $R_U(t)$ to be $R_S(t)=D^{-1}T_S(t)D$ and $R_U(t)=D^{-1}T_U(t)D$. This gives a decomposition of $R(t)$ into two parts with almost exactly the same properties as $T_S(t)$ and $T_U(t)$. The stable subspace S of $T(t)$ is mapped into the stable subspace $D^{-1}S$ of $R(t)$, and the unstable subspace of $T(t)$ is mapped into the unstable subspace $D^{-1}U$ of $R(t)$. In particular, we get the following description of the unstable subspace $D^{-1}U$ of $R(t)$: An element (φ,f) belongs to $D^{-1}U$ if and only if

$$\dot{\varphi}(t)+\mu*\varphi(t)=0, \qquad t<0,$$

and

$$f(t)=-\int_{(t,\infty)}[d\mu(s)]\varphi(t-s), \qquad t\geqq 0.$$

Observe that although φ may be C^∞ smooth, f will not in general be even continuous (f is continuous if φ is continuous and μ has no point masses).

Acknowledgments. The author wants to thank the referees for their very appropriate criticism of the original presentation, and for providing us with a part of the motivation to rewrite [32]. One of the questions posed by one of the referees, i.e. the question whether the ordinary initial function and the forcing function semigroups are equivalent in the infinite delay case, has been given a partial answer in [32].

REFERENCES

[1] V. Barbu and S. I. Grossman, *Asymptotic behavior of linear integrodifferential systems*, Trans. Amer. Math. Soc., 173 (1972), pp. 277–288.

[2] C. Bernier and A. Manitius, *On semigroups in* $\mathbf{R}\times L^p$ *corresponding to differential equations with delays*, Canad. J. Math., XXX (1978), pp. 897–914.

[3] J. A. Burns and T. L. Herdman, *Adjoint semigroup theory for a class of functional differential equations*, this Journal, 7 (1976), pp. 729–745.

[4] J. A. Burns, T. Herdman and H. W. Stech, *Linear functional differential equations as semigroups on product spaces*, this Journal, 14 (1983), pp. 98–116.

[5] M. C. DELFOUR, *Status of the state space theory of linear hereditary differential systems with delays in state and control variables*, in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, Berlin, 1980, pp. 83–96.

[6] M. C. DELFOUR AND A. MANITIUS, *The structural operator F and its role in the theory of retarded systems*, I, II, J. Math. Anal. Appl., 73 (1980), pp. 466–490; 74 (1980), pp. 359–381.

[7] O. DIEKMANN, *Volterra integal equations and semigroups of operators*, Report TW 197/80, Stichting Mathematisch Centrum, Amsterdam, 1980.

[8] O. DIEKMANN AND S. A. VAN GILS, *Invariant manifolds of Volterra integral equations of convolution type*, J. Differential Equations, to appear.

[9] S. A. VAN GILS, *Some studies in dynamical system theory*: I *Volterra integral equations of convolution type*, II *Hopf bifurcation and symmetry*, Ph. D. Thesis, Technische Hogeschool Delft, Delft, 1984.

[10] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, Berlin, 1977.

[11] ———, *Functional differential equations with infinite delays*, J. Math. Anal. Appl., 48 (1974), pp. 276–283.

[12] J. K. HALE AND J. KATO, *Phase space for retarded equations with infinite delay*, Funkcial. Ekvac., 21 (1978), pp. 11–41.

[13] J. K. HALE AND K. R. MEYER, *A Class of Functional Equations of Neutral Type*, Mem. Amer. Math. Soc. 76, 1967.

[14] D. HENRY, *The adjoint of a linear functional differential equation and boundary value problems*, J. Differential Equations, 9 (1971), pp. 55–66.

[15] ———, *Linear autonomous neutral functional differential equations*, J. Differential Equations, 15 (1974), p. 106–128.

[16] G. S. JORDAN, O. J. STAFFANS AND R. L. WHEELER, *Local analyticity in weighted L^1-spaces and applications to stability problems for Volterra equations*, Trans. Amer. Math. Soc., 274 (1982), pp. 749–782.

[17] F. KAPPEL, *Laplace-transform methods and linear autonomous functional-differential equations*, Berichte der Mathematische-statistischen Sektion im Forschungszentrum Graz, Report 64, 1976, pp. 1–62.

[18] ———, *Linear autonomous functional differential equations in the state space C^**, Technical Report 34, Technische Universität Graz, Graz, 1984.

[19] A. MANITIUS, *Completeness and F-completeness of eigenfunctions associated with retarded functional differential equations*, J. Differential Equations, 35 (1980), pp. 1–29.

[20] R. K. MILLER, *Nonlinear Volterra Integral Equations*, Benjamin, Menlo Park, CA, 1971.

[21] ———, *Linear Volterra integrodifferential equations as semigroups*, Funkcial. Ekvac., 17 (1974), pp. 39–55.

[22] R. K. MILLER AND G. R. SELL, *Volterra Integral Equations and Topological Dynamics*, Mem. Amer. Math. Soc., 102, 1970.

[23] T. NAITO, *Adjoint equations of autonomous linear functional differential equations with infinite retardations*, Tohoku Math. J., 28 (1976), pp. 135–143.

[24] ———, *On autonomous linear functional differential equations with infinite retardations*, J. Differential Equations, 21 (1976), pp. 297–315.

[25] ———, *On linear autonomous retarded equations with an abstract phase space for infinite delay*, J. Differential Equations, 33 (1979), pp. 74–91.

[26] R. S. PHILLIPS, *The adjoint semi-group*, Pacific J. Math., 5 (1955), pp. 269–283.

[27] D. SALAMON, *Control and Observation of Neutral Systems*, Research Notes in Mathematics 91, Pitman, London, 1984.

[28] O. J. STAFFANS, *On a neutral functional differential equation in a fading memory space*, J. Differential Equations, 50 (1983), pp. 183–217.

[29] ———, *A neutral* FDE *with stable D-operator is retarded*, J. Differential Equations, 49 (1983), pp. 208–217.

[30] ———, *The null space and the range of a convolution operator in a fading memory space*, Trans. Amer. Math. Soc., 281(1984), pp. 361–388.

[31] ———, *Some well posed functional equations which generate semigroups*, J. Differential Equations, to appear.

[32] ———, *Semigroups generated by a convolution equations*, in Infinite Dimensional Systems, Proceedings, Retzhob, 1983, F. Kappel and W. Schappacher, eds., Springer-Verlag, Berlin, 1984.

[33] ———, *Semigroups generated by a neutral functional differential equation*, this Journal, to appear.

[34] H. W. STECH, *On the adjoint theory for autonomous linear functional differential equations with unbouded delays*, J. Differential Equations, 27 (1978), pp. 421–443.

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

DENSE SETS AND FAR FIELD PATTERNS IN ELECTROMAGNETIC WAVE PROPAGATION*

DAVID COLTON† AND RAINER KRESS‡

Abstract. It is shown that the electric far field patterns corresponding to the scattering of entire incident fields by a bounded perfectly conducting obstacle are dense in the space of square integrable tangential vector fields defined on the boundary of the unit ball if and only if there does not exist a Maxwell eigenfunction that is an electromagnetic Herglotz pair, i.e. a solution $\{E, H\}$ of Maxwell's equations defined in all of space such that

$$\lim_{r\to\infty}\frac{1}{r}\iint_{|x|\leqq r}\left(|E(x)|^2+|H(x)|^2\right)dx<\infty.$$

1. Introduction. A basic task in the investigation of the inverse scattering problem for time-harmonic acoustic and electromagnetic waves is the study of the class of far field patterns corresponding to the scattering of entire incident fields by a bounded obstacle. Indeed if $\mathbf{T}$ denotes the operator mapping the incident field and scattering obstacle onto the far field pattern, then the inverse scattering problem is to construct $\mathbf{T}^{-1}$ defined on the range of $\mathbf{T}$, and the determination of this range is nothing more than the description of the class of far field patterns. Unfortunately, little is known concerning this class except for the well-known fact that the far field patterns are entire functions of their independent (complex) variables for each positive fixed value of the wave number [3], i.e. the range of T is not all of $L^2(\partial\Omega)$ where Ω is the unit ball. We note that this implies that the inverse scattering problem is an improperly posed problem since the far field patterns are, in practice, determined from inexact measurements.

Recently Colton [1] and Colton and Kirsch [2] have investigated the case of acoustic scattering and asked the question if the class of far field patterns corresponding to a fixed scattering obstacle and all entire incident fields is dense in $L^2(\partial\Omega)$. The rather surprising answer to this question is that if the impedance of the scattering obstacle is positive, then the far field patterns are dense in $L^2(\partial\Omega)$, whereas if the scattering obstacle is sound-soft (i.e. Dirichlet boundary data) or sound-hard (i.e. Neumann boundary data) then the far field patterns are dense in $L^2(\partial\Omega)$ if and only if there does not exist an (interior) eigenfunction that is an entire Herglotz wave function, i.e. a solution u of the Helmholtz equation defined in all of space such that

$$\lim_{r\to\infty}\frac{1}{r}\iint_{|x|\leqq r}|u(x)|^2\,dx<\infty.$$

This phenomenon is rather unusual since in a wide variety of improperly posed problems in mathematical physics the range of the operator which one wants to invert

* Received by the editors November 22, 1983, and in revised form April 5, 1984.

† Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19711. The research of this author was supported in part by the U.S. Air Force Office of Scientific Research under grant 81-0103.

‡ Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Göttingen, West Germany.

is dense in (but not equal to) the Banach space in which the measurements are being made. Hence, the inverse scattering problem is peculiar even in the class of improperly posed problems. Furthermore, not only is the property of the far field patterns being dense very sensitive on the shape of the domain, but from physical considerations the interior eigenvalues should in fact have nothing to do with the exterior scattering problem at all.

The purpose of this paper is to extend the above results for acoustic scattering to the electromagnetic case. In particular we shall show that the far field patterns of the electric fields corresponding to the scattering of entire incident fields by a bounded perfectly conducting obstacle are dense in the space of square integrable tangential vector fields defined on $\partial\Omega$ if and only if there does not exist a Maxwell eigenfunction that is an electromagnetic Herglotz pair, i.e. a solution $\{E,H\}$ of Maxwell's equations defined in all of space such that

$$\lim_{r\to\infty}\frac{1}{r}\iint_{|x|\leq r}\left(|E(x)|^2+|H(x)|^2\right)dx<\infty.$$

This result will be established by first constructing an appropriate complete set of functions defined on the boundary of the scattering obstacle, and then establishing an integral representation for the electric field of an electromagnetic Herglotz pair. Although we only prove necessary and sufficient conditions for the electric far field patterns to be dense, from the symmetry of Maxwell's equations we can easily deduce an analogous result for the magnetic far field patterns.

In the analysis which follows we denote the scalar product of two vectors by $(\cdot,\cdot)$, the vector product by $[\cdot,\cdot]$, and the triple product of three vectors by $(\cdot,\cdot,\cdot)$. By an entire solution to the vector Helmholtz equation or Maxwell's equations we mean a solution that is defined in all of Euclidean three-space $\mathbb{R}^3$.

2. Complete sets of tangential vector fields. We begin by defining our notation. Let j_n denote the spherical Bessel function of order n and h_n the spherical Hankel function of the first kind of order n. Let

$$Y_n^m(\theta,\varphi)=\left[\frac{(2n+1)(n-|m|)!}{4\pi(n+|m|)!}\right]^{1/2}P_n^{|m|}(\cos\theta)e^{im\varphi},$$

$n=0,1,2,\cdots,m=-n,\cdots,n$, denote the spherical harmonic of order n where P_n^m is the associated Legendre polynomial. Define

$$u_n^m(x):=j_n(kr)Y_n^m(\theta,\varphi),$$
$$v_n^m(x):=h_n(kr)Y_n^m(\theta,\varphi),$$

where (r,θ,φ) are the spherical coordinates of $x\in\mathbb{R}^3$ and $k>0$ is the wave number. Let $e_j, j=1,2,3$, denote the Cartesian unit coordinate vectors and define

$$E_{j,n}^m(x):=\operatorname{curl}\operatorname{curl}e_ju_n^m(x),$$

$j=1,2,3,\ n=0,1,2,\cdots,m=-n,\cdots,n$. Note that the $E_{j,n}^m$ are entire solutions of the vector Helmholtz equation. Since they are divergence free, the pair $\{E_{j,n}^m,H_{j,n}^m\}$, where

$$H_{j,n}^m:=\frac{1}{ik}\operatorname{curl}E_{j,n}^m,$$

is a solution of Maxwell's equations

$$\operatorname{curl} E - ikH = 0, \qquad \operatorname{curl} H + ikE = 0.$$

The set consisting of all functions $E_{j,n}^m$ will be denoted by

$$\mathscr{E} := \left\{ E_{j,n}^m, j = 1,2,3,\ n = 0,1,2,\cdots, m = -n,\cdots,n \right\}.$$

Now let D denote a bounded region in $\mathbb{R}^3$ with the boundary ∂D consisting of a finite number of disjoint, closed, bounded surfaces belonging to class C^2. The exterior domain $\mathbb{R}^3 \setminus \overline{D}$ is assumed to be connected, whereas D itself may have more than one component. We assume that the unit normal ν to ∂D is directed into the exterior of D. Let

$$X := \left\{ a : \partial D \to \mathbb{C}^3 | (\nu, a) = 0, a \in L^2(\partial D) \right\}$$

denote the Hilbert space of square integrable tangential fields defined on the boundary ∂D.

Before proceeding to the main results of this section, we first present a brief discussion of the jump relations for vector potentials with square integrable densities. To do so we begin by stating the corresponding results for continuous densities. Consider the vector potential

$$A(x) := \int_{\partial D} \Phi(x,y) a(y)\, ds(y), \qquad x \in \mathbb{R}^3 \setminus \partial D$$

where

$$\Phi(x,y) := \frac{1}{4\pi} \frac{e^{ik|x-y|}}{|x-y|}, \qquad x \neq y,$$

is the fundamental solution to the Helmholtz equation and a is a continuous tangential field defined on ∂D. We then have the following result.

LEMMA 2.1. *For the vector potential A with continuous tangential density a we have*

$$\lim_{\substack{h \to 0 \\ h > 0}} \left[\nu(x), \operatorname{curl} A(x + h\nu(x)) - \operatorname{curl} A(x - h\nu(x)) \right] = a(x), \tag{2.1}$$

$$\lim_{\substack{h \to 0 \\ h > 0}} \left[\nu(x), \operatorname{curl}\operatorname{curl} A(x + h\nu(x)) - \operatorname{curl}\operatorname{curl} A(x - h\nu(x)) \right] = 0, \tag{2.2}$$

uniformly for $x \in \partial D$.

Proof. For a proof of (2.1) see [3, Thm. 2.26]. For (2.2), as in the case of [3, Thm. 2.21], it suffices to carry out the proof for $k = 0$. Setting $x_\pm := x \pm h\nu(x)$, $x \in \partial D$, we observe that curl curl$A$ = grad div$A$ for $k = 0$ and use this to obtain

$$4\pi \operatorname{curl}\operatorname{curl} A(x_\pm) = \int_{\partial D} \left\{ 3 \frac{(x_\pm - y, a(y))}{|x_\pm - y|^5} (x_\pm - y) - \frac{a(y)}{|x_\pm - y|^3} \right\} ds(y).$$

In particular, if $a = [\nu, c]$ where c is constant, we see by Gauss' theorem that

$$\operatorname{div} A(x) = -\int_{\partial D} \left(\nu(y), c, \operatorname{grad}_y \Phi(x,y) \right) ds(y) = 0, \qquad x \in \mathbb{R}^3 \setminus \partial D.$$

Hence we have that

$$4\pi\left[\nu(x),\operatorname{curl}\operatorname{curl}A(x_\pm)\right]$$

$$=\int_{\partial D}\left\{3\frac{(x_\pm-y,\nu(y),b(y)-b(x))}{|x_\pm-y|^5}[\nu(x),x_\pm-y]\right.$$

$$\left.-\frac{[\nu(x),[\nu(y),b(y)-b(x)]]}{|x_\pm-y|^3}\right\}ds(y)$$

where we have set $b:=[a,\nu]$. The proof of the lemma can now be completed as in [3, Thm. 2.21].

By using Lemma 2.1, the proof of the following lemma can now be carried out analogously to the results of Kersten [5] for the case of scalar potentials with square integrable densities.

LEMMA 2.2. *For the vector potential A with square integrable tangential density a we have*

$$\lim_{\substack{h\to 0\\ h>0}}\int_{\partial D}\left|[\nu(x),\operatorname{curl}A(x+h\nu(x))-\operatorname{curl}A(x-h\nu(x))]-a(x)\right|^2 ds(x)=0,$$

$$\lim_{\substack{h\to 0\\ h>0}}\int_{\partial D}\left|[\nu(x),\operatorname{curl}\operatorname{curl}A(x+h\nu(x))-\operatorname{curl}\operatorname{curl}A(x-h\nu(x))]\right|^2 ds(x)=0.$$

We are now in a position to prove our first theorem.

THEOREM 2.1. *Let $\eta>0$. Then*

$$W:=\{c:=[[\operatorname{curl}E,\nu],\nu]+i\eta[\nu,E],\ E\in\mathscr{E}\}$$

is complete in X.

Proof. Let $a\in X$ such that

$$\int_{\partial D}(\bar a,c)\,ds=0$$

for all $c\in W$, i.e.

$$\int_{\partial D}\{-(\bar a,\operatorname{curl}E)+i\eta(\bar a,\nu,E)\}\,ds=0 \tag{2.3}$$

for all $E\in\mathscr{E}$. The theorem will be proved if we can show that $a=0$. To this end we first write

$$E^m_{j,n}=\operatorname{curl}\operatorname{curl}e_ju^m_n=-\Delta e_ju^m_n+\operatorname{grad}\operatorname{div}e_ju^m_n=k^2e_ju^m_n+\operatorname{grad}\frac{\partial u^m_n}{\partial x_j}$$

and hence

$$\operatorname{curl}E^m_{j,n}=k^2\operatorname{curl}e_ju^m_n=k^2\left[\operatorname{grad}u^m_n,\ e_j\right].$$

We can now rewrite (2.3) as

$$(2.4)\qquad \int_{\partial D}\left(k^2\left[e_j,\operatorname{grad}u_n^m\right]+i\eta k^2[\nu,e_j]u_n^m+i\eta\left[\nu,\operatorname{grad}\frac{\partial u_n^m}{\partial y_j}\right],\bar{a}\right)ds(y)=0.$$

Define the vector field F by

$$F(x):=\operatorname{curl}\int_{\partial D}\bar{a}(y)\Phi(x,y)\,ds(y)$$
$$+\frac{i\eta}{k^2}\operatorname{curl}\operatorname{curl}\int_{\partial D}[\nu(y),\bar{a}(y)]\,\Phi(x,y)\,ds(y).$$

We use the identities

$$\operatorname{curl}_x\bar{a}(y)\Phi(x,y)=\left[\operatorname{grad}_x\Phi(x,y),\bar{a}(y)\right]=\left[\bar{a}(y),\operatorname{grad}_y\Phi(x,y)\right]$$

and

$$\begin{aligned}\left(e_j,\operatorname{grad}_x\operatorname{div}_x[\nu(y),\bar{a}(y)]\,\Phi(x,y)\right)&=\frac{\partial}{\partial x_j}\operatorname{div}_x[\nu(y),\bar{a}(y)]\,\Phi(x,y)\\&=\frac{\partial}{\partial x_j}(\operatorname{grad}_x\Phi(x,y),\nu(y),\bar{a}(y))\\&=-\frac{\partial}{\partial x_j}(\operatorname{grad}_y\Phi(x,y),\nu(y),\bar{a}(y))\\&=-\left(\operatorname{grad}_y\frac{\partial\Phi}{\partial x_j},\nu(y),\bar{a}(y)\right)\\&=\left(\operatorname{grad}_y\frac{\partial\Phi}{\partial y_j},\nu(y),\bar{a}(y)\right)\end{aligned}$$

to calculate

$$\begin{aligned}k^2\left(e_j,F(x)\right)=k^2&\int_{\partial D}\left(e_j,\bar{a}(y),\operatorname{grad}_y\Phi(x,y)\right)ds(y)\\&+i\eta k^2\int_{\partial D}\left(e_j,\nu(y),\bar{a}(y)\right)\Phi(x,y)\,ds(y)\\&+i\eta\int_{\partial D}\left(\nu(y),\bar{a}(y),\operatorname{grad}_y\frac{\partial\Phi(x,y)}{\partial y_j}\right)ds(y).\end{aligned}$$

Making use of the expansion

$$\Phi(x,y)=ik\sum_{n=0}^{\infty}\sum_{m=-n}^{n}v_n^m(x)u_n^m(y),$$

which together with its term by term derivatives is uniformly convergent for $|x|>|y|$, we see that

$$
\begin{aligned}
k^2\big(e_j,F(x)\big)=ik\sum_{n=0}^{\infty}\sum_{m=-n}^{n} v_n^m(x)\Big\{ & k^2\int_{\partial D}\big(e_j,\bar{a}(y),\operatorname{grad}u_n^m(y)\big)\,ds(y)\\
&+i\eta k^2\int_{\partial D}\big(e_j,\nu(y),\bar{a}(y)\big)u_n^m(y)\,ds(y)\\
&+i\eta\int_{\partial D}\left(\nu(y),\bar{a}(y),\operatorname{grad}\frac{\partial u_n^m(y)}{\partial y_j}\right)ds(y)\Big\}
\end{aligned}
$$

for $|x|$ sufficiently large. From (2.4) we can now conclude that $F(x)=0$ for $|x|$ sufficiently large, and by the analyticity of solutions to the Helmholtz equation we see that $F(x)=0$ for $x\in\mathbb{R}^3\setminus\bar{D}$. From Lemma 2.2 we now see that on ∂D

$$
\begin{aligned}
-[\nu,F_-]&=\bar{a},\\
-[\nu,\operatorname{curl}F_-]&=i\eta[\nu,\bar{a}],
\end{aligned}
$$

and with the aid of Gauss' theorem we have

$$
\begin{aligned}
i\eta\int_{\partial D}|a|^2ds&=\int_{\partial D}\big(\nu,\bar{F}_-,\operatorname{curl}F_-\big)\,ds\\
&=\iint_D \operatorname{div}[\bar{F},\operatorname{curl}F]\,dx\\
&=\iint_D\left\{|\operatorname{curl}F|^2-(\bar{F},\operatorname{curl}\operatorname{curl}F)\right\}dx\\
&=\iint_D\left\{|\operatorname{curl}F|^2-k^2|F|^2\right\}dx.
\end{aligned}
$$

Since the left-hand side is purely imaginary and the right-hand side is real, it follows that $a=0$, and this completes the proof of the theorem.

Remark. Note that Theorem 2.1 is not valid if $\eta=0$. To see this let D be the unit ball. Then consider

$$F_n^m(x):=\operatorname{curl}\operatorname{curl}xu_n^m(x).$$

Straightforward calculations show that

$$F_n^m(x)=k^2xu_n^m(x)+\operatorname{grad}\big[u_n^m(x)+\big(x,\operatorname{grad}u_n^m(x)\big)\big]$$

and

$$\operatorname{curl}F_n^m(x)=k^2\big[\operatorname{grad}u_n^m(x),x\big].$$

Therefore, if k is a zero of j_n for some n, then

$$[\nu,\operatorname{curl}F_n^m]=k^2j_n(k)\operatorname{grad}Y_n^m=0\quad\text{on }\partial D$$

and

$$[\nu,F_n^m]=kj_n'(k)[\nu,\operatorname{grad}Y_n^m]\neq 0\quad\text{on }\partial D$$

for all $m=-n,\cdots,n$. We apply Gauss' theorem to obtain

$$\int_{\partial D}([[\operatorname{curl}E,\nu],\nu],[\nu,F_n^m])\,ds=\int_{\partial D}\{(\nu,\operatorname{curl}E,F_n^m)-(\nu,\operatorname{curl}F_n^m,E)\}\,ds$$

$$=\int_D \operatorname{div}\{[\operatorname{curl}E,F_n^m]-[\operatorname{curl}F_n^m,E]\}\,dx=0$$

since $E\in\mathscr{E}$ and F_n^m solve the vector Helmholtz equation. Hence, W is not complete if k is a zero of j_n (in the case $\eta=0$ and ∂D the unit ball).

We now consider the set T of all solutions $\{E,H\}$ of boundary value problems of the form

$$\begin{aligned}&\operatorname{curl}E-ikH=0\\&\operatorname{curl}H+ikE=0\end{aligned}\quad \text{in } \mathbb{R}^3\setminus\bar{D},$$

$$[\nu,E]=0 \qquad \text{on } \partial D,$$

where $E=E^i+E^s$, $H=H^i+H^s$ such that $E^i\in\mathscr{E}$, $H^i:=(1/ik)\operatorname{curl}E^i$, and

$$\left[H^s,\frac{x}{|x|}\right]-E^s=o\left(\frac{1}{|x|}\right),\qquad |x|\to\infty.$$

Then we have the following theorem.

THEOREM 2.2. *The set*

$$V:=\{b:=[[H,\nu],\nu],\ \{E,H\}\in T\}$$

is complete in X.

Proof. We first derive a uniquely solvable integral equation for $b=[[H,\nu],\nu]$ by combining the magnetic and electric field equations (cf. [3, §4.8]). Define integral operators $\mathbf{M}$, $\mathbf{M}'$ and $\mathbf{N}$ by

$$(\mathbf{M}a)(x):=2\int_{\partial D}[\nu(x),\operatorname{curl}_x\{\Phi(x,y)a(y)\}]\,ds(y),\qquad x\in\partial D,$$

$$\mathbf{M}'a:=[\nu,\mathbf{M}[\nu,a]],$$

$$(\mathbf{N}a)(x):=2\left[\nu(x),\operatorname{curl}_x\operatorname{curl}_x\int_{\partial D}\Phi(x,y)[\nu(y),a(y)]\,ds(y)\right],\qquad x\in\partial D,$$

where a is a Hölder continuous tangential vector field defined on ∂D which in the case of the operator N in addition is assumed to have a Hölder continuous surface divergence $\operatorname{Div}[\nu,a]$. We note that the operators $\mathbf{M}$ and $\mathbf{M}'$ are adjoint with respect to the pairing

$$\langle a,b\rangle:=\int_{\partial D}(a,b)\,ds,$$

and the operator $\mathbf{N}$ is self adjoint (cf. [3, p. 63]). From the representation theorems for solutions to Maxwell's equations we can now deduce (cf. [3, §4.8]) that

$$[\nu,E^s]-\mathbf{M}[\nu,E^s]-\frac{1}{ik}\mathbf{N}[\nu,[\nu,H^s]]=0,$$

$$-[\nu,E^i]-\mathbf{M}[\nu,E^i]-\frac{1}{ik}\mathbf{N}[\nu,[\nu,H^i]]=0.$$

If we now add these two equations and use the boundary condition $[\nu, E]=0$ we see that

$$\mathbf{N}b = -2ik[\nu, E^i]. \tag{2.5}$$

Similarly, from the magnetic field equations

$$[\nu, H^s] - \mathbf{M}[\nu, H^s] + \frac{1}{ik}\mathbf{N}[\nu, [\nu, E^s]] = 0,$$

$$-[\nu, H^i] - \mathbf{M}[\nu, H^i] + \frac{1}{ik}\mathbf{N}[\nu, [\nu, H^i]] = 0,$$

we see that

$$b + \mathbf{M}'b = 2[\nu, [\nu, H^i]]. \tag{2.6}$$

We now combine (2.5) and (2.6) to obtain the combined magnetic and electric field equation

$$b + \mathbf{M}'b + \frac{i\eta}{k^2}\mathbf{N}b = \frac{2}{ik}\{[[\operatorname{curl} E^i, \nu], \nu] + i\eta[\nu, E^i]\} \tag{2.7}$$

which for $\eta > 0$ is uniquely solvable (cf. [3, Thm. 4.47]).

Now let $a \in X$ such that

$$\int_{\partial D} (\bar{a}, b)\, ds = 0$$

for all $b \in V$. Our aim is to show that $a = 0$. To this end we see from (2.7) that

$$\int_{\partial D}\left(\bar{a}, \left(\mathbf{I} + \mathbf{M}' + \frac{i\eta}{k^2}\mathbf{N}\right)^{-1} c\right) ds = 0$$

for all $c \in W$. But this implies that

$$\int_{\partial D}\left(\left(\mathbf{I} + \mathbf{M} + \frac{i\eta}{k^2}\mathbf{N}\right)^{-1} \bar{a}, c\right) ds = 0$$

for all $c \in W$, and by Theorem 2.1, we can conclude that

$$\left(\mathbf{I} + \mathbf{M} + \frac{i\eta}{k^2}\mathbf{N}\right)^{-1} \bar{a} = 0,$$

which implies that $a = 0$. (Here we have made use of the fact that $(\mathbf{I} + \mathbf{M} + (i\eta/k^2)\mathbf{N})^{-1}$ is defined and injective on X. This result is shown in the space of Hölder continuous tangential fields in [3, Thm. 4.44] through the use of a regularizing technique for the singular operator $\mathbf{I} + \mathbf{M} + i(\eta/k^2)\mathbf{N}$ and the same method can also be used to show the result for the space X).

3. Far field patterns.

Before we can present our results on the denseness of electric far field patterns in the space of square integrable tangential fields defined on the boundary of the unit ball, we need to introduce a special class of entire solutions of Maxwell's equations which are the electromagnetic analogue of Herglotz wave functions for the scalar Helmholtz equation (cf. [4]).

DEFINITION 3.1. Let $\{E,H\}$ be a solution of Maxwell's equations. Then $\{E,H\}$ will be called an *electromagnetic Herglotz pair* if

$$\lim_{r\to\infty}\frac{1}{r}\iint_{|x|\leqq r}\left(|E(x)|^2+|H(x)|^2\right)dx<\infty.$$

Remark. It is easily verified that $E_{j,n}^m$ and $H_{j,n}^m := (1/ik)\operatorname{curl} E_{j,n}^m$ form an electromagnetic Herglotz pair.

Motivated by the results of Hartman and Wilcox [4] for the scalar Helmholtz equation, we begin by obtaining an integral representation for the electric field of an electromagnetic Herglotz pair.

THEOREM 3.1. *Let $\{E,H\}$ be an entire solution of Maxwell's equations. Then $\{E,H\}$ is an electromagnetic Herglotz pair if and only if E has the representation*

$$E(x)=\frac{1}{4\pi}\int_{\partial\Omega}a(\hat{y})e^{ik(x,\hat{y})}ds(\hat{y})$$

where a is a square integrable tangential field defined on the boundary $\partial\Omega$ of the unit ball Ω.

Proof. Let $\{E,H\}$ be an electromagnetic Herglotz pair. Then E and H are both solutions of the vector Helmholtz equation and hence have the expansions

$$E(x)=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}i^n a_{nm}u_n^m(x),$$

$$H(x)=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}i^n b_{nm}u_n^m(x),$$

where a_{nm} and b_{nm} are constant vectors and the series are uniformly convergent on compact subsets of $\mathbb{R}^3$. Since $\{E,H\}$ satisfies the Herglotz conditions, it follows from the results of [4] that

$$\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(|a_{nm}|^2+|b_{nm}|^2\right)<\infty.$$

For $\hat{x}\in\partial\Omega$ we can now define $b\in L^2(\partial\Omega)$ by

$$b(\hat{x}):=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}b_{nm}Y_n^m(\theta,\varphi)$$

where (θ,φ) are the spherical coordinates of $\hat{x}$, and from the addition formula

$$e^{ik(x,\hat{y})}=4\pi\sum_{n=0}^{\infty}\sum_{m=-n}^{n}i^n u_n^m(x)\overline{Y}_n^m(\theta',\varphi')$$

where (θ',φ') are the spherical coordinates of $\hat{y}\in\partial\Omega$, we see that

$$H(x)=\frac{1}{4\pi}\int_{\partial\Omega}b(\hat{y})e^{ik(x,\hat{y})}ds(\hat{y}).$$

Since from Maxwell's equation $E=-(1/ik)\operatorname{curl} H$ we can compute

$$-ikE(x)=\operatorname{curl} H(x)=\operatorname{curl}\frac{1}{4\pi}\int_{\partial\Omega} b(\hat{y})e^{ik(x,\hat{y})}ds(\hat{y})\tag{3.1}$$

$$=\frac{ik}{4\pi}\int_{\partial\Omega}[\hat{y},b(\hat{y})]\,e^{ik(x,\hat{y})}ds(\hat{y}),$$

i.e.

$$E(x)=\frac{1}{4\pi}\int_{\partial\Omega} a(\hat{y})e^{ik(x,\hat{y})}ds(\hat{y})\tag{3.2}$$

where $a(\hat{y}):=-[\hat{y},b(\hat{y})]$ is a square integrable tangential field defined on $\partial\Omega$.

Conversely, suppose $\{E,H\}$ is a solution of Maxwell's equations such that E has the representation (3.2) for some square integrable tangential field a defined on $\partial\Omega$. Then each component of E can be represented in the form (3.2) for a scalar square integrable function a defined on $\partial\Omega$, and hence from the results of [4] for scalar Herglotz wave functions we can now conclude that

$$\lim_{r\to\infty}\frac{1}{r}\iint_{|x|\leqq r}|E(x)|^2dx<\infty.$$

From Maxwell's equations we have that $H=(1/ik)\operatorname{curl} E$, and hence from a calculation analogous to that in (3.1) we see that H can also be represented in the form (3.2) for some square integrable tangential field a. Hence we can conclude that

$$\lim_{r\to\infty}\frac{1}{r}\iint_{|x|\leqq r}|H(x)|^2dx<\infty,$$

i.e. $\{E,H\}$ is an electromagnetic Herglotz pair.

We now turn our attention to the class of electric far field patterns corresponding to the scattering of an entire incident field by a perfect conductor. To this end let D be as defined in §2 of this paper and let the incident field $\{E^i,H^i\}$ be an entire solution of Maxwell's equations. Then the exterior Maxwell boundary value problem is to find $E=E^i+E^s$ and $H=H^i+H^s$ such that$E$, $H\in C^1(\mathbb{R}^3\backslash\bar{D})\cap C(\mathbb{R}^3\backslash D)$and

$$\begin{matrix}\operatorname{curl} E-ikH=0\\ \operatorname{curl} H+ikE=0\end{matrix}\quad\text{in }\mathbb{R}^3\backslash\bar{D},\tag{3.3a}$$

$$[\nu,E]=0\quad\text{on }\partial D,\tag{3.3b}$$

$$\left[H^s,\frac{x}{|x|}\right]-E^s=o\left(\frac{1}{|x|}\right),\qquad |x|\to\infty,\tag{3.3c}$$

where $\{E^s,H^s\}$ denotes the scattered field. The existence of a unique solution to the boundary value problem (3.3) is well known (cf. [3]). From [3, pp. 110, 114], we see that E^s has the representation

$$E^s(x)=-\frac{1}{ik}\operatorname{curl}\operatorname{curl}\int_{\partial\Omega}[\nu(y),H(y)]\,\Phi(x,y)\,ds(y),\qquad x\in\mathbb{R}^3\backslash\bar{D}$$

and using the relations [3, p. 113]

$$\Phi=\frac{1}{4\pi|x|}e^{ik\{|x|-(\hat{x},y)\}}+O\left(\frac{1}{|x|^2}\right),$$

$$\operatorname{curl}\operatorname{curl}e\Phi=k^2\Phi\{e-(e,\hat{x})\hat{x}\}+O\left(\frac{1}{|x|^2}\right),$$

where e is a constant vector, we see that

$$E^s(x)=\frac{ike^{ik|x|}}{4\pi|x|}\int_{\partial\Omega}e^{-ik(\hat{x},y)}\{[\nu,H]-([\nu,H],\hat{x})\hat{x}\}\,ds(y)+O\left(\frac{1}{|x|^2}\right).$$

DEFINITION 3.2. The function

$$F(\hat{x}):=\frac{1}{4\pi}\int_{\partial\Omega}e^{-ik(\hat{x},y)}\{[\nu,H]-([\nu,H],\hat{x})\hat{x}\}\,ds(y)$$

is called the *far field pattern* of the electric field corresponding to the boundary value problem (3.3).

Note that if F is the electric far field pattern corresponding to the boundary value problem (3.3), then $(\hat{x},F)=0$, i.e. F is tangential to the unit sphere.

We now want to establish the main result of this paper, that is to determine necessary and sufficient conditions for the class of electric far field patterns corresponding to all entire incident fields to be dense in the Hilbert space

$$Y:=\{\,a:\partial\Omega\to\mathbf{C}^3|(\hat{x},a)=0,\,a\in L^2(\partial\Omega),\,\hat{x}\in\partial\Omega\}.$$

In order to state our theorem we need the following definition.

DEFINITION 3.3. The positive number k is called a *Maxwell eigenvalue* provided there exists a nontrivial solution of the boundary value problem

$$\left.\begin{array}{l}\operatorname{curl}E-ikH=0\\ \operatorname{curl}H+ikE=0\end{array}\right.\quad\text{in } D,$$

$$[\nu,E]=0\quad\text{on }\partial D.$$

Such a nontrivial pair $\{E,H\}$ is called a *Maxwell eigenfunction*.

We note that the set of Maxwell eigenvalues is countable with its only accumulation point being the point at infinity (cf. [6]).

THEOREM 3.2. *Consider the set $\mathcal{F}_k$ of electric far field patterns corresponding to all entire incident fields and a given domain D. Then*

a) *If k is not a Maxwell eigenvalue, then $\mathcal{F}_k$ is dense in Y.*

b) *If k is a Maxwell eigenvalue, then $\mathcal{F}_k$ is dense in Y if and only if none of the Maxwell eigenfunctions is an electromagnetic Herglotz pair.*

Proof. Suppose there exists $a\in Y$ such that

$$\int_{\partial\Omega}(\bar{a},F)\,ds=0$$

for all $F \in \mathscr{F}_k$. Then

$$(3.4) \qquad \int_{\partial\Omega} (\bar{a}, F)\, ds = \int_{\partial\Omega} \left([\nu, H], \frac{1}{4\pi} \int_{\partial\Omega} \overline{a(\hat{x}) e^{-ik(\hat{x},y)}}\, ds(\hat{x}) \right) ds(y)$$

$$= \int_{\partial\Omega} ([\nu, H], E_0)\, ds(y),$$

where

$$(3.5) \qquad E_0(y) := \frac{1}{4\pi} \int_{\partial\Omega} \overline{a(\hat{x})} e^{-ik(\hat{x},y)}\, ds(\hat{x}).$$

Note that E_0 as defined by (3.5) is an entire solution of the vector Helmholtz equation, and since

$$\operatorname{div} E_0 = -\frac{ik}{4\pi} \int_{\partial\Omega} \left(\overline{a(\hat{x})}, \hat{x}\right) e^{-ik(\hat{x},y)}\, ds(\hat{x}) = 0,$$

we have that E_0 and $H_0 := (1/ik) \operatorname{curl} E_0$ satisfies Maxwell's equations [3, p. 112]. Using the vector identity $([[\nu, H], \nu], [E_0, \nu]) = ([\nu, H], E_0)$, we can rewrite (3.4) in the form

$$(3.6) \qquad \int_{\partial\Omega} (\bar{a}, F)\, ds = -\int_{\partial\Omega} ([[H, \nu], \nu], [E_0, \nu])\, ds.$$

Hence from (3.6) and Theorem 2.2 we now see that if

$$(3.7) \qquad \int_{\partial\Omega} (\bar{a}, F)\, ds = 0$$

for all $F \in \mathscr{F}_k$, then $[E_0, \nu] = 0$ on ∂D, i.e. $\{E_0, H_0\}$ is either a Maxwell eigenfunction or E_0 (and hence H_0) is identically zero.

If k is not a Maxwell eigenvalue and (3.7) is valid, then E_0 is identically zero, and from (3.5) and the analysis in Theorem 3.1 we see that a is identically zero, i.e. $\mathscr{F}_k$ is dense in Y. This proves part a) of the theorem. If k is a Maxwell eigenvalue and (3.7) is valid, then we see that either a is identically zero or $E_1 := \bar{E}_0$ and $H_1 := (1/ik) \operatorname{curl} \bar{E}_0$ constitute a Maxwell eigenfunction, and from Theorem 3.1 we see that $\{E_1, H_1\}$ is an electromagnetic Herglotz pair. This proves part b) of the theorem.

Example. Let D be the unit ball. Then from the remark after Theorem 2.1 we see that if $j_n(k) = 0$ then $E := \operatorname{curl} F_n^m$, $H := (1/ik) \operatorname{curl} E$, $m = -n, \cdots, n$, constitute Maxwell eigenfunctions which are easily seen to be electromagnetic Herglotz pairs. Hence in this case $\mathscr{F}_k$ is not dense in Y.

REFERENCES

[1] D. Colton, *Far field patterns for the impedance boundary value problem in acoustic scattering*, Applic. Anal., 16 (1983), pp. 131–139.

[2] D. Colton and A. Kirsch, *Dense sets and far field patterns in acoustic wave propagation*, this Journal, 15 (1984), pp. 996–1006.

[3] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.

[4] P. Hartman and C. Wilcox, *On solutions of the Helmholtz equation in exterior domains*, Math. Z., 75 (1961), pp. 228–255.

[5] H. Kersten, *Grenz- und Sprungrelationen für Potentiale mit quadrat-summierbarer Flächenbelegung*, Resultate der Math., 3 (1982), pp. 17–24.

[6] C. Müller and H. Niemeyer, *Greensche Tensoren und asymptotische Gesetze der elektromagnetischen Hohlraumschwingungen*, Arch. Rat. Mech. Anal., 1 (1961), pp. 305–358.

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

ROGERS' LINEARIZATION FORMULA FOR THE CONTINUOUS q-ULTRASPHERICAL POLYNOMIALS AND QUADRATIC TRANSFORMATION FORMULAS*

GEORGE GASPER[†]

Abstract. A simple proof is given for the Rogers' linearization formula for the continuous q-ultraspherical polynomials. This formula is then used to derive several quadratic transformation formulas.

1. Introduction.

Almost 90 years ago, in his work [11]–[13] on the now famous Rogers–Ramanujan identities, Rogers [13, p. 29] used an induction argument to prove the linearization formula

(1.1)

$$C_n(x;\beta|q)C_m(x;\beta|q)=\sum_{k=0}^{\min(m,n)}\frac{(q;q)_{m+n-2k}(\beta;q)_{m-k}(\beta;q)_{n-k}}{(q;q)_k(q;q)_{m-k}(q;q)_{n-k}}\cdot\frac{(\beta;q)_k(\beta^2;q)_{m+n-k}(1-\beta q^{m+n-2k})}{(\beta q;q)_{m+n-k}(\beta^2;q)_{m+n-2k}(1-\beta)}C_{m+n-2k}(x;\beta|q),$$

where, following Askey and Ismail [2], the continuous q-ultraspherical polynomials are defined for $x=\cos\theta$ by

$$C_n(\cos\theta;\beta|q)=\frac{(\beta;q)_n}{(q;q)_n}e^{in\theta}\,{}_2\phi_1\left[\begin{matrix}q^{-n},\beta\\ \beta^{-1}q^{1-n}\end{matrix};q,q\beta^{-1}e^{-2i\theta}\right]. \tag{1.2}$$

Here $(a;q)_0=1$, $(a;q)_n=(1-a)(1-aq)\cdots(1-aq^{n-1})$ for $n=1,2,\cdots$, and a ${}_{r+1}\phi_r$ basic hypergeometric series is defined by

$${}_{r+1}\phi_r\left[\begin{matrix}a_1,\cdots,a_{r+1}\\ b_1,\cdots,b_r\end{matrix};q,z\right]=\sum_{k=0}^{\infty}\frac{(a_1;q)_k\cdots(a_{r+1};q)_k}{(b_1;q)_k\cdots(b_r;q)_k}\frac{z^k}{(q;q)_k} \tag{1.3}$$

whenever the series converges. We also let $(a;q)_\infty=(1-a)(1-aq)(1-aq^2)\cdots$ when $|q|<1$. The series in (1.3) is said to be balanced (Saalschützian) if $qa_1a_2\cdots a_{r+1}=b_1b_2\cdots b_r$ and to be well-poised if $qa_1=a_2b_1=a_3b_2=\cdots=a_{r+1}b_r$.

Additional proofs of (1.1) were not published until 1981 when Bressoud [6] used a nontrivial inductive type argument to prove it in an equivalent multiple series form (which, as in the Rogers proof, required knowing the linearization coefficients explicitly beforehand) and Rahman [10] gave an impressive computation of the linearization coefficients (as ${}_{10}\phi_9$ series) for the more general continuous q-Jacobi polynomials that is quite lengthy even in the ultraspherical case. Rahman's results were employed by the author [8] to give a convolution structure associated with the linearization, a Wiener–Lévy type theorem, and positivity of a generalized translation operator. Askey

* Received by the editors November 1, 1983.

[†] Department of Mathematics, Northwestern University, Evanston, Illinois 60201. This work was supported in part by the National Science Foundation under grant MCS-8002507 A01.

and Ismail used (1.1) to give an inverse formula [2, (4.19)] that was subsequently used in Gasper and Rahman [9] to derive an ${}_8\phi_7$ series representation for the Poisson kernel $P_t(x,y;\ \beta|q)$ for the continuous q-ultraspherical polynomials from which the positivity of this kernel follows immediately for $-1<t<1$, $-1\leqq x, y\leqq 1$ when $-1<q<1$ and $0\leqq\beta<1$. Applications of (1.1) to q-ultraspherical functions of the second kind and to sieved ultraspherical polynomials were recently given in Askey, Koornwinder and Rahman [3] and in Al-Salam, Allaway and Askey [1], respectively.

These applications finally convinced the author to publish the following simple computational proof of (1.1) which he discovered several years ago (see [10, p. 961]) while searching for a proof as simple as that given by Bailey [4] for Legendre polynomials. We shall also use (1.1) to derive several quadratic transformation formulas involving series that are neither balanced nor well-poised.

2. Proof of (1.1). From Heine's formula [5, 8.4(2)]

$$(2.1)\qquad {}_2\phi_1\left[\begin{matrix}a,b\\ c\end{matrix};q,z\right]=\frac{(abz/c;\ q)_\infty}{(z;\ q)_\infty}\ {}_2\phi_1\left[\begin{matrix}c/a,c/b\\ c\end{matrix};q,\ \frac{abz}{c}\right]$$

it follows that if $|q|<|\beta|<1$ and $x=\cos\theta$, then

$$(2.2)\quad C_m(x;\ \beta|q)=\frac{(\beta;\ q)_m(\beta e^{-2i\theta};\ q)_\infty e^{im\theta}}{(q;\ q)_m(q\beta^{-1}e^{-2i\theta};\ q)_\infty}\ {}_2\phi_1\left[\begin{matrix}q\beta^{-1},\beta^{-2}q^{1-m}\\ \beta^{-1}q^{1-m}\end{matrix};q,\beta e^{-2i\theta}\right],$$

and hence

$$(2.3)\quad C_n(x;\ \beta|q)C_m(x;\ \beta|q)=A_{m,n}\sum_{r=0}^{n}\frac{(q^{-n};\ q)_r(\beta;\ q)_r}{(q;\ q)_r(\beta^{-1}q^{1-n};\ q)_r}(q\beta^{-1}e^{-2i\theta})^r$$

$$\cdot\sum_{s=0}^{\infty}\frac{(q\beta^{-1};\ q)_s(\beta^{-2}q^{1-m};\ q)_s}{(q;\ q)_s(\beta^{-1}q^{1-m};\ q)_s}(\beta e^{-2i\theta})^s$$

$$=A_{m,n}\sum_{j=0}^{\infty}\frac{(q\beta^{-1};\ q)_j(\beta^{-2}q^{1-m};\ q)_j}{(q;\ q)_j(\beta^{-1}q^{1-m};\ q)_j}(\beta e^{-2i\theta})^j$$

$$\cdot{}_4\phi_3\left[\begin{matrix}\beta q^{m-j},\ \beta,\ q^{-n},\ q^{-j}\\ \beta^2q^{m-j},\ \beta q^{-j},\ \beta^{-1}q^{1-n}\end{matrix};q,\ q\right],$$

where

$$A_{m,n}=\frac{(\beta;\ q)_m(\beta;\ q)_n(\beta e^{-2i\theta};\ q)_\infty}{(q;\ q)_m(q;\ q)_n(q\beta^{-1}e^{-2i\theta};\ q)_\infty}e^{i(m+n)\theta}.$$

At this point it is crucial to observe that since the above ${}_4\phi_3$ is balanced it follows from Watson's [5, 8.5(2)] transformation formula

$${}_4\phi_3\left[\begin{matrix}x,y,z,q^{-n}\\ u,v,w\end{matrix};q,q\right]$$

$$=\frac{(u/z;\ q)_n(u/y;\ q)_n}{(u;\ q)_n(u/yz;\ q)_n}$$

$$\cdot{}_8\phi_7\left[\begin{matrix}a,\ q\sqrt{a},\ -q\sqrt{a},\ w/x,\ v/x,\ y,\ z,\ q^{-n}\\ \sqrt{a},\ -\sqrt{a},\ v,\ w,\ zu^{-1}q^{1-n},\ yu^{-1}q^{1-n},\ yzq/u\end{matrix};q,\ \frac{vwq^n}{yz}\right],$$

where $a=yzu^{-1}q^{-n}=vwx^{-1}q^{-1}$, that

(2.4)

$$
{}_4\Phi_3\left[\begin{matrix}\beta q^{m-j},\beta,q^{-n},q^{-j}\\ \beta^2q^{m-j},\beta q^{-j},\beta^{-1}q^{1-n}\end{matrix};q,q\right]
=\frac{(\beta^{-2}q^{1-m-n};q)_j(\beta^{-1}q^{1-m};q)_j}{(\beta^{-1}q^{1-m-n};q)_j(\beta^{-2}q^{1-m};q)_j}
$$

$$
\cdot{}_8\Phi_7\left[\begin{matrix}c,q\sqrt{c},-q\sqrt{c},\beta^{-2}q^{1+j-m-n},q^{-m},\beta,q^{-n},q^{-j}\\ \sqrt{c},-\sqrt{c},\beta q^{-j},\beta^{-1}q^{1-n},\beta^{-2}q^{1-m-n},\beta^{-1}q^{1-m},\beta^{-1}q^{1+j-m-n}\end{matrix};q,\frac{q}{\beta}\right]
$$

with $c=\beta^{-1}q^{-m-n}$. Hence, from (2.3) and (2.4),

(2.5)

$$
C_n(x;\beta|q)C_m(x;\beta|q)
$$

$$
=A_{m,n}\sum_{j=0}^{\infty}\frac{(q\beta^{-1};q)_j(\beta^{-2}q^{1-m-n};q)_j}{(q;q)_j(\beta^{-1}q^{1-m-n};q)_j}(\beta e^{-2i\theta})^j
$$

$$
\cdot{}_8\phi_7\left[\begin{matrix}c,q\sqrt{c},-q\sqrt{c},q^{-n},\beta^{-2}q^{1+j-m-n},q^{-j},\beta,q^{-m}\\ \sqrt{c},-\sqrt{c},\beta^{-1}q^{1-m},\beta q^{-j},\beta^{-1}q^{1+j-m-n},\beta^{-2}q^{1-m-n},\beta^{-1}q^{1-n}\end{matrix};q,\frac{q}{\beta}\right]
$$

$$
=A_{m,n}\sum_{k=0}^{\min(m,n)}\frac{(c;q)_k(q\sqrt{c};q)_k(-q\sqrt{c};q)_k(q^{-n};q)_k(\beta;q)_k(q^{-m};q)_k}{(q;q)_k(\sqrt{c};q)_k(-\sqrt{c};q)_k(\beta^{-1}q^{1-m};q)_k}
$$

$$
\cdot\frac{(c;q)_k(q\sqrt{c};q)_k(-q\sqrt{c};q)_k(q^{-n};q)_k(\beta;q)_k(q^{-m};q)_k}{(\beta^{-2}q^{1-m-n};q)_k(\beta^{-1}q^{1-n};q)_k}
$$

$$
\cdot\frac{(\beta^{-2}q^{1-m-n};q)_{2k}}{(\beta^{-1}q^{1-m-n};q)_{2k}}(q\beta^{-1}e^{-2i\theta})^k\,{}_2\phi_1\left[\begin{matrix}q\beta^{-1},\beta^{-2}q^{1+2k-m-n}\\ \beta^{-1}q^{1+2k-m-n}\end{matrix};q,\beta e^{-2i\theta}\right]
$$

$$
=\sum_{m=0}^{\min(m,n)}\frac{(q;q)_{m+n-2k}(\beta;q)_{m-k}(\beta;q)_{n-k}(\beta;q)_k(\beta^2;q)_{m+n-k}}{(q;q)_k(q;q)_{m-k}(q;q)_{n-k}(\beta q;q)_{m+n-k}(\beta^2;q)_{m+n-2k}}
$$

$$
\cdot\frac{1-\beta q^{m+n-2k}}{1-\beta}C_{m+n-2k}(x;\beta|q)
$$

by changing the order of summation and using (2.2). This completes the proof since the restrictions that $|q|<|\beta|<1$ and $-1\leqq x\leqq 1$ can now be dropped.

3. Quadratic transformation formulas. From the above proof it follows that (1.1) is actually equivalent to the transformation formula (2.4). If instead of using both (1.2) and (2.2), we just use the series representation in (1.2) for all three polynomials in (1.1) and compare the coefficient of $e^{i(m+n-2j)\theta}$ on both sides of (1.1), we find that (1.1) is also equivalent to the transformation formulas

(3.1a)

$$
{}_4\phi_3\left[\begin{matrix}\beta, q^{-j}, q^{-m}, \beta q^{n-j} \\ \beta^{-1}q^{1-j}, \beta^{-1}q^{1-m}, q^{1+n-j}\end{matrix}; q, \frac{q^2}{\beta^2}\right]
= \frac{(q^{-m-n};q)_j(\beta^{-1}q^{1-n};q)_j}{(q^{-n};q)_j(\beta^{-1}q^{1-m-n};q)_j}
$$

$$
\cdot {}_9\phi_8\left[\begin{matrix}\beta, q^{-j}, q^{-m}, q^{-n}, \beta^{-1}q^{-m-n}, q^{j-m-n}: \\ \beta^{-1}q^{1-j}, \beta^{-1}q^{1-m}, \beta^{-1}q^{1-n}, \beta^{-2}q^{1-m-n}, \beta^{-1}q^{1+j-m-n}:\end{matrix}\right.
$$

$$
\left.\begin{matrix}\beta^{-2}q^{1-m-n}, \beta^{-2}q^{2-m-n}, \beta^{-1}q^{2-m-n} \\ q^{1-m-n}, q^{-m-n}, \beta^{-1}q^{-m-n}\end{matrix}; q, q^2, \frac{q}{\beta}\right]
$$

for $j=0,1,\cdots,n$ and

(3.1b)

$$
{}_4\phi_3\left[\begin{matrix}\beta, q^{-n}, q^{j-m-n}, \beta q^{j-n} \\ \beta^{-1}q^{1-n}, \beta^{-1}q^{1+j-m-n}, q^{1+j-n}\end{matrix}; q, \frac{q^2}{\beta^2}\right]
= \frac{(\beta^{-1}q^{1-j};q)_n(q^{m+1};q)_n}{(q^{-j};q)_n(\beta q^{m};q)_n}\left(\frac{\beta}{q}\right)^n
$$

$$
\cdot {}_9\phi_8\left[\begin{matrix}\beta, q^{-j}, q^{-m}, q^{-n}, \beta^{-1}q^{-m-n}, q^{j-m-n}: \\ \beta^{-1}q^{1-j}, \beta^{-1}q^{1-m}, \beta^{-1}q^{1-n}, \beta^{-2}q^{1-m-n}, \beta^{-1}q^{1+j-m-n}:\end{matrix}\right.
$$

$$
\left.\begin{matrix}\beta^{-2}q^{1-m-n}, \beta^{-2}q^{2-m-n}, \beta^{-1}q^{2-m-n} \\ q^{1-m-n}, q^{-m-n}, \beta^{-1}q^{-m-n}\end{matrix}; q, q^2, \frac{q}{\beta}\right]
$$

for $j=n, n+1,\cdots,n+m$. In (3.1a) and (3.1b) the bi-basic hypergeometric series are defined by

$$
{}_{r+s+1}\phi_{r+s}\left[\begin{matrix}a_1,\cdots,a_{r+1}: & b_1,\cdots,b_s \\ c_1,\cdots,c_r: & d_1,\cdots,d_s\end{matrix}; q, q', z\right]
= \sum_{k=0}^{\infty}\frac{(a_1;q)_k\cdots(a_{r+1};q)_k(b_1;q')_k\cdots(b_s;q')_k}{(c_1;q)_k\cdots(c_r;q)_k(d_1;q')_k\cdots(d_s;q')_k}\frac{z^k}{(q;q)_k}.
$$

Note that (3.1b) can be obtained from (3.1a) by replacing j, n, and m by n, j, and $m+n-j$, respectively. By analytic continuation m and n can be arbitrary (complex valued) parameters in (3.1a) when $j=0,1,2,\ldots$. If the coefficient of the bi-basic ${}_9\phi_8$ in (3.1a) is replaced by its equivalent form $(q^{n+1};q)_m(\beta q^{n-j};q)_m/(q^{1+n-j};q)_m(\beta q^n;q)_m$, then (3.1a) holds for arbitrary complex values of j and n when $m=0,1,2,\cdots$.

Since the above ${}_4\phi_3$ series are neither balanced nor well-poised, it is natural to investigate what structure they do have. From the way we have written these ${}_4\phi_3$ series, it is obvious that β times each denominator term is q times a numerator term, a property that we shall call "level". In general, the ${}_{r+1}\phi_r$ series in (1.3) will be said to be a level series if $a_1 = qa_2/b_1 = qa_3/b_2 = \cdots = qa_{r+1}/b_r$. When the order of summation in a terminating level series is reversed then the resulting series is also a level series. If we rewrite the ${}_4\phi_3$ in (3.1a) in the form

$$ {}_4\phi_3\left[\begin{matrix} \beta q^{n-j}, \beta, q^{-j}, q^{-m} \\ q^{1+n-j}, \beta^{-1}q^{1-m}, \beta^{-1}q^{1-j} \end{matrix}; q, \frac{q^2}{\beta^2}\right], $$

then it is of the type

$$ {}_4\phi_3\left[\begin{matrix} a, b, c, d \\ aq/b, \alpha/c, \alpha/d \end{matrix}; q, z\right] $$

which, following Whipple [14, p. 270], we shall call a half-poised series. Turning to the above (identical) bi-basic ${}_9\phi_8$ series, observe that the terms corresponding to basee q have the "level" property, while the terms corresponding to base q^2 have the property that $b_1d_1 = b_2d_2 = b_3d_3 = \beta^{-2}q^{2-2m-2n}$, giving series which are split between being level and half-poised. By using $(a;\ q^2)_n = (\sqrt{a};\ q)_n(-\sqrt{a};\ q)_n$, we can also write these bi-basis ${}_9\phi_8$ series as ${}_{12}\phi_{11}$ series in base q that are well-poised (with $a_1 = \beta^{-1}q^{-m-n}$) and, in fact, very well-poised (i.e., they also have the property that $q\sqrt{a_1} = a_2 = -a_3 = qb_1 = -qb_2$). Explicitly, by analytic continuation, it follows from (3.1a) and (3.1b) that

(3.2)

$$ \begin{aligned} &{}_4\phi_3\left[\begin{matrix} a, b, c, d \\ bq/a, cq/a, dq/a \end{matrix}; q, \frac{q^2}{a^2}\right] \\ &\quad = \frac{(a/d;\, q)_\infty (bq/d;\, q)_\infty (cq/d;\, q)_\infty (abc/d;\, q)_\infty}{(q/d;\, q)_\infty (ab/d;\, q)_\infty (ac/d;\, q)_\infty (bcq/d;\, q)_\infty} \\ &\quad\cdot {}_{12}\phi_{11}\left[\begin{matrix} bc/d, q\sqrt{bc/d}, -q\sqrt{bc/d}, a, b, c, ab/d, \\ \sqrt{bc/d}, -\sqrt{bc/d}, bcq/ad, cq/d, bq/d, cq/a, \end{matrix}\right. \\ &\qquad\qquad \left.\begin{matrix} ac/d, \sqrt{bcq/ad}, -\sqrt{bcq/ad}, q\sqrt{bc/ad}, -q\sqrt{bc/ad} \\ bq/a, \sqrt{abcq/d}, -\sqrt{abcq/d}, \sqrt{abc/d}, -\sqrt{abc/d} \end{matrix}; q, \frac{q}{a}\right], \end{aligned} $$

where a, b, or c is of the form q^{-n} with $n = 0, 1, 2, \cdots$ (so that both series terminate).

By replacing a, b, c, d in (3.2) by a^a, q^b, q^c, q^d, respectively, and letting $q \to 1$ we obtain the following quadratic transformation formula between a ${}_4F_3$ and ${}_9F_8$ hypergeometric series:

$$ \begin{aligned} (3.3)\quad &{}_4F_3\left[\begin{matrix} a, b, c, d \\ 1+b-a, 1+c-a, 1+d-a \end{matrix}; 1\right] \\ &\quad = \frac{\Gamma(1-d)\Gamma(a+b-d)\Gamma(a+c-d)\Gamma(1+b+c-d)}{\Gamma(a-d)\Gamma(1+b-d)\Gamma(1+c-d)\Gamma(a+b+c-d)} \\ &\quad\cdot {}_9F_8\left[\begin{matrix} b+c-d, 1+(b+c-d)/2, a, b, c, \\ (b+c-d)/2, 1+b+c-a-d, 1+c-d, 1+b-d, \end{matrix}\right. \\ &\qquad\qquad \left.\begin{matrix} a+b-d, a+c-d, (1+b+c-a-d)/2, 1+(b+c-a-d)/2 \\ 1+c-a, 1+b-a, (1+a+b+c-d)/2, (a+b+c-d)/2 \end{matrix}; 1\right], \end{aligned} $$

where a, b, or c is a negative integer.

Since $C_n(x;\ q^\lambda|q)$ tends to the ultraspherical polynomial $C_n^\lambda(x)$ as $q\to 1$, additional quadratic transformation formulas can be obtained from the limiting $q\to 1$ ultraspherical polynomial case of (1.1)

$$C_m^\lambda(x)C_n^\lambda(x)=\sum_{k=0}^{\min(m,n)}\frac{(m+n-2k)!(\lambda)_{m-k}(\lambda)_{n-k}(\lambda)_k(2\lambda)_{m+n-k}}{k!(m-k)!(n-k)!(\lambda+1)_{m+n-k}(2\lambda)_{m+n-2k}}\cdot\frac{\lambda+m+n-2k}{\lambda}C^\lambda_{m+n-2k}(x), \tag{3.4}$$

where $(a)_0=1$ and $(a)_k=a(a+1)\cdots(a+k-1)=\Gamma(k+a)/\Gamma(a)$, by using the following power series representations [7, p. 176]

$$C_n^\lambda(x)=\frac{(2\lambda)_n}{n!}\,{}_2F_1\left[\begin{matrix}-n,\ n+2\lambda\\ \lambda+\frac{1}{2}\end{matrix};\frac{1-x}{2}\right], \tag{3.5}$$

$$C_n^\lambda(x)=\frac{(\lambda)_n}{n!}(2x)^n\,{}_2F_1\left[\begin{matrix}-\frac{1}{2}n,\ \frac{1}{2}-\frac{1}{2}n\\ 1-n-\lambda\end{matrix};\frac{1}{x^2}\right], \tag{3.6}$$

$$C_{2n}^\lambda(x)=\frac{(2\lambda)_{2n}}{(2n)!}\,{}_2F_1\left[\begin{matrix}-n,\ n+\lambda\\ \lambda+\frac{1}{2}\end{matrix};1-x^2\right], \tag{3.7}$$

$$C_{2n+1}^\lambda(x)=\frac{(2\lambda)_{2n+1}}{(2n+1)!}x\,{}_2F_1\left[\begin{matrix}-n,\ n+\lambda+1\\ \lambda+\frac{1}{2}\end{matrix};1-x^2\right]. \tag{3.8}$$

Using (3.5), (3.6), (3.7), and both (3.7) and (3.8) in (3.4) yield respectively,

$$\begin{aligned}&{}_4F_3\left[\begin{matrix}-j,\ \frac{1}{2}-\lambda-j,\ -m,\ m+2\lambda\\ 1+n-j,\ 1-2\lambda-n-j,\ \lambda+\frac{1}{2}\end{matrix};1\right]\\ &=\frac{(j+n+2\lambda)_m(\lambda)_m(n+1)_m}{(1+n-j)_m(2\lambda)_m(n+\lambda)_m}\\ &\cdot{}_9F_8\left[\begin{matrix}-\lambda-m-n,1-\frac{1}{2}(\lambda+m+n),\ -m,\ -n,\lambda,\ \frac{1}{2}(j-m-n),\\ -\frac{1}{2}(\lambda+m+n),1-\lambda-n,1-\lambda-m,\ 1-2\lambda-m-n,\end{matrix}\right.\\ &\qquad\left.\begin{matrix}\frac{1}{2}(1+j-m-n),\ -\lambda+\frac{1}{2}(1-m-n),1-\lambda-\frac{1}{2}(m+n)\\ 1-\lambda-\frac{1}{2}(j+m+n),\ -\lambda+\frac{1}{2}(1-j-m-n),\ \frac{1}{2}(1-m-n),\ -\frac{1}{2}(m+n)\end{matrix};1\right],\end{aligned} \tag{3.9}$$

(3.10)

$${}_4F_3\left[\begin{matrix} -j, \lambda+n-j, -\frac{1}{2}m, \frac{1}{2}-\frac{1}{2}m \\ \frac{1}{2}+\frac{1}{2}n-j, 1+\frac{1}{2}n-j, 1-\lambda-m \end{matrix};1\right]$$

$$=\frac{(\lambda+n-j)_m(n+1)_m}{(1+n-2j)_m(n+\lambda)_m}\,{}_8F_7\left[\begin{matrix} -\lambda-n-m, 1-\frac{1}{2}(\lambda+m+n), -m, -n, -j, \lambda, -\lambda+\frac{1}{2}(1-m-n), 1-\lambda-\frac{1}{2}(m+n) \\ -\frac{1}{2}(\lambda+m+n), 1-\lambda-n, 1-\lambda-m, 1-\lambda+j-m-n, 1-2\lambda-m-n, \frac{1}{2}(1-m-n), -\frac{1}{2}(m+n) \end{matrix};-1\right],$$

(3.11)

$${}_4F_3\left[\begin{matrix} -j, \frac{1}{2}-\lambda-j, -m, m+\lambda \\ 1+n-j, 1-\lambda-n-j, \lambda+\frac{1}{2} \end{matrix};1\right]$$

$$=\frac{(\lambda+j+n)_m(n+1)_m(\lambda)_{2m}(2n+2\lambda)_{2m}}{(1+n-j)_m(n+\lambda)_m(2\lambda)_{2m}(2n+\lambda)_{2m}}$$

$$\cdot\,{}_7F_6\left[\begin{matrix} -\lambda-2m-2n, 1-\frac{1}{2}\lambda-m-n, -2m, -2n, j-m-n, \lambda, 1-\lambda-m-n \\ -\frac{1}{2}\lambda-m-n, 1-\lambda-2n, 1-\lambda-2m, 1-\lambda-j-m-n, 1-2\lambda-2m-2n, -m-n \end{matrix};1\right],$$

(3.12)

$${}_4F_3\left[\begin{matrix} -j, \frac{1}{2}-\lambda-j, -m, m+\lambda+1 \\ 1+n-j, 1-\lambda-n-j, \lambda+\frac{1}{2} \end{matrix};1\right]$$

$$=\frac{(\lambda+j+n)_{m+1}(n+1)_m(\lambda)_{2m+1}(2n+2\lambda)_{2m+1}}{(1+n-j)_m(n+\lambda)_{m+1}(2\lambda)_{2m+1}(2n+\lambda)_{2m+1}}$$

$$\cdot\,{}_7F_6\left[\begin{matrix} -1-\lambda-2m-2n, \frac{1}{2}-\frac{1}{2}\lambda-m-n, -2m-1, -2n, \lambda, -j-m-n, -\lambda-m-n \\ -\frac{1}{2}-\frac{1}{2}\lambda-m-n, 1-\lambda-2n, -\lambda-2m, -2\lambda-2m-2n, -\lambda-j-m-n, -m-n \end{matrix};1\right],$$

where, e.g., j and n may assume complex values when $m=0,1,2,\cdots$.

For completeness it should be noted that, analogous to the use of (2.1) in the proof of (1.1), we can also apply Euler's transformation formula

$$(3.13)\qquad {}_2F_1\left[{a,b \atop c};z\right]=(1-z)^{c-a-b}{}_2F_1\left[{c-a,c-b \atop c};z\right]$$

to each of the ${}_2F_1$ series in (3.5), (3.6), (3.7), and the substitute each series along with its transformed series into (3.4) to derive, respectively, the transformation formulas

(3.14)

$$\begin{aligned}
&{}_4F_3\left[\begin{matrix} -j,\ \frac{1}{2}-\lambda-j,\ -m, m+2\lambda \\ \frac{1}{2}+\lambda+n-j,\ \frac{1}{2}-\lambda-n-j, \lambda+\frac{1}{2}\end{matrix};1\right]\\
&=\frac{(\frac{1}{2}+\lambda+n+j)_m(\lambda)_m(n+2\lambda)_m}{(\frac{1}{2}+\lambda+n-j)_m(2\lambda)_m(n+\lambda)_m}\\
&\cdot {}_7F_6\left[\begin{matrix} -\lambda-m-n, 1-\frac{1}{2}(\lambda+m+n),\ -m,\ -n,\lambda,\ \frac{1}{2}\left(\frac{1}{2}-\lambda+j-m-n\right),\ \frac{1}{2}\left(\frac{3}{2}-\lambda+j-m-n\right) \\ -\frac{1}{2}(\lambda+m+n),\ 1-\lambda-n, 1-\lambda-m, 1-2\lambda-m-n,\ \frac{1}{2}\left(\frac{3}{2}-\lambda-j-m-n\right),\ \frac{1}{2}\left(\frac{1}{2}-\lambda-j-m-n\right)\end{matrix};1\right],
\end{aligned}$$

(3.15)

$$\begin{aligned}
&{}_4F_3\left[\begin{matrix} -j, \lambda+n-j,\ -\frac{1}{2}m,\ \frac{1}{2}-\frac{1}{2}m \\ \lambda+\frac{1}{2}n-j,\ \frac{1}{2}+\lambda+\frac{1}{2}n-j, 1-\lambda-m\end{matrix};1\right]\\
&=\frac{(\lambda+n-j)_m(n+2\lambda)_m}{(2\lambda+n-2j)_m(n+\lambda)_m}\\
&\cdot {}_6F_5\left[\begin{matrix} -\lambda-m-n, 1-\frac{1}{2}(\lambda+m+n),\ -m,\ -n,\ -j,\lambda \\ -\frac{1}{2}(\lambda+m+n), 1-\lambda-n, 1-\lambda-m, 1-\lambda+j-m-n, 1-2\lambda-m-n\end{matrix};-1\right],
\end{aligned}$$

and

(3.16)

$$
{}_4F_3\left[\begin{matrix} -j,\ \frac{1}{2}-\lambda-j,\ -m,\ m+\lambda \\ \frac{1}{2}+n-j,\ \frac{1}{2}-\lambda-n-j,\ \lambda+\frac{1}{2} \end{matrix};1\right]
$$

$$
=\frac{\left(\frac{1}{2}+\lambda+n+j\right)_m\left(\frac{1}{2}+n\right)_m(\lambda)_{2m}(2n+2\lambda)_{2m}}{\left(\frac{1}{2}+n-j\right)_m\left(\frac{1}{2}+\lambda+n\right)_m(2\lambda)_{2m}(2n+\lambda)_{2m}}
$$

$$
\cdot\,{}_7F_6\left[\begin{matrix} -\lambda-2m-2n,\ 1-\frac{1}{2}\lambda-m-n,\ -2m,\ -2n,\ \lambda,\ \frac{1}{2}-\lambda-m-n,\ \frac{1}{2}+j-m-n \\ -\frac{1}{2}\lambda-m-n,\ 1-\lambda-2n,\ 1-\lambda-2m,\ 1-2\lambda-2m-2n,\ \frac{1}{2}-m-n,\ \frac{1}{2}-\lambda-j-m-n \end{matrix};1\right].
$$

Note that the ${}_4F_3$ series on the left of (3.12) and (3.14)–(3.16) are balanced.

The above very well-poised ${}_6F_5(-1)$ and ${}_7F_6(1)$ series can also be written as multiples of ${}_3F_2(1)$ and balanced ${}_4F_3(1)$ series, respectively, by applying the formulas [5, 4.4(2)] and [5, 4.4(5)]. In particular, this gives from (3.11), (3.12), (3.14), (3.15), and (3.16), respectively, that

(3.17)

$$
{}_4F_3\left[\begin{matrix} -j,\ \frac{1}{2}-\lambda-j,\ -m,\ m+\lambda \\ 1+n-j,\ 1-\lambda-n-j,\ \lambda+\frac{1}{2} \end{matrix};1\right]
$$

$$
=\frac{(\lambda+j+n)_m(n+1)_m}{(1+n-j)_m(n+\lambda)_m}\,{}_4F_3\left[\begin{matrix} -j,\ -2m,\ -2n,\ \lambda \\ 1-\lambda-j-m-n,\ -m-n,\ 2\lambda \end{matrix};1\right],
$$

(3.18)

$$
{}_4F_3\left[\begin{matrix} -j,\ \frac{1}{2}-\lambda-j,\ -m,\ m+\lambda+1 \\ 1+n-j,\ 1-\lambda-n-j,\ \lambda+\frac{1}{2} \end{matrix};1\right]
$$

$$
=\frac{(\lambda+j+n)_{m+1}(n+1)_m}{(1+n-j)_m(n+\lambda)_{m+1}}\,{}_4F_3\left[\begin{matrix} -j,\ -2m-1,\ -2n,\ \lambda \\ -\lambda-j-m-n,\ -m-n,\ 2\lambda \end{matrix};1\right],
$$

(3.19)

$$
{}_4F_3\left[\begin{matrix} -j,\ \frac{1}{2}-\lambda-j,\ -m,\ m+2\lambda \\ \frac{1}{2}+\lambda+n-j,\ \frac{1}{2}-\lambda-n-j,\ \lambda+\frac{1}{2} \end{matrix};1\right]
$$

$$
=\frac{\left(\frac{1}{2}+\lambda+n+j\right)_m}{\left(\frac{1}{2}+\lambda+n-j\right)_m}\,{}_4F_3\left[\begin{matrix} -j,\ -m,\ -n,\ \lambda \\ \frac{1}{2}\left(\frac{1}{2}-\lambda-j-m-n\right),\ \frac{1}{2}\left(\frac{3}{2}-\lambda-j-m-n\right),\ 2\lambda \end{matrix};1\right],
$$

(3.20)

$$
{}_4F_3\left[\begin{matrix} -j,\ \lambda+n-j,\ -\frac{1}{2}m,\ \frac{1}{2}-\frac{1}{2}m \\ \lambda+\frac{1}{2}n-j,\ \frac{1}{2}+\lambda+\frac{1}{2}n-j,\ 1-\lambda-m \end{matrix};1\right]
$$

$$
=\frac{(\lambda+n-j)_m(n+2\lambda)_m}{(2\lambda+n-2j)_m(\lambda)_j}\,{}_3F_2\left[\begin{matrix} -m,\ -n,\ 1-2\lambda+j-m-n \\ 1-2\lambda-m-n,\ 1-\lambda+j-m-n \end{matrix};1\right],
$$

(3.21)

$$
{}_4F_3\left[\begin{matrix} -j,\ \frac{1}{2}-\lambda-j,\ -m,\ m+\lambda \\ \frac{1}{2}+n-j,\ \frac{1}{2}-\lambda-n-j,\ \lambda+\frac{1}{2} \end{matrix};1\right]
$$

$$
=\frac{\left(\frac{1}{2}+\lambda+n+j\right)_m\left(\frac{1}{2}+n\right)_m}{\left(\frac{1}{2}+n-j\right)_m\left(\frac{1}{2}+\lambda+n\right)_m}\,{}_4F_3\left[\begin{matrix} -j,\ -2m,\ -2n,\ \lambda \\ \frac{1}{2}-m-n,\ \frac{1}{2}-\lambda-j-m-n,\ 2\lambda \end{matrix};1\right].
$$

Formulas (3.14)–(3.21) hold for arbitrary complex values of j and n when $m=0,1,2,\cdots$. Most of the above formulas can undoubtedly also be derived by using known transformation formulas, but in this section our main goal was only to point out transformation formulas that easily follow from (1.1).

Several summation formulas (mostly known) follow from the formulas in this section by considering cases in which one of the series reduces to only a few terms or is summable by known summation formulas. However, here we shall only point out that from the $j=\frac{1}{2}-\lambda$ cases of (3.9) and (3.14) it follows on setting $a=-\lambda-m-n$ and $d=1+a-\lambda$ that

(3.22)

$$
{}_7F_6\left[\begin{matrix} a,\ 1+\frac{1}{2}a,\ \frac{d}{2},\ \frac{1}{2}+\frac{1}{2}d,\ 1+a-d,\ 1+2a-d+m,\ -m \\ \frac{1}{2}a,\ 1+a-\frac{1}{2}d,\ \frac{1}{2}+a-\frac{1}{2}d,\ d,\ d-a-m,\ 1+a+m \end{matrix};1\right]
$$

$$
=\frac{(1+a)_m(2+2a-2d)_m}{(1+a-d)_m(1+2a-d)_m},
$$

which, like a similar formula in Bailey [5, p. 98, Ex. 8], is not a special case of Dougall's formula [5, 4.3(5)].

Acknowledgment. I wish to thank the referee for suggesting some improvements in the presentation of these results.

REFERENCES

[1] W. Al-Salam, W. R. Allaway and R. Askey, *Sieved ultraspherical polynomials*, Trans. Amer. Math. Soc., 234 (1984), pp. 39–55.

[2] R. Askey and M. Ismail, *A generalization of ultraspherical polynomials*, Studies in Pure Mathematics, P. Erdös, ed., Birkhäuser, Basel, 1983, pp. 55–78.

[3] R. Askey, T. Koornwinder and Mizan Rahman, *An integral of products of ultraspherical functions and a q-extension*, to appear.

[4] W. N. Bailey, *On the product of two Legendre polynomials*, Proc. Camb. Phil. Soc., 29 (1933), pp. 173–177.

[5] ———, *Generalized Hypergeometric Series*, Cambridge Univ. Press, Cambridge, 1935.

[6] D. M. Bressoud, *Linearization and related formulas for q-ultraspherical polynomials*, this Journal, 12 (1981), pp. 161–168.

[7] A. Erdélyi et al., *Higher Transcendental Functions, Vol.* 2, McGraw-Hill, New York, 1953.

[8] G. Gasper, *A convolution structure and positivity of a generalized translation operator for the continuous q-Jacobi polynomials*, Conference on Harmonic Analysis in Honor of Antoni Zygmund, Wadsworth International Group, Belmont, CA, 1983, pp. 44–59.

[9] G. Gasper and Mizan Rahman, *Positivity of the Poisson kernel for the continuous q-ultraspherical polynomials*, this Journal, 14 (1983), pp. 409–420.

[10] Mizan Rahman, *The linearization of the product of continuous q-Jacobi polynomials*, Canad. J. Math., 23 (1981), pp. 961–987.

[11] L. J. Rogers, *On the expansion of some infinite products*, Proc. London Math. Soc., 24 (1893), pp. 337–352.

[12] ———, *Second memoir on the expansion of certain infinite products*, Proc. London Math. Soc., 25 (1894), pp. 318–343.

[13] ———, *Third memoir on the expansion of certain infinite products*, Proc. London Math. Soc., 26 (1895), pp. 15–32.

[14] F. J. W. Whipple, *Some transformations of generalized hypergeometric series*, Proc. London Math. Soc., 26 (1927), pp. 257–272.

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

012

ASYMPTOTIC EXPANSION OF THE FIRST ELLIPTIC INTEGRAL*

B. C. CARLSON† AND JOHN L. GUSTAFSON‡

Abstract. Asymptotic formulas with error bounds are obtained for the first elliptic integral near its logarithmic singularity. It is convenient to start from the more general problem of expanding the integral over the positive t-axis of $[(t+x)(t+y)(t+z)(t+w)]^{-1/2}$. The method of Mellin transforms gives an asymptotic expansion that converges uniformly if $0<\max\{x,y\}/\min\{z,w\}\leqq r<1$. Each term of the series contains a Legendre polynomial and the derivative of a Legendre function with respect to its degree; this derivative involves nothing worse than a logarithm. A simple bound for the relative error of the Nth partial sum is obtained from Wong's formula for the remainder, aided by Chebyshev's integral inequality. Error bounds are given also for more accurate asymptotic formulas containing a complete elliptic integral. Formulas for the standard elliptic integral of the first kind are obtained in the case $w=\infty$. The method of Mellin transforms and Wong's formula are discussed in an appendix.

1. Introduction and summary.

Legendre's first elliptic integral,

$$F(\varphi,k)=\int_0^{\varphi}(1-k^2\sin^2\theta)^{-1/2}d\theta,$$

has a logarithmic singularity at $\varphi=\pi/2$, $k=1$. The asymptotic behavior of the complete case is well-known:

$$K(k)=F(\pi/2,k)\sim\log(4/k'),\qquad k'=(1-k^2)^{1/2},\quad k\to 1-,$$

where log denotes the natural logarithm. This is the case $\varphi=\pi/2$ of an asymptotic formula that is less widely known:

$$F(\varphi,k)\sim\log\frac{4}{\Delta+\cos\varphi},\qquad \Delta=(1-k^2\sin^2\varphi)^{1/2},\quad k\to 1-,\quad \varphi\to(\pi/2)-.$$

The latter formula is implicit in a series expansion given by Kaplan [10, (3)] and has been derived by two other methods in [11, (5.2)] and [3, (9.2–10)]. In the present paper we shall obtain error bounds:

$$(1.1)\qquad (1-\theta)K(k)=\log(4/k'),\qquad 0<\theta<\tfrac{1}{4}k'^2,\quad 0\leqq k<1,$$

$$(1.2)\qquad (1-\theta)F(\varphi,k)=(\sin\varphi)\log\frac{4}{\Delta+\cos\varphi},\qquad 0<\theta<\tfrac{1}{4}(\Delta^2+\cos^2\varphi),$$

where $0\leqq k\leqq 1$, $0\leqq\varphi\leqq\pi/2$, and $k\sin\varphi<1$. As $k\sin\varphi\to 1$ the upper bound for θ is asymptotically best possible.

*Received by the editors October 1, 1983, and in revised form March 19, 1984. This work was begun at the University of Maryland, where the first author was a visitor at the Institute for Physical Science and Technology, with the support of AROD contract DAAG 29-80-C-0032.

†Ames Laboratory and Department of Mathematics, Iowa State University, Ames, Iowa 50011. Ames Laboratory is operated for the U. S. Department of Energy by Iowa State University under contract no. W-7405-ENG-82. This work was supported by the Director of Energy Research, Office of Basic Energy Sciences.

‡Now at Floating Point Systems, Inc., P. O. Box 23489, Portland, Oregon 97223.

The asymptotic formula is the first term of a uniformly convergent series whose Nth partial sum approximates $F(\varphi,k)$ with a relative error bounded by $(\frac{1}{2})_N\Delta^{2N}/N!$, where $(a)_N=a(a+1)\cdots(a+N-1)$. In deriving this series by the method of Mellin transforms, we shall use the symmetric elliptic integral

$$R_F(x,y,z)=\frac{1}{2}\int_0^\infty [(t+x)(t+y)(t+z)]^{-1/2}dt, \tag{1.3}$$

which is homogeneous of degree $-\frac{1}{2}$,

$$R_F(\lambda x,\lambda y,\lambda z)=\lambda^{-1/2}R_F(x,y,z), \tag{1.4}$$

and is normalized so that

$$R_F(x,x,x)=x^{-1/2}. \tag{1.5}$$

For other properties of R_F see the references listed in [4]; for a Fortran program see [6]. The connection with Legendre's integral is

$$F(\varphi,k)=(\sin\varphi)R_F(\cos^2\varphi,\Delta^2,1), \qquad K(k)=R_F(0,1-k^2,1). \tag{1.6}$$

The logarithmic singularity of R_F occurs when two of its arguments are 0 or (by homogeneity) when one argument, say z, is infinite. Assuming $0\leqq x\leqq z$, $0\leqq y\leqq z$, and $x+y>0$, we shall derive the uniformly convergent expansion

$$\begin{aligned}R_F(x,y,z)=\frac{1}{2}z^{-1/2}\sum_{n=0}^{\infty}\Big[&-R_n\Big(\frac{1}{2},\frac{1}{2};x,y\Big)L_n\Big(\frac{1}{2},\frac{1}{2};z^{-1},0\Big)\\&-L_n\Big(\frac{1}{2},\frac{1}{2};x,y\Big)R_n\Big(\frac{1}{2},\frac{1}{2};z^{-1},0\Big)\Big],\end{aligned} \tag{1.7}$$

where R_n and L_n are expressible in terms of a Legendre function and its derivative with respect to the degree,

$$R_\nu\Big(\frac{1}{2},\frac{1}{2};x,y\Big)=(xy)^{\nu/2}P_\nu\Big(\frac{x+y}{2(xy)^{1/2}}\Big), \qquad L_\nu=\frac{\partial R_\nu}{\partial\nu}. \tag{1.8}$$

Both functions are symmetric in x and y. If n is a nonnegative integer, R_n is a homogeneous polynomial,

$$R_n\Big(\frac{1}{2},\frac{1}{2};x,y\Big)=\sum_{m=0}^{n}\frac{(\frac{1}{2})_m(\frac{1}{2})_{n-m}}{m!(n-m)!}x^my^{n-m}, \qquad R_n\Big(\frac{1}{2},\frac{1}{2};z^{-1},0\Big)=\frac{(\frac{1}{2})_n}{n!}z^{-n}. \tag{1.9}$$

The function L_ν, a Dirichlet average of $x^\nu\log x$, is discussed in [5], where it is shown that

$$-L_n\Big(\frac{1}{2},\frac{1}{2};z^{-1},0\Big)=\frac{(\frac{1}{2})_n}{n!}z^{-n}\Big[\log z+\psi(1+n)-\psi\Big(\frac{1}{2}+n\Big)\Big]. \tag{1.10}$$

Here ψ is the logarithmic derivative of the gamma function, satisfying

$$\begin{aligned}&\psi(1)-\psi\Big(\frac{1}{2}\Big)=2\log 2,\\&\psi(1+n)-\psi\Big(\frac{1}{2}+n\Big)=2\Big[\log 2-1+\frac{1}{2}-\cdots-\frac{1}{2n-1}+\frac{1}{2n}\Big], \qquad n\geqq 1.\end{aligned} \tag{1.11}$$

The quantity $L_n(\frac{1}{2},\frac{1}{2};x,y)$ is more complicated except for low values of n:

$$(1.12)\qquad \begin{aligned} L_0\left(\frac{1}{2},\frac{1}{2};x,y\right)&=2\log\left[\frac{1}{2}(x^{1/2}+y^{1/2})\right],\\ L_1\left(\frac{1}{2},\frac{1}{2};x,y\right)&=\frac{1}{2}(x+y)L_0\left(\frac{1}{2},\frac{1}{2};x,y\right)+\frac{1}{2}(x^{1/2}-y^{1/2})^2. \end{aligned}$$

By differentiating the recurrence relation for Legendre functions, a recurrence relation (2.24) for L_n can be obtained [5, (3.4)] which shows that

$$(1.13)\qquad L_n\left(\frac{1}{2},\frac{1}{2};x,y\right)=R_n\left(\frac{1}{2},\frac{1}{2};x,y\right)L_0\left(\frac{1}{2},\frac{1}{2};x,y\right)+\lambda_n\left(\frac{1}{2},\frac{1}{2};x,y\right),$$

where λ_n is a homogeneous polynomial of degree $2n$ in $x^{1/2}$ and $y^{1/2}$.

Furthermore we shall obtain error bounds for the truncated series,

$$(1.14)$$

$$\begin{aligned} (1-\theta_N)R_F(x,y,z)=\frac{1}{2}z^{-1/2}\sum_{n=0}^{N-1}\Bigg[&-R_n\left(\frac{1}{2},\frac{1}{2};x,y\right)L_n\left(\frac{1}{2},\frac{1}{2};z^{-1},0\right)\\ &-L_n\left(\frac{1}{2},\frac{1}{2};x,y\right)R_n\left(\frac{1}{2},\frac{1}{2};z^{-1},0\right)\Bigg], \end{aligned}$$

$$0<\theta_N<R_N\left(\frac{1}{2},\frac{1}{2};x,y\right)R_N\left(\frac{1}{2},\frac{1}{2};z^{-1},0\right)\leqq\frac{(\frac{1}{2})_N}{N!}\left(\frac{\max\{x,y\}}{z}\right)^N,$$

$$N\geqq 1,\quad 0\leqq x\leqq z,\quad 0\leqq y\leqq z,\quad x+y>0.$$

The case $N=1$ is

$$(1.15)\qquad (1-\theta_1)R_F(x,y,z)=z^{-1/2}\log\frac{4z^{1/2}}{x^{1/2}+y^{1/2}},\qquad 0<\theta_1<\frac{x+y}{4z},$$

which implies (1.2) by way of (1.6). The case $N=2$ is

$$(1.16)\quad (1-\theta_2)R_F(x,y,z)=z^{-1/2}\left[\left(1+\frac{x+y}{4z}\right)\log\frac{4z^{1/2}}{x^{1/2}+y^{1/2}}-\frac{x+y-x^{1/2}y^{1/2}}{4z}\right],$$

$$0<\theta_2<\frac{3}{8}\left(\frac{\max\{x,y\}}{z}\right)^2.$$

The upper bound given for θ_1 is the case $N=1$ of the upper bound given for θ_N as a product of R-polynomials; both of these bounds are asymptotically best possible as $(x+y)/z\to 0$. Positive lower bounds are given in (3.27), (3.42), and (3.44). Equations (1.15) and (1.16) agree with [7, pp. 19, 26], where a different bound of comparable accuracy is given for the absolute error $\theta_1 R_F$ but no bound is given for $\theta_2 R_F$.

The complete case of (1.14) is

(1.17)

$$(1-\theta_N)K(k)=\frac{1}{2}\sum_{n=0}^{N-1}\frac{(\frac{1}{2})_n}{n!}\left[-L_n\left(\frac{1}{2},\frac{1}{2};k'^2,0\right)-k'^{2n}L_n\left(\frac{1}{2},\frac{1}{2};1,0\right)\right]$$

$$=\sum_{n=0}^{N-1}\frac{(\frac{1}{2})_n(\frac{1}{2})_n}{n!n!}\left[\log\frac{1}{k'}+\psi(1+n)-\psi\left(\frac{1}{2}+n\right)\right]k'^{2n},$$

$$0<\theta_N<\frac{(\frac{1}{2})_N(\frac{1}{2})_N}{N!N!}k'^{2N},\quad 0\leqq k<1,\quad k'=(1-k^2)^{1/2}.$$

The terms of the series are well known [2,(900.06)], but the upper bound for θ_N is new and is asymptotically best possible as $k\to 1-$.

The occurrence of the pair $(z^{-1},0)$ in (1.7) suggests a more general expansion containing (z^{-1},w^{-1}). Indeed there is such an expansion of

$$(1.18)\qquad I(x,y,z,w)=(zw)^{1/2}\int_0^\infty[(t+x)(t+y)(t+z)(t+w)]^{-1/2}dt,$$

and its proof is more symmetrical than that for R_F. Hence we shall treat $I(x,y,z,w)$ first and consider later the case

$$(1.19)\qquad 2z^{1/2}R_F(x,y,z)=I(x,y,z,\infty).$$

The generalization of (1.14) is

(1.20)

$$(1-\theta_N)I(x,y,z,w)=\sum_{n=0}^{N-1}\left[-R_n\left(\frac{1}{2},\frac{1}{2};x,y\right)L_n\left(\frac{1}{2},\frac{1}{2};z^{-1},w^{-1}\right)\right.$$

$$\left.-L_n\left(\frac{1}{2},\frac{1}{2};x,y\right)R_n\left(\frac{1}{2},\frac{1}{2};z^{-1},w^{-1}\right)\right],$$

$$0<\theta_N<R_N\left(\frac{1}{2},\frac{1}{2};x,y\right)R_N\left(\frac{1}{2},\frac{1}{2};z^{-1},w^{-1}\right),\qquad N\geqq 1.$$

We define

$$(1.21)\qquad \rho=\max\{x,y\}/\min\{z,w\}.$$

As $N\to\infty$ the partial sum converges uniformly to I if $0<\rho\leqq r<1$, where r is any number in $(0,1)$. As $\rho\to 0$ the upper bound for θ_N is asymptotically best possible.

More accurate but less elementary approximations to I and R_F can be obtained by separating all logarithmic terms with the help of (1.13) and summing them to get a complete elliptic integral times a logarithm. The resulting expansion of I is given in (2.20) with error bounds in (3.37). First and second approximations for R_F are given in (3.46) and (3.47). The first approximation for Legendre's integral is

$$(1.22)\qquad F(\varphi,k)=\frac{2}{\pi}K(k')\log\frac{4}{\Delta+\cos\varphi}-\varepsilon,$$

$$\frac{\Delta^2\sin\varphi}{8}<\varepsilon<\frac{\Delta^2\log 2}{k^2\sin\varphi}.$$

In an appendix we shall derive the asymptotic expansion of

$$\int_0^\infty f(t)h(\lambda t)\,dt, \qquad \lambda\to+\infty,$$

where f and h are algebraically dominated. The proof differs in some respects from that of Wong [13]. Also, we shall give an alternative proof of Wong's formula for the remainder from the contour-integral representation due to Handelsman and Lew [8].

After this paper was submitted, we learned of a recent paper by Wong [15] containing a somewhat different form of the asymptotic expansion of $R_F(x,y,z)$ with bounds for the absolute error. Although $z^{1/2}R_F(x,y,z)$ and the terms of its asymptotic expansion depend only on the ratios x/z and y/z, the error bound obtained from [15, (4.32)] depends also on the value of z. If z is large compared to x and y but small compared to unity, the error bound may be excessively large.

2. The series with remainder. Let x, $y\in[0,\infty)$ and z, $w\in(0,\infty]$. It is assumed that x and y are not both 0, and z and w are not both infinite. If z or w is infinite, (1.18) is taken to mean

$$(2.1)\quad \begin{aligned} I(x,y,z,w)&=\int_0^\infty\left[(t+x)(t+y)(1+tz^{-1})(1+tw^{-1})\right]^{-1/2}dt\\ &=\int_0^\infty\left[(1+tx)(1+ty)(1+t^{-1}z^{-1})(1+t^{-1}w^{-1})\right]^{-1/2}t^{-1}dt,\end{aligned}$$

where t has been replaced by $1/t$ in the second integral. (The limit as $w\to\infty$ or $x\to0$ may be taken under the integral sign by the Lebesgue theorem of dominated convergence.) Thus I has the form

$$(2.2)\qquad I(x,y,z,w)=\int_0^\infty f(t)h(t)\,dt,$$

where

$$(2.3)\qquad f(t)=t^{-1}(1+tx)^{-1/2}(1+ty)^{-1/2}\sim\sum_{n=0}^\infty f_nt^{n-1},\qquad t\to0+,$$

$$(2.4)\qquad h(t)=(1+t^{-1}z^{-1})^{-1/2}(1+t^{-1}w^{-1})^{-1/2}\sim\sum_{n=0}^\infty h_nt^{-n},\qquad t\to+\infty.$$

Comparison with the generating function of R-polynomials [3, (6.6–1)] shows that

$$(2.5)\qquad f_n=(-1)^nR_n\left(\frac12,\frac12;x,y\right),\qquad h_n=(-1)^nR_n\left(\frac12,\frac12;z^{-1},w^{-1}\right),$$

where R_n is the homogeneous polynomial (1.9).

We assume for the moment that both x and y are positive. In the strip $-1<\operatorname{Re}s<0$ the Mellin transform of f is represented by an absolutely convergent integral,

$$(2.6)\qquad M[f;1-s]=\int_0^\infty t^{-s}f(t)\,dt=\int_0^\infty t^{-s-1}(1+tx)^{-1/2}(1+ty)^{-1/2}dt.$$

The integral can be evaluated in terms of an R-function by [3, Ex. 6.8–8]:

$$(2.7)\qquad M[f;1-s]=B(-s,1+s)R_s\left(\frac12,\frac12;x,y\right)=\frac{-\pi}{\sin\pi s}R_s\left(\frac12,\frac12;x,y\right),$$

where B is the beta function. Since R_s is entire in s if x and y are strictly positive [3, Cor. 6.3–4], the last member of (2.7) shows that $M[f;1-s]$ has a meromorphic continuation with simple poles at the integers. Its Laurent expansion about a nonnegative integer n is

$$M[f;1-s]=\frac{-f_n}{s-n}+F_n+O(s-n),\qquad s\to n, \tag{2.8}$$

where f_n is given by (2.5). The constant term is

$$F_n=(-1)^{n+1}L_n\left(\frac{1}{2},\frac{1}{2};x,y\right), \tag{2.9}$$

where the function L_ν, discussed in [5], is defined for any complex ν by

$$L_\nu\left(\frac{1}{2},\frac{1}{2};x,y\right)=\frac{\partial}{\partial\nu}R_\nu\left(\frac{1}{2},\frac{1}{2};x,y\right). \tag{2.10}$$

The function R_ν is expressible in terms of a Legendre function by (1.8) (see [3, (6.8–18)]), while L_ν involves both a Legendre function and its derivative with respect to the degree.

Next we verify that (2.7), and therefore (2.8) and (2.9), still hold if exactly one of x and y is 0 and $\operatorname{Re}s>-\frac{1}{2}$. Say $y=0$ for the sake of definiteness. The integral in (2.6) converges absolutely in the strip $-\frac{1}{2}<\operatorname{Re}s<0$ and is essentially a beta function,

$$M[f;1-s]=x^sB(-s,\tfrac{1}{2}+s). \tag{2.11}$$

It has a meromorphic continuation with simple poles at the nonnegative integers and the negative half-odd-integers. Gauss' theorem for a hypergeometric function with unit argument [3, (8.3–4)] yields

$$B(-s,1+s)R_s\left(\frac{1}{2},\frac{1}{2};x,0\right)=B\left(-s,\frac{1}{2}+s\right)R_s\left(\frac{1}{2};x\right)=x^sB\left(-s,\frac{1}{2}+s\right),\qquad \operatorname{Re}s>-\frac{1}{2},\quad s\neq 0,1,2,\cdots. \tag{2.12}$$

Thus (2.11) is the same as (2.7) with $y=0$ provided that $\operatorname{Re}s>-\frac{1}{2}$. (Note that R_s is no longer entire in s if one of its arguments vanishes, although (2.12) shows that it is still holomorphic for $\operatorname{Re}s>-\frac{1}{2}$.)

The Mellin transform of h is

$$\begin{aligned} M[h;s]&=\int_0^\infty t^{s-1}h(t)\,dt\\ &=\int_0^\infty t^{s-1}(1+t^{-1}z^{-1})^{-1/2}(1+t^{-1}w^{-1})^{-1/2}dt\\ &=\int_0^\infty t^{-s-1}(1+tz^{-1})^{-1/2}(1+tw^{-1})^{-1/2}dt, \end{aligned} \tag{2.13}$$

where t has been replaced by $1/t$ in the last integral. This is the same as (2.6) with (x,y) replaced by (z^{-1},w^{-1}). Even if z or w is infinite, we conclude immediately that the following equations hold for $\operatorname{Re} s > -\frac{1}{2}$:

$$\begin{aligned}(2.14)\qquad M[h;s] &= \frac{-\pi}{\sin \pi s} R_s\left(\frac{1}{2},\frac{1}{2};z^{-1},w^{-1}\right)\\ &= \frac{-h_n}{s-n} + H_n + O(s-n), \quad s \to n, \quad n = 0,1,2,\cdots,\\ H_n &= (-1)^{n+1} L_n\left(\frac{1}{2},\frac{1}{2};z^{-1},w^{-1}\right),\end{aligned}$$

where h_n is given by (2.5).

We are now ready to apply the expansion formula (A.18) in the Appendix. It would only spoil the symmetry to single out z or w as a large parameter at this stage, and so we put $\lambda = 1$. Comparison of (2.3) and (2.4) with (A.2) and (A.3) shows that the sets A and B are the set of nonnegative integers. Absolute convergence of the Mellin transforms (A.4) is assured by choosing c so that $-\frac{1}{2} < c < 0$. Using Wong's form (A.8) of the remainder, we may choose $\sigma = N$ where N is any positive integer (see the last sentence of the Appendix). Then we find

$$\begin{aligned}(2.15)\qquad I(x,y,z,w) &= \int_0^\infty f(t)h(t)\,dt\\ &= -\sum_{n=0}^{N-1} \operatorname*{Res}_{s=n}\{M[f;1-s]M[h;s]\} + r_N\\ &= -\sum_{n=0}^{N-1} \operatorname*{Res}_{s=n}\left\{\left(\frac{-f_n}{s-n} + F_n\right)\left(\frac{-h_n}{s-n} + H_n\right)\right\} + r_N\\ &= \sum_{n=0}^{N-1} (f_n H_n + F_n h_n) + r_N\\ &= \sum_{n=0}^{N-1}\left[-R_n\left(\frac{1}{2},\frac{1}{2};x,y\right)L_n\left(\frac{1}{2},\frac{1}{2};z^{-1},w^{-1}\right)\right.\\ &\qquad \left. -L_n\left(\frac{1}{2},\frac{1}{2};x,y\right)R_n\left(\frac{1}{2},\frac{1}{2};z^{-1},w^{-1}\right)\right] + r_N,\end{aligned}$$

where

$$(2.16)\qquad r_N = \int_0^\infty \varphi_N(t)\psi_N(t)\,dt, \qquad N \geqq 1.$$

Defining

$$(2.17)\qquad \rho = \max\{x,y\}/\min\{z,w\},$$

we shall show in the next section that $r_N \to 0$ as $N \to \infty$ if $0 < \rho < 1$.Then I is represented by a convergent infinite series that has an alternative form obtained by substituting (1.13):

$$\begin{aligned}(2.18)\qquad I(x,y,z,w) &= \left[-L_0\left(\frac{1}{2},\frac{1}{2};x,y\right) - L_0\left(\frac{1}{2},\frac{1}{2};z^{-1},w^{-1}\right)\right]\\ &\quad \cdot \sum_{n=0}^{\infty} R_n\left(\frac{1}{2},\frac{1}{2};x,y\right)R_n\left(\frac{1}{2},\frac{1}{2};z^{-1},w^{-1}\right) - \sum_{n=0}^{\infty} s_n,\end{aligned}$$

(2.19)

$$s_n = R_n\left(\frac{1}{2},\frac{1}{2};x,y\right)\lambda_n\left(\frac{1}{2},\frac{1}{2};z^{-1},w^{-1}\right)+\lambda_n\left(\frac{1}{2},\frac{1}{2};x,y\right)R_n\left(\frac{1}{2},\frac{1}{2};z^{-1},w^{-1}\right).$$

Note that $\lambda_0 = s_0 = 0$ and that s_n is a polynomial in the square roots of x, y, z^{-1}, w^{-1}. The series of products of R-polynomials in (2.18) can be summed by Meixner's formula [3, (6.11–3)]. Using also (1.12), we find

$$\text{(2.20)}\quad I(x,y,z,w)$$

$$= 2R_K\left[\left(1-\frac{x}{z}\right)\left(1-\frac{y}{w}\right),\left(1-\frac{x}{w}\right)\left(1-\frac{y}{z}\right)\right]\log\frac{4}{\left(x^{1/2}+y^{1/2}\right)\left(z^{-1/2}+w^{-1/2}\right)}$$

$$-\sum_{n=0}^{N-1} s_n - \delta_N,$$

$$\text{(2.21)}\qquad \delta_N = \sum_{n=N}^{\infty} s_n.$$

The complete elliptic integral R_K satisfies

$$\text{(2.22)}\quad R_K(p,q) = \frac{2}{\pi}R_F(0,p,q) = \frac{2}{\pi}q^{-1/2}K\left(\left(1-\frac{p}{q}\right)^{1/2}\right) = \frac{2}{\pi}p^{-1/2}K\left(\left(1-\frac{q}{p}\right)^{1/2}\right).$$

Its numerical value can be computed quickly by Gauss's algorithm [3, (6.10–8)],

$$\text{(2.23)}\qquad R_K(X^2,Y^2) = \frac{1}{M(X,Y)},$$

where $M(X, Y)$ is the arithmetic-geometric mean of X and Y.

In the next section we shall determine bounds for r_N/I and r_N in (2.15) and for δ_N in (2.20). Both expansions are more useful for finding approximations with error bounds than for numerical calculation to high accuracy because L_N and λ_n are computed from an inhomogeneous recurrence relation [5, (3.4)]:

$$\text{(2.24)}\quad (n+1)L_{n+1} - \left(n+\frac{1}{2}\right)(x+y)L_n + nxyL_{n-1} = -R_{n+1} + (x+y)R_n - xyR_{n-1}$$

$$= \frac{(x-y)^2}{n(n+1)}\frac{\partial^2}{\partial x\,\partial y}R_{n+1},$$

where $L_n = L_n(\frac{1}{2},\frac{1}{2};x,y)$ and $R_n = R_n(\frac{1}{2},\frac{1}{2};x,y)$. In the case $n=0$ one needs $R_{-1} = (xy)^{-1/2}$ since the last member of (2.24) is indeterminate. The inhomogeneous recurrence relation still holds if L is replaced by λ because R satisfies the homogeneous recurrence relation. Since $\lambda_0 = 0$ by (1.13), it follows from (2.24) that $\lambda_1 = \frac{1}{2}(x+y) - (xy)^{1/2}$ and hence by induction that λ_n is a homogeneous polynomial of degree $2n$ in $x^{1/2}$ and $y^{1/2}$ containing $(x^{1/2}-y^{1/2})^2$ as a factor.

3. Error bounds and convergence. The functions in the integrand of (2.16) are given by (A.9) and (A.10) for every positive integer N:

$$\text{(3.1)}\qquad \varphi_N(t) = f(t) - \sum_{n=0}^{N-1} f_n t^{n-1}, \qquad \psi_N(t) = h(t) - \sum_{n=0}^{N-1} h_n t^{-n}.$$

Equation (2.3) implies

$$(3.2)\qquad f(t)\sim t^{-1},\quad \varphi_N(t)\sim f_N t^{N-1},\quad t\to 0+.$$

Noting that $(-1)^N f_N>0$ by (2.5), we shall show, for every $t>0$ and $N\geqq 1$, that

$$(3.3)\qquad 0<(-1)^N\varphi_N(t)\leqq(-1)^N f_N t^N f(t).$$

That is, the ratio of $(-1)^N\varphi_N$ to f is majorized for all $t>0$ by the ratio of their asymptotic formulas (3.2). Analogous inequalities hold for ψ_N.

A proof starts from the integral representations [3, (6.6–6), (5.7–1)]

$$(3.4)\qquad tf(t)=(1+tx)^{-1/2}(1+ty)^{-1/2}=\int_0^1\frac{d\mu(u)}{1+t[ux+(1-u)y]},$$

$$(-1)^n f_n=R_n\left(\frac{1}{2},\frac{1}{2};x,y\right)=\int_0^1[ux+(1-u)y]^n d\mu(u),$$

$$d\mu(u)=\pi^{-1}u^{-1/2}(1-u)^{-1/2}du,\qquad \int_0^1 d\mu(u)=1.$$

It follows that

$$(3.5)\qquad t\varphi_N(t)=tf(t)-\sum_{n=0}^{N-1}f_n t^n=(-t)^N\int_0^1\frac{[ux+(1-u)y]^N}{1+t[ux+(1-u)y]}d\mu(u).$$

We may apply Chebyshev's inequality [9, Thm. 236] to the last integral because v^N is an increasing function of v while $1/(1+tv)$, $t>0$, is a decreasing function of v on the positive real line. The result is

$$0<(-1)^N t\varphi_N(t)\leqq t^N\int_0^1[ux+(1-u)y]^N d\mu(u)\cdot\int_0^1\frac{d\mu(u)}{1+t[ux+(1-u)y]},$$

with equality if and only if $x=y$. This proves that

$$0<(-1)^N t\varphi_N(t)\leqq t^N(-1)^N f_N tf(t),$$

which is equivalent to (3.3).

Similarly (2.4) implies

$$(3.6)\qquad h(t)\sim 1,\quad \psi_N(t)\sim h_N t^{-N},\quad t\to+\infty.$$

The same procedure that led to (3.5) now leads to

$$(3.7)\qquad \psi_N(t)=(-t)^{-N}\int_0^1\frac{[uz^{-1}+(1-u)w^{-1}]^N}{1+t^{-1}[uz^{-1}+(1-u)w^{-1}]}d\mu(u).$$

For every $t>0$ and $N\geqq 1$ Chebyshev's inequality yields

$$(3.8)\qquad 0<(-1)^N\psi_N(t)\leqq(-1)^N h_N t^{-N}h(t)$$

with equality if and only if $z=w$. That is, the ratio of $(-1)^N\psi_N(t)$ to $h(t)$ is majorized for all $t>0$ by the ratio of their asymptotic formulas (3.6).

Combining (3.3) and (3.8), we find

$$(3.9)\qquad 0<\varphi_N(t)\psi_N(t)\leqq f_N h_N f(t)h(t)$$

with equality if and only if $x=y$ and $z=w$. Integrating over all positive t, we obtain an inequality for the remainder (2.16),

$$0<r_N \leqq f_N h_N I(x,y,z,w). \tag{3.10}$$

We define the fractional remainder θ_N by

$$r_N=\theta_N I(x,y,z,w), \tag{3.11}$$

and so

$$0<\theta_N \leqq f_N h_N. \tag{3.12}$$

In the rest of this section we shall use frequently the abbreviations

$$R_n(x,y)=R_n\left(\frac{1}{2},\frac{1}{2};x,y\right), \qquad L_n(x,y)=L_n\left(\frac{1}{2},\frac{1}{2};x,y\right), \tag{3.13}$$

$$\xi=\max\{x,y\}, \quad \zeta=\min\{z,w\}, \quad \rho=\frac{\max\{x,y\}}{\min\{z,w\}}=\frac{\xi}{\zeta}. \tag{3.14}$$

Two inequalities for R_n [3, (6.2–24, 25)] are

$$R_n(x,y)\leqq \xi^n, \qquad R_n(x,y)\leqq \frac{\left(\frac{1}{2}\right)_n}{n!}(x+y)^n, \tag{3.15}$$

the first being an equality if and only if $x=y$ and the second if and only if $xy=0$ or $n=1$. The first inequality is the sharper one for large n if $xy\neq 0$. From (3.12) and (2.5) we find

$$0<\theta_N \leqq R_N(x,y)R_N(z^{-1},w^{-1})\leqq \rho^N \tag{3.16}$$

with equalities if and only if $x=y$ and $z=w$. If $0<\rho<1$ then $\theta_N\to 0$ and $r_N\to 0$ as $N\to\infty$, whence

(3.17)

$$I(x,y,z,w)=\sum_{n=0}^{\infty}\left[-R_n(x,y)L_n(z^{-1},w^{-1})-L_n(x,y)R_n(z^{-1},w^{-1})\right], \qquad 0<\rho<1.$$

Next we show that this series converges uniformly if $0<\rho\leqq r<1$ where r is any number in $(0,1)$. Since the remainder after N terms is $r_N=\theta_N I$, it is necessary to majorize I as well as θ_N. For this purpose we put $N=1$ in (2.15) and use (1.12) to obtain

$$\begin{gathered}(1-\theta_1)I(x,y,z,w)=2\log\frac{4}{(x^{1/2}+y^{1/2})(z^{-1/2}+w^{-1/2})},\\ 0<\theta_1\leqq (x+y)(z^{-1}+w^{-1})/4\leqq\rho,\end{gathered} \tag{3.18}$$

with equalities if and only if $x=y$ and $z=w$. Since $\xi^{1/2}\leqq x^{1/2}+y^{1/2}\leqq 2\xi^{1/2}$, with equality on the left if and only if $xy=0$, (3.18) implies

$$\log\frac{1}{\rho}<I<\frac{\log(16/\rho)}{1-\rho}. \tag{3.19}$$

The right-hand inequality is strict because it combines two inequalities with mutually exclusive conditions of equality. From (3.16) and (3.19) we get inequalities for $r_N=\theta_N I$:

$$0<r_N<\frac{\rho^N}{1-\rho}\log\frac{16}{\rho}. \tag{3.20}$$

Since $0\leqq\rho\log(1/\rho)\leqq 1/e$ if $0<\rho\leqq 1$, the condition $0<\rho\leqq r<1$ insures that

$$0<r_N<\frac{1}{1-r}\left(r^N\log 16+\frac{r^{N-1}}{e}\right).$$

The upper bound is independent of x,y,z,w and tends to 0 as $N\to\infty$, proving uniform convergence.

To derive a strictly positive lower bound for r_N, we replace the denominator of the integrand in (3.5) by $1+t\xi$, obtaining

$$(-1)^N\varphi_N(t)\geqq(-1)^N f_N t^{N-1}/(1+t\xi) \tag{3.21}$$

with equality if and only if $x=y$. Similarly (3.7) implies

$$(-1)^N\psi_N(t)\geqq(-1)^N h_N t^{-N}/(1+t^{-1}\zeta^{-1}) \tag{3.22}$$

with equality if and only if $z=w$. Wong's formula (2.16) gives

$$r_N=\int_0^\infty \varphi_N\psi_N\,dt\geqq f_N h_N\int_0^\infty\frac{dt}{(1+t\xi)(t+\zeta^{-1})}=f_N h_N\frac{\log(1/\rho)}{1-\rho} \tag{3.23}$$

with equality if and only if $x=y$ and $z=w$. A lower bound for $\theta_N=r_N/I$ then follows from (3.19):

$$\theta_N>f_N h_N\frac{\log(1/\rho)}{\log(16/\rho)}. \tag{3.24}$$

In summary we conclude from (3.16), (3.18), and (3.23) that the absolute error $r_N=\theta_N I$ in (2.15) satisfies

$$\begin{aligned}\frac{\log(1/\rho)}{1-\rho}&\leqq\frac{r_N}{R_N(x,y)R_N(z^{-1},w^{-1})}\\&\leqq\frac{2}{1-(x+y)(z^{-1}+w^{-1})/4}\log\frac{4}{(x^{1/2}+y^{1/2})(z^{-1/2}+w^{-1/2})}\\&<\frac{\log(16/\rho)}{1-\rho}\end{aligned} \tag{3.25}$$

with equalities if and only if $x=y$ and $z=w$. By (3.12) and (3.24) the relative error in

(3.26)

$$(1-\theta_N)I(x,y,z,w)=\sum_{n=0}^{N-1}\left[-R_n(x,y)L_n(z^{-1},w^{-1})-L_n(x,y)R_n(z^{-1},w^{-1})\right]$$

satisfies

$$\frac{\log(1/\rho)}{\log(16/\rho)} < \frac{\theta_N}{R_N(x,y)R_N(z^{-1},w^{-1})} \leqq 1 \tag{3.27}$$

with equality if and only if $x=y$ and $z=w$. It follows from these inequalities that as $\rho \to 0$,

$$I \sim \log\frac{1}{\rho}, \quad r_N \sim R_N(x,y)R_N(z^{-1},w^{-1})\log\frac{1}{\rho}, \quad \theta_N \sim R_N(x,y)R_N(z^{-1},w^{-1}). \tag{3.28}$$

Thus the upper bound for θ_N in (1.20) is asymptotically best possible. It is evident from (3.4) that

$$R_n(x,y) \geqq R_n(\xi,0) = \frac{(\frac{1}{2})_n}{n!}\xi^n \tag{3.29}$$

with equality if and only if $xy=0$. Combining this with (3.15), we find

$$\left(\frac{(\frac{1}{2})_N}{N!}\right)^2 \rho^N \leqq R_N(x,y)R_N(z^{-1},w^{-1}) \leqq \rho^N \tag{3.30}$$

with equality on the left if and only if $xy=0$ and $z^{-1}w^{-1}=0$ and equality on the right if and only if $x=y$ and $z=w$. Wallis' formula [3,(2.5–6)] implies

$$\frac{1}{\pi(N+\frac{1}{2})} < \left(\frac{(\frac{1}{2})_N}{N!}\right)^2 < \frac{1}{\pi N}. \tag{3.31}$$

The expansion (2.20) has a remainder δ_N that contains no logarithms and so is expected to be smaller than r_N if $\rho \ll 1$. In the absence of a representation of δ_N like Wong's formula for r_N, we must bound the individual terms s_n in (2.21) and carry out the summation. It is helpful to rewrite (2.19) in the form

$$s_n = R_n(x,y)R_n(z^{-1},w^{-1})\left[\frac{\lambda_n(x,y)}{R_n(x,y)} + \frac{\lambda_n(z^{-1},w^{-1})}{R_n(z^{-1},w^{-1})}\right]. \tag{3.32}$$

We find from [5, Thm. 7.1 and following Remark] that L_n/R_n increases with n and approaches the logarithm of the largest argument as $n \to \infty$. By (1.13) λ_n/R_n also increases with n, and

$$\frac{\lambda_1(x,y)}{R_1(x,y)} \leqq \frac{\lambda_n(x,y)}{R_n(x,y)} \leqq \log\xi - L_0(x,y) = 2\log\frac{2\xi^{1/2}}{x^{1/2}+y^{1/2}}, \qquad n \geqq 1, \tag{3.33}$$

with equalities for $n>1$ if and only if $x=y$. Since $\lambda_1(x,y) = \frac{1}{2}(x^{1/2}-y^{1/2})^2$ by (1.12), we have

$$\frac{\lambda_1(x,y)}{R_1(x,y)} + \frac{\lambda_1(z^{-1},w^{-1})}{R_1(z^{-1},w^{-1})} = 2\left(1 - \frac{x^{1/2}y^{1/2}}{x+y} - \frac{z^{1/2}w^{1/2}}{z+w}\right). \tag{3.34}$$

Also, from the integral representation (3.4) of R_n and Chebyshev's inequality [9, Thm. 236], we deduce that

$$R_{n-N}(x,y)R_N(x,y) \leqq R_n(x,y) \leqq \xi^{n-N}R_N(x,y), \qquad n \geqq N, \tag{3.35}$$

with equalities for $n > N$ if and only if $x=y$. Combined with (3.32), these inequalities show that

(3.36)

$$\left(1-\frac{x^{1/2}y^{1/2}}{x+y}-\frac{z^{1/2}w^{1/2}}{z+w}\right)R_{n-N}(x,y)R_{n-N}(z^{-1},w^{-1})$$

$$\leqq \frac{s_n}{2R_N(x,y)R_N(z^{-1},w^{-1})} \leqq \rho^{n-N}\log\frac{4\rho^{1/2}}{(x^{1/2}+y^{1/2})(z^{-1/2}+w^{-1/2})},$$

where $n \geqq N \geqq 1$ and the equalities hold for $n>1$ if and only if $x=y$ and $z=w$. Summation on n in accordance with (2.21) and use of Meixner's formula [3,(6.11-3)] yield

(3.37)

$$\left(1-\frac{x^{1/2}y^{1/2}}{x+y}-\frac{z^{1/2}w^{1/2}}{z+w}\right)R_K\left[\left(1-\frac{x}{z}\right)\left(1-\frac{y}{w}\right),\left(1-\frac{x}{w}\right)\left(1-\frac{y}{z}\right)\right]$$

$$\leqq \frac{\delta_N}{2R_N(xy)R_N(z^{-1},w^{-1})} \leqq \frac{1}{1-\rho}\log\frac{4\rho^{1/2}}{(x^{1/2}+y^{1/2})(z^{-1/2}+w^{-1/2})}$$

with equalities if and only if $x=y$ and $z=w$. This implies

$$0 \leqq \delta_N \leqq \frac{\rho^N \log 16}{1-\rho}, \tag{3.38}$$

which shows that, as $N \to \infty$, $\delta_N \to 0$ uniformly for $0<\rho \leqq r < 1$. As expected, comparison with (3.28) and (3.30) proves that $\delta_N/r_N \to 0$ as $\rho \to 0$. By (2.23) the R_K-function in (3.37) satisfies $1 < R_K \leqq (1-\rho)^{-1}$ since $M(X,Y)$ is a strictly increasing function of X and Y and since $M(X,X)=X$.

Finally we consider the case $w^{-1}=0$, in which

$$I(x,y,z,\infty)=2z^{1/2}R_F(x,y,z) \quad \text{and} \quad \rho=\max\{x,y\}/z, \tag{3.39}$$

according to (2.1) and (1.3). An interesting difference from the case of finite w is that the expansion (3.17) is valid if $\rho=1$ when $w=\infty$. (For example, if $x=y=z$ and $w=\infty$, both sides of (3.17) are 2; but if $x=y=z=w$, the left side is 1 and the right side is 0.) Since $z \neq w$, (3.16) becomes

$$0<\theta_N<R_N(x,y)R_N(z^{-1},0) \leqq \frac{(\frac{1}{2})_N}{N!}\rho^N. \tag{3.40}$$

Because $(\frac{1}{2})_N/N! \to 0$ as $N \to \infty$, it suffices to assume $0<\rho \leqq 1$ to insure convergence of the infinite series (1.7).

To show that the convergence is uniform for $0<\rho\leqq 1$, we put $w^{-1}=0$ in (3.25) to get

$$\frac{\log(1/\rho)}{1-\rho}<\frac{N!z^N r_N}{\left(\frac{1}{2}\right)_N R_N(x,y)}<\frac{2}{1-(x+y)/4z}\log\frac{4z^{1/2}}{x^{1/2}+y^{1/2}}<\frac{\log(16/\rho)}{1-\rho/2}, \tag{3.41}$$

where the first member is taken to be 1 if $\rho=1$. It follows that

$$0<r_N<\frac{\left(\frac{1}{2}\right)_N}{N!}\frac{\rho^N\log(16/\rho)}{1-\rho/2}.$$

Since $\rho^N\log(1/\rho)\leqq 1/Ne$ for $0<\rho\leqq 1$, we see that

$$0<r_N<\frac{\left(\frac{1}{2}\right)_N}{N!}2\left(\log 16+\frac{1}{Ne}\right).$$

The upper bound is independent of x,y,z and tends to 0 as $N\to\infty$, proving uniform convergence.

Equations (3.18) and (3.27) yield

$$\begin{gathered}(1-\theta_1)z^{1/2}R_F(x,y,z)=\log\frac{4z^{1/2}}{x^{1/2}+y^{1/2}},\\ \frac{\log(1/\rho)}{\log(16/\rho)}<\frac{4z\theta_1}{x+y}<1,\qquad \rho=\max\{x,y\}/z,\end{gathered} \tag{3.42}$$

which implies (even with a lower bound of 0 for θ_1)

$$\frac{1}{2}\log\frac{4}{\rho}<z^{1/2}R_F(x,y,z)<\frac{\log(16/\rho)}{2-\rho}. \tag{3.43}$$

Equations (2.15), (3.27), and (3.30) give a closer approximation:

$$\begin{gathered}(1-\theta_2)z^{1/2}R_F(x,y,z)=\left(1+\frac{x+y}{4z}\right)\log\frac{4z^{1/2}}{x^{1/2}+y^{1/2}}-\frac{1}{4z}\left(x+y-x^{1/2}y^{1/2}\right),\\ \frac{9\rho^2}{64}\frac{\log(1/\rho)}{\log(16/\rho)}<\theta_2<\frac{3\rho^2}{8}.\end{gathered} \tag{3.44}$$

Putting $w=\infty$ does not change the factor $1-\rho$ in the denominator in (3.37), and so our condition for uniform convergence of the series (2.21) remains $0<\rho\leqq r<1$. Since $z\neq w$ there are now strict inequalities in (3.37), of which a simplified but less precise version is

$$\left(\frac{\left(\frac{1}{2}\right)_N}{N!}\right)^2\rho^N<\delta_N<\frac{\left(\frac{1}{2}\right)_N}{N!}\frac{\rho^N\log 16}{1-\rho}. \tag{3.45}$$

Since $s_0=0$ and R_K is homogeneous of degree $-\frac{1}{2}$, the cases $N=1$ and $N=2$ of (2.20) yield

$$(3.46)\qquad R_F(x,y,z)=R_K(z-x,z-y)\log\frac{4z^{1/2}}{x^{1/2}+y^{1/2}}-\frac{1}{2}z^{-1/2}\delta_1,$$

$$\frac{\rho}{4}<\delta_1<\frac{\rho\log 4}{1-\rho},\qquad \rho=\max\{x,y\}/z,$$

$$(3.47)\qquad \delta_1=\frac{x+y-x^{1/2}y^{1/2}}{2z}+\delta_2,\qquad \frac{9\rho^2}{64}<\delta_2<\frac{3\rho^2\log 2}{2(1-\rho)}.$$

For example, these give respectively $0.3158<R_F(0.5,1,100)<0.3165$ and $0.3163943<R_F(0.5,1,100)<0.3163989$, the true value being $R_F(0.5,1,100)=0.31639714\ldots$.

Appendix. The method of Mellin transforms. The asymptotic expansion of

$$(\text{A}.1)\qquad \int_0^\infty f(t)h(\lambda t)\,dt,\qquad \lambda\to+\infty,$$

where f and h have asymptotic power series shown below in (A.2) and (A.3), has been obtained in terms of Mellin transforms by Handelsman and Lew [8], Soni and Soni [12], Bleistein and Handelsman [1], and Wong [13] [14] [15]. Wong represents the remainder by an integral over the real line [13, (2.14)] that is very useful for obtaining error bounds. Bleistein and Handelsman admit a larger class of functions and give the general term of the expansion in a convenient form in the logarithmic case [1, Ex. 4.16]. With assumptions adequate for the purposes of this paper, we present a proof similar in spirit to Wong's but different in some other respects. Also, we sketch the Handelsman–Lew–Bleistein proof and deduce our form of the expansion and Wong's form of the remainder from theirs (contrary to the expectation in the first paragraph of [13]). Part of what follows is not used in the body of the paper but may be helpful in clarifying the relation between two versions of the method of Mellin transforms.

Let f and h be real functions, locally integrable on $(0,\infty)$, with asymptotic expansions

$$(\text{A}.2)\qquad f(t)\sim\sum_{\alpha\in A}f_\alpha t^{\alpha-1},\qquad t\to 0+,$$

$$(\text{A}.3)\qquad h(t)\sim\sum_{\beta\in B}h_\beta t^{-\beta},\qquad t\to+\infty,$$

where f_α and h_β are constant coefficients and A and B are denumerable sets of real numbers with no finite cluster point. We assume there exists a real number c such that the Mellin transforms

$$(\text{A}.4)\qquad \int_0^\infty t^{-s}f(t)\,dt\quad\text{and}\quad\int_0^\infty t^{s-1}h(t)\,dt$$

converge absolutely if $\operatorname{Re}s=c$. (For weaker assumptions see [1, §4.5].) This implies that $c<\alpha$ for every $\alpha\in A$ and that the Mellin transform of f converges absolutely and is holomorphic in the vertical strip $c<\operatorname{Re}s<\alpha_1$, where α_1 is the least element of A. It implies further that the Mellin transform of f can be continued analytically (as proved in Lemma A.2 below) to the half-plane $\operatorname{Re}s>c$, where it is analytic except for a simple pole at every $\alpha\in A$ with residue $-f_\alpha$. We denote the analytically continued Mellin transform by $M[f;1-s]$ and define F_α to be the constant term in the Laurent expansion about α,

$$M[f;1-s]=\frac{-f_\alpha}{s-\alpha}+F_\alpha+O(s-\alpha), \qquad s\to\alpha\in A. \tag{A.5}$$

Likewise the second integral in (A.4) is holomorphic in the nonempty strip $c<\operatorname{Re}s<\beta_1$ (where β_1 is the least element of B), its analytic continuation $M[h;s]$ is meromorphic in the half-plane $\operatorname{Re}s>c$, and

$$M[h;s]=\frac{-h_\beta}{s-\beta}+H_\beta+O(s-\beta), \qquad s\to\beta\in B. \tag{A.6}$$

THEOREM A. *Let λ be positive and σ be real. The integral* (A.1), *if it converges, satisfies*

$$\int_0^\infty f(t)h(\lambda t)\,dt=\sum_{\alpha\in A\setminus B}^{\alpha<\sigma} f_\alpha M[h;\alpha]\lambda^{-\alpha}+\sum_{\beta\in B\setminus A}^{\beta<\sigma} h_\beta M[f;1-\beta]\lambda^{-\beta} \tag{A.7}$$

$$+\sum_{\gamma\in A\cap B}^{\gamma<\sigma}(f_\gamma h_\gamma\log\lambda+f_\gamma H_\gamma+F_\gamma h_\gamma)\lambda^{-\gamma}+r_\sigma,$$

$$r_\sigma=\int_0^\infty\left[f(t)-\sum_{\alpha\in A}^{\alpha<\sigma}f_\alpha t^{\alpha-1}\right]\left[h(\lambda t)-\sum_{\beta\in B}^{\beta<\sigma}h_\beta\lambda^{-\beta}t^{-\beta}\right]dt. \tag{A.8}$$

Before proving the theorem, we establish some properties of the factors in the integrand of the remainder r_σ. For any real σ we define

$$\varphi_\sigma(t)=f(t)-\sum_{\alpha\in A}^{\alpha<\sigma}f_\alpha t^{\alpha-1}, \tag{A.9}$$

$$\psi_\sigma(t)=h(t)-\sum_{\beta\in B}^{\beta<\sigma}h_\beta t^{-\beta}. \tag{A.10}$$

Note that $\varphi_\sigma=f$ if $\sigma\leqq\alpha_1$ and $\psi_\sigma=h$ if $\sigma\leqq\beta_1$.

LEMMA A.1. *Given $\sigma>\alpha_1$ let α' and α'' be the greatest and least elements, respectively, of A such that $\alpha'<\sigma\leqq\alpha''$. Then*

$$\int_0^\infty t^{-s}\varphi_\sigma(t)\,dt$$

converges absolutely in the strip $\alpha'<\operatorname{Re}s<\alpha''$. If $\sigma>\beta_1$ and if β' and β'' are the greatest and least elements of B such that $\beta'<\sigma\leqq\beta''$, then

$$\int_0^\infty t^{s-1}\psi_\sigma(t)\,dt$$

converges absolutely in the strip $\beta'<\operatorname{Re}s<\beta''$.

Proof. As $t\to0+$, $\varphi_\sigma(t)\sim f_{\alpha''}t^{\alpha''-1}$ by (A.2) and (A.9). From the assumption that

$$\int_0^\infty|t^{-s}f(t)|\,dt<\infty \quad \text{if } \operatorname{Re}s=c,$$

it follows that $f(t)=o(t^{c-1})$, $t\to+\infty$. Since $\alpha'\geqq c$ we see that $\varphi_\sigma(t)\sim -f_{\alpha'}t^{\alpha'-1}$ as $t\to+\infty$. Similarly, $\psi_\sigma(t)\sim h_{\beta''}t^{-\beta''}$ as $t\to+\infty$, and $\psi_\sigma(t)\sim h_{\beta'}t^{-\beta'}$ as $t\to 0+$. The asserted convergence properties follow at once.

A curious result, to be proved in Lemma A.2, is suggested by the following example. If n is a positive integer, consider

$$f(t)=(1+t)^{-1}\sim 1-t+t^2-\cdots,\qquad t\to 0+,$$

$$\varphi_{n+1}(t)=(1+t)^{-1}-1+t-\cdots-(-t)^{n-1}=(-t)^n(1+t)^{-1},$$

$$\int_0^\infty t^{-s}f(t)\,dt=\frac{\pi}{\sin\pi s},\qquad 0<\operatorname{Re}s<1,$$

$$\int_0^\infty t^{-s}\varphi_{n+1}(t)\,dt=\frac{(-1)^n\pi}{\sin\pi(s-n)}=\frac{\pi}{\sin\pi s},\qquad n<\operatorname{Re}s<n+1.$$

Thus f and φ_{n+1} have the same analytically continued Mellin transform,

$$M[f;1-s]=M[\varphi_{n+1};1-s]=\frac{\pi}{\sin\pi s},$$

although the defining integrals converge in disjoint strips and the functions differ by a polynomial in t. The explanation, in terms of the generalized Mellin transform discussed by Bleistein and Handelsman [1, p. 115], is that the generalized Mellin transform of a polynomial is 0 although the defining integral of the ordinary Mellin transform diverges for every s. Another interpretation is provided by regularization [14, p. 424]. Since we assume the integrals (A.4) converge absolutely if $\operatorname{Re}s=c$, ordinary Mellin transforms suffice in proving the following lemma (cf. [12, (4.2)] [14, p. 426] [15, p. 157]).

LEMMA A.2. *$M[f;1-s]$ is meromorphic in the half-plane $\operatorname{Re}s>c$, where its only singularities are a simple pole at every $\alpha\in A$ with residue $-f_\alpha$. Likewise, $M[h;s]$ is meromorphic for $\operatorname{Re}s>c$ and has a simple pole at every $\beta\in B$ with residue $-h_\beta$. If φ_σ and ψ_σ are defined by* (A.9) *and* (A.10), *then $M[\varphi_\sigma;\cdot]=M[f;\cdot]$ and $M[\psi_\sigma;\cdot]=M[h;\cdot]$ for every real σ.*

Proof. We shall prove only the statements relating to f since the proof for h is entirely similar. If σ is real, α_1 is the least element of A, and s is in the strip $c<\operatorname{Re}s<\alpha_1$, then

$$\text{(A.11)}\qquad \begin{aligned} M[f;1-s]&=\int_0^\infty t^{-s}f(t)\,dt\\ &=\int_0^1 t^{-s}\left[\varphi_\sigma(t)+\sum_{\alpha\in A}^{\alpha<\sigma}f_\alpha t^{\alpha-1}\right]dt+\int_1^\infty t^{-s}f(t)\,dt\\ &=\int_0^1 t^{-s}\varphi_\sigma(t)\,dt-\sum_{\alpha\in A}^{\alpha<\sigma}\frac{f_\alpha}{s-\alpha}+\int_1^\infty t^{-s}f(t)\,dt.\end{aligned}$$

The sum of powers can be integrated because $\alpha\in A$ implies $\alpha\geqq\alpha_1>\operatorname{Re}s$. Since $\varphi_\sigma(t)=O(t^{\sigma-1})$ as $t\to 0+$, the first term in the last line is holomorphic in the half-plane $\operatorname{Re}s<\sigma$. The second term is meromorphic with simple poles and residues exhibited, and the third term is holomorphic in the half-plane $\operatorname{Re}s>c$. The sum of the three terms provides the meromorphic continuation to the strip $c<\operatorname{Re}s<\sigma$. By taking σ arbitrarily large, the first part of the lemma is proved.

In the second part, since $\varphi_\sigma = f$ if $\sigma \leqq \alpha_1$, we may assume $\sigma > \alpha_1$. Hence $\sigma \in (\alpha', \alpha'']$ in the notation of Lemma A.1. If s is in the strip$\alpha' < \mathrm{Re}\, s < \alpha''$, then

$$
\begin{aligned}
M[\varphi_\sigma; 1-s] &= \int_0^\infty t^{-s}\varphi_\sigma(t)\,dt \\
&= \int_0^1 t^{-s}\varphi_\sigma(t)\,dt + \int_1^\infty t^{-s}\left[f(t) - \sum_{\alpha \in A}^{\alpha < \sigma} f_\alpha t^{\alpha-1}\right] dt \\
&= \int_0^1 t^{-s}\varphi_\sigma(t)\,dt + \int_1^\infty t^{-s} f(t)\,dt - \sum_{\alpha \in A}^{\alpha < \sigma} \frac{f_\alpha}{s-\alpha}.
\end{aligned} \tag{A.12}
$$

The sum of powers can be integrated because $\alpha \in A$ and $\alpha < \sigma$ imply $\alpha \leqq \alpha' < \mathrm{Re}\, s$. The last line is the same as the last line of (A.11).

Proof of Theorem A. We consider first the case where A and B are disjoint. Let the half-open interval $[\mu, \nu)$ contain exactly one point $\alpha \in A$ and no point in B. Then

$$
\begin{gathered}
\varphi_\mu(t) = \varphi_\nu(t) + f_\alpha t^{\alpha-1}, \qquad \psi_\mu(t) = \psi_\nu(t), \\
\varphi_\mu(t)\psi_\mu(\lambda t) - \varphi_\nu(t)\psi_\nu(\lambda t) = f_\alpha t^{\alpha-1}\psi_\nu(\lambda t).
\end{gathered}
$$

The integral of the right side over $(0, \infty)$ converges absolutely by Lemma A.1 since we may identify σ with ν and s with α to obtain $\beta' < \mu \leqq s < \sigma \leqq \beta''$. Hence

$$
\begin{aligned}
\int_0^\infty [\varphi_\mu(t)\psi_\mu(\lambda t) - \varphi_\nu(t)\psi_\nu(\lambda t)]\,dt &= f_\alpha \int_0^\infty t^{\alpha-1}\psi_\nu(\lambda t)\,dt = f_\alpha \lambda^{-\alpha}\int_0^\infty t^{\alpha-1}\psi_\nu(t)\,dt \\
&= f_\alpha \lambda^{-\alpha} M[\psi_\nu; \alpha] = f_\alpha M[h; \alpha]\lambda^{-\alpha}.
\end{aligned} \tag{A.13}
$$

We have used Lemma A.2 in the last step. Similarly, if $[\mu, \nu)$ contains exactly one point $\beta \in B$ and no point in A, we find

$$
\begin{aligned}
\int_0^\infty [\varphi_\mu(t)\psi_\mu(\lambda t) - \varphi_\nu(t)\psi_\nu(\lambda t)]\,dt &= h_\beta \int_0^\infty (\lambda t)^{-\beta}\varphi_\nu(t)\,dt \\
&= h_\beta \lambda^{-\beta} M[\varphi_\nu; 1-\beta] = h_\beta M[f; 1-\beta]\lambda^{-\beta}.
\end{aligned} \tag{A.14}
$$

Since Theorem A is trivial if $\sigma \leqq c$, we assume $\sigma > c$. If A and B are disjoint, we can partition $[c, \sigma)$ into a union of half-open intervals, each containing exactly one point of $A \cup B$. To each interval corresponds an integral of the form (A.13) or (A.14), and the sum of all the integrals telescopes to

$$
\begin{aligned}
&\int_0^\infty [\varphi_c(t)\psi_c(\lambda t) - \varphi_\sigma(t)\psi_\sigma(\lambda t)]\,dt \\
&\qquad = \sum_{\alpha \in A}^{\alpha < \sigma} f_\alpha M[h; \alpha]\lambda^{-\alpha} + \sum_{\beta \in B}^{\beta < \sigma} h_\beta M[f; 1-\beta]\lambda^{-\beta}.
\end{aligned} \tag{A.15}
$$

Since $\varphi_c = f$ and $\psi_c = h$, this proves the theorem when A and B are disjoint.

If A and B have points in common, we apply a small perturbation to $h(t)$ to make A and B disjoint. Any two points, say $\alpha \in A$ and $\beta \in B$, which coalesce when the perturbation is removed, make a joint contribution to the right side of (A.15) that becomes

$$\lim_{\beta \to \alpha} \{ f_\alpha M[h;\alpha]\lambda^{-\alpha} + h_\beta M[f;1-\beta]\lambda^{-\beta} \}$$

$$= \lim_{\beta \to \alpha} \left\{ f_\alpha \left[\frac{-h_\beta}{\alpha-\beta} + H_\beta + O(\alpha-\beta) \right] \lambda^{-\alpha} + h_\beta \left[\frac{-f_\alpha}{\beta-\alpha} + F_\alpha + O(\beta-\alpha) \right] \lambda^{-\beta} \right\}$$

$$= \lim_{\beta \to \alpha} \left\{ f_\alpha h_\beta \frac{\lambda^{-\alpha} - \lambda^{-\beta}}{\beta-\alpha} + f_\alpha H_\beta \lambda^{-\alpha} + F_\alpha h_\beta \lambda^{-\beta} \right\},$$

where we have used (A.5) and (A.6). Since $h_\beta \to h_\alpha$ and $H_\beta \to H_\alpha$ as the perturbation is removed, the above limit is

$$(f_\alpha h_\alpha \log \lambda + f_\alpha H_\alpha + F_\alpha h_\alpha)\lambda^{-\alpha}.$$

This completes the proof of Theorem A.

Alternative proof of Theorem A. Bleistein and Handelsman [1, §4.4 and Ex. 4.16] give a shorter but less elementary proof, adapted from [8], which uses Parseval's theorem for Mellin transforms, the residue theorem, and the asymptotic behavior of Mellin transforms along vertical lines in the complex plane. We shall sketch their proof without verifying the properties of Mellin transforms. Since the integrals (A.4) are assumed to converge absolutely if $\operatorname{Re} s = c$, Parseval's theorem yields

$$\text{(A.16)} \qquad \int_0^\infty f(t) h(\lambda t)\, dt = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \lambda^{-s} M[f;1-s] M[h;s]\, ds.$$

The factor λ^{-s} comes from

$$\text{(A.17)} \qquad \int_0^\infty t^{s-1} h(\lambda t)\, dt = \lambda^{-s} \int_0^\infty t^{s-1} h(t)\, dt = \lambda^{-s} M[h;s].$$

If $\sigma > c$ and $\sigma \notin A \cup B$, we may apply the residue theorem to a rectangle with vertices $c \pm iT$, $\sigma \pm iT$ because $M[f;1-s]$ and $M[h;s]$ are meromorphic in the half-plane $\operatorname{Re} s > c$ by Lemma A.2. The contributions from the top and bottom of the rectangle vanish as $T \to \infty$ because both Mellin transforms tend to zero [1, §4.3] as $s \to \infty$ along any vertical line in the half-plane $\operatorname{Re} s > c$. Hence the integrals along the two vertical sides of the rectangle differ in the limit by the sum of residues in the strip $c < \operatorname{Re} s < \sigma$:

$$\text{(A.18)} \qquad \int_0^\infty f(t) h(\lambda t)\, dt = - \sum_{s \in A \cup B}^{s < \sigma} \operatorname{Res}\{ \lambda^{-s} M[f;1-s] M[h;s] \} + r_\sigma,$$

$$\text{(A.19)} \qquad r_\sigma = \frac{1}{2\pi i} \int_{\sigma-i\infty}^{\sigma+i\infty} \lambda^{-s} M[f;1-s] M[h;s]\, ds.$$

To see that (A.18) and (A.7) agree, consider for example a point $\gamma \in A \cap B$. Then (A.5) and (A.6) show that

$$\lambda^{-s}M[f;1-s]M[h;s]=\lambda^{-\gamma}\left[1-(s-\gamma)\log\lambda+O\left((s-\gamma)^2\right)\right]$$

$$\cdot\left[\frac{-f_\gamma}{s-\gamma}+F_\gamma+O(s-\gamma)\right]\left[\frac{-h_\gamma}{s-\gamma}+H_\gamma+O(s-\gamma)\right]$$

$$=\lambda^{-\gamma}\left[\frac{f_\gamma h_\gamma}{(s-\gamma)^2}-\frac{f_\gamma h_\gamma\log\lambda+f_\gamma H_\gamma+F_\gamma h_\gamma}{s-\gamma}+O(1)\right],\qquad s\to\gamma.$$

The negative of the residue is $(f_\gamma h_\gamma \log\lambda + f_\gamma H_\gamma + F_\gamma h_\gamma)\lambda^{-\gamma}$, in agreement with (A.7). The residue at a point belonging to either A or B but not both is calculated similarly.

Equation (A.8) is more useful than (A.19) for obtaining error bounds. To see that the two expressions for the remainder are equal if σ is not in A or B, we use Lemma A.2 to rewrite (A.19) as

$$r_\sigma=\frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty}\lambda^{-s}M[\varphi_\sigma;1-s]M[\psi_\sigma;s]\,ds.$$

By Lemma A.1 the defining integrals of $M[\varphi_\sigma;1-s]$ and $M[\psi_\sigma;s]$ converge absolutely on the path of integration, where Re $s=\sigma$. Thus Parseval's theorem yields

$$r_\sigma=\int_0^\infty \varphi_\sigma(t)\psi_\sigma(\lambda t)\,dt$$

in agreement with (A.8). Note that r_σ, defined as a function of σ by (A.8), is continuous from the left at points of $A\cup B$, while the integral in (A.19) is not well defined at such points and is infinite as a point $\sigma\in A\cap B$, where the integrand has a double pole.

Acknowledgments. This work was stimulated by conversations with Roderick Wong, whose continued interest has been encouraging. We thank Harry McMaken for a reference and very helpful remarks.

REFERENCES

[1] N. Bleistein and R. A. Handelsman, *Asymptotic Expansion of Integrals*, Holt, Rinehart, and Winston, New York, 1975.

[2] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Scientists*, 2nd ed., Springer-Verlag, New York, 1971.

[3] B. C. Carlson, *Special Functions of Applied Mathematics*, Academic Press, New York, 1977.

[4] ______, *Short proofs of three theorems on elliptic integrals*, this Journal, 9 (1978), pp. 524–528.

[5] ______, *Dirichlet averages of $x^t \log x$*, in preparation.

[6] B. C. Carlson and Elaine M. Notis, *Algorithms for incomplete elliptic integrals*, ACM Trans. Math. Software, 7 (1981), pp. 398–403.

[7] J. L. Gustafson, Ph. D. thesis, Iowa State Univ., Ames, 1982.

[8] R. A. Handelsman and J. S. Lew, *Asymptotic expansion of a class of integral transforms with algebraically dominated kernels*, J. Math. Anal. Applic., 35 (1971), pp. 405–433.

[9] G. H. Hardy, J. E. Littlewood and G. Pólya, *Inequalities*, 2nd ed., Cambridge Univ. Press, Cambridge, 1959.

[10] E. L. Kaplan, *Auxiliary table for the incomplete elliptic integrals*, J. Math. and Phys., 27 (1948), pp. 11–36.

[11] W. J. NELLIS AND B. C. CARLSON, *Reduction and evaluation of elliptic integrals*, Math. Comp., 20 (1966), pp. 223–231.

[12] K. SONI AND R. P. SONI, *Slowly varying functions and asymptotic behavior of a class of integral transforms* I, J. Math. Anal. Applic., 49 (1975), pp. 166–179.

[13] R. WONG, *Explicit error terms for asymptotic expansions of Mellin convolutions*, J. Math. Anal. Applic., 72 (1979), pp. 740–756.

[14] ———, *Error bounds for asymptotic expansions of integrals*, SIAM Rev., 22 (1980), pp. 401–435.

[15] ———, *Applications of some recent results in asymptotic expansions*, Congressus Numerantium, 37 (1983), pp. 145–182.

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

013

ON SIEVED ORTHOGONAL POLYNOMIALS I: SYMMETRIC POLLACZEK ANALOGUES*

MOURAD E. H. ISMAIL[†]

Abstract. Two sieved analogues of the Pollaczek polynomials are introduced and the weight functions for the new polynomials are computed. Various asymptotic and explicit formulas are derived. Generating functions are also included.

AMS-MOS subject classifications (1980). Primary 33A65, 34B25; secondary 42C05, 34B25

Key words. orthogonal polynomials, continued fractions, Jacobi matrices, weight function, Pollaczek polynomials

1. Introduction. The ultraspherical polynomials are generated by the recurrence relation

$$(1.1)\quad 2(n+\lambda)xC_n^\lambda(x)=(n+1)C_{n+1}^\lambda(x)+(n+2\lambda-1)C_{n-1}^\lambda(x),\qquad n>0,$$

and the initial conditions

$$(1.2)\qquad C_0^\lambda(x)=1,\qquad C_1^\lambda(x)=2\lambda x.$$

Rogers [17], [18] studied the polynomials $\{C_n(x;\beta|q)\}$ defined by

$$(1.3)\quad \begin{aligned} &2x(1-\beta q^n)C_n(x;\beta|q)\\ &\quad=(1-q^{n+1})C_{n+1}(x;\beta|q)+(1-\beta^2q^{n-1})C_{n-1}(x;\beta|q),\qquad n>0,\end{aligned}$$

$$(1.4)\qquad C_0(x;\beta|q)=1,\ C_1(x;\beta|q)=2x(1-\beta)/(1-q).$$

He solved the connection coefficient and the linearization coefficient problems for these polynomials and used them to prove the well-known Rogers–Ramanujan identities of the theory of partitions; see e.g. Andrews [5]. The Rogers polynomials are called the continuous q-ultraspherical polynomials. They generalize the ultraspherical polynomials in the sense

$$(1.5)\qquad \lim_{q\to1}C_n(x;q^\lambda|q)=C_n^\lambda(x).$$

The weight function for these polynomials was computed very recently by Askey and Wilson [9] and Askey and Ismail [6], [7]. Pollaczek and Szegö generalized the ultraspherical polynomials in a different direction. Szegö [20] introduced the polynomials $\{P_n^\lambda(x;a,b)\}$ defined via

$$(1.6)\quad \begin{aligned} &2[x(n+\lambda+a)+b]P_n^\lambda(x;a,b)\\ &\quad=(n+1)P_{n+1}^\lambda(x;a,b)+(n+2\lambda-1)P_{n-1}^\lambda(x;a,b),\qquad n>0,\end{aligned}$$

* Received by the editors December 22, 1983. This work was partially supported by the National Science Foundation under grant MCS 8313931.

[†] Department of Mathematics, Arizona State University, Tempe, Arizona 85287.

with

$$P_0^\lambda(x;a,b)=1, \qquad P_1^\lambda(x;a,b)=2[b+x(\lambda+a)]. \tag{1.7}$$

Pollaczek [14] studied the case $\lambda=\frac{1}{2}$ earlier. The polynomials $\{P_n^\lambda(x;a,b)\}$ are now called the Pollaczek polynomials. Pollaczek's memoir [15] contains a good survey of his methods.

Al-Salam, Allaway and Askey [3] set

$$q=s\omega_k, \quad \beta=s^{\lambda k}, \quad \omega_k:=\exp\left(\frac{2\pi i}{k}\right), \tag{1.8}$$

$$C_n(x;\beta|q)=(\beta^2;q)_n c_n(x;\beta|q)/(q;q)_n, \tag{1.9}$$

and noted that

$$\lim_{q\to 1} c_n(x;\beta|q)=c_n^\lambda(x;k),$$

exists and that the $c_n^\lambda(x;k)$'s satisfy

$$c_0^\lambda(x;k)=1, \qquad c_1^\lambda(x;k)=x, \tag{1.10}$$

$$\begin{aligned}&2xc_n^\lambda(x;k)=c_{n+1}^\lambda(x;k)+c_{n-1}^\lambda(x;k), \qquad n\neq mk,\\ &2x(m+\lambda)c_{mk}^\lambda(x;k)=(m+2\lambda)c_{mk+1}^\lambda(x;k)+mc_{mk-1}^\lambda(x;k), \qquad m>0.\end{aligned} \tag{1.11}$$

Another set of polynomials similarly results from letting

$$q=s\omega_k, \qquad \beta=s^{\lambda k+1}\omega_k, \tag{1.12}$$

and

$$B_n^\lambda(x;k)=\lim_{s\to 1} C_n\left(x;s^{\lambda k+1}\omega_k|s\omega_k\right). \tag{1.13}$$

Al-Salam, Allaway and Askey refer to both polynomials as sieved ultraspherical polynomials.

In this work we carry this programme one step further. We derive a sieved analogue of the symmetric Pollaczek polynomials ($b=0$ in (1.6) and (1.7)). The starting point is to discover a three-parameter generalization of the two-parameter (β and q) family of polynomials $\{C_n(x;\beta|q)\}$. The second step is to consider the limiting case when q lies on the unit circle and choose the remaining parameters appropriately. The third and most important step is to find the measure (distribution function) that these polynomials are orthogonal with respect to. The appropriate symmetric three-parameter family that generalizes the continuous q-ultraspherical polynomials is

$$F_0(x;\alpha,\beta|q)=1, \qquad F_1(x;\alpha,\beta|q)=2x(1-\alpha)/(1-q), \tag{1.14}$$

$$\begin{aligned}&2x(1-\alpha q^n)F_n(x;\alpha,\beta|q)\\ &\quad=(1-q^{n+1})F_{n+1}(x;\alpha,\beta|q)+(1-\beta q^{n-1})F_{n-1}(x;\alpha,\beta|q), \qquad n>0.\end{aligned} \tag{1.15}$$

These polynomials have been studied by Askey and Ismail [8]. They used a different normalization and their polynomials are random walk polynomials of a birth and death process (see [8] for details). We now observe that

$$\lim_{s\to 1} F_n\left(x;\omega_k s^{k\lambda+ka+1};\omega_k^2 s^{2\lambda k+2}|s\omega_k\right)$$

defines a new set of polynomials. Let $B_n^\lambda(x;a;k)$ denote the above limit. It is easy to see that the recurrence relation (1.15) becomes

(1.16)

$$\begin{aligned} &2xB_n^\lambda(x;a,k)=B_{n+1}^\lambda(x;a,k)+B_{n-1}^\lambda(x;a,k), \qquad n+1\neq km,\\ &2x(\lambda+a+m)B_{mk-1}^\lambda(x;a;k)=mB_{mk}^\lambda(x;a;k)+(2\lambda+m)B_{mk-2}^\lambda(x;a;k), \qquad m>0, \end{aligned}$$

while the initial conditions (1.14) become

$$B_0^\lambda(x;a;k)=1, \qquad B_1^\lambda(x;a;k)=2x. \tag{1.17}$$

We are assuming $k>1$ because the case $k=1$ of (1.16) is the symmetric Pollaczek polynomials. Similarly

$$\lim_{s\to 1}(s\omega_k;s\omega_k)_n F_n(x;s^{k\lambda+ka},s^{2k\lambda};\omega_k s)/(s^{2k\lambda};\omega_k s)_n \tag{1.18}$$

exists, where

$$(\sigma;p)_0:=1, (\sigma;p)_n=(1-\sigma)(1-\sigma p)\cdots(1-\sigma p^{n-1}), \qquad n>0. \tag{1.19}$$

Denote the polynomials in (1.18) by $c_n^\lambda(x;a,k)$. They satisfy

$$c_0^\lambda(x;a;k)=1, \qquad c_1^\lambda(x;a;k)=x(\lambda+a)/\lambda, \tag{1.20}$$

and

$$\begin{aligned} &2xc_n^\lambda(x;a;nk)=c_{n+1}^\lambda(x;a;k)+c_{n-1}^\lambda(x;a;k), \qquad n\neq km,\\ &2x(\lambda+a+m)c_{mk}^\lambda(x)=(2\lambda+m)c_{mk+1}^\lambda(x;a;k)+mc_{mk-1}^\lambda(x;a;k). \end{aligned} \tag{1.21}$$

The relationships (1.20) and (1.21) can be proved as follows. Set

$$F_n(x;\alpha;\beta|q)=(\beta;q)_n\eta_n(x)/(q;q)_n. \tag{1.22}$$

The η_n's also depend on α and β. The substitution of the F_n's as in (1.22) in the relationships (1.14) and (1.15) implies

$$\eta_0(x)=1, \qquad \eta_1(x)=2x(1-\alpha)/(1-\beta) \tag{1.23}$$

and

$$2x(1-\alpha q^n)\eta_n(x)=(1-\beta q^n)\eta_{n+1}+(1-q^n)\eta_{n-1}(x), \qquad n>0. \tag{1.24}$$

Now, setting $\alpha=s^{k\lambda+ka}$, $\beta=s^{2\lambda k}$, $q=\omega_k s$ and letting $s\to 1$ in (1.22) and (1.23) establish (1.20) and (1.21). Observe that when $a=0$ the $B_n^\lambda(x;a,k)$'s and $c_n^\lambda(x;a,k)$'s reduce to the $B_n^\lambda(x;k)$'s and $c_n^\lambda(x;a)$'s of Al-Salam, Allaway and Askey [3]. Al-Salam, Allaway and Askey called $\{c_n(x;k)\}$ and $\{B_n^\lambda(x;k)\}$ the sieved ultraspherical polynomials of the first and second kinds respectively. We shall follow this terminology and call $\{c_n^\lambda(x;a,k)\}$ the sieved polynomials of the first kind and call $\{B_n^\lambda(x;a,k)\}$ the sieved polynomials of the second kind. We need to caution the reader about a possible confusion this terminology might cause. The polynomials of the second kind are not the numerator polynomials in the corresponding continued fraction, Chihara [10] and Askey and Ismail [8].

In §2 we derive generating functions for the sieved polynomials of the first and second kinds. These are the limiting cases of the generating functions for the F_n's. We also obtain explicit representations for the sieved polynomials $c_n^\lambda(x;a,k)$ and $B_n^\lambda(x;a,k)$ as trigonometric polynomials. In §3 we use a heuristic argument to discover very good candidates for the distribution functions (positive measures) that $\{c_n^\lambda(x;a,k)\}$ and $\{B_n^\lambda(x;a,k)\}$ are orthogonal with respect to. Rigorous proofs are included in §5. In §4 we investigate the numerator polynomials associated with the c_n^λ's and the B_n^λ's. The asymptotic behavior of the numerator polynomials is determined and the associated continued fraction is computed. In §5 we use Markov's theorem and some of the asymptotic results of §§3 and 4 to compute the distribution function. It turns out that the discrete spectrum is empty, that is the distribution function is absolutely continuous, when $a \geq 0$ and $\lambda > -\frac{1}{2}$. The discrete spectrum is countably infinite when $a<0$, $\lambda > -\frac{1}{2}$ and $\lambda + a + 1 > 0$. Note that this gives a new proof of the orthogonality relations for the sieved ultraspherical polynomials in Al-Salam, Allaway and Askey [3]. Our proof uses Markov's theorem and asymptotic analysis, the techniques used successfully by Pollaczek [15] and later by Askey and Ismail [8]. Our results are equivalent to computing the spectral measure of a self-adjoint bounded Jacobi matrix. We refer the interested reader to Akhiezer and Glazman [1] and Akhiezer [2] for the connection between spectral properties of Jacobi matrices and orthogonal polynomials.

2. Generating functions. For completeness we include a derivation of a generating function for the F_n's. Set

$$F(x,t) := \sum_0^\infty t^n F_n(x;\alpha,\beta\,|\,q). \tag{2.1}$$

Multiply (1.15) by t^{n+1} and add the resulting equations for $n=1,2,\cdots$, to obtain the functional equation

$$F(x,t) = \frac{1-2\alpha xt+\beta t^2}{1-2xt+t^2} F(x,qt). \tag{2.2}$$

We also used (1.14). We now iterate the functional equation (2.2), that is successively replace t by qt and obtain

$$F(\cos\theta,t) = \frac{(t/\gamma;q)_n (t/\Delta;q)_n}{(te^{-i\theta};q)_n (te^{i\theta};q)_n} F(\cos\theta, tq^n), \tag{2.3}$$

where

$$1-2\alpha xt+\beta t^2 = (1-t/\gamma)(1-t/\Delta). \tag{2.4}$$

Letting $n \to \infty$ in (2.3) gives

$$F(\cos\theta,t) = \frac{(t/\gamma;q)_\infty (t/\Delta;q)_\infty}{(te^{i\theta};q)_\infty (te^{-i\theta};q)_\infty}, \tag{2.5}$$

where $(\sigma;q)_\infty$ is $\Pi_0^\infty(1-\sigma q^n)$.

We now show how to obtain generating functions for our sieved polynomials from (2.5). Recall the q-binomial theorem, Slater [19, p. 92], Andrews [5, Thm. 2.1]

$$(\sigma z;q)_\infty/(z;q)_\infty = \sum_0^\infty (\sigma;q)_n z^n/(q;q)_n. \tag{2.6}$$

We now consider the B_n^λ's of (1.16) and (1.17), so choose

$$\alpha=\omega_k s^{k\lambda+ka+1},\quad \beta=\omega_k^2 s^{2k\lambda+2},\quad q=\omega_k s. \tag{2.7}$$

We choose γ and Δ in (2.4) as

$$\gamma=\left(x\alpha+\sqrt{x^2\alpha^2-\beta}\right)/\beta,\quad \Delta=\left(x\alpha-\sqrt{x^2\alpha^2-\beta}\right)/\beta,\quad x=\cos\theta, \tag{2.8}$$

then use (2.6) to obtain

$$(t/\gamma;q)_\infty/(te^{-i\theta};q)_\infty=\sum_0^\infty(e^{i\theta}/\gamma;q)_n(te^{-i\theta})^n/(q;q)_n. \tag{2.9}$$

When α, β and q are as in (2.7), a typical term in the quotient of finite products $(e^{i\theta}/\gamma;q)_n/(q;q)_n$ is $[1-q^je^{i\theta}/\gamma]/[1-q^{j+1}]$. It is easy to see that (2.7) and (2.8) give

$$\frac{1}{\gamma}=\omega_k s^{\lambda k+1}\left[s^{ka}\cos\theta-i\sqrt{1-s^{2ak}\cos^2\theta}\,\right]. \tag{2.10}$$

Therefore

$$\lim_{s\to1}\frac{1-q^je^{i\theta}/\gamma}{1-q^{j+1}}=1\quad \text{if } k\nmid j+1.$$

If $j+1=mk$ the above limit is $[\lambda+m+ia\cot\theta]/m$. This proves

$$\lim_{s\to1}\frac{(e^{i\theta}/\gamma;q)_n}{(q;q)_n}=\frac{(1+\lambda+ia\cot\theta)_m}{m!},\qquad n+1=mk+l,\quad 0\le l<k,$$

under the assumption (2.7), where

$$(\sigma)_0=1,\qquad (\sigma)_n=\sigma(\sigma+1)\cdots(\sigma+n-1),\quad n>0.$$

This calculation and (2.9) imply

$$\begin{aligned}\lim_{s\to1}\frac{(t/\gamma;q)_\infty}{(te^{-i\theta};q)_\infty}&=\lim_{s\to1}\left\{\sum_{\substack{0\le l<k\\ m\ge0}}\frac{(e^{i\theta}/\gamma;q)_{mk+l}}{(q;q)_{mk+l}}(te^{-i\theta})^{mk+l}\right\}\\&=\sum_{m=0}^\infty\frac{(\lambda+1+ia\cot\theta)_m}{m!}(te^{-i\theta})^{mk}\sum_{l=0}^{k-1}(te^{-i\theta})^l\\&=\sum_{m=0}^\infty\frac{(\lambda+1+ia\cot\theta)_m}{m!}(te^{-i\theta})^{mk}\left[\frac{1-t^ke^{-ik\theta}}{1-te^{-i\theta}}\right]\\&=(1-te^{-i\theta})^{-1}(1-t^ke^{-ik\theta})^{-\lambda-ia\cot\theta},\end{aligned}$$

in view of the binomial theorem. This and (2.5) establish the generating function

$$\begin{aligned}&\sum_0^\infty B_n^\lambda(\cos\theta;a;k)t^n \\&\quad=(1-2t\cos\theta+t^2)^{-1}(1-t^ke^{-ik\theta})^{-\lambda-ia\cot\theta}(1-t^ke^{ik\theta})^{-\lambda+ia\cot\theta}\end{aligned} \tag{2.11}$$

The generating function (2.11) enables us to obtain explicit representations for our B_n^λ's. It also yields another formula relating the B_n^λ's to the Pollaczek polynomials.

Recall the generating function

$$\frac{1}{1-2t\cos\theta+t^2}=\sum_{n=0}^{\infty}U_n(\cos\theta)t^n, \tag{2.12}$$

Rainville [16, p. 301], and

$$\sum_0^{\infty}P_n^{\lambda}(\cos\theta;a,1)t^n=\left(1-te^{-i\theta}\right)^{-\lambda-ia\cot\theta}\left(1-te^{i\theta}\right)^{-\lambda+ia\cot\theta}, \tag{2.13}$$

where the P_n^{λ}'s are the Pollaczek polynomials as defined by (1.6) and (1.7) see Chihara [10, p. 184]. We now expand the first factor on the right side of (2.11) in powers of t as in (2.12), then expand the rest in powers of t^k using (2.13). This establishes, upon equating the coefficients of various powers of t, the identity

$$B_n^{\lambda}(\cos\theta;0;k)=\sum_{j=0}^{\lfloor n/k\rfloor}P_j^{\lambda}(\cos k\theta;0,1)U_{n-kj}(\cos\theta), \tag{2.14}$$

where $\lfloor\sigma\rfloor$ means the largest integer less than or equal to σ. We can also factor $(1-2t\cos\theta+t^2)^{-1}$ as $(1-te^{i\theta})^{-1}(1-te^{-i\theta})^{-1}$ then use the binomial theorem to expand the right side of (2.11). Simple manipulations lead to

$$B_n^{\lambda}(\cos\theta;a;k)=\sum_{j,l,m,r}\exp[i\theta(j-l+kr-km)]\frac{(\lambda+ia\cot\theta)_m}{m!}\frac{(\lambda-ia\cot\theta)_r}{r!}$$

where the sum is taken over $j,l,m,n\geq 0$ such that $n=l+j+km+kr$ and $l,j<k$. Therefore

(2.15)

$$B_n^{\lambda}(\cos\theta;a;k)=\sum_{\substack{j,r,m\geq 0\\ j+km+kr\leq n}}\frac{(\lambda+ia\cot\theta)_m}{m!}\frac{(\lambda-ia\cot\theta)_r}{r!}\exp[i\theta(2j+2kr-n)],$$

with $j<k$. We now proceed with the c_n^{λ}'s. Recall that in this case we set

$$\alpha=s^{k\lambda+ka},\quad \beta=s^{2k\lambda},\quad q=s\omega_k \tag{2.16}$$

then let $s\to 1$. We need a slight modification of the generating function (2.5) in order to obtain a full generating function for the c_n^{λ}'s by letting $s\to 1$. Clearly (2.2) implies

$$F(\cos\theta,t)-\alpha F(\cos\theta,qt)=\frac{\left[1-\alpha+(\beta-\alpha)t^2\right]}{1-2xt+t^2}F(\cos\theta,qt),$$

that is

$$\sum_0^{\infty}\frac{(1-\alpha q^n)}{1-\alpha}F_n(\cos\theta;\alpha,\beta|q)t^n=\left[1+\frac{\beta-\alpha}{1-\alpha}t^2\right]\frac{(qt/\gamma;q)_\infty(qt/\Delta;q)_\infty}{(te^{-i\theta};q)_\infty(te^{i\theta};q)_\infty}. \tag{2.17}$$

Starting with (2.17) and essentially repeating the argument used to derive (2.11) from (2.5) we obtain

(2.18)

$$\sum_{n=0}^{\infty} b_n c_n^{\lambda}(\cos\theta;a;k)t^n=\left[1-\frac{\lambda-a}{\lambda+a}t^2\right](1-2t\cos\theta+t^2)^{-1}(1-t^ke^{-ik\theta})^{-\lambda-ia\cot\theta}$$
$$\cdot(1-t^ke^{ik\theta})^{-\lambda+ia\cot\theta}$$

where

$$b_n:=\{(\lambda+1+a)_{\lfloor n/k\rfloor}(2\lambda)_{\lceil n/k\rceil}\}/\{(1)_{\lfloor n/k\rfloor}(\lambda+a)_{\lceil n/k\rceil}\},\tag{2.19}$$

and $\lceil\sigma\rceil$ means the smallest integer greater than or equal to σ.

If we combine (2.18) and (2.11), we get

$$b_n c_n^{\lambda}(x;a;k)=B_n^{\lambda}(x;a;k)-\frac{\lambda-a}{\lambda+a}B_{n-2}^{\lambda}(x;a;k),\qquad n>1.\tag{2.20}$$

Two more representations for $c_n^{\lambda}(x;q;k)$ can be obtained from (2.14), (2.15) and (2.20). Note that (2.11) and (2.12) imply

$$B_n^{\lambda}(\cos\theta;a;k)=U_n(x),\qquad n=0,1,\cdots,k-1.\tag{2.21}$$

This also follows from (2.14).

We now derive two families of generating functions. Our proofs rely on the following lemma.

LEMMA 2.1. *If*

$$P(t)=\sum_0 P_nt^n,\tag{2.22}$$

then

$$\sum_{j=0}^{k-1}P(t\omega_k^j)\omega_k^{-lj}=k\sum_{n=0}^{\infty}P_{kn+l}t^{l+kn},\qquad l=0,1,\cdots,k-1.\tag{2.23}$$

Proof. Clearly

$$\sum_{j=0}^{k-1}\omega_k^{jb}=\begin{cases}k & \text{if } k\mid b,\\ (1-\omega_k^{kb})/(1-\omega_k^b) & \text{if } k\nmid b.\end{cases}$$

Thus

$$\sum_{j=0}^{k-1}\omega_k^{jb}=\begin{cases}k & \text{if } k\mid b,\\ 0 & \text{if } k\nmid b.\end{cases}\tag{2.24}$$

This leads to

$$\sum_{j=0}^{k-1}\omega_k^{-jl}P(t\omega_k^j)=\sum_{j=0}^{k-1}\omega_k^{-jl}\sum_0^{\infty}P_nt^n\omega_k^{jn}=\sum_n P_nt^n\sum_{j=0}^{k-1}\omega_k^{j(n-l)},$$

which when combined with (2.24) implies (2.23).

THEOREM 2.2. *We have*

$$(2.25)\quad \sum_{n=0}^{\infty} B^{\lambda}_{l+kn}(\cos\theta;a;k)t^n=\left(1-te^{ik\theta}\right)^{-\lambda-1+ia\cot\theta}\left(1-te^{-ik\theta}\right)^{-\lambda-1-ia\cot\theta}$$
$$\cdot\{U_l(\cos\theta)+tU_{k-l-2}(\cos\theta)\}$$

and

(2.26)

$$\sum_{n=0}^{\infty}\frac{(2\lambda)_{n+1}}{(\lambda+a)n!}c^{\lambda}_{l+kn}(\cos\theta;a;k)t^n$$
$$=\left(1-te^{ik\theta}\right)^{-\lambda-1+ia\cot\theta}\left(1-te^{-ik\theta}\right)^{-\lambda-1-ia\cot\theta}$$
$$\cdot\left[\frac{2\lambda}{\lambda+a}\{U_l(\cos\theta)+tU_{k-l-2}(\cos\theta)\}-\frac{\lambda-a}{\lambda+a}(1-2t\cos k\theta+t^2)\delta_{l,0}\right.$$
$$\left.+2\frac{a-\lambda}{\lambda+a}\cos\theta\{U_{l-1}(\cos\theta)+tU_{k-l-1}(\cos\theta)\}\right]$$

where $l=0,1,\cdots,k-1$ *and* $U_{-1}(x)$ *is interpreted as zero.*

Proof. We first prove (2.25). From Lemma 2.1 we get

$$k\sum_{n=0}^{\infty}B^{\lambda}_{l+kn}(\cos\theta;a;k)t^{l+kn}$$
$$=\sum_{j=0}^{k-1}\omega_k^{-jl}\left(1-2t\omega_k^{j}\cos\theta+t^2\omega_k^{2j}\right)^{-1}$$
$$\cdot\left(1-t^ke^{-ik\theta}\right)^{-\lambda-ia\cot\theta}\left(1-t^ke^{ik\theta}\right)^{-\lambda+ia\cot\theta}$$
$$=\left(1-t^ke^{-ik\theta}\right)^{-\lambda-ia\cot\theta}\left(1-t^ke^{ik\theta}\right)^{-\lambda+ia\cot\theta}\sum_{j=0}^{k-1}\omega_k^{-jl}\sum_{n=0}^{\infty}U_n(\cos\theta)t^n.$$

Applying Lemma 2.1 to the right side of the above equality establishes

$$(2.27)\quad \sum_{n=0}^{\infty}B^{\lambda}_{l+kn}(\cos\theta;a;k)t^n=\left(1-t^ke^{-ik\theta}\right)^{-\lambda-ia\cot\theta}\left(1-t^ke^{ik\theta}\right)^{-\lambda+ia\cot\theta}$$
$$\cdot\sum_{n=0}^{\infty}U_{l+kn}(\cos\theta)t^n.$$

The observation

$$\sum_{n=0}^{\infty}t^nU_{l+kn}(\cos\theta)=\sum_{n=0}^{\infty}t^n\{e^{i(l+kn+1)\theta}-e^{-i(kn+l+1)\theta}\}/(2i\sin\theta)$$
$$=\left\{\frac{e^{i(l+1)\theta}}{1-te^{ik\theta}}-\frac{e^{-i(l+1)\theta}}{1-te^{-ik\theta}}\right\}/(2i\sin\theta),$$

and (2.27) prove (2.25).

Formula (2.26) can be proved along the same lines. The interested reader can easily fill in the details of the proof.

Using (2.13), (2.25) and (2.26) we obtain

$$\begin{aligned}(2.28)\qquad B^{\lambda}_{l+kn}(\cos\theta;0;k) &= U_l(\cos\theta)P^{\lambda+1}_n(\cos k;0;1)\\ &\quad + U_{k-l-2}(\cos\theta)\cdot P^{\lambda+1}_{n-1}(\cos k\theta;0,1)\end{aligned}$$

and a similar formula for the polynomials of the first kind when $a=0$. Recall that

$$P^{\lambda}_n(x,0,1)=C^{\lambda}_n(x).$$

The generating functions (2.25) and (2.26) reduce to new generating functions of the sieved ultraspherical polynomials when $a=0$.

Al-Salam and Chihara [4] determined all pairs of orthogonal polynomials $\{p_n(x)\}$ and $\{q_n(x)\}$ such that their convolution

$$Q_n(x,y)=\sum_{j=0}^{n} p_j(x)q_{n-j}(y), \qquad n\geq 0.$$

defines $\{Q_n(x,y)\}$ as an orthogonal polynomial set in x for all y. They were motivated by two examples involving Hermite and Laguerre polynomials, namely

$$\begin{aligned} &p_n(x)=q_n(x)=H_n(\sqrt{2}x)/n!, && Q_n(x,y)=2^{-n}H_n(x+y)/n!,\\ &p_n(x)=L^{\alpha}_n(x),\ q_n(x)=L^{\beta}_n(x), && Q_n(x,y)=L^{\alpha+\beta+1}_n(x+y).\end{aligned}$$

A related question, which is still open, is to characterize all orthogonal polynomials $\{p_n(x)\}$, $\{q_n(x)\}$, $\{Q_n(x)\}$ that satisfy

$$(2.29)\qquad Q_n(x)=\sum_{j=0}^{n} p_j(x)q_{n-j}(x).$$

This seems to be a much harder question. A slightly more general question is to replace the convolution (2.29) by

$$(2.30)\qquad Q_n(\cos\theta)=\sum_{j=0}^{\lfloor n/k\rfloor} p_j(\cos k\theta)q_{n-kj}(\cos\theta),$$

and ask the same question. Formla (2.14) is an example of (2.30).

3. Limiting relations. Recall that if $\{p_n(x)\}$ is a sequence of polynomials satisfying

$$(3.1)\qquad p_{n+1}(x)=(A_nx+B_n)p_n(x)-C_np_{n-1}(x), \qquad n>0,$$

$$(3.2)\qquad p_0(x)=1,\quad p_1(x)=A_0x+B_0,\quad A_0\neq 0$$

and the positivity condition

$$(3.3)\qquad A_{n-1}A_nC_n>0, \qquad n>0,$$

then there exists a positive measure $d\psi$ such that all the moments $\int t^n d\psi$ exist and

$$(3.4)\qquad \int_{-\infty}^{\infty} p_n(x)p_m(x)\,d\psi(x)=\lambda_n\delta_{m,n}$$

with

$$\lambda_n = \frac{A_0}{A_n} C_1 \cdots C_n \lambda_0, \qquad n > 0. \tag{3.5}$$

We shall normalize ψ by $\lambda_0 = 1$. The function ψ is called the distribution function and is normalized by $\psi(-\infty)=0$, $\psi(x)=\frac{1}{2}[\psi(x+0)+\psi(x-0)]$ and, of course $\int d\psi = 1$. Nevai [10, pp. 141, 143] proved that if

$$\sum_{n=1}^{\infty} |B_n/A_n| + \left|\{C_n/A_nA_{n-1}\}^{1/2} - \frac{\gamma}{2}\right| < \infty \tag{3.6}$$

then

$$d\psi = \psi'\,dx + d\psi_j,$$

where ψ' is continuous and positive on $(-\gamma,\gamma)$, vanishes outside $[-\gamma,\gamma]$, and ψ_j is a jump function, constant inside $(-\gamma,\gamma)$. He also proved the limiting relation

$$\limsup_n \left\{ p_n^2(x)\psi'(x)\sqrt{\gamma^2-x^2} \right\}/\lambda_n = \frac{\pi}{2}, \tag{3.7}$$

for almost all $x \in (-\gamma,\gamma)$.

The condition (3.6) is not satisfied when $p_n(x)$ is $B_n^\lambda(x;a;k)$ or $c_n^\lambda(x;a;k)$. The function ψ', as it will out, vanishes $k-1$ times in $(-1,1)$. In both cases $\gamma=1$. The asymptotic formula (3.7) still holds for both sets of polynomials. As a matter of fact, this is how I discovered what ψ is. So, we now compute the left side of (3.7). We shall use Darboux's method.

THEOREM 3.1 (Darboux's method). *Assume that $f(z)=\sum_0^\infty f_n z^n$ is an analytic function in $|z|<r$ and that $g(z)=\sum_0^\infty g_n z^n$ is a comparison function, that is, $f-g$ is continuous $|z|=r$ and $g(z)$ is analytic in $|z|<r$. Then*

$$f_n = g_n + o(r^{-n}). \tag{3.8}$$

Olver [11, §8.9] proves Theorem 3.1 from the Riemann – Lebesgue lemma.

Our first asymptotic results are the following.

THEOREM 3.2. *We have*

$$B_n^\lambda(\cos\theta;a;k) \approx \frac{(2k)^{-\lambda} n^\lambda \exp(a\cot\theta(\pi/2 - k\theta + l\pi))}{|\Gamma(\lambda+1-ia\cot\theta)| |\sin\theta| |\sin(k\theta)|^\lambda} \cos[\varepsilon_n(\theta)], \tag{3.9}$$

$$as\ n \to \infty, \quad 0<\theta<\pi, \quad l\pi < k\theta < (l+1)\pi,$$

with $l = 0,1,\cdots$ and

$$\varepsilon_n(\theta) := (n+k\lambda+1)\theta - \frac{\pi}{2}(\lambda+1) - a\cot\theta \ln\left(\frac{n}{2k}\right) + \arg\Gamma(\lambda+1+ai\cot\theta) \tag{3.10}$$
$$- \pi l\lambda - \arg\left[(\sin k\theta)^{ai\cot\theta}\right],$$

and

(3.11)

$c_n^\lambda(\cos\theta; a; k)$

$$\approx \frac{\sqrt{2}n^\lambda \exp(a\cot\theta(\pi/2 - k\theta + l\pi))\left[\lambda^2 + a^2 + (a^2 - \lambda^2)\cos 2\theta\right]^{1/2}}{|\lambda + a||\Gamma(\lambda + 1 - ia\cot\theta)|\sin\theta\left[2k|\sin(k\theta)|\right]^\lambda}\cos[\phi + \varepsilon_n(\theta)]$$

as $n \to \infty$, *where* $\varepsilon_n(\theta)$ *is as in* (3.10), $0 < \theta < \pi$, $l\pi < k\theta < (l+1)\pi$ *and*

$$\phi = \arg\left[1 - (\lambda - a)e^{-2i\theta}/(\lambda + a)\right]. \tag{3.12}$$

Proof. The dominant term in a comparison function for (2.11) is

$$(1 - te^{i\theta})^{-1-\lambda+ia\cot\theta}k^{-\lambda+ia\cot\theta}(1 - e^{-2i\theta})^{-1}(1 - e^{-2ik\theta})^{-\lambda-ia\cot\theta}$$
$$+ \text{a similar term with } \theta \text{ replaced by } -\theta.$$

Therefore

(3.13)

$$B_n(\cos\theta; a; k) \approx \mathrm{Re}\left\{\frac{k^{-\lambda+ia\cot\theta}(\lambda + 1 - ai\cot\theta)_n \exp[i\theta(n + 1 + k\lambda + iak\cot\theta)]}{2n!\exp[i(\pi/2)(\lambda + 1 + ia\cot\theta)](2\sin k\theta)^{\lambda + ia\cot\theta}\sin\theta}\right\}.$$

Recall that

$$\frac{(\sigma)_n}{n!} = \frac{\Gamma(\sigma + n)}{\Gamma(\sigma)\Gamma(n+1)} \approx \frac{n^{\sigma-1}}{\Gamma(\sigma)} \quad \text{as } n \to \infty.$$

The above relationship and (3.13) yield

$$B_n(\cos\theta; a; k) \approx \mathrm{Re}\left\{\frac{(2k)^{-\lambda}\exp((\pi/2 - k\theta + l\pi)a\cot\theta)n^{\lambda - ia\cot\theta}}{2\sin\theta|\sin k\theta|^\lambda|\Gamma(\lambda + 1 - ai\cot\theta)|}\right.$$
$$\cdot\exp\left[i\theta(n + 1 + k\lambda) + ia\cot\theta\ln(k/2)\right.$$
$$\left.\left. - i\frac{\pi}{2}(\lambda + 2l + 1) - ia\cot\theta\ln(|\sin k\theta|) - i\arg\Gamma(\lambda + 1 - ai\cot\theta)\right]\right\}$$

which, after some simplification, reduces to (3.9).

The proof of (3.11) is similar and will be omitted. This completes the proof of Theorem 3.2.

In the case of $\{B_n^\lambda(x; a; k)\}$, the coefficients in (3.1) and (3.2) are given by

$$B_n = 0, \quad A_n = \begin{cases} 2, & n \neq mk - 1, \\ \dfrac{2(\lambda + a + m)}{m}, & n = mk - 1, \end{cases} \quad C_n = \begin{cases} 1, & n \neq mk - 1, \\ \dfrac{2\lambda + m}{m}, & n = mk - 1, \end{cases} \tag{3.14}$$

$\gamma = 1$, so (3.6) diverges like $\ln n$ and (3.18) and (3.21) give

$$\lambda_n \approx (2\lambda + 1)_{(n/k)}/(n/k)! = \frac{\Gamma(2\lambda + 1 + n/k)}{\Gamma(2\lambda + 1)\Gamma(1 + n/k)},$$

hence

$$\lambda_n \approx \left(\frac{n}{k}\right)^{2\lambda} \frac{1}{\Gamma(2\lambda+1)}, \tag{3.15}$$

since $\Gamma(a+n)/\Gamma(b+n) \approx n^{a-b}$ and λ_0 is normalized to be 1. Nevai's result (3.7), in view of (3.9) and (3.15) suggests

$$\psi'(x) = \frac{2^{2\lambda+1}}{\pi}(1-x^2)^{\lambda+1/2}|U_{k-1}(x)|^{2\lambda}|\Gamma(\lambda+1-ia\cot\theta)|^2 \frac{1}{\Gamma(2\lambda+1)} \cdot \exp[a\cot\theta(2k\theta-\pi-2l\pi)], \tag{3.16}$$

$x=\cos\theta\varepsilon(-1,1)$. In the next section we shall prove that ψ is absolutely continuous and ψ' is indeed given by (3.16) under the normalization $\lambda_0=1$. When $a=0$, (3.16) reduces to a result mentioned in Al-Salam, Allaway and Askey [3]. We shall also prove that (3.7) holds when $P_n(x)$ is $c_n^\lambda(x;a;k)$. This raises the question of the validity of (3.7) when

$$\sup_{n>0} \frac{1}{\ln n}\left\{ \sum_{j=1}^{n} |B_j/A_j| + \left|\sqrt{C_j/A_jA_{j-1}} - \frac{\gamma}{2}\right| \right\} < \infty. \tag{3.17}$$

Condition (3.17) seems to be sufficient for the validity of the asymptotic relationship (3.7).

We now determine the asymptotic behavior of the polynomials B_n^λ and c_n^λ in the complex plane cut along $[-1,1]$. If $x \notin [-1,1]$ then the quantities $e^{\pm i\theta} = x \pm i\sqrt{1-x^2}$ have different absolute values. It is easy to see that

$$|e^{i\theta}| > 1 > |e^{-i\theta}|, \quad \operatorname{Im} x > 0, \quad \operatorname{Re} x \notin [-1,1], \quad x=\cos\theta, \tag{3.18}$$

$$|e^{i\theta}| < 1 < |e^{-i\theta}|, \quad \operatorname{Im} x < 0, \quad \operatorname{Re} x \notin [-1,1], \quad x=\cos\theta, \tag{3.19}$$

provided that x lies in a neighborhood of the real axis. A proof similar to our proof of Theorem 3.2 establishes the following result.

THEOREM 3.3. *The asymptotic results*

$$B_n^\lambda(x;a;k) \approx \frac{(1-e^{-2i\theta})^{-1}(1-e^{-2i\theta k})^{-\lambda-ia\cot\theta} k^{-\lambda+ia\cot\theta} n^{\lambda-ia\cot\theta}}{\Gamma(\lambda+1-ia\cot\theta)\exp(-in\theta)}, \tag{3.20}$$

$$b_n c_n^\lambda(x;a;k) \approx \frac{[\lambda+a-(\lambda-a)e^{-2i\theta}]}{\lambda+a} B_n^\lambda(x;a;k) \tag{3.21}$$

holds as $n \to \infty$ and x and θ are as in (3.18) *provided that $\lambda+1-ia\cot\theta \neq 0, -1, -1, \cdots$. In the lower half plane the corresponding asymptotic formulas follow from* (3.20) *and* (3.21) *by replacing θ by $-\theta$.*

When

$$\lambda+1-ia\cot\theta = -j, \qquad j=0,1,2,\cdots, \tag{3.22}$$

then θ must be purely imaginary and (3.18) holds if and only if $x>1$. In this case (3.22) gives

$$\frac{ax}{\sqrt{x^2-1}} = -(\lambda+j+1). \tag{3.23}$$

This has a solution $x>1$ if and only if

$$a(\lambda+j+1)<0. \tag{3.24}$$

If (3.22) and (3.24) hold, then the generating function (2.11) becomes an analytic function of t in $|t|<|e^{i\theta}|$ and (3.20) must be replaced by

$$B_n^\lambda(x;a;k)\simeq\frac{(1-e^{2i\theta})^{-1}(1-e^{2ik\theta})^{j+1}}{\Gamma(\lambda+1+ia\cot\theta)\exp(in\theta)}\left(\frac{n}{k}\right)^{\lambda+ia\cot\theta} \tag{3.25}$$

Similarly, the asymptotic behavior of $c_n^\lambda(x;a;k)$ when x satisfies (3.23) can be determined.

4. The numerator polynomials. Let $\{p_n(x)\}$ and $\{p_n^*(x)\}$ be two solutions of the three-term recurrence relation (3.1) and assume that $\{p_n(x)\}$ satisfy the initial conditions (3.2), i.e.

$$p_0(x)=1,\quad p_1(x)=A_0x+B_0,\quad A_0\neq 0, \tag{4.1}$$

and $\{p_n^*(x)\}$ satisfy

$$p_0^*(x)=0,\qquad p_1^*(x)=A_0. \tag{4.2}$$

The p_n's are called the denominator polynomials of the continued fraction

$$\chi(x):=\frac{A_0}{A_0x+B_0-}\,\frac{C_1}{A_1x+B_1-}\,\frac{C_2}{A_2x+B_2-}\cdots \tag{4.3}$$

while the p_n^* are its numerator polynomials. In fact $p_n^*(x)/p_n(x)$ is the nth convergent of the continued fraction (4.3). When the support of the measure $d\psi$, see (3.4), is bounded, then Markov's theorem asserts

$$\chi(x)=\lim_{n\to\infty}\frac{p_n^*(x)}{p_n(x)}=\int_{-\infty}^{\infty}\frac{d\psi(u)}{x-u}, \tag{4.4}$$

is valid in the x complex plane cut along the support of $d\psi$. In the present section we study the polynomials $c_n^{\lambda *}(x;a;k)$ and $B_n^{\lambda *}(x;a;k)$. We derive generating functions for these polynomials and use the generating functions to determine the asymptotic behavior of $c_n^{\lambda *}$ and $B_n^{\lambda *}$ for large n and fixed x. The purpose is to combine these asymptotic results with the asymptotic results of §3 then compute $\chi(x)$ from (4.4) and then determine the distribution function from the inversion formula for the Stieltjes transform

$$\psi(t_2)-\psi(t_1)=\lim_{\varepsilon\to 0^+}\frac{1}{2\pi i}\int_{t_1}^{t_2}[\chi(u-i\varepsilon)-\chi(u+i\varepsilon)]\,du. \tag{4.5}$$

Recall that the F_n's satisfy the recursion relation (1.15); hence F_n^* also satisfies (1.15) and, in view of (1.4), (4.1) and (4.2), the initial conditions

$$F_0^*(x;\alpha;\beta|q)=0,\qquad F_1^*(x;\alpha,\beta|q)=2(1-\alpha)/(1-q). \tag{4.6}$$

Now multiply (1.15), with F_n replaced by F_n^*, by t^{n+1} and add the results for $n=1,\cdots$. This and (4.6) give

$$F^*(x,t)=\frac{2(1-\alpha)t}{1-2xt+t^2}+\frac{1-2x\alpha t+\beta t^2}{1-2xt+t^2}F^*(x,qt), \tag{4.7}$$

where

$$F^*(x,t):=\sum_{n=0}^{\infty}F_n^*(x;\alpha;\beta|q)t^n. \tag{4.8}$$

The solution of the functional equation (4.7) is

$$F^*(\cos\theta,t)=2t(1-\alpha)\sum_0^{\infty}\frac{(t/\gamma;q)_n(t/\Delta;q)_n}{\left(te^{-i\theta};q\right)_{n+1}\left(te^{i\theta};q\right)_{n+1}}q^n, \tag{4.9}$$

where γ and Δ are as in (2.4). Using the observation

$$(\sigma;q)_n=(\sigma;q)_\infty/(\sigma q^n;q)_\infty,$$

and (2.5) we can express $F^*(\cos\theta,t)$ as

$$F^*(\cos\theta,t)=2t(1-\alpha)F(\cos\theta,t)\sum_0^{\infty}\frac{\left(te^{-i\theta}q^{n+1};q\right)_\infty\left(te^{i\theta}q^{n+1};q\right)_\infty}{(q^nt/\gamma;q)_\infty(q^nt/\Delta;q)_\infty}q^n.$$

The q-binomial theorem (2.6) and the above representation lead to

$$F^*(\cos\theta,t)=2t(1-\alpha)F(\cos\theta,t)\sum_{m,j=0}^{\infty}\frac{\left(\gamma qe^{-i\theta};q\right)_j\left(\Delta qe^{i\theta};q\right)_m}{(q;q)_j(q;q)_m}\left(\frac{t}{\gamma}\right)^j\left(\frac{t}{\Delta}\right)^m \cdot\sum_{n=0}^{\infty}q^{(m+j+1)n}.$$

Therefore we have

(4.10)

$$F^*(\cos\theta,t)=2tF(\cos\theta,t)\sum_{m,j=0}^{\infty}\frac{\left(\gamma q^2e^{-i\theta};q\right)_j\left(\Delta q^2e^{i\theta};q\right)_m}{(q;q)_j(q;q)_m}\left(\frac{t}{\gamma}\right)^j\left(\frac{t}{\Delta}\right)^m \cdot\frac{\left(1-\gamma qe^{-i\theta}\right)}{\left(1-\gamma e^{-i\theta}q^{j+1}\right)}\frac{\left(1-\Delta qe^{i\theta}\right)}{\left(1-\Delta e^{i\theta}q^{m+1}\right)}\frac{1-\alpha}{1-q^{m+j+1}}.$$

The numerator polynomials of the sieved polynomials of the second kind arise when q, α and β are chosen as in (2.7) and $s\to1$. It is easy to see that

$$\lim_{s\to1}\frac{1-\gamma qe^{-i\theta}q^r}{1-q^r}=\begin{cases}1 & \text{if } k\nmid r,\\ \dfrac{-\lambda-ia\cot\theta+r/k}{r/k} & \text{if } k|r,\end{cases}$$

$$\lim_{s\to1}\frac{1-\Delta qe^{i\theta}q^r}{1-q^r}=\begin{cases}1 & \text{if } k\nmid r,\\ \dfrac{-\lambda+ia\cot\theta+r/k}{r/k} & \text{if } k|r,\end{cases}$$

$$\lim_{s\to1}\frac{1-\gamma qe^{-i\theta}}{1-\gamma e^{-i\theta}q^{1+kr}}=-\frac{\lambda+ia\cot\theta}{-\lambda-ia\cot\theta+r},\qquad \lim_{s\to1}\frac{1-\Delta qe^{i\theta}}{1-\Delta e^{i\theta}q^{1+kr}}=\frac{-\lambda+ia\cot\theta}{-\lambda+ia\cot\theta+r}.$$

As $s \to 1$ the terms in the sum and in (4.10) will vanish except in three cases

$$\text{(I)}\ k|j \text{ and } k|m, \quad \text{(II)}\ k|j \text{ and } k|j+m+1, \quad \text{(III)}\ k|m \text{ and } k|m+j+1.$$

We let

$$\lim_{s\to 1} \frac{F^*(\cos\theta,t)}{F(\cos\theta,t)} = \sum\nolimits_1(\theta) + \sum\nolimits_2(\theta) + \sum\nolimits_3(\theta),$$

where k,j belong to cases I, II, III in Σ_1, Σ_2, Σ_3, respectively. Using the limiting relationships following (4.10), it is easy to see

$$\sum\nolimits_1(\theta) = 2t(1-t^k e^{-ik\theta})^{\lambda+ia\cot\theta}(1-t^k e^{ik\theta})^{\lambda-ia\cot\theta},$$

$$\sum\nolimits_2(\theta) = 2t(te^{i\theta})^{k-1}\sum_{j,m}\frac{(-\lambda-ia\cot\theta)_j}{j!}\,\frac{(1-\lambda+ia\cot\theta)_m}{m!}$$

$$\cdot(te^{-i\theta})^{kj}(te^{i\theta})^{mk}\frac{(\lambda-ia\cot\theta)}{j+m+1},$$

$$\sum\nolimits_3(\theta) = \sum\nolimits_2(-\theta).$$

In Σ_2 we replace $t^{k(m+j+1)}/(j+m+1)$ by $\int_0^{t^k} u^{j+m}\,du$ to obtain the integral representation

(4.11)

$$\sum\nolimits_2(\theta) = 2e^{i(k-1)\theta}(\lambda-ia\cot\theta)\int_0^{t^k}(1-ue^{-ik\theta})^{\lambda+ia\cot\theta}(1-ue^{ik\theta})^{\lambda-1-ia\cot\theta}\,du$$

and a similar integral representation for Σ_3. Both representations are valid when $\lambda > 0$. Recall that

$$\text{(4.12)} \qquad \lim_{s\to 1} F(\cos\theta,t) = \sum_0^{\infty} B_n^{\lambda}(\cos\theta;a;k)t^n =: B(\cos\theta,t),$$

say. Thus

$$\text{(4.13)} \quad B(\cos\theta,t) = (1-2t\cos\theta+t^2)^{-1}(1-t^k e^{-ik\theta})^{-\lambda-ia\cot\theta}(1-t^k e^{ik\theta})^{-\lambda+ia\cot\theta},$$

see (2.11). Therefore

$$\text{(4.14)} \quad \sum_{n=0}^{\infty} B_n^{\lambda *}(\cos\theta;a;k)t^n = \frac{2t}{1-2xt+t^2} + B(\cos\theta,t)\left\{\sum\nolimits_2(\theta)+\sum\nolimits_2(-\theta)\right\}.$$

We now apply Darboux's method to (4.14). The result when $\lambda > 0$ is

(4.15)

$$B_n^{\lambda *}(\cos\theta;a;k)$$

$$\approx 2B_n^{\lambda}(\cos\theta;a;k)\Big[(\lambda-ia\cot\theta)e^{i(k-1)\theta}$$

$$\cdot\int_0^{e^{\overline{-ik\theta}}}(1-ue^{-ik\theta})^{\lambda+ia\cot\theta}(1-ue^{ik\theta})^{\lambda-1-ia\cot\theta}\,du$$

$$+(\lambda+ia\cot\theta)e^{-i(k-1)\theta}\int_0^{e^{\overline{-ik\theta}}}(1-ue^{ik\theta})^{\lambda-ia\cot\theta}(1-ue^{-ik\theta})^{\lambda-1+ia\cot\theta}\,du\Big],$$

where $x=\cos\theta$, $\operatorname{Im} x>0$. When $\operatorname{Im} x<0$ the above formula holds with $e^{-ik\theta}$ replaced by $e^{ik\theta}$ only in the upper limit of the integrals. The above integral is a Hadamard integral, see [8] and [15].

We now essentially repeat the above analysis for the numerator polynomials of the first kind. Recall that in this case

$$\alpha=s^{k\lambda+ka},\quad \beta=s^{2k\lambda},\quad q=s\omega_k \tag{4.16}$$

and we let

$$\gamma=s^{-k\lambda}\{s^{ka}\cos\theta+i\sqrt{1-s^{2ka}\cos\theta}\},\qquad \Delta=s^{-k\lambda}\{s^{ka}\cos\theta-i\sqrt{1-s^{2ka}\cos^2\theta}\}. \tag{4.17}$$

The analysis till (4.10) remains valid but we rewrite (4.10) in the form

$$F^*(\cos\theta,t)=2tF(\cos\theta,t)\sum_{m,j=0}^{\infty}\frac{(\gamma qe^{-i\theta};q)_j(\Delta qe^{i\theta};q)_m}{(q;q)_j(q;q)_m}\left(\frac{t}{\gamma}\right)^j\left(\frac{t}{\Delta}\right)^m\frac{1-\alpha}{1-q^{m+j+1}}. \tag{4.18}$$

One can easily apply (4.16) and (4.17) to show that

$$\lim_{s\to1}(1-\gamma e^{-i\theta}q^r)/(1-q^r)=\lim_{s\to1}(1-\Delta e^{i\theta}q^r)/(1-q^r)=1\quad\text{if } k\nmid r,$$

and

$$\lim_{s\to1}\frac{1-\gamma e^{-i\theta}q^{kr}}{1-q^{kr}}=\frac{r-\lambda-ia\cot\theta}{r},\qquad \lim_{s\to1}\frac{1-\Delta e^{i\theta}q^{kr}}{1-q^{kr}}=\frac{r-\lambda+ia\cot\theta}{r}.$$

Furthermore

$$\lim\frac{1-\alpha}{1-q^{m+j+1}}=\begin{cases}0 & \text{if } k\nmid j+m+1,\\ (\lambda+a)/r & \text{if } j+m+1=kr.\end{cases}$$

Using the above calculations, we see that the result of letting $s\to1$ in (4.18) is

$$\lim_{s\to1}\frac{F^*(\cos\theta,t)}{F(\cos\theta,t)}=2k(\lambda+a)\sum_{k|j+m+1}\frac{(1-\lambda-ia\cot\theta)_{\lfloor j/k\rfloor}(1-\lambda+ia\cot\theta)_{\lfloor m/k\rfloor}}{(1)_{\lfloor j/k\rfloor}(1)_{\lfloor j/k\rfloor}(m+j+1)}\cdot t^{j+m+1}e^{i(m-j)\theta}.$$

Set

$$j=kj_1+l,\quad m=km_1+k-l-1,\quad 0\le l<k,$$

so $j_1\ge0$ and $m_1\ge0$. Thus

$$\lim_{s\to1}\frac{F^*(\cos\theta,t)}{F(\cos\theta,t)}=2(\lambda+a)\sum_{j_1,m_1=0}^{\infty}\frac{(1-\lambda-ia\cot\theta)_{j_1}(1-\lambda+ia\cot\theta)_{m_1}}{j_1!m_1!(m_1+j_1+1)}t^{(m_1+j_1+1)k}\cdot e^{-i\theta}e^{ik\theta(m_1-j_1+1)}\sum_{l=0}^{k-1}e^{-2il\theta},$$

that is

(4.19)

$$\lim_{s\to 1}\frac{F^*(\cos\theta,t)}{F(\cos\theta,t)}=2(\lambda+a)\frac{\sin k\theta}{\sin\theta}\int_0^{t^k}\left(1-ue^{-ik\theta}\right)^{\lambda-1+ia\cot\theta}\left(1-ue^{ik\theta}\right)^{\lambda-1-ia\cot\theta}du.$$

Applying Darboux's method to the above generating function establishes

$$c_n^{\lambda*}(\cos\theta;a;k)\approx 2(\lambda+a)c_n^{\lambda}(\cos\theta;k;a)\frac{\sin k\theta}{\sin\theta} \tag{4.20}$$

$$\cdot\int_0^{\overline{e^{-ik\theta}}}\left(1-ue^{-ik\theta}\right)^{\lambda-1+ia\cot\theta}\left(1-ue^{ik\theta}\right)^{\lambda-1-ia\cot\theta}du,$$

where, as before, $x=\cos\theta$, $\operatorname{Im}x>0$. If $\operatorname{Im}x<0$, $e^{-ik\theta}$ in the upper limit of the integral should be replaced by $e^{ik\theta}$.

Using the integral representation

$$F(a,b;c;z)=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}dt, \tag{4.21}$$

$|z|<1$, $\operatorname{Re}(c)>\operatorname{Re}(b)>0$, Rainville [16, p. 47] we can express Σ_2 as a multiple of a hypergeometric function. This enables us to express the right side of the generating function (4.14) and (4.19) in terms of hypergeometric functions. In the next section we use the integral representations directly to compute the distribution function. Using hypergeometric functions would have complicated the problem by forcing us to use analytic continuation and contiguous formulas, see Erdélyi et al. [11], to achieve the same results.

5. Orthogonality relations and continued fractions. The support of the measure $d\psi$ in (3.4) is bounded when $\{B_n/A_n\}$ and $\{C_{n+1}/A_nA_{n+1}\}$ are bounded sequences, Chihara [10, p. 109]. Both sequences are bounded in the case of the sieved polynomials of the first and second kinds. Thus, Markov's theorem is applicable and (4.4) holds. Set

$$\chi_1(x):=\lim_{n\to\infty}B_n^{\lambda*}(x;a;k)/B_n^{\lambda}(x;a;k), \tag{5.1}$$

$$\chi_2(x):=\lim c_n^{\lambda*}(x;a;k)/c_n^{\lambda}(x;a;k). \tag{5.2}$$

Clearly (4.15) implies

$$\chi_1(\cos\theta)=2(\lambda-ia\cot\theta)e^{i(k-1)\theta}\int_0^{\overline{e^{-ik\theta}}}\left(1-ue^{-ik\theta}\right)^{\lambda+ia\cot\theta}\left(1-ue^{ik\theta}\right)^{\lambda-1-ia\cot\theta}du$$

$$+2(\lambda+ia\cot\theta)e^{-i(k-1)\theta} \tag{5.3}$$

$$\cdot\int_0^{\overline{e^{-ik\theta}}}\left(1-ue^{ik\theta}\right)^{\lambda-ia\cot\theta}\left(1-ue^{-ik\theta}\right)^{\lambda-1+ia\cot\theta}du,$$

$\operatorname{Im}(\cos\theta)>0$. An application of (4.21) expresses $\chi_1(x)$ in terms of hypergeometric functions and makes it clear that $\chi_1(x)$ has no poles when $a\geq 0$, so the discrete part of $d\psi$, if any, may occur on the support of the absolutely continuous component. The inversion formula (4.5) implies

$$2\pi i\psi'(x)=\chi_1(x-i0)-\chi_1(x+i0).$$

Therefore

(5.4)

$$\pi i\psi'(x)=(\lambda-ia\cot\theta)e^{i(k-1)\theta}\int_{e^{-ik\theta}}^{e^{ik\theta}}(1-ue^{-ik\theta})^{\lambda+ia\cot\theta}(1-ue^{ik\theta})^{\lambda-1-ia\cot\theta}du$$

$$+(\lambda+ia\cot\theta)e^{-i(k-1)\theta}\int_{e^{-ik\theta}}^{e^{ik\theta}}(1-ue^{ik\theta})^{\lambda+ia\cot\theta}(1-ue^{-ik\theta})^{\lambda-1-ia\cot\theta}du.$$

The integrals in (5.4) are beta integrals. The change of variable $u=e^{-ik\theta}+v(e^{ik\theta}-e^{-ik\theta})$ reduces the first term on the right side of (5.4) to

$$-e^{-i\theta}(\lambda-ia\cot\theta)(1-e^{-2ik\theta})^{\lambda+ia\cot\theta}(1-e^{2ik\theta})^{\lambda-ia\cot\theta}$$

$$\cdot\int_0^1 v^{\lambda-1-ia\cot\theta}(1-v)^{\lambda+ia\cot\theta}dv,$$

that is

$$-2^{2\lambda}e^{-i\theta}|\sin k\theta|^{2\lambda}|\Gamma(\lambda+1+ia\cot\theta)|^2\exp[a(2k\theta-\pi-2\pi l)\cos\theta]/\Gamma(2\lambda+1),$$

where $l\pi\le k\theta<(l+1)\pi$, $l=0,1,\cdots$. Similarly, the second term on the right side of (5.4) is

$$2^{2\lambda}e^{i\theta}|\sin k\theta|^{2\lambda}|\Gamma(\lambda+1+ia\cot\theta)|^2\exp[a(2k\theta-\pi-2\pi l)\cot\theta]/\Gamma(2\lambda+1).$$

This and (5.4) give

(5.5)

$$W^\lambda(\cos\theta;a;k)=\sin\theta|\sin k\theta|^{2\lambda}|\Gamma(\lambda+1+ia\cot\theta)|^2\exp[a(2k\theta-\pi-2\pi l)\cot\theta],$$

where

$$W^\lambda(\cos\theta;a;k):=\pi\Gamma(2\lambda+1)2^{-2\lambda-1}\psi'(x).$$

Recall that if $\{P_n(x)\}$ is orthogonal with respect to a unique distribution function then x_0 is a mass point if and only if $\Sigma_0^\infty P_n^2(x_0)/\lambda_n$ diverges; Akhiezer [1, p. 69], λ_n is as in (3.4). In the case under consideration

$$\lambda_n\approx Cn^{2\lambda},\qquad C\neq 0, \tag{5.6}$$

follows from (3.15). Furthermore (3.9) and (5.6) show that the series

$$\sum_1^\infty\{B_n^\lambda(x;a;k)\}^2n^{-2\lambda} \tag{5.7}$$

diverges for $x\in(-1,1)$, hence $(-1,1)$ carries no discrete masses. We next consider the points ±1. In the case of the Pollaczek polynomials $P_n^\lambda(x;a,0)$, λ_n is $2^{1-2\lambda}\pi\Gamma(2\lambda+n)/\{n!(n+a+\lambda)\}$, Chihara [10] or Szegö [21], so $\lambda_n\approx C_1n^{2\lambda-2}$, $C_1\neq 0$ and the points ±1 did not support discrete masses. Therefore

$$\sum_1^\infty P_n^{\lambda+1}(\pm 1,a,0)n^{-2\lambda}$$

diverges and (2.25) establishes the divergence of the series in (5.7) at $x = \pm 1$. This can be also proved directly from (2.11) and the aysmptotic properties of Laguerre polynomials. Finally, (3.14) and (3.5) give, after simple manipulations,

$$\lambda_n = \frac{(\lambda + a + 1)_{\lfloor n/k \rfloor}(2\lambda + 1)_{\lfloor (n+1)/k \rfloor}}{(1)_{\lfloor n/k \rfloor}(\lambda + a + 1)_{\lfloor (n+1)/k \rfloor}}. \tag{5.8}$$

The positivity condition (3.3) is satisfied if and only if $\lambda > -\frac{1}{2}$, $\lambda + a + 1 > 0$. This establishes the orthogonality relation

$$\int_{-1}^{1} B_n^\lambda(x;a;k) B_m^\lambda(x;a;k) W^\lambda(x;a;k)\,dx = \frac{\pi\Gamma(2\lambda+1)}{2^{2\lambda+1}}\lambda_n \delta_{m,n}, \tag{5.9}$$

where W^λ is as in (5.5), $\lambda > 0$, $a \geq 0$ and λ_n is given by (5.8). When $a = 0$, (5.9) reduces to the orthogonality relation mentioned in [3]. The case $0 \geq \lambda > -\frac{1}{2}$ is more complicated and will not be treated here.

We now compute $\chi_2(x)$ and the weight function that the c_n^λ's are orthogonal with respect to. Clearly (4.20) and (5.2) yield

$$\chi_2(x) = 2(\lambda + a)\frac{\sin k\theta}{\sin\theta}\int_0^{e^{-ik\theta}} (1 - ue^{-ik\theta})^{\lambda - 1 + ia\cot\theta}(1 - ue^{ik\theta})^{\lambda - 1 - ia\cot\theta}\,du, \tag{5.10}$$

if $x = \cos\theta$, $\operatorname{Im} x > 0$. If $\operatorname{Im} x < 0$, the $e^{-ik\theta}$ appearing in the upper limit of the integral should be replaced by $e^{ik\theta}$. Here again with the help of the integral representation (4.21) one can identify $\chi_2(x)$ as a hypergeometric function and observe that it has no poles in the complex plane cut along $[-1,1]$. Let $\sigma(x)$ be the corresponding distribution function so

$$\chi_2(x) = \int_{-\infty}^{\infty} \frac{d\sigma(u)}{x - u} \tag{5.11}$$

and (4.5) yields

$$2\pi i\sigma'(x) = \chi_2(x - i0) - \chi_2(x + i0);$$

hence

$$2\pi i\sigma'(x) = 2(\lambda + a)\frac{\sin k\theta}{\sin\theta}\int_{e^{-ik\theta}}^{e^{ik\theta}} (1 - ue^{-ik\theta})^{\lambda - 1 + ia\cot\theta}(1 - ue^{ik\theta})^{\lambda - 1 - ia\cot\theta}\,du.$$

As before, we make the change of variable $u = e^{-ik\theta} + (e^{ik\theta} - e^{-ik\theta})v$ and obtain

$$w^\lambda(x;a;k) = \frac{|\sin k\theta|^{2\lambda}}{\sin\theta}|\Gamma(\lambda + ia\cot\theta)|^2 \exp[a(2k\theta - \pi - 2\pi l)\cot\theta], \tag{5.12}$$

where

$$w^\lambda(x;a;k) := \pi\frac{\Gamma(2\lambda)}{\lambda + a}2^{-2\lambda+1}\sigma'(x). \tag{5.13}$$

In the notation of (3.1),

$$A_n = 2, \quad C_n = 1 \quad \text{if } k \nmid n,$$

$$A_{mk} = \frac{2(\lambda + a + m)}{2\lambda + m}, \qquad C_{mk} = \frac{m}{2\lambda + m},$$

where we used (1.21). The λ_n's in this case, see (3.5) are given by

$$\lambda_{mk+l}=\frac{(\lambda+a)m!}{(2\lambda)_{m+1}} \quad \text{if } l\neq 0, \qquad \lambda_{mk}=\frac{(\lambda+a)m!}{(\lambda+a+m)(2\lambda)_m}.$$

In other words

$$\lambda_n=\frac{(\lambda+a)_{\lceil m/k\rceil}(1)_{\lfloor n/k\rfloor}}{(\lambda+a+1)_{\lfloor n/k\rfloor}(2\lambda)_{\lceil n/k\rceil}}. \tag{5.14}$$

Clearly $\lambda_n \approx Cn^2$, $C\neq 0$. As in the case of the B_n^λ's one can show that

$$\sum_0^\infty \{c_n^\lambda(x;a;k)\}^2/\lambda_n=\infty$$

for $x\in[-1,1]$, so the discrete support is empty. Thus, we have

$$\int_{-1}^{1} c_n^\lambda(x;a;k)c_m^\lambda(x;a;k)w^\lambda(x;a;k)\,dx=\frac{\pi\Gamma(2\lambda)}{(\lambda+a)}2^{-2\lambda+1}\lambda_n\delta_{m,n}, \tag{5.15}$$

λ_n is given by (5.14) $\lambda>0$, $\lambda+a>0$ and $w^\lambda(x;a;k)$ is defined by (5.12).

We now discuss the cases when the distribution function has a nonvanishing discrete part. The masses occur at the singularities of the corresponding continued fractions. The polynomials $\{B_n^\lambda(x;a;k)\}$ and $\{c_n^\lambda(x;a;k)\}$ are symmetric so the distribution function must be even and it suffices to consider the singularities of $\chi_1(x)$ and $\chi_2(x)$ in the half plane $\operatorname{Re} x\geq 0$. The points $x=\pm 1$ are singularities of $\chi_1(x)$ and $\chi_2(x)$ but, as we saw earlier, do not support discrete masses. The additional singularities of $\chi_1(x)$, if any, will coincide with the roots of (3.22), see (5.3), and these singularities occur if and only if (3.24) holds. When $a\geq 0$, (3.24) is violated and the discrete spectrum becomes empty. On the other hand, when $a<0$ (3.23) and the symmetry of the polynomials show that the discrete masses are located at $\pm x_j$, with

$$x_j=(\lambda+j+1)\left\{(\lambda+j+1)^2-a^2\right\}^{-1/2}, \qquad j=0,1,\cdots.$$

Since the x_j's are simple poles of $\chi_1(x)$ the mass J_j at $\pm x_j$ is the residue of $\chi_1(x)$ at $x=x_j$. The residues J_j can be easily computed.

The singularities of $\chi_2(x)$ can be similarly analyzed. It is easy to show, from (5.14) that the positivity condition (3.3) holds if and only if $\lambda>-1/2$, $\lambda+a>-1$, $\lambda(\lambda+a)>0$. The poles of $\chi_2(x)$ are solutions of

$$\lambda-ia\cot\theta=-j, \qquad j=0,1,\cdots. \tag{5.16}$$

The above equation has a solution satisfying $e^{i\theta}>1>e^{-i\theta}$ if and only if

$$a(\lambda+j)<0, \tag{5.17}$$

in which case the solutions are $\pm y_j$ with

$$y_j=(\lambda+j)\left\{(\lambda+j)^2-a^2\right\}^{-1/2}, \qquad j=0,1,\cdots. \tag{5.18}$$

The condition (5.17) identifies the set of jumps of $\sigma(x)$ as

I empty when $a\geq 0$, $\lambda\geq 0$,

II $\{\pm y_0\}$ when $\lambda>0$, $a<0$.

In all the above cases I and II it is further assumed that $\lambda + a > 0$. Here again, we shall not treat the case $\lambda < 0$; those cases will be investigated in a future work.

Acknowledgments. I am greatly indebted to Richard Askey and Waleed Al-Salam for giving me a preprint of their work [3]. The critical comments of Richard Askey and Jairo Charris of this work were extremely helpful.

REFERENCES

[1] N. I. Akhiezer and I. M. Glazman, *Theory of Linear Operators in Hilbert Space, Vol.* 2, Frederick Ungar, New York, 1961.

[2] N. I. Akhiezer, *The Classical Moment Problem*, Hafner, New York, 1976.

[3] W. Al-Salam, Wm. Allaway and R. Askey, *Sieved ultraspherical polynomials*, Trans. Amer. Math. Soc., 234 (1984), pp. 39–55.

[4] W. Al-Salam and T. Chihara, *Convolutions of orthogonal polynomials*, this Journal, 7 (1976), pp. 16–28.

[5] G. E. Andrews, *The Theory of Partitions*, Addison-Wesley, Reading, MA, 1977.

[6] R. Askey and M. E. H. Ismail, *The Rogers q-ultraspherical polynomials*, in Approximation Theory III, E. Cheney, ed., Academic Press, New York, 1980, pp. 175–182.

[7] ______, *A generalization of the ultraspherical polynomials*, in Studies in Pure Mathematics, P. Erdös, ed., Birkhaüser, Basel, 1983, pp. 55–78.

[8] ______, *Recurrence relations, continued fractions and orthogonal polynomials*, Memoirs Amer. Math. Soc., 300, 1984.

[9] R. Askey and J. Wilson, *Some basic hypergeometric orthogonal polynomials*, Memoirs Amer. Math. Soc., 319, 1985.

[10] T. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[11] A. Erdélyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[12] P. Nevai, *Orthogonal polynomials*, Memoirs Amer. Math. Soc., 213, 1979.

[13] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[14] F. Pollaczek, *Sur une génèralisation des polynômes de Legendre*, C. R. Acad. Sc., Paris, 228 (1949), pp. 1363–1365.

[15] ______, *Sur une génèralisation des polynômes de Jacobi*, Memorial des Sciences Mathèmatiques, 131, 1956.

[16] E. D. Rainville, *Special Functions*, Macmillan, New York, 1960.

[17] L. J. Rogers, *Second memoir on the expansion of certain infinite products*, Proc. London Math. Soc., 25 (1894), pp. 318–342.

[18] ______, *Third memoir on the expansion of certain infinite products*, Proc. London Math. Soc., 26 (1895), pp. 15–32.

[19] L. J. Slater, *Generalized Hypergeometric Functions*, Cambridge Univ. Press, Cambridge, 1966.

[20] C. Szegö, *On certain special sets of orthogonal polynomials*, Proc. Amer. Math. Soc., 1 (1950), pp. 731–737, reprinted in Collected Papers, Vol. 3, Birkhaüser, Boston, pp. 225–231.

[21] ______, *Orthogonal polynomials*, 4th edition, American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 16, No. 5, September 1985

014

AN ELEMENTARY PROOF OF LOCAL INVERTIBILITY FOR GENERALIZED AND ATTENUATED RADON TRANSFORMS*

ANDREW MARKOE† AND ERIC TODD QUINTO‡

Abstract. There is a great deal of current interest in inverting attenuated Radon transforms which occur in single photon emission tomography. These transforms are special cases of generalized Radon transforms R_μ which are defined by integrating a function over lines with respect to given positive C^2 measures μ.

As a positive result, we show R_μ is locally invertible. However, on the negative side, we present counterexamples to show that some smoothness assumptions on the measures are crucial for invertibility. The second example shows that limited angle tomography is not possible in general, even with somewhat smoother measures.

Introduction. One contribution of this paper is a proof of local invertibility for generalized Radon transforms on lines. This is an improvement on a result in the folklore of Fourier integral operators that this transform is locally invertible for positive smooth measures on lines in R^2. The improvements are that the measures need only be C^2 and positive instead of C^∞ and positive, and that the proof is elementary. We hope the techniques used here may lead to a global invertibility proof. Secondly we show by two counterexamples that some smoothness is essential for the measures. Example 1 gives a strictly positive bounded measure μ for which R_μ is not invertible, and Example 2 provides a nicer measure for which "limited angle tomography" would not be possible.

The attenuated Radon transform (Natterer [5]) of single photon emission tomography is a generalized Radon transform whose inversion would be of.great practical value in diagnostic medicine. For constant attenuation, invertibility is known (Bellini et al. [1], Markoe [4], Quinto [7], Tretiak–Metz [11]). However invertibility is unknown in general.

Let $\cdot$ be the standard inner product on R^2 and let $|\cdot|$ be the norm. For $\theta\in[0,2\pi]$ let $\bar{\theta}=(\cos\theta,\sin\theta)$ and $\theta^\perp=(-\sin\theta,\cos\theta)$. For $(\theta,s)\in[0,2\pi]\times R$, $L(\theta,s)=\{x\in R^2|\ x\cdot\bar{\theta}=s\}$ is the line perpendicular to $\bar{\theta}$ and s units from the origin. Let $\mu(x,\theta)$ be a C^2 function on $R^2\times[0,2\pi]$. We will always assume functions of θ are 2π periodic along with their derivatives. Let $f\in L^p_c(R^2)$, that is f is an L^p function of compact support. We define the generalized Radon transform R_μ by

$$R_\mu f(\theta,s)=\int_{-\infty}^{\infty} f(s\bar{\theta}+t\theta^\perp)\mu(s\bar{\theta}+t\theta^\perp,\theta)\,dt. \tag{1}$$

This is simply the integral of f over $L(\theta,s)$ in the measure $\mu(x,\theta)$ times Lebesgue measure on the line. Generalized Radon transforms on $(n-1)$-dimensional hyperplanes in R^n are defined in a similar manner [7]. Certain classes of generalized Radon transforms have been inverted [3], [6], [7], but even for $\mu>0$ there is no general inversion theorem.

*Received by the editors August 18, 1983, and in revised form June 6, 1984.

† Department of Mathematics, Rider College, Lawrenceville, New Jersey 08648.

‡ Department of Mathematics, Tufts University, Medford, Massachusetts 02155. The work of this author was partially supported by the National Science Foundation under grant MCS 82-01627.

Results. Our main theorem is a step on the way to inverting attenuated and generalized Radon transforms.

THEOREM. *Let μ be a strictly positive C^2 function of $(x,\theta)\in R^2\times[0,2\pi]$ of period 2π in θ along with its derivatives. Let $x_0\in R^2$ and $2<p$. Then there is a nonempty neighborhood, U_{x_0}, of x_0 such that R_μ is injective on domain $L^p(U_{x_0})$.*

The theorem is proved by a perturbation argument; the closer R_μ is to a classical Radon transform, the larger U_{x_0} is (see (8)).

For $\mu\in C^\infty$ this theorem can be proven on the domain of distributions of compact support by showing that R_μ is an elliptic Fourier integral operator (Guillemin–Sternberg [2], see also Quinto [6]), but the proof below uses only elementary analysis. This theorem generalizes directly to the X-ray transform on R^n.

Proof. First, there is no loss of generality in assuming μ is periodic of period π in θ. (Define $\bar{\mu}(x,\theta)=\frac{1}{2}(\mu(x,\theta)+\mu(x,\theta+\pi))$ then $R_{\bar{\mu}}$ satisfies the hypotheses of the theorem and is π periodic. Invertibility of $R_{\bar{\mu}}$ implies invertibility of R_μ.) Now assume $x_0=0$ and that the neighborhood $U_{x_0}\subset B(1)$ where $B(r)$ is the ball of radius r centered at 0. Then μ can be smoothly altered so that $\mu(x,\theta)=1$ for $|x|>2$ but μ is unchanged for $|x|\leqq 1$. For $g\in L^p_{\text{loc}}([0,2\pi]\times R)$ define

$$R^0_\mu g(x)=\int_0^{2\pi}\frac{g(\theta,x\cdot\bar{\theta})}{2\mu(x,\theta)}\,d\theta.$$

To a tomographer R^0_μ is a weighted back projection.

One proves R_μ: $L^p(B(1))\to L^p([0,2\pi]\times R)$ and R^0_μ: $L^p([0,2\pi]\times R)\to L^p_{\text{loc}}(R^2)$ for $1\leqq p\leqq\infty$ are continuous by using the definitions and Holder's inequality (Rudin [8]).

For $y\in R^2-0$ let $\arg y\in[0,2\pi)$ be the angle between the vector y and the positive x-axis, $y=|y|\arg y$.

A calculation using the fact that μ is a periodic as well as a polar integration on R^2 shows that for $f\in L^p_c(R^2)$

$$R^0_\mu R_\mu f(x)=\int_{R^2}\frac{f(x+y)\mu(x+y,\arg y+\pi/2)}{|y|\mu(x,\arg y+\pi/2)}\,dy. \tag{2}$$

Rewriting this we see

$$R^0_\mu R_\mu f(x)=f*\left(\frac{1}{|x|}\right)+\int_{R^2}\frac{f(x+y)M(x,y,\arg y)}{|y|}\,dy, \tag{3}$$

where

$$M(x,y,\theta)=\frac{\mu(x+y,\theta+\pi/2)-\mu(x,\theta+\pi/2)}{\mu(x,\theta+\pi/2)}.$$

Let $Kf(x)$ be the second integral in (3).

Define a norm on $C^1_c(R^2)$ by

$$\|f\|_{1,p}=\sum_{j=1}^{2}\left\|\frac{\partial f}{\partial x_j}\right\|_{L^p}.$$

This can be extended as a seminorm to distributions with L^p first derivatives.

The following two lemmas are the keys to the proof.

LEMMA 1. *For each $1<p<\infty$ there is a positive constant c_p such that for every $f\in L^p(B(1))$, $c_p\|f\|_{L^p}\leqq\|f*(1/|x|)\|_{1,p}$.*

LEMMA 2. *Let $2<p<\infty$ and let $c>0$. Then there exists a $\delta=\delta(p,c,\mu)$ such that for all $f\in L^p(B(\delta))$, $\|Kf\|_{1,p}\leqq c\|f\|_{L^p}$.*

The lemmas will be proved momentarily. Now, the lemmas in conjunction with (3) prove that for $c<c_p$ and $\delta=\delta(c,p,\mu)$, $\|R^0_\mu R_\mu f\|_{1,p}\geqq(c_p-c)\|f\|_{L^p}$ for $f\in L^p(B(\delta))$. This proves the theorem for $U_{x_0}=B(\delta)$ and $\delta<\delta(c_p,p,\mu)$ (see (8)).

It is interesting to note that the radius δ can be estimated by (8) in terms of the derivatives of μ and norms of certain Riesz operators. The closer R_μ is to a classical Radon transform ($\mu\equiv$ constant, $B\equiv 0$ in (8)) the larger δ is.

Proof of Lemma 1. Let R_jf be the standard Riesz transform of f, $j=1,2$ (Stein [9, p. 572]). Then $(\partial/\partial x_j)(f*(1/|x|))=LR_jf$ for some $L\neq 0$ [9, (20), p. 126]. For $f\in L^p(B(1))$, the derivative is understood distributionally. Now the equation $R_1^2+R_2^2=-\mathrm{id}$ [10, (2.9) p. 224] combined with the L^p continuity of R_j [10, Thm. 2.6, pp. 223–4] finish the proof of the lemma.

Proof of Lemma 2. The hypotheses on μ guarantee a positive uniform lower bound on μ as well as uniform upper bounds on μ and its derivatives of order less than or equal to two. This and the mean value theorem provide a constant B such that

$$\left|\frac{\partial M}{\partial x_j}\right|\leqq B,\quad\left|\frac{\partial M}{\partial y_j}\right|\leqq B,\quad |M|\leqq B|y|,\quad\left|\frac{\partial M}{\partial\theta}\right|\leqq B|y|,\tag{4}$$

for all (x,y,θ) and $j=1,2$. Now assume $f\in C_c^1(B(1))$. To prove the lemma we compute $\|Kf\|_{1,p}$ by bounding each first derivative of Kf by a multiple of $|f|*1/|x|$. First

$$\frac{\partial}{\partial x_j}Kf(x)=\int_{R^2}\frac{\partial}{\partial x_j}\frac{M(x,y,\arg y)}{|y|}f(x+y)\,dy$$
$$+\int_{R^2}\frac{M(x,y,\arg y)}{|y|}\frac{\partial}{\partial x_j}f(x+y)\,dy.$$

The derivative can be brought inside the integral since the integrand is bounded uniformly by an L^1 function in y. Then $(\partial/\partial x_j)f(x+y)$ can be replaced by $(\partial/\partial y_j)f(x+y)$ because of symmetry in $x+y$. An integration by parts combined with an elementary chain rule calculation yields

$$\frac{\partial}{\partial x_j}Kf(x)=\int_{R^2}\left[\frac{\partial}{\partial x_j}M(x,y,\arg y)-\frac{\partial}{\partial y_j}M(x,y,\arg y)\right.$$
$$\left.+\frac{y_j}{|y|^2}M(x,y,\arg y)-\frac{\partial}{\partial\theta}M(x,y,\arg y)\frac{(-1)^jy_{3-j}}{|y|^2}\right]\frac{f(x+y)}{|y|}\,dy.\tag{5}$$

Using the bounds from (4) in (5) and taking L^p norms proves that, for $f\in C_c^1(R^2)$,

$$\|Kf\|_{1,p}\leqq 8B\left\||f|*\frac{1}{|x|}\right\|_{L^p}.\tag{6}$$

Finally, the Hardy–Littlewood–Sobolev theorem [9, p 119] and Holder's inequality [8] prove, for $p>2$ and $f\in L^p(B(\delta))$

$$\left\| |f| * \frac{1}{|x|} \right\|_{L^p} \leqq d_p \|f\|_{L^r} \leqq d_p \pi^{1/2}\delta \|f\|_{L^p}, \tag{7}$$

where $r=2p/(2+p)$ and d_p is the L^r-L^p operator norm of this Riesz potential. Now (6) and (7) prove Lemma 2 for $f\in C_c^1(B(\delta))$ when

$$\delta=\delta(c,p,\mu)=\frac{c}{\left(8Bd_p\pi^{1/2}\right)}. \tag{8}$$

Since $C_c^1(B(\delta))$ is dense in $L^p(B(\delta))$, the lemma is true for $f\in L^p(B(\delta))$.

The basic argument above should work for the Radon transform on hyperplanes in R^n, however [6, (28)] would be used to calculate (2) and different norms would be used.

A Radon transform R_μ is *rotation invariant* if for each orthogonal transformation u and all f, $R_\mu(f\circ u^{-1})(u\theta,s)=R_\mu f(\theta,s)$, where $u\theta=\arg(u\bar{\theta})$.

Example 1. A strictly positive bounded $\mu(x,\theta)$ is constructed such that R_μ is rotation invariant, but R_μ is not invertible on $L_c^\infty(R^2)$.

Construction. Define $T_j=\{x|(j-1)/j\leqq|x|<j/(j+1)\}$. Let $|s|<1$. Now let $j_0(s)$ be the unique index such that the line $L(\theta,s)$ satisfies $T_{j_0}(s)\cap L(\theta,s)\neq\varnothing$ but $T_j\cap L(\theta,s)=\varnothing$ for $j<j_0(s)$. For $j\geqq j_0(s)$ define $a_j(s)=l(L(\theta,s)\cap T_j)$, where l is Lebesgue measure on the line. Note that neither j_0 nor a_j depends on θ as each T_j is rotation invariant. Let $|x|<1$ and $\theta\in[0,2\pi]$. Let $j_0=j_0(x\cdot\bar{\theta})$. Then for some $k\in\{0,1,2,\cdots\}$, $x\in T_{j_0+2k}\cup T_{j_0+2k+1}$. Define

$$\mu(x,\theta)=\begin{cases}1/2 & \text{if } x\in T_{j_0+2k},\\ a_{j_0+2k}(x\cdot\bar{\theta})/a_{j_0+2k+1}(x\cdot\bar{\theta}) & \text{if } x\in T_{j_0+2k+1}.\end{cases} \tag{9}$$

Also define $\mu=1$ for $|x|\geqq 1$ and all θ.

It is clear that μ is positive and rotation invariant. Here is an argument to show that μ is bounded. Let $j\in\mathbb{N}$, for $s\in[0,j/(j+1)]$ elementary calculus shows the maximum of $a_j(s)/a_{j+1}(s)$ occurs at $s=(j-1)/j$. Then it is straightforward to show $\lim_{j\to\infty}a_j((j-1)/j)/a_{j+1}((j-1)/j)=(\sqrt{2}-1)^{-1}$. This proves there is an $M>0$ such that for all j and all $s\in[-j/(j+1),j/(j+1)]$, $|a_j(s)/a_{j+1}(s)|\leqq M$. Examining the definition of μ, (9), proves μ is bounded.

Now define $f(x)=(-2)^{-j}$ on T_j and $f(x)=0$ for $|x|\geqq 1$. Clearly f is L_c^∞ and nontrivial. We now show that $R_\mu f=0$. Clearly for $|s|\geqq 1$, $R_\mu f=0$. Let $|s|<1$ and $j_0=j_0(s)$ then

$$\begin{aligned}R_\mu f(\theta)&=\int_{L(\theta,s)}f(x)\mu(x,\theta)\,dl(x)\\&=\sum_{k=0}^{\infty}\int_{T_{j_0+2k}\cap L(\theta,s)}f(x)\mu(x,\theta)\,dl(x)+\int_{T_{j_0+2k+1}\cap L(\theta,s)}f(x)\mu(x,\theta)\,dl(x)\\&=\sum_{k=0}^{\infty}\Big[(-2)^{-(j_0+2k)}\left(\frac{1}{2}\right)l\left(T_{j_0+2k}\cap L(\theta,s)\right)\\&\qquad+(-2)^{-(j_0+2k+1)}\left(a_{j_0+2k}(s)/a_{j_0+2k+1}(s)\right)l\left(T_{j_0+2k+1}\cap L(\theta,s)\right)\Big]\end{aligned}$$

as seen by pairing adjacent T_j. But this last sum is clearly zero.

Remarks. Although the f constructed in Example 1 is not smooth, it could easily be modified to be smooth. Just keep the rotation invariance and taper f off smoothly to zero on the union of the boundaries of T_j. Of course the definition of μ would have to be modified somewhat to account for the tapering of f. In any case, we omit the details, and merely note the interest in a smooth nontrivial function in the null space of a positive rotation invariant Radon transform.

The union of the boundaries of the T_j give an idea of the simplest possible structure of the zero set of a nontrivial function in the null space of a general Radon transform.

If μ is smooth and positive and R_μ is rotation invariant then R_μ is invertible on the domain of compactly supported distributions [7]. Hence counterexamples like Example 1 must have non-$C^\infty\mu$. Moreover, the result from [7] can be used to prove that, for any C^2 positive μ, R_μ is invertible when restricted to compactly supported radial functions. In contrast, in Example 1, the function f for which $R_\mu f=0$ is radial.

Note that the function $\mu(x,\theta)$ in Example 1 gets arbitrarily close to zero for some (x,θ) for which $x\cdot\bar{\theta}$ is near $j/(j+1)$ for each j.

Example 2. A positive $\mu(x,\theta)$ that is bounded and bounded away from zero is constructed such that limited angle tomography is not possible for R_μ. Specifically, a function $f\in L_c^\infty(R^2)$ is constructed so that $R_\mu f(\theta,s)=0$ for $\theta\in[-\pi/4,\pi/4]$ and all s.

Construction. Let A be the triangle with vertices $(-2,0)$, $(0,1)$, $(2,0)$ and let B be the triangle with vertices $(-2,0)$, $(0,-1)$, $(2,0)$. For $\theta\in[-\pi/4,\pi/4]$ and $x=(a,b)\in R^2$ define

$$(10)\qquad \mu(x,\theta)=\begin{cases}1 & \text{if } b<0,\\[4pt] \dfrac{l(B\cap L(\theta,x\cdot\bar{\theta}))}{l(A\cap L(\theta,x\cdot\bar{\theta}))} & \text{if } b\geqq 0 \text{ and } -2\cos\theta<x\cdot\bar{\theta}<2\cos\theta.\end{cases}$$

By elementary plane geometry, the ratio in (10) is constant for fixed θ and $L(\theta,x\cdot\bar{\theta})$ near $(2,0)$ or $(-2,0)$. Therefore μ can be defined for all x and $\theta\in[-\pi/4,\pi/4]$ to be bounded, bounded away from zero and discontinuous only along the x-axis, $b=0$. Now μ can easily be extended to $R^2\times[0,2\pi]$ to have these properties.

Let $f=\chi_A-\chi_B$. Then, by the definition of μ, $R_\mu f(\theta,s)=0$ for $\theta\in[-\pi/4,\pi/4]$ and all s.

These examples make clear the necessity of some smoothness restrictions on μ for R_μ to be invertible and somewhat more stringent restrictions on μ for "limited angle tomography" to be possible.

Acknowledgments. Professor Quinto would like to thank Frank Natterer, Alfred Louis and Uwe Heike for their kind hospitality during his stay at the Institut für Numerische und Instrumentelle Mathematik at Universität Münster as well as Don Solmon and David Finch for valuable conversations. The authors greatly appreciate the referee's advice in simplifying the proofs of Lemmas 1 and 2.

Note added in proof. Very recently Jan Boman has proven injectivity for R_μ with positive real analytic measures μ on lines as well as for Radon transforms with real analytic measures on certain other real analytic curves (to appear in Proc. Conference on the Constructive Theory of Functions, Varma, Bulgaria). He has an example of a positive $C^\infty\mu$ for which R_μ is not invertible. This example is neither an attenuated Radon transform nor an averaged attenuated transform in the sense of [7, (4.2)].

REFERENCES

[1] S. Bellini, M. Piacentini and C. Cafforio, *Compensation of tissue absorption in emission tomography*, IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-27 (1979), pp. 213–218.

[2] V. Guillemin and S. Sternberg, *Geometric Asymptotics*, American Mathematical Society, Providence, RI, 1977.

[3] A. Hertle, *On the injectivity of the attenuated Radon transform*, Proc. Amer. Math. Soc., 92 (1984), pp. 201–205.

[4] A. Markoe, *Fourier inversion of the attenuated X-ray transform*, this Journal, 15 (1984), pp. 718–722.

[5] F. Natterer, *On the inversion of the attenuated Radon transform*, Numer. Math., 32 (1979), pp. 431–438.

[6] E. T. Quinto, *The dependence of the generalized Radon transform on defining measures*, Trans. Amer. Math. Soc., 257 (1980), pp. 331–346.

[7] ———, *The invertibility of rotation invariant Radon transforms*, J. Math. Anal. Appl., 91 (1983), pp. 510–522; *Erratum*, 94 (1983), pp. 602–603.

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

[9] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton Univ. Press, Princeton, NJ, 1970.

[10] E. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton Univ. Press, Princeton, NJ, 1971.

[11] O. J. Tretiak and C. L. Metz, *The exponential Radon transform*, SIAM J. Appl. Math., 39 (1980), pp. 341–354.

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

DEGENERATE HOPF BIFURCATION AND NERVE IMPULSE*

ISABEL SALGADO LABOURIAU[†]

Abstract. It has been established by other authors that the clamped Hodgkin and Huxley equations for the nerve impulse have two branches of periodic solutions arising through Hopf bifurcation. In this paper these solution branches are shown to join, using singularity theory methods developed by M. Golubitsky and W. Langford (J. Differential Equations, 41 (1981), pp. 375–415). The equations are perturbed by varying parameters like temperature and average membrane permeability to certain ions. A hidden organizing centre for the equations is obtained, and its unfolding provides a topological description of the periodic orbits that bifurcate from the equilibrium solution.

1. Introduction. It has been known since the nineteenth century that the activity of nerve cells is accompanied by electrical changes. External factors, like the activity of other nerve cells, or sensory stimulation, can induce fluctuations on the electric potential across the cell membrane. If depolarization reaches a threshold level, a large perturbation is generated and travels as a wave along the axon (a cytoplasmatic outgrowth of the nerve cell). This propagated disturbance is called an *action potential.*

In 1952, Hodgkin and Huxley introduced a new experimental technique for investigating the electrical activity in isolated giant axons of squid [8], [9], [10], [11]. Because of their large diameter these axons can be threaded lengthwise with electrodes of low resistivity compared to the axon's protoplasm. In this way, spatial variations of current are eliminated over a length of axon, and the electric potential can be measured or controlled, with appropriate electronics. When current is applied to the "clamped" length of axon, it responds to suprathreshold stimulation with a stationary voltage pulse. The concentration of ions in the saline solution surrounding the axon can be varied, and in this way Hodgkin and Huxley established that these pulses, called stationary action potentials, appear as a consequence of variations in the membrane permeability to certain ions, especially sodium (Na^+) and potassium (K^+). The experiments are described by a system of nonlinear differential equations, known in the literature as the clamped Hodgkin and Huxley equations [12], [18]. These equations are presented in §2, below.

At the molecular level, the mechanisms of selective membrane permeability to Na^+ and K^+ are only incompletely understood. We quote Rinzel [18]: "The major Hodgkin and Huxley contributions were the separation of [the total ionic currents] into its individual ionic currents, [...] demonstration of their independence, and derivation of empirical expressions for them and the ionic conductances." In the formulation of the quantitative model it was found to be more convenient to make the ionic conductances depend on dummy variables instead of using them directly. The sodium flow was thus separated into a fast sodium activation and a slow Na^+ inactivation, in the notation of §2, the variables M and H, respectively. For the K^+ conductance a single slow equation (for the activation variable N) was found to be sufficient.

These empirical expressions fit the experimental results remarkably well. Quoting Rinzel [18] again: "Even though the Hodgkin and Huxley model is based upon a restricted set of data (voltage clamp) and for a single perparation, its qualitative

* Received by the editors December 2, 1983, and in revised form June 11, 1984.

[†] Mathematics Institute, University of Warwick, England. Present address, Departamento de Matemâtica Aplicada, Universidade do Porto, Rua das Taipas, 135, 4000 Porto, Portugal.

features are consistent with the classical signaling phenomena [...]. Among these features are propagation of a single impulse and trains of impulses, threshold properties for their initiation, appropriate dependence of propagation characteristics and thresholds upon temperature and other parameters, also subthreshold behaviour with a linear regime for small signals. These characteristics were demonstrated primarily by numerical calculations." Moreover, this qualitative model has been successfully applied, with modifications in the parameters, to nerve cells in other animals as well as to other excitable tissue, like muscles. For more details on the Hodgkin and Huxley model, see [1] and [18].

Under some circumstances "clamped" axons respond to maintained stimulation with a train of stationary pulses that lasts until the stimulus is withdrawn [5], [14]. Hassard [6] used the classical Hopf bifurcation theorem [15] to show that the Hodgkin and Huxley equations have small amplitude periodic solutions bifurcating from the steady state solution. The applied current I is used as a bifurcation parameter, and two families of periodic solutions were found (see §3 below). These results were improved by Rinzel and Miller [19], who developed a numerical method for tracing unstable periodic solution branches. Their findings are described in §3.

In this paper we use singularity theory techniques to study the way periodic solutions bifurcate from the equilibrium solution of the Hodgkin and Huxley equations. We compute the invariants that arise in Golubitsky and Langford's classification of degenerate Hopf bifurcations [2], thus establishing the existence of two temperatures where the Hodgkin and Huxley equations are contact equivalent to generalized Hopf bifurcation germs of codimension 1 (§§4, 5).

Following a suggestion of Ian Stewart, we perturb the system, in an attempt to force the two degenerate points to coalesce. We find that this can be done by varying the value of the average sodium permeability $\bar{g}_{\mathrm{Na}}$. A topological explanation for the bifurcation diagrams described in [19] is obtained by computing some of Golubitsky and Langford's invariants for the perturbed system, and studying its unfolding. We show that the two periodic solution branches first described in [6] join into a single loop, and we obtain evidence for the existence of bifurcation diagrams not previously described.

2. Nerve impulse equations.

The Hodgkin and Huxley equations [10] relate the difference of electric potential across the cell membrane (V) and the ionic conductances (M, N, and H), to the stimulus intensity (I), and temperature (T), as follows:

$$\text{H1}\qquad \begin{aligned} \frac{dV}{dt} &= -G(V,M,N,H)-I,\\ \frac{dM}{dt} &= \phi(T)[(1-M)\alpha_M(V)-M\beta_M(V)],\\ \frac{dN}{dt} &= \phi(T)[(1-N)\alpha_N(V)-N\beta_N(V)],\\ \frac{dH}{dt} &= \phi(T)[(1-H)\alpha_H(V)-H\beta_H(V)], \end{aligned}$$

or, writing $U=(V,M,N,H)$, $dU/dt=\Gamma(U,I,T)$.

Temperature enters the equations as $\phi(T)=3^{(T-6.3)/10}$, and the function G is given by

$$G(V,M,N,H)=\bar{g}_{\mathrm{Na}}M^3H(V-V_{\mathrm{Na}})+\bar{g}_{\mathrm{K}}N^4(V-V_{\mathrm{K}})+\bar{g}_L(V-V_L).$$

The constants

$$\bar{g}_{\mathrm{Na}}=120, \quad \bar{g}_{\mathrm{K}}=36, \quad \bar{g}_L=0.3,$$
$$V_{\mathrm{Na}}=-115, \quad V_{\mathrm{K}}=12, \quad V_L=10.599$$

were obtained from experimental data, and have the dimensions of conductance/cm^2 for the $\bar{g}$s, and millivolts for the V's. The functions α_J and β_J, $J=M, N, H$, are given by

$$\alpha_M(V)=\psi\left(\frac{V+25}{10}\right), \qquad \beta_M(V)=4e^{(V/18)},$$
$$\alpha_N(V)=\psi\left(\frac{V+10}{10}\right)0.1, \qquad \beta_N(V)=0.125e^{(V/80)},$$
$$\alpha_H(V)=0.007e^{(V/20)}, \qquad \beta_H(V)=(1+e^{(V+30)/10})^{-1}$$

where $\psi(0)=1$, and $\psi(x)=x/e^x-1$ for $x\neq 0$. Notice that $\alpha_J(V)+\beta_J(V)\neq 0$ for all V and J.

For any choice of the parameters I and T, a steady-state solution, $(V_*, M_*, N_*, H_*, I, T)$, of H1 must satisfy

$$J_*=\frac{\alpha_J(V_*)}{\alpha_J(V_*)+\beta_J(V_*)}=j_\infty(V_*), \qquad j=M,N,H$$

as well as $f(V_*)=G(V_*, m_\infty(V_*), n_\infty(V_*), h_\infty(V_*))=-I$. For the values of I, $\bar{g}_{\mathrm{ion}}$, V_{ion} used in this paper, f is monotonically increasing (see [13]) and therefore invertible. It is convenient to change coordinates in H1 so as to have the origin of $\mathbb{R}^4$ as the steady-state. The new variables are:

$$\lambda=f^{-1}(-I)=V_*, \qquad v=V-\lambda, \qquad u=(v,m,n,h),$$
$$m=M-m_\infty(\lambda), \qquad n=N-n_\infty(\lambda), \qquad h=H-h_\infty(\lambda).$$

In these coordinates the equations become $du/dt=\gamma(u,\lambda,T)$ or

$$\text{HH} \qquad \begin{aligned} \frac{dv}{dt}&=C(v,m,n,h,\lambda),\\ \frac{dj}{dt}&=\phi(T)\big[(1-j-j_\infty(\lambda))\alpha_j(v+\lambda)-(j-j_\infty(\lambda))\beta_J(v+\lambda)\big],\end{aligned}$$

with $j=m,n,h$ and

$$C(u,\lambda)=\bar{g}_{\mathrm{Na}}\left[m_\infty^3(\lambda)h_\infty(\lambda)(\lambda-V_{\mathrm{Na}})-(m+m_\infty(\lambda))^3(h+h_\infty(\lambda))(\lambda+v-V_{\mathrm{Na}})\right]$$
$$+\bar{g}_{\mathrm{K}}\left[n_\infty^4(\lambda)(\lambda-V_{\mathrm{K}})-(n-n_\infty(\lambda))^4(\lambda+v-V_{\mathrm{K}})\right]-\bar{g}_L v.$$

This change of variables reduces the problem to the form used in [2] for the classification of degenerate Hopf bifurcations (see §4). It also has the advantage of eliminating one error factor in numerical computations, as it is no longer necessary to compute f^{-1}, but, since λ decreases when I grows, all our pictures are mirror images of those obtained by other authors.

3. Preliminary results. The eigenvalues of $d\gamma(0,\lambda,T)/du$ were computed numerically for several values of λ and T within physiologically significant range, using analytical expressions for the partial derivatives of γ and NAG[1] subroutines for the linear algebra. For each value of T we found two real eigenvalues and a complex conjugate pair,

$$\sigma(\lambda,T)+i\theta(\lambda,T).$$

For temperatures greater than 29°C the complex eigenvalues have strictly negative real parts (see Fig. 1), while below 28.5°C the complex pair crosses the imaginary axis twice, transversely, confirming the findings of Hassard [6] for temperatures 6.3°C and 0°C.

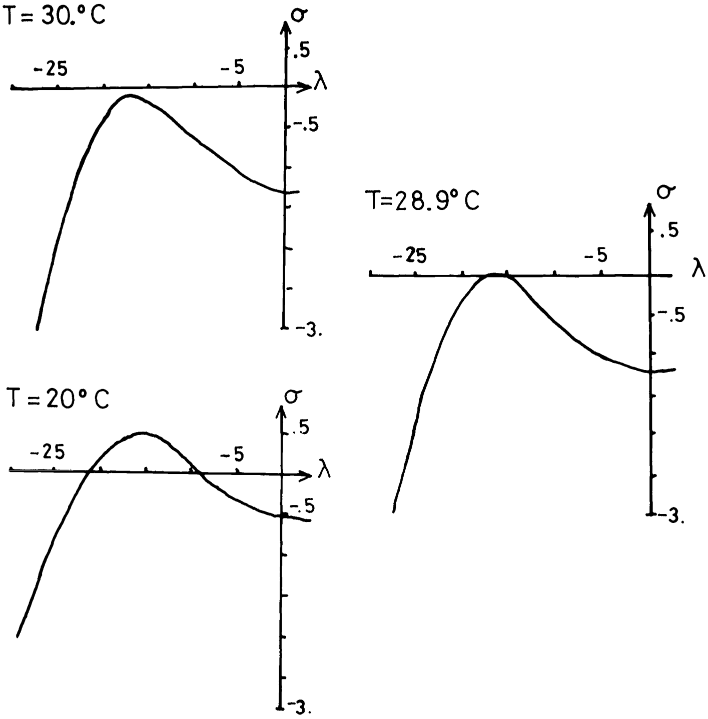


FIG. 1. *The real part $\sigma(\lambda,T)$ of the complex eigenvalues of $d\gamma(0,\lambda,T)/du$ plotted against λ.*

Therefore, for $T<28.5$°C, the equations HH satisfy the hypotheses of the Hopf theorem [15]. This establishes the existence of two families of periodic solutions bifurcating from the equilibrium $(0,\lambda,T)$ at the points where the eigenvalues are purely imaginary. The direction of bifurcation can be decided from the sign of a coefficient (μ_2 in the Hopf theorem) computed using the derivatives of γ at the bifurcation points. Hassard [6] showed that for temperatures $T=6.3$°C and 0°C, μ_2 is negative at both bifurcation points (see Fig. 2). Rinzel and Miller [19] obtained numerically a single periodic solution branch (Fig. 2) that contains the two local bifurcations described by Hassard.

[1] National Algorithm Library.

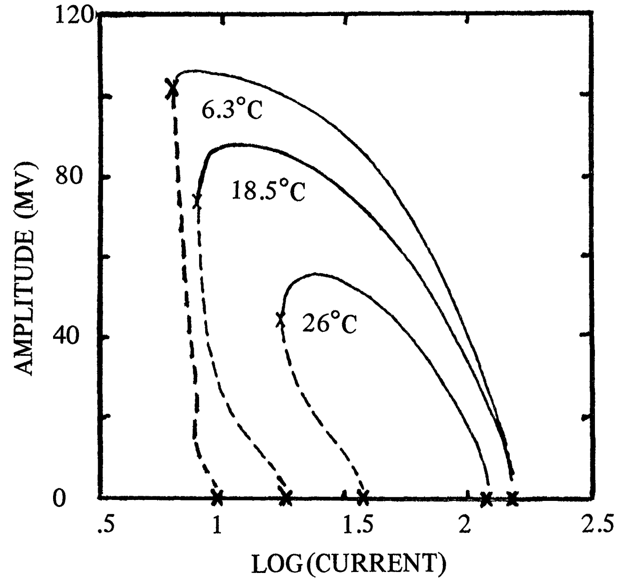


FIG. 2. *Amplitude* (V) *of periodic solutions of* H1 *as functions of applied current* (I) *for three temperatures. Dashed portions correspond to unstable limit cycles. Points of Hopf bifurcation and knees on amplitude curves are indicated by x. Reproduced from* [19].

For some temperature T_c between 28.5°C and 29°C, the curves of Fig. 1 must have zero as their maximum value, since the maximum changes sign between those temperatures. In other words, one of the λ-parametrized curves of eigenvalues on the complex plane is tangent to the imaginary axis. At this temperature Hopf's hypothesis of transverse crossing is violated, but a generalization of the Hopf theorem can be used to describe what happens for $T > T_c$.

4. Generalized Hopf bifurcation. In the remaining sections, we study the local behaviour of parametrized families of differential equations satisfying some (but not all) of the hypotheses in the Hopf theorem. The equivalence class of all maps that agree on some neighbourhood of a point p is called a *germ* at p. Operations like addition, composition, and differentiation, are defined on germs by performing the analogous operation on their representatives. In what follows we shall often make no distinction between a germ and its representatives.

We call the germ at $(0,\lambda_c)$ of a parametrized family of differential equations $\dot{x}=F(x,\lambda)$ a *generalized Hopf bifurcation germ* when $F(0,\lambda)\equiv 0$ and the derivative $d_xF(0,\lambda)$ has a pair of simple complex eigenvalues $\sigma(\lambda)+i\theta(\lambda)$ crossing the imaginary axis at λ_c, i.e. $\sigma(\lambda_c)=0$ and $\theta(\lambda_c)\neq 0$. We also require the derivative $d_xF(0,\lambda_c)$ to have no other eigenvalues of the form $ik\theta(\lambda_c)$ for k integer (nonresonance condition). This is part of the hypotheses in the Hopf theorem. When no ambiguity can arise, we call such germs *Hopf bifurcations*.

Two generalized Hopf bifurcation germs $\dot{x}=F(x,\lambda)$ and $\dot{y}=G(x,\lambda)$ at $(0,\lambda_1)$ and $(0,\lambda_2)$ respectively, are *contact equivalent* if there are smooth germs of changes of coordinates $X(x,\lambda)$ and of parameter $\Lambda(\lambda)$ deforming G into F, up to multiplication by an invertible matrix $T(x,\lambda)$:

$$F(x,\lambda)=T(x,\lambda)\cdot G(X(x,\lambda),\Lambda(\lambda)). \tag{4.1}$$

The coordinate changes X and Λ must also satisfy:

$$X(0,\lambda)\equiv 0, \qquad \Lambda(\lambda_1)=\lambda_2, \tag{4.2}$$

and

$$\det D_x X(0,\lambda_1)\neq 0, \qquad \frac{d\Lambda}{d\lambda}(\lambda_1)>0. \tag{4.3}$$

Locally, two contact equivalent Hopf bifurcation germs have the same number of periodic orbits for corresponding values of the parameter λ. Moreover, their amplitude graphs (like those of Fig. 2) have the same qualitative features. A discussion of the adequacy of this equivalence relation for the study of bifurcation problems can be found in [3] or in [21].

Golubitsky and Langford [2] use singularity theory techniques to classify and characterize (i.e., give conditions for occurrence of) generalized Hopf bifurcation germs under this equivalence. The problem is reduced to the study of zeros of a function g: $\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ of the form $g(x,\lambda)=xa(x^2,\lambda)$. Contact equivalent Hopf bifurcation germs are transformed into contact equivalent functions of this form, and periodic solutions into zeros of g. All the derivatives of $a(z,\lambda)$, with $z=x^2$, can be computed from those of F, in the original problem, and explicit formulae are given in [2] for some of them.

In the reduced form, the classical Hopf theorem becomes the case when both partial derivatives:

$$a_\lambda=\frac{\partial a}{\partial\lambda}(0,\lambda_c) \quad\text{and}\quad a_z=\frac{\partial a}{\partial z}(0,\lambda_c) \quad\text{with } z=x^2 \tag{4.4}$$

are nonzero. The coefficient μ_2, that describes the direction of bifurcation in the Hopf theorem, is given by

$$\mu_2=\frac{-a_z}{a_\lambda}. \tag{4.5}$$

These germs belong to the two contact equivalence classes represented by

$$g_\varepsilon(x,\lambda)=xa(x^2,\lambda)=x(x^2+\varepsilon\lambda) \tag{4.6}$$

with $\varepsilon=\pm 1$. The germs g_ε are *structurally stable* in the sense that any small perturbation by a smooth function is contact equivalent to g_ε. Using results of [4], it can be seen that this property reflects the structural stability of the family of differential equations $\dot{x}=F(x,\lambda)$.

Bifurcation problems where either a_z or a_λ vanish are called *degenerate*—these are the cases when one of the hypotheses in the classical Hopf theorem fails. If the only degeneracy is a nontransverse crossing of the imaginary axis, i.e.:

$$\sigma'(\lambda_c)=-a_\lambda=0 \quad\text{and}\quad a_z\neq 0\neq a_{\lambda\lambda}, \tag{4.7}$$

then the problem is contact equivalent, after reduction, to:

$$\tilde{g}_\varepsilon(x,\lambda)=x^3+\varepsilon\lambda^2 x \quad\text{with } \varepsilon=\pm 1. \tag{4.8}$$

The problems $\tilde{g}_\varepsilon$ are not structurally stable. For fixed $\alpha\neq 0$, the germs

$$\tilde{G}_\varepsilon(x,\lambda,\alpha)=x^3+\varepsilon\lambda^2 x+\varepsilon\alpha x \tag{4.9}$$

are not contact equivalent to $\tilde{g}_\varepsilon$, as can be seen in Fig. 3, where we show all the bifurcation problems associated to each of the $\tilde{G}_\varepsilon$. A Hopf bifurcation equivalent to $\tilde{g}_+$ can be perturbed into one that has periodic solutions near the origin, or into one having no small amplitude periodic solutions at all.

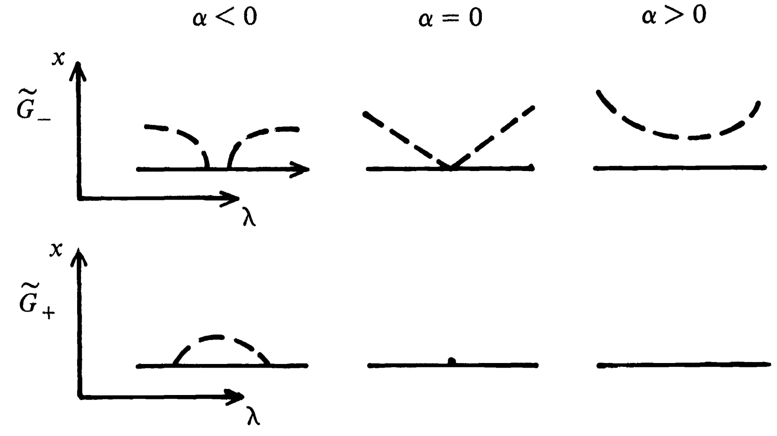


FIG. 3. *The least degenerate model of two Hopf bifurcations coalescing, redrawn from* [2]. *The ordinate x stands for amplitude, and dashed lines represent the periodic solutions, obtained rotating the picture around the rest state* (*solid lines*).

A parametrized family of perturbations, like $\tilde{G}_\varepsilon$, of a germ $g(x,\lambda)$, is called an *unfolding* of g. The *dimension* of an unfolding is the number of real parameters used. In example (4.6), the addition of the term $\varepsilon\alpha x$ stabilizes the germ $\tilde{g}_\varepsilon$, in the sense that any smooth perturbation of $\tilde{g}_\varepsilon$ is contact equivalent to a germ $\tilde{g}(x,\lambda)=\tilde{G}_\varepsilon(x,\lambda,\alpha)$, for fixed α. In general, an n-dimensional unfolding $G(x,\lambda,\alpha)$ of a germ g that stabilizes g in this sense, is called *versal*. The *codimension* of g is the least integer n for which a n-dimensional versal unfolding of g exists. The codimension of a germ can be computed without recourse to stability arguments (see [21], [2], [3]). When G is a versal unfolding of g, whose dimension equals the codimension of g, we say that G is a *universal unfolding* of g.

5. Another codimension 1 problem. We have computed $a_z(0,\lambda_c,T)$ at the two bifurcation points of HH for several values of T between 0°C and 29°C. This was done with a FORTRAN program that used Golubitsky and Langford's [2] formulae, and analytical expressions for the derivatives of $\gamma(u,\lambda)$. Some of the results are shown in Table 4.

Around the temperature T_c where the bifurcation points coalesce, a_z is positive. For the first bifurcation we have

$$\mu_2=\frac{-a_z}{a_\lambda}=\frac{a_z}{\alpha'(\lambda)}>0 \tag{5.1}$$

whereas at the second one $\mu_2<0$, and therefore the behaviour is described by the germs (4.9)—see Fig. 5. At the temperature T_c, HH is contact equivalent to $\tilde{g}_+$, and the point $(0,\lambda_c,T_c)$ is called an *organizing centre* [21], [22] for the system HH—the local dynamics at this point determines the behaviour of the periodic solution branch, with the

TABLE 4

λ	$I(\lambda)$	$T(\lambda)$	$a_z \times 10^4$
-12.07	39.08	26.322	-42.59
-14.40	56.43	28.288	-20.59
-15.56	67.00	28.763	-7.16
-16.14	72.80	28.851	-0.33
-16.16	73.00	28.853	-0.10
-16.22	73.62	28.855	0.60
-16.26	74.04	28.8566	1.06
-16.32	74.67	28.8579	1.75
-16.38	75.30	28.8580	2.44
-16.46	76.15	28.8566	3.61
-18.08	95.03	28.399	20.11
-20.02	122.14	26.085	32.17

$T(\lambda)$ = temperature where there is a Hopf bifurcation at λ. $a_z = a_z(0, \lambda, T(\lambda))$

temperature T playing the role of an unfolding parameter. From the discussion of the preceding section we know that for T in some interval (T_c, T') there will be no small amplitude periodic solutions near the constant solution $u=0$. For T in some interval (T'', T_c) we have something similar to the findings of Rinzel and Miller [19]: a single periodic solution branch bifurcating from equilibrium and rejoining it.

The analysis above does not apply at a lower temperature T_1, where one of the bifurcation points satisfies $a_z(0, \lambda_1, T_1) = 0$. The problem HH at (λ_1, T_1) is contact equivalent to:

$$f_\varepsilon(x, \lambda) = x^5 + \varepsilon\lambda x, \qquad \varepsilon = \pm 1 \tag{5.2}$$

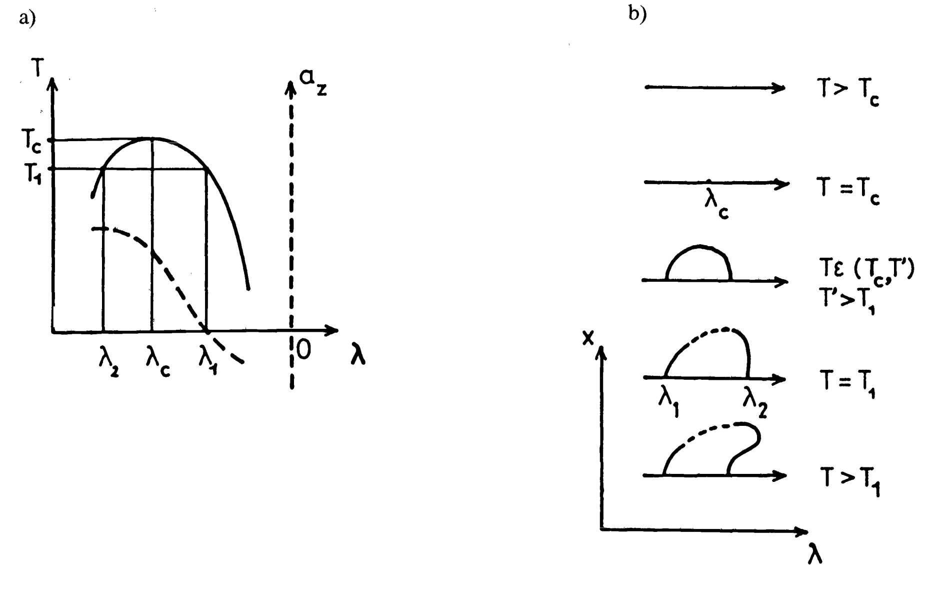


FIG. 5. *Bifurcation diagrams for the Hodgkin and Huxley equations.* a) *Schematic representation of* $T(\lambda)$ = *temperature where there is a Hopf bifurcation at* λ (*solid line*), *and* $a_z(0, \lambda, T(\lambda))$ (*dashed line*). *b*) *Corresponding bifurcation diagrams, with x standing for amplitude. Dashed lines correspond to hypothetical joining.*

a representative of the class of generalized Hopf bifurcation germs satisfying:

$$a_z=0, \quad a_\lambda \neq 0, \quad a_{zz} \neq 0 \quad \text{with } \varepsilon=\operatorname{sign}(a_\lambda). \tag{5.3}$$

The codimension of f_ε is 1 and it unfolds as

$$F_\varepsilon(x,\lambda,\alpha)=x^5+2\alpha x^3+\varepsilon\lambda x. \tag{5.4}$$

The bifurcation diagrams associated to F_ε are shown in Fig. 6, where it can be seen that F_ε always has a single periodic solution branch. For $\alpha<0$, a characteristic knee appears in this branch, following the change in the sign of μ_2.

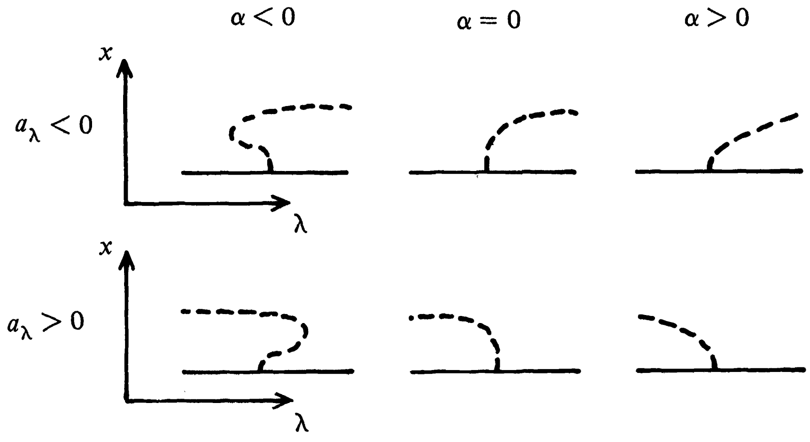


FIG. 6. *Bifurcation diagrams present on the codimension* 1 *problems* F_ε. *Conventions as in Fig.* 3.

The point $(0,\lambda_1,T_1)$ is thus a second organizing centre for HH—the change of sign of a_z helps to explain why a similar knee appears in Rinzel and Miller's picture (cf. Fig. 2). This analysis, however, applies only to a neighbourhood of $(0,\lambda_1,T_1)$ not containing the second (nondegenerate) Hopf bifurcation point present at the same temperature. If we could put the two analyses together, as in Fig. 5, the result would be remarkably similar to Fig. 2.

6. The perturbed Hodgkin and Huxley equations. In an attempt to bring the two organizing centres together into a highly degenerate point, we perturbed the equations HH, by varying the values of the average ion permeabilities $\bar{g}_{\mathrm{Na}}$ and $\bar{g}_{\mathrm{K}}$. The numerical proximity of the two points in question suggests this procedure as a natural way of studying the transition from one local behavior to the other. The more degenerate organizing centre thus obtained is called a *hidden organizing centre*, since the local dynamics of the perturbed equations around this point contains all the information about the change in the direction of bifurcation discussed in the preceding sections. Hidden organizing centres are discussed in [20] and [22].

Variations in the average ion permeabilities $\bar{g}_{\mathrm{ion}}$ did not change the pattern of two Hopf bifurcations below a critical temperature $T_c(\bar{g}_{\mathrm{ion}})$ and a single one at T_c, as in Fig. 1. Figure 7 shows $a_z(0,\lambda_c(\bar{g}_{\mathrm{Na}}),T_c(\bar{g}_{\mathrm{Na}}))$ for several values of the average sodium permeability. A search by the golden section [17] was carried out to determine the critical point $(\lambda_c(\bar{g}_{\mathrm{Na}}),T_c(\bar{g}_{\mathrm{Na}}))$ with increasing precision, until the computed values of a_z agreed to within four significant figures in successive computations. The result is that

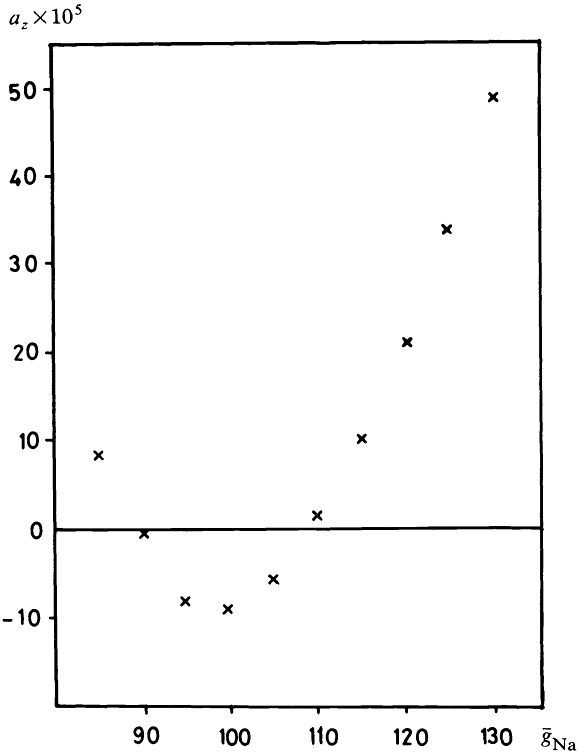


FIG. 7. $a_z(0,\lambda_c(g_{Na}), T_c(g_{Na}))$ *for the perturbed Hodgkin and Huxley equations.*

for some value $\bar{g}$ *of* g_{Na} *between* 105 *and* 110, *i.e. within* 10% *of the original value of* 120, *there is a single generalized Hopf bifurcation at* $(0,\lambda_c(\bar{g}), T_c(\bar{g}))$, *with*

$$a_z(0,\lambda_c(\bar{g}),T_c(\bar{g}))=0.$$

This is the hidden organizing centre.

If, in the process of bringing the two points together, we have not introduced any further degeneracy into the problem, the result is contact equivalent to some member of a one-parameter family of topologically equivalent codimension three germs, represented by:

$$(6.1) \qquad h_\varepsilon(x,\lambda,b)=x^5+2b\lambda x^3+\varepsilon\lambda^2 x, \qquad \varepsilon=\pm 1, \quad b\neq 0, \pm 1.$$

The family as a whole has codimension 2, universal unfolding

$$(6.2) \qquad H_\varepsilon(x,\lambda,b,\alpha,\beta)=h_\varepsilon(x,\lambda,b)+x[\operatorname{sign}(b)\beta\lambda+\alpha],$$

and is defined by the conditions:

$$(6.3) \qquad \begin{array}{ll} a_z=0=a_\lambda, & a_{zz}\neq 0\neq a_{\lambda\lambda}, \\ b=\dfrac{a_{z\lambda}}{|a_{zz}\cdot a_{\lambda\lambda}|^{1/2}}\neq 0, \ \pm 1, & \varepsilon=\operatorname{sign}(a_{zz}a_{\lambda\lambda}). \end{array}$$

The bifurcation diagrams for H_ε are shown in Figs. 8 and 9, together with the regions in (α,β,b)-space where they appear. See also [16].

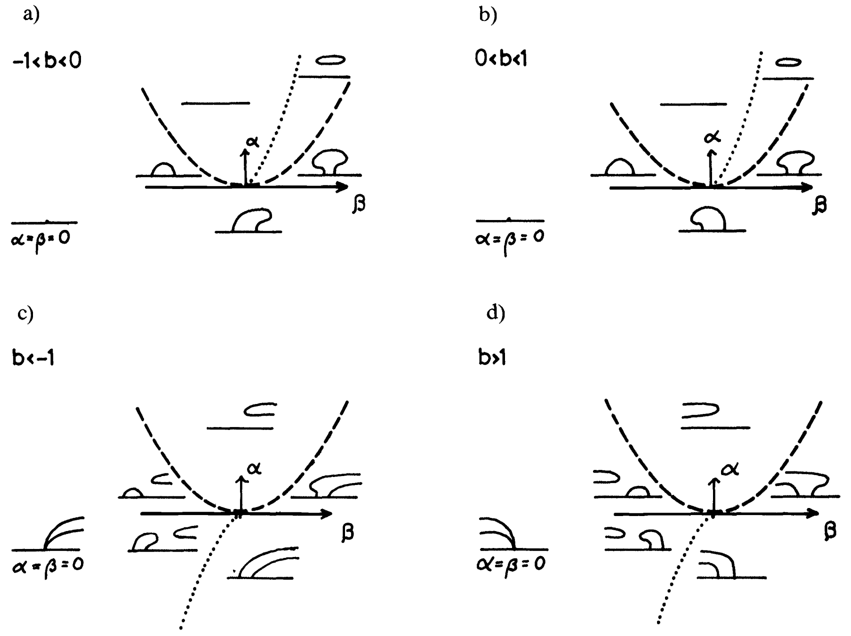


FIG. 8. *Universal unfolding of h_+, showing the regions of (α,β)-space corresponding to each bifurcation diagram. These regions are delimited by the curves*: $\sim\alpha=0$, $\sim\alpha=\beta^2/4(b-1)$, $\alpha=\beta^2/4$.

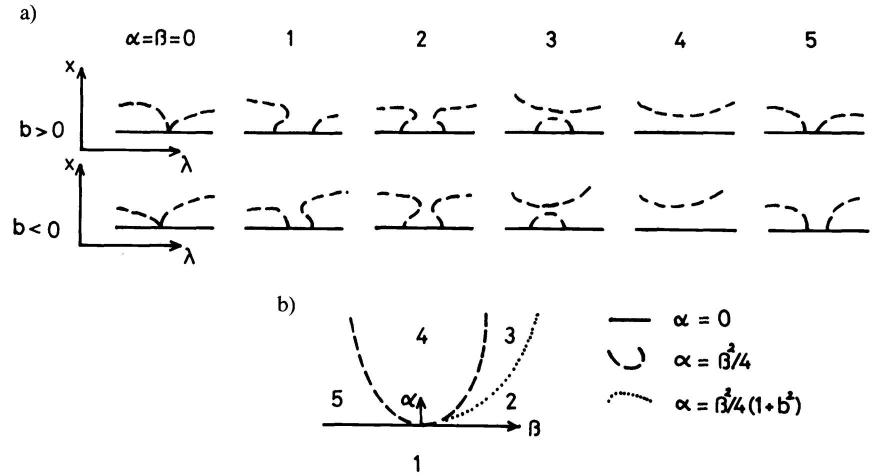


FIG. 9. *Bifurcation diagrams for H_-, redrawn from* [2]. a) *Conventions as in Fig.* 3. b) *Regions in* (α,β)*-space, corresponding to the diagrams in* a).

7. Conclusion. By perturbing the Hodgkin and Huxley equations we can establish that the original equations are contact equivalent to members of the family of germs $H(\cdots,b,\alpha,\beta)$ of (6.2). In this way, HH can be represented as a T-parametrized curve in one of the components of α-β-b-space. Changes in the values of the parameters deform this curve, and when we set $\bar{g}_{\mathrm{Na}}=\bar{g}$ it goes through a point $(\alpha,\beta,b)=(0,0,b)$ corresponding to the hidden organizing centre.

We can check the nondegeneracy conditions for contact equivalence to h_ε, using numerical estimates of the paremters b and ε, and of the second order derivatives of a. This is currently being done.

It is easy to predict from the results of [19] and from the bifurcation diagrams of Figs. 8 and 9, that the value of b has to be negative, since it determines the direction of bifurcation. The actual value of b is not irrelevant, because the germs (6.2) are not contact equivalent. Nevertheless, from a topological point of view (i.e. if we allow changes of coordinates that are continuous, but not necessarily differentiable) germs (6.2) can be grouped in four cases, two of them with b negative:

Case 1. $-1<b<0$. All the tree bifurcation diagrams discussed in §5 are present, thus providing a topological explanation of Fig. 2. The curve corresponding to HH should lie outside the shaded area in Fig. 10a. The perturbed curve ($\bar{g}_{\mathrm{Na}}=\bar{g}$) goes through the origin of the (α,β)-plane, and further perturbation could make it cross the shaded area, introducing two new bifurcation diagrams.

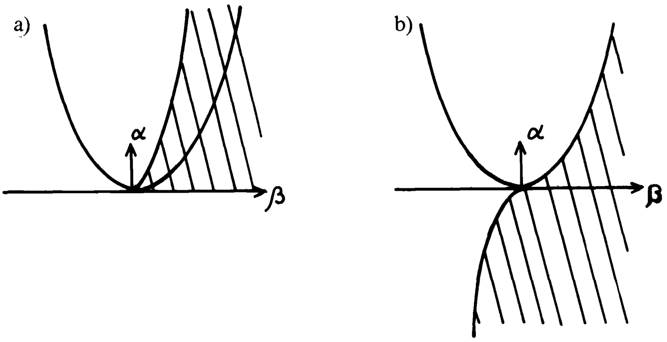


FIG. 10. *The system* HH *can be represented as a curve parametrized by* T *in* (α,β)*-space. Shaded areas correspond to diagrams not found by Rinzel* [19], a) $-1<b<0$, b) $b<-1$.

Case 2. $b<-1$. All the bifurcation diagrams of §5 are present, followed by a branch of periodic solutions whose amplitude is bounded away from zero. Such a branch cannot be found using the classical Hopf theorem, and for this reason it would easily be missed in a numerical tracing of the periodic solutions. The set of values of λ for which these periodic solutions exist, must contain an interval of infinite length for all choices of T and $\bar{g}_{\mathrm{Na}}$ (cf. Fig. 8c). This last feature, however, might be an artifact of the local analysis.

8. Discussion. The two cases outlined in the last section correspond to very different experimental results. If b is less than -1, the presence of the isolated solution branch implies that the linearly stable solution $u=0$ loses asymptotic stability. In this case, small perturbations of the equilibrium solution can have marked consequences, even if it is impossible to reach the stable periodic orbits. For a discussion of an analogous case, see the first chapter of [7].

The average permeabilities $\bar{g}_{\mathrm{ion}}$ would be our target for experiments, if b is found to be in the interval $(-1,0)$. Although they are expected to vary from cell to cell, and between different species, these permeabilities are difficult to control experimentally without introducing unknown factors. The experimental consequences of variations in $\bar{g}_{\mathrm{K}}$ are discussed in [14]. The test, in our case, would be to find cells with the behaviour that corresponds to the shaded area in Fig. 10a.

No matter what the value of b is, we have obtained a topological explanation of the results in [19]. Moreover, we can expect to find periodic solutions not described in

[19], like the isola in Fig. 8a, and the mushroom shape corresponding to part of the shaded area in Fig. 10b. The last diagram appears to be present in current clamp experiments with low Ca^{++} concentrations ([18] and references therein). However, in neither case is the second knee in Fig. 2 explained. This suggests the presence of an even more degenerate germ nearby.

Acknowledgments. For many useful discussions concerning this work, I should like to thank Ian Stewart. Thanks are also due to Martin Golubitsky for introducing me to the subject of degenerate Hopf bifurcation. Financial support was provided by the Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq) of Brazil.

REFERENCES

[1] J. CRONIN, *Mathematics of Cell Electrophysiology*, Marcel Dekker, New York, 1981.

[2] M. GOLUBITSKY AND W. F. LANGFORD, *Classification and unfoldings of degenerate Hopf bifurcations*, J. Differential Equations, 41 (1981), pp. 375–415.

[3] M. GOLUBITSKY AND D. SCHAEFFER, *A theory for imperfect bifurcation via singularity theory*, Comm. Pure Appl. Math., 32 (1979), pp. 21–98.

[4] L. C. GUIMARÃES, *Contact equivalence and bifurcation theory*, in Functional Differential Equations and Bifurcation, Lecture Notes in Mathematics 799, Springer, New York, 1980, pp. 140–151.

[5] R. GUTTMAN, S. LEWIS AND J. RINZEL, *Control of repetitive firing in squid axon membrane as a model for a neuroneoscillator*, J. Physiol., 305 (1980), pp. 377–395.

[6] B. HASSARD, *Bifurcation of periodic solutions of the Hodgkin–Huxley model for the squid giant axon*, J. Theoret. Biol., 71 (1978), pp. 401–420.

[7] B. HASSARD, N. D. KAZARINOFF AND Y-H. WAN, *Theory and Applications of Hopf Bifurcation*, London Mathematical Society, London, 1981.

[8] A. L. HODGKIN, A. F. HUXLEY AND B. KATZ, *Measurement of current-voltage relations in the membrane of the squid giant axon of Loligo*, J. Physiol., 116 (1952), pp. 449–472.

[9] A. L. HODGKIN AND A. F. HUXLEY, *Currents carried by sodium and potassium ions through the membrane of the giant axon in Loligo*, J. Physiol., 116 (1952), pp. 449–472.

[10] ———, *The components of membrane conductance in the giant axon of Loligo*, J. Physiol., 116 (1952), pp. 473–496.

[11] ———, *The dual effect of membrane potential on sodium conductance in the giant axon of Loligo*, J. Physiol., 116 (1952), pp. 497–506.

[12] ———, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiol., 117 (1952), pp. 500–504.

[13] A. HOLDEN, *Hopf bifurcation and repetitive activity of excitable cells*, in Mathematics in Medicine and Biology, Proceedings, Bari, Italy, 1983.

[14] A. HOLDEN, P. G. HAYDON AND W. WINLOW, *Multiple equilibria and exotic behaviour in excitable membranes*, Biol. Cybern., 46 (1983), pp. 167–172.

[15] E. HOPF, *Abzweigung einer periodischen Losung von einer stationaren Losung eines Differentialsystems*, Ber. Verh. Sachs. Akad. Wiss. Leipsig Math.-Nat., 94 (1942), pp. 3–22. Several versions of the same result appear in [7].

[16] I. S. LABOURIAU, *Note on the unfolding of degenerate Hopf bifurcation germs*, J. Differential Equations, to appear.

[17] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

[18] J. RINZEL, *Integration and propagation of neuroelectric signals*, in Studies in Mathematical Biology, S. A. Levin, ed., MAA Studies in Mathematics 15, Mathematics Association of America, Washington, DC, 1978.

[19] J. RINZEL AND R. N. MILLER, *Numerical calculation of stable and unstable periodic solutions to the Hodgkin-Huxley equations*, Math. Biosci, 49 (1980), pp. 27–59.

[20] D. SCHAEFFER, *Qualitative analysis of a model for boundary effects in the Taylor problem*, Math. Proc. Camb. Phil. Soc., 87 (1980), pp. 307–337.

[21] I. STEWART, *Applications of catastrophe theory to the physical sciences*, Physica D, 2 (1981), pp. 245–305.

[22] E. C. ZEEMAN, *Bifurcation, catastrophe and turbulence*, in New Directions in Applied Mathematics, Proceedings, Case Western Univ., Cleveland, OH, to appear.

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

NONGENERIC HOPF BIFURCATIONS IN FUNCTIONAL DIFFERENTIAL EQUATIONS*

HARLAN W. STECH†

Abstract. An algorithm is presented to prove the existence and determine stability of Hopf bifurcations in systems of functional differential equations. It is then applied to equations with "small" higher order terms and a study of bifurcations simultaneous to critical linear parts. The paper concludes with a thorough treatment of scalar integro-differential equations.

1. Introduction. In recent papers [9], [10] the author has considered the problem of determining the stability type of generic and particular nongeneric Hopf bifurcations in functional differential equations. The technique derived is based on the method of Lyapunov–Schmidt and reduces this problem to the evaluation of certain coefficients in an associated scalar bifurcation equation. Explicit formulae sufficient for resolving generic and first order nongeneric Hopf bifurcations are presented.

Our goal here is to consider the case of nongeneric Hopf bifurcations in further detail. In §2, we distill from [10] an algorithm for the computation of the bifurcation equation to any finite order. We then consider two situations in which the results of [10] provide no useful information. Section 3 is devoted to equations with "small" nonlinearities, while §4 presents a technique for proving the existence and stability of Hopf bifurcations existing simultaneously with a "center" for the linearized problem. This section also presents an elementary method for deciding the local stability properties of an equilibrium given that the linearization is critical. Section 4 provides a complete classification of all first order nongeneric Hopf bifurcations in a class of scalar integro-differential equations.

2. The method of Lyapunov–Schmidt. For $n \geqq 1$, $\mathbb{R}^n$ and $\mathbb{C}^n$ denote the usual Euclidean n-spaces of real and complex column vectors with norm $|\cdot|$. If $\xi = \mathrm{col}(\xi_1, \xi_2, \cdots, \xi_n)$ and $\zeta = (\zeta_1, \zeta_2, \cdots, \zeta_n) \in [\mathbb{C}^n]^T$, we define $\zeta \cdot \xi = \Sigma_i \zeta_i \xi_i$. The space $C = C[-1,0]$ is the usual Banach space of continuous $\mathbb{R}^n$-valued functions under the supremum norm $\|\cdot\|$. $C^*[-1,0]$ is its dual. For $y: [-1, a) \to \mathbb{C}^n$, $a > 0$ and $0 \leqq t < a$ we define $y_t: [-1,0] \to \mathbb{C}^n$ by $y_t(s) = y(t+s)$, $s \in [-1,0]$.

The system under study is

$$\dot{y}(t) = L(\alpha) y_t + H(\alpha; y_t) = \int_{-1}^{0} d\eta(\alpha; s) y_t(s) + H(\alpha; y_t), \tag{2.1}$$

where α is a parameter in some real Banach space $\mathscr{A}$ (e.g., see §5). Here, y is $\mathbb{R}^n$-valued and $\eta(\alpha; \cdot)$ is a real $n \times n$ matrix-valued function whose rows are in C^*. The functionals L and H are assumed sufficiently smooth so as to allow the computations that follow. This requires that L and H be continuous in $(\alpha, \psi) \in \mathscr{A} \times C$ and, for fixed $\alpha \in \mathscr{A}$, $H(\alpha; \psi)$ be k times continuously Fréchet differentiable in ψ; $k \geqq 7$. Accordingly, we

*Received by the editors June 22, 1984.

† Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

have the expansion

$$H(\alpha;\psi)=\sum_{j=2}^{k-1} H_j(\alpha;\psi^j)+\mathcal{O}\left(\|\psi\|^k\right), \tag{2.2}$$

where H_j are α-dependent, symmetric, bounded j-linear forms on C. Following [10], for $\psi\in C$ with derivatives $\psi^{(i)}\in C$; $i=2,\cdots,k-2$, we assume $L(\alpha)\psi$, $H(\alpha;\ \psi)$, and $H_j(\alpha;\ \psi^j)$ are C^{k-2} functions of $\alpha\in\mathscr{A}$.

The linearized stability of the zero solution is determined by the zeros of the characteristic equation $0=\det\Delta(\alpha;\ \lambda)$, where

$$\Delta(\alpha;\lambda)\equiv\lambda I-\int_{-1}^{0} d\eta(\alpha;s)e^{\lambda s}. \tag{2.3}$$

We assume that at $\alpha=\alpha_0$ there is a pair of simple complex-conjugate characteristic values $\pm i\omega_0$, $\omega_0>0$. All other characteristic values are assumed to satisfy $\mathrm{Re}\{\lambda\}<-\delta$ for some $\delta>0$, for all α in a neighborhood of α_0. (Equivalently, $\pm i\omega_0$ are simple, and the only characteristic values with $\mathrm{Re}\{\lambda\}\geqq 0$ at $\alpha=\alpha_0$.) It follows that there is a unique family of characteristic values (simple) $\lambda(\alpha)\equiv\mu(\alpha)+i\omega(\alpha)$ defined for all α near α_0 satisfying $\lambda(\alpha_0)=i\omega_0$. We denote by $\xi=\xi(\alpha)$ and $\xi^*=\xi^*(\alpha)$ any nonzero solutions of $\Delta(\alpha;\ \lambda(\alpha))\xi=\xi^*\Delta(\alpha;\ \lambda(\alpha))=0$, respectively. Both are unique up to scalar multiples. Furthermore, for $\Delta'(\alpha;\ \lambda)\equiv(\partial/\partial\lambda)\Delta(\alpha;\ \lambda)$, one has $\xi^*(\alpha)\Delta'(\alpha;\ \lambda(\alpha))\xi(\alpha)\neq 0$. Accordingly, we may define $\hat{\xi}=\xi^*/[\xi^*\Delta'\xi]$.

The fundamental solution $r=r(\alpha;\ \cdot)$ associated with the linearization of (2.1) solves $r(t)=0$ for $t<0$, $r(0)=I$ and

$$\dot{r}(t)=\int_{-1}^{0} d\eta(\alpha;s)r(t+s)$$

for $t>0$. Moreover, the Laplace transform of r is given by $\mathscr{L}(r)(\lambda)=\underline{\Delta}^{-1}(\alpha;\ \lambda)$, where Δ is given by (2.3). Associated with the characteristic roots $\lambda(\alpha)$, $\overline{\lambda(\alpha)}$ is a decomposition of $r(t)$ as $r(t)=r_Q(t)+r_P(t)$, where $|r_Q(t)|\leqq Me^{-\delta t}$ for $t\geqq 0$ and $r_P(t)\cdot h=2\mathrm{Re}\{[\hat{\xi}(\alpha)\cdot h]\xi(\alpha)e^{\lambda(\alpha)t}\}$ for all $t\in\mathbb{R}$ and $h\in\mathbb{R}^n$.

This decomposition induces a decomposition in the variation of constants formula for (2.1). When the method of Lyapunov–Schmidt is applied to the limiting equations obtained from this decomposition, one obtains the following. See [10] for details.

LEMMA 2.1. *Let y be any $2\pi/\nu$-periodic solution of* (2.1) *and define*

$$z(s)=\hat{\xi}\cdot\left[y(s)+\int_0^1 e^{-\lambda(\alpha)u}\left[\int_{-1}^{0} d\eta_v(\alpha;v-u)y(s+v)\right]du\right], \tag{2.4}$$

$$c=\frac{\nu}{2\pi}\int_0^{2\pi/\nu} e^{-\nu is}z(s)\,ds, \tag{2.5}$$

and $w(s)=z(s)-ce^{\nu is}$. Then $(y,\ w,\ c)$ solves the system

$$y(s)=2\mathrm{Re}\{c\varphi(s)+w(s)\xi\}+\int_0^{\infty} r_Q(u)H(\alpha;y_{s-u})\,du,\qquad s\leqq 0, \tag{2.6a}$$

$$\dot{w}(s)=\lambda(\alpha)w(s)+[\hat{\xi}\cdot H(\alpha;y_s)]^0,\quad s\leqq 0,\qquad \int_0^{2\pi/\nu} e^{-\nu is}w(s)\,ds=0, \tag{2.6b}$$

$$0=[\lambda(\alpha)-i\nu]c+\frac{\nu}{2\pi}\int_0^{2\pi/\nu} e^{-\nu iu}\hat{\xi}\cdot H(\alpha;y_u)\,du, \tag{2.6c}$$

where $\varphi(s)\equiv\xi(\alpha)e^{\nu is}$, *and*

$$[f(s)]^0\equiv f(s)-\frac{\nu}{2\pi}\int_0^{2\pi/\nu}e^{\nu i(s-u)}f(u)\,du.$$

Conversely, if (y,w,c) *is a solution of* (2.6) *with* (y,w) $2\pi/\nu$*-periodic, then* y *solves* (2.1).

We remark that the smallness and smoothness hypotheses on H are not used in the proof of this lemma. In fact, the result holds for continuous $H:\mathscr{A}\times\mathbb{R}\times C\to\mathbb{R}^n$ with $H(\cdot;\, t+2\pi/\nu,\,\cdot)=H(\cdot;\, t,\,\cdot)$. However, if the regularity and smallness hypotheses are satisfied, then, for fixed c sufficiently small there is a unique $2\pi/\nu$-periodic solution (y,w) of (2.6a, b). Both y and w are c^{k-1} functions of α, c, ν, and s. Substituting y into (2.6c), the problem of finding small periodic solutions is equivalent to finding small $c\neq 0$ solving (2.6c). Without loss of generality, we may take c real and positive, as other solutions correspond to shifts in phase. It is not difficult to show that the function defined by the right side of (2.6c) is an odd function of c. Thus, the real and imaginary parts may be written as

$$\text{(2.7a)}\qquad 0=\mu(\alpha)c+\operatorname{Re}\left\{\frac{\nu}{2\pi}\int_0^{2\pi/\nu}\hat{\xi}\cdot H(\alpha;\, y_u)e^{-\nu iu}du\right\},$$

$$\text{(2.7b)}\qquad 0=\omega(\alpha)-\nu+\operatorname{Im}\left\{\frac{\nu}{2\pi}\int_0^{2\pi/\nu}\frac{1}{c}\hat{\xi}\cdot H(\alpha;\, y_u)e^{-\nu iu}du\right\}.$$

The implicit function theorem (or simple iteration) shows that (2.7b) has a unique solution $\nu=\nu(\alpha;\ c)$ for all (α,c) near $(\alpha_0,0)$. The resulting "reduced" bifurcation equation

$$\text{(2.8)}\qquad 0=g(\alpha;\ c)=\mu(\alpha)c+K_3(\alpha)c^3+K_5(\alpha)c^5+\cdots$$

obtained by substituting into (2.7a) is then a real, α-dependent equation in c whose small roots correspond in a 1-1 manner with small periodic solutions $y(t)=2\operatorname{Re}\{c\varphi(t)\}+\mathcal{O}(c^2)$ of (2.1) with period near $2\pi/\omega_0$. (In the case of "classical" Hopf bifurcation, $\alpha\in\mathbb{R}$, and it is assumed that $\mu'(\alpha_0)\neq 0$. The implicit function theorem shows there exists a unique family of periodic solutions bifurcating from $y=0$ at $\alpha=\alpha_0$.)

The following theorem (see [10]) relates the stability type of any such periodic solution to the real equation

$$\text{(2.9)}\qquad \dot{c}=g(\alpha;\ c).$$

THEOREM 2.2. *There exists* $\varepsilon>0$ *such that for all* α *near* α_0, *and each* $c(\alpha)$ *solving* (2.8) *with* $|c(\alpha)|<\varepsilon$, *the associated periodic solution of* (2.1) *is orbitally asymptotically stable* (*unstable*) *if* $c=c(\alpha)$ *is asymptotically stable* (*unstable*) *as an equilibrium solution of* (2.9).

In fact, a more precise result holds: There is a local two-dimensional, α-dependent, center manifold invariant for and containing all small periodic solutions of (2.1). Any small equilibrium $c(\alpha)$ of (2.9) and the associated periodic solution of (2.1) share the same semistability properties (see [3] and [10]).

This theorem shows the importance of being able to approximate $g(\alpha;\ c)$ for c near zero. As shown above, this may be accomplished by computing the coefficients in the expansion $G(\omega;\ c,\nu)=(\lambda(\alpha)-i\nu)c+M_3(\alpha;\ \nu)c^3+M_5(\omega;\ \nu)c^5+\cdots$, where G denotes the right side of (2.6c). To calculate these through order c^{m+1}, one first computes the coefficients in the expansion $y(t)=2\text{Re}\{c\varphi(t)\}+\sum_{l=2}^{m} y^{(l)}(t)c^l$ by alternately using (2.6a) and (2.6b). It is not difficult to show that $y^{(l)}(t)$ has the form

$$y^{(l)}(t)=A_{l,l}e^{l\nu it}+A_{l,l-2}e^{(l-2)\nu it}+\cdots+A_{l,-l}e^{-l\nu it},$$

where $\overline{A_{l,j}}=A_{l,-j}$. (We define $y^{(1)}(t)=2\text{Re}\{c\varphi(t)\}=A_{1,1}e^{\nu it}+\bar{A}_{1,1}e^{-\nu it}$, with $A_{1,1}=\xi(\alpha)$.) These coefficients can be obtained through repeated application of the following lemma. (Recall $A_{l,-1}=\bar{A}_{l,1}$.)

LEMMA 2.3. *For $l\geqq 2$, if the coefficient of c^l in*

$$\sum_{j=2}^{l} H_j\left(\alpha;\left[\sum_{m=1}^{l-1} y_t^{(m)}c^m\right]^j\right)$$

is $\sum_j B_{l,j}(\alpha;\nu)e^{j\nu it}$, then

$$\text{(2.10)}\quad A_{l,j}(\alpha;\nu)=\begin{cases}\Delta^{-1}(\alpha;j\nu i)B_{l,j}(\alpha;\nu) & \textit{for } j\neq\pm 1,\\ \left(\Delta^{-1}(\alpha;\nu i)-\dfrac{1}{\nu i-\lambda(\alpha)}\xi[\hat{\xi}\cdot]\right)B_{l,1}(\alpha;\nu) & \textit{for } j=1.\end{cases}$$

Note that the singularity in the expression $\Delta^{-1}(\alpha;\lambda)-(1/(\lambda-\lambda(\alpha)))\xi[\hat{\xi}\cdot h]$ at $\lambda=\lambda(\alpha)$ is removable. In fact, for $h\in\mathbb{C}^n$, $\Delta^{-1}(\alpha;\lambda)h-(1/(\lambda-\lambda(\alpha)))\xi[\hat{\xi}\cdot h]$ is analytic in λ in a neighborhood of $\lambda(\alpha)$, with expansion

$$\begin{aligned}\Delta^{-1}(\alpha;\lambda)h&-\frac{1}{\lambda-\lambda(\alpha)}\xi[\hat{\xi}\cdot h]\\ &=d-[\hat{\xi}\Delta'(\alpha;\lambda(\alpha))d]\xi-\frac{1}{2}[\hat{\xi}\Delta''(\alpha;\lambda(\alpha))\xi][\hat{\xi}\cdot h]\xi+\mathcal{O}(\lambda-\lambda(\alpha)),\end{aligned}$$

where $d\in\mathbb{C}^n$ is any solution of

$$\Delta(\alpha;\lambda(\alpha))d=h-\Delta'(\alpha;\lambda(\alpha))\xi\cdot[\hat{\xi}\cdot h].$$

See [10] for details.

Through order c^5, one accordingly obtains $M_k(\alpha;\nu)=\hat{\xi}\cdot N_k(\alpha;\nu)$, where

$$N_3(\alpha;\nu)=3H_3(\alpha;\varphi^2,\bar{\varphi})+2H_2(\alpha;\bar{\varphi},A_{2,2}e^{2\nu i\cdot})+2H_2(\alpha;\varphi,A_{2,0}),$$

with $A_{2,2}$, $A_{2,0}$ the unique solutions of

$$\Delta(\alpha;2\nu i)A_{2,2}=H_2(\alpha;\varphi^2)$$

and

$$\Delta(\alpha;0)A_{2,0}=2H_2(\alpha;\varphi,\bar{\varphi}).$$

Similarly, at $\alpha=\alpha_0$,

$$\begin{aligned}N_5(\alpha;\nu)=&2H_2(\alpha;\varphi,A_{4,0})+2H_2(\alpha;\bar{\varphi},A_{4,2}e^{2\nu i\cdot})+2H_2(\alpha;A_{2,2}e^{2\nu i\cdot},\bar{A}_{3,1}e^{-\nu i\cdot})\\&+2H_2(\alpha;\bar{A}_{2,2}e^{-2\nu i\cdot},A_{3,3}e^{3\nu i\cdot})+2H_2(\alpha;A_{2,0},A_{3,1}e^{\nu i\cdot})+3H_3(\alpha;\varphi^2,\bar{A}_{3,1}e^{-\nu i\cdot})\\&+6H_3(\alpha;\varphi,\bar{\varphi},A_{3,1}e^{\nu i\cdot})+3H_3(\alpha;\bar{\varphi}^2,A_{3,3}e^{3\nu i\cdot})+6H_3(\alpha;\bar{\varphi},A_{2,2}e^{2\nu i\cdot},A_{2,0})\\&+6H_3(\alpha;\varphi,A_{2,2}e^{2\nu i\cdot},\bar{A}_{2,2}e^{-2\nu i\cdot})+3H_3(\alpha;\varphi,(A_{2,0})^2)+12H_4(\alpha;\varphi,\bar{\varphi}^2,A_{2,2}e^{2\nu i\cdot})\\&+12H_4(\alpha;\varphi^2,\bar{\varphi},A_{2,0})+4H_4(\alpha;\varphi^3,\bar{A}_{2,2}e^{-2\nu i\cdot})+10H_5(\alpha;\varphi^3,\bar{\varphi}^2),\end{aligned}$$

with $A_{3,3}$, $A_{3,1}$, $A_{4,2}$ and $A_{4,0}$ the unique solutions of

$$\begin{aligned}\Delta(\alpha;3\nu i)A_{3,3}&=H_3(\alpha;\varphi^3)+2H_2(\alpha;A_{2,2}e^{2\nu i\cdot},\varphi),\\A_{3,1}&=d-[\xi\Delta'd]-\frac{1}{2}[\hat{\xi}\Delta''\xi]M_3\xi,\end{aligned}$$

where d is any solution of $\Delta(\alpha;\lambda(\alpha))d=N_3-(\Delta'\xi)M_3$,

$$\begin{aligned}\Delta(\alpha;2\nu i)A_{4,2}=&2H_2(\alpha;\varphi,A_{3,1}e^{\nu i\cdot})+2H_2(\alpha;\bar{\varphi},A_{3,3}e^{3\nu i\cdot})\\&+2H_2(\alpha;A_{2,2}e^{2\nu i\cdot},A_{2,0})+6H_3(\alpha;\varphi,\bar{\varphi},A_{2,2}e^{2\nu i\cdot})\\&+3H_3(\alpha;\varphi^2,A_{2,0})+4H_4(\alpha;\varphi^3,\bar{\varphi}),\end{aligned}$$

and

$$\begin{aligned}\Delta(\alpha;0)A_{4,0}=&2H_2(\alpha;\varphi,\bar{A}_{3,1}e^{-\nu i\cdot})+2H_2(\alpha;\bar{\varphi},A_{3,1}e^{i\nu\cdot})\\&+H_2(\alpha;(A_{2,0})^2)+2H_2(\alpha;A_{2,2}e^{2\nu i\cdot},\bar{A}_{2,2}e^{-2\nu i})\\&+3H_3(\alpha;\varphi^2,\bar{A}_{2,2}e^{-2\nu i\cdot})+3H_3(\alpha;\bar{\varphi}^2,A_{2,2}e^{2\nu i\cdot})\\&+6H_3(\alpha;\varphi,\bar{\varphi},A_{2,0})+6H_4(\alpha;\varphi^2,\bar{\varphi}^2).\end{aligned}$$

Finally,

$$K_3(\alpha)=\operatorname{Re}\{M_3(\alpha;\omega(\alpha))\}\tag{2.11}$$

and

$$K_5(\alpha_0)=\operatorname{Re}\{M_5(\alpha_0;\omega_0)\}+\frac{\partial}{\partial\nu}\operatorname{Re}\{M_3(\alpha_0;\omega_0)\}\cdot\operatorname{Im}\{M_3(\alpha_0;\omega_0)\}.\tag{2.12}$$

3. Small nonlinearities. If H in (2.1) has the expansion

$$H(\alpha;\cdot)=H_\ell(\alpha;\cdot)+H_{\ell+1}(\alpha;\cdot)+\cdots\tag{3.1}$$

for $\ell>5$, then $K_3\equiv K_5\equiv 0$, and the reduced bifurcation equation must be computed to higher accuracy before the bifurcation structure of (2.1) can be understood. In this section, we use the algorithm of the previous section to compute (2.9) through order $c^{2\ell-1}$. Our results are then applied to a class of integro-differential equations.

Assume (3.1) with $\ell \geqq 3$. More precisely, we define $\mathscr{j}$ to be the smallest odd index j and $\mathscr{k}$ to be the smallest even index j for which $H_j(\alpha_0;\ \cdot) \not\equiv 0$, and assume that (3.1) holds for all α near α_0, with ℓ-$\min\{\mathscr{j}, \mathscr{k}\}$. For $j \geqq 3$ odd, we define $j^* = (j-1)/2$.

Observe from (2.6a, b) and (3.1) that $y(t) = cy^{(1)}(t) + c^\ell y^{(\ell)}(t) + \mathcal{O}(c^{\ell+1})$, and therefore $H(\alpha;\ y_t) = H(\alpha;\ cy_t^{(1)}) + \mathcal{O}(c^{2\ell-1})$. Thus $y^{(j)}$ for $j = \ell, \ell+2, \cdots, 2\ell-2$ can be obtained from the expansion of $H(\alpha;\ cy_t^{(1)})$. For $\ell \leqq j \leqq 2\ell-2$,

$$B_{j,j-2m}(\alpha;\ \nu) = \binom{j}{m} H_j(\alpha;\ \varphi^{j-m}, \bar{\varphi}^m), \qquad 0 \leqq m \leqq j,$$

and the associated $A_{j,j-2m}$ can be computed by Lemma 2.3. To obtain the required expansion of G, we substitute into (2.6c):

(3.2)

$$\begin{aligned} 0 &= [\lambda(\alpha) - i\nu]c + \frac{\nu}{2\pi}\int_0^{2\pi/\nu} e^{-\nu i u}\hat{\xi} H\left(\alpha;\ cy_u^{(1)} + \sum_{j=\ell}^{2\ell-2} c^j y_u^{(j)}\right) du + \mathcal{O}(c^{3\ell-2}) \\ &= [\lambda(\alpha) - i\nu]c + \hat{\xi}\cdot B_{\ell,1}c^\ell + \hat{\xi}\cdot B_{\ell+1,1}c^{\ell+1} + \cdots + \hat{\xi}\cdot B_{2\ell-2,1}c^{2\ell-2} \\ &\quad + \hat{\xi}\cdot\left[B_{2\ell-1,1} + \sum_{m=0}^{\ell-1} \ell\binom{\ell-1}{m} H_\ell(\alpha;\ \varphi^m, \varphi^{\ell-1-m}, A_{\ell,\ell-2m}e^{(\ell-2m)\nu i\cdot})\right]c^{2\ell-1} + \mathcal{O}(c^{2\ell}) \\ &= (\lambda(\alpha) - i\nu)c + \sum_{m=\ell}^{2\ell-1} M_m(\alpha;\ \nu)c^m + \mathcal{O}(c^{2\ell}). \end{aligned}$$

In fact, based on our knowledge of y through order $c^{2\ell-2}$, one can compute (2.6c) through order $c^{3\ell-3}$. We omit the calculations since they are similar to that of calculating $M_{2\ell-1}$-only more complicated.

From the imaginary part of (3.2),

$$\nu(\alpha;\ c) = \omega(\alpha) + \sum_{m=\ell}^{2\ell-1} \operatorname{Im}\{\hat{\xi}\cdot B_{m,1}(\alpha;\ \omega(\alpha))\}c^{m-1} + \mathcal{O}(c^{2\ell-2}), \tag{3.3}$$

allowing an expansion of (2.8) through order $c^{3\ell-3}$. Through order $c^{2\ell-1}$, this reads

$$\begin{aligned} 0 &= \mu(\alpha)c + \sum_{m=\ell}^{2\ell-1} \operatorname{Re}\{M_m(\alpha;\ \omega(\alpha))\}c^m \\ &\quad + \operatorname{Re}\left\{\hat{\xi}\cdot\frac{\partial}{\partial\nu}B_{\ell,1}(\alpha, \nu)\right\}\cdot\operatorname{Im}\{\hat{\xi}\cdot B_{\ell,1}(\alpha;\ \omega(\alpha))\}c^{2\ell-1} + \mathcal{O}(c^{2\ell}) \\ &= \mu(c) + K_{\mathscr{m}}(\alpha)c^{\mathscr{m}} + K_{\mathscr{m}+2}(\alpha)c^{\mathscr{m}+2} + \cdots + K_{2\ell-1}(\alpha)c^{2\ell-1} + \mathcal{O}(c^{2\ell}) \\ &= g(\alpha;\ c), \end{aligned} \tag{3.4}$$

where $\mathscr{m} = \min\{\mathscr{j}, 2\ell-1\}$. Observe that $\nu \to \varphi(\cdot)$ is a smooth function into $C([-1,0];\ \mathbb{C}^n)$, with $(d/d\nu)[\varphi(\cdot)](s) = is\varphi(s)$. The partial derivative in (3.4) is easily computed.

THEOREM 3.1. *Assume that at least one of the terms $K_j(\alpha_0)$; $m \leqq j \leqq 2\ell - 1$ is nonzero. Denote by $\imath$ the smallest such index. Then* (2.1) (*with* (3.1)) *has no more than* $1 + ((\imath - m)/2)$ *distinct families of periodic orbits bifurcating from* $y = 0$ *at* $\alpha = \alpha_0$. *Moreover, if* $g(\alpha;\ c(\alpha)) = 0$ *and* $c(\alpha) \to 0$ *as* $\alpha \to \alpha_0$, *then either*

$$c(\alpha)^{\imath - 1} \leqq \frac{\imath - m + 6}{2}\left|\frac{\mu(\alpha)}{K_\imath(\alpha)}\right|$$

or

$$c(\alpha)^{\imath - j} \leqq \frac{\imath - m + 6}{2}\left|\frac{K_j(\alpha)}{K_\imath(\alpha)}\right|$$

for some $j \in \{m, m+2, \cdots, \imath - 2\}$. *If* $\imath = m$, *then* $\mu(\alpha)c(\alpha)^{1-\imath} \to -K_\imath(\alpha_0)$ *as* $\alpha \to \alpha_0$.

Proof. The first assertion follows by the mean value theorem. The growth estimates on $c(\alpha)$ are obtained by minor modifications of the proof of [1, Lemma 7.2.1]. □

The following example illustrates these results.

Example 3.2. Consider the scalar integrodifferential equation

$$\dot{y}(t) = -\alpha \int_{-1}^{0} g(y(t+s))\, d\eta(s), \tag{3.5}$$

where $\alpha > 0$, η is of bounded variation on $[-1, 0]$, and g has the expansion $g(y) \equiv y + h(y) = y + g_\ell y^\ell + g_{\ell+1} y^{\ell+1} + \cdots$. The linearized equation has characteristic function $\Delta(\alpha;\ \lambda) = \lambda + \alpha \int_{-1}^{0} e^{\lambda s}\, d\eta(s)$. We assume that at $\alpha = \alpha_0 > 0$ there are simple characteristic values $\lambda = \pm i\omega_0$, $\omega_0 > 0$, with all other values having negative real parts. In particular, this implies $\int_{-1}^{0} d\eta(s) > 0$ since otherwise there is a real, nonnegative characteristic value for all $\alpha > 0$. Let $\lambda(\alpha)$ be as in §2.

The cosine transform of $d\eta$ is defined for $\nu \in \mathbb{R}$ by

$$f(\eta;\ \nu) = \int_{-1}^{0} \cos(\nu s)\, d\eta(s).$$

Then $f(\eta;\ 0) > 0$ and, from the real part of $\Delta(\alpha_0;\ i\omega_0) = 0$, $f(\eta;\ \omega_0) = 0$. The following lemma shows that the orders of the zeros of $\mu(\alpha)$ at α_0 and $f(\eta;\ \nu)$ at ω_0 are the same.

LEMMA 3.3. *Let* $f(\eta;\ \nu) = f_p \cdot (\nu - \omega_0)^p + \mathscr{O}((\nu - \omega_0)^{p+1})$ *and* $\lambda(\alpha) = \mu(\alpha) + i\omega(\alpha) = \mu_q \cdot (\alpha - \alpha_0)^q + \mathscr{O}((\alpha - \alpha_0)^{q+1}) + i[\omega_0 + \omega_1(\alpha - \alpha_0) + \mathscr{O}((\alpha - \alpha_0)^2)]$, *where* $p, q \geqq 1$ *and* f_p *and* μ_q *are not zero. Then* $p = q$. *Moreover* (*set* $p = q = m$),

(i) *If* $m = 1$, *then* $\mu_1 = -\omega_0 f_1 / |\Delta'(\alpha_0;\ i\omega_0)|^2$,

$$\alpha_0 \omega_1 = \omega_0 \left[1 + \alpha_0 \int_{-1}^{0} \cos(\omega_0 s)\, s\, d\eta(s)\right] \Big/ |\Delta'(\alpha_0;\ i\omega_0)|^2.$$

(ii) *If* $m > 1$, *then* $\mu_m = -(\alpha_0 \omega_1)^{m+1} f_m / \omega_0$, *where*

$$\alpha_0 \omega_1 = \omega_0 \Big/ \left[1 + \alpha_0 \int_{-1}^{0} \cos(\omega_0 s)\, s\, d\eta(s)\right].$$

Proof. For (i), simply expand $(d/d\alpha)(\Delta(\alpha;\ \lambda(\alpha))) \equiv 0$. For (ii), substitute the indicated expansion for $\lambda(\alpha)$ into $\Delta(\alpha;\ \lambda(\alpha)) = 0$. (Observe that for $m > 1$, $\Delta'(\alpha_0;\ i\omega_0) = [1 + \alpha_0 \int_{-1}^{0} \cos(\omega_0 s) s\, d\eta(s)]$ is nonzero by the simplicity of $i\omega_0$.) Details are omitted. □

To calculate $g(\alpha;\ c)$, we take $\xi = \xi^* = 1$ and obtain $\hat{\xi} = 1/\Delta'(\alpha;\ \lambda(\alpha))$. The bifurcation equation (2.6c) is most directly treated by substituting the higher order terms from (3.5) into (2.6c), reversing the order of integration, and using the $2\pi/\nu$-periodicity of y.

One obtains the bifurcation equation

$$(3.6)\quad 0=(\lambda(\alpha)-i\nu)c-\frac{\alpha}{\Delta'(\alpha;\lambda(\alpha))}\int_{-1}^{0}e^{\nu is}\,d\eta(s)\cdot\frac{\nu}{2\pi}\int_{0}^{2\pi/\nu}e^{-\nu is}h(y(s))\,ds.$$

By our previous discussions, $y(t)=y^{(1)}(t)c+y^{(\ell)}(t)c^{\ell}+\mathcal{O}(c^{\ell+1})$ with coefficients of $y^{(\ell)}$ given by

$$A_{\ell,\ell-2m}(\alpha;\nu)=$$

$$\binom{\ell}{m}g_{\ell}[(\ell-2m)\nu i-\Delta(\alpha;(\ell-2m)\nu i)]\begin{cases}\Delta^{-1}(\alpha;(\ell-2m)\nu i), & \ell-2m\neq\pm 1,\\ \Delta^{-1}(\alpha;\nu i) & \\ \quad-\dfrac{1}{(\nu i-\lambda(\alpha))\Delta'(\alpha;\lambda(\alpha))}, & \ell-2m=1.\end{cases}$$

From (3.6),

(3.7)

$$0=[\lambda(\alpha)-i\nu]c+\frac{i\nu-\Delta(\alpha;i\nu)}{\Delta'(\alpha;\lambda(\alpha))}\left[\binom{\mathscr{j}}{\mathscr{j}^*}g_{\mathscr{j}}c^{\mathscr{j}}+\binom{\mathscr{j}+2}{\mathscr{j}^*+1}g_{\mathscr{j}+2}c^{\mathscr{j}+2}+\cdots\right.$$
$$+\binom{2\ell-1}{\ell}g_{2\ell-1}c^{2\ell-1}+\ell g_{\ell}\sum_{m=0}^{\ell-1}\binom{\ell-1}{m}A_{\ell,\ell-2m}(\alpha_0;\omega_0)c^{2\ell-1}$$
$$\left.+\mathcal{O}\big(c^{2\ell-1}(|\alpha-\alpha_0|+|\nu-\omega_0|)\big)+\mathcal{O}(c^{2\ell})\right].$$

Recall that $2\mathscr{j}^*+1=\mathscr{j}$. Note that if $\mathscr{j}>2\mathscr{k}-1=2\ell-1$ then the terms preceding the summation symbol are all $\mathcal{O}(c^{2\ell})$. We proceed under the assumption $\mathscr{j}<\mathscr{k}$; the case $\mathscr{k}<\mathscr{j}$ is similar.

Observe that $[i\omega(\alpha)-\Delta(\alpha;\lambda(\alpha))]/\Delta'(\alpha;\lambda(\alpha))=[\lambda(\alpha)+\mathcal{O}(\mu(\alpha))]/\Delta'(\alpha;\lambda(\alpha))=\alpha\lambda'(\alpha)+\mathcal{O}(\mu(\alpha))$. Thus, through order $c^{2\mathscr{j}-3}$ the reduced bifurcation equation reads

$$(3.8)\qquad 0=\mu(\alpha)c+\alpha\mu'(\alpha)\left[\binom{\mathscr{j}}{\mathscr{j}^*}g_{\mathscr{j}}c^{\mathscr{j}}+\cdots+\binom{2\mathscr{j}-3}{\mathscr{j}-1}g_{2\mathscr{j}-3}c^{2\mathscr{j}-3}\right]$$
$$+\mathcal{O}\big(c^{\mathscr{j}}|\mu(\alpha)|\big)+\mathcal{O}(c^{2\mathscr{j}-1}).$$

Case 1: $\mu'(\alpha_0)\neq 0$. The implied function theorem implies that there is a unique family $y(\alpha;\cdot)$ of periodic orbits bifurcating from $y=0$ at $\alpha=\alpha_0$. For $g_{\mathscr{j}}<0$ (>0) the bifurcation is supercritical (subcritical). Theorem 2.2 shows that the orbits are orbitally asymptotically stable (unstable) for $\mu'(\alpha_0)g_{\mathscr{j}}<0$ (>0). The family grows at the rate $\|y_t(\alpha;\cdot)\|\sim\text{const}\cdot|\alpha-\alpha_0|^{1/(\mathscr{j}-1)}$. If the coefficients $g_{\mathscr{j}},\cdots,g_{2\mathscr{j}-1}$ are allowed to depend on α, one can construct for each j, $1\leqq j\leqq(\mathscr{j}+1)/2$, an equation (3.5) with precisely j bifurcating families.

Case 2: $\mu'(\alpha_0)=0$. Equation (3.8) must be computed through order $c^{2\mathscr{j}-1}$. By Lemma 3.3, $\Delta'(\alpha_0;i\omega_0)$ is real. This implies $(\partial/\partial\nu)\mathrm{Re}\{M_{\mathscr{j}}(\alpha_0;\omega_0)\}=0$, and the reduced bifurcation equation reads

(3.9)

$$
\begin{aligned}
0&=\mu(\alpha)c+\left[\alpha\mu'(\alpha)+\mathcal{O}(\mu(\alpha))\right]\left[\binom{j}{j^*}g_jc^j+\cdots+\binom{2j-1}{j}g_{2j-1}c^{2j-1}\right]\\
&\quad+\operatorname{Re}\left\{\frac{\omega_0 i}{\Delta'(\alpha_0;i\omega_0)}\cdot g_j\sum_{m=0}^{j-1} j\binom{j-1}{m}A_{j,j-2m}(\alpha_0;\omega_0)\right\}c^{2j-1}\\
&\quad+\mathcal{O}(c^{2j})+\mathcal{O}\left(c^{2j-1}|\alpha-\alpha_0|\right)\\
&=\mu(\alpha)c+\alpha\mu'(\alpha)\binom{j}{j^*}g_jc^j-\frac{\omega_0 g_j}{\Delta'(\alpha_0;i\omega_0)}\\
&\qquad\cdot\sum_{m=0}^{j^*}(j-2m)\binom{j}{m}\operatorname{Im}\{A_{j,j-2m}(\alpha_0;\omega_0)\}c^{2j-1}+\cdots\\
&=\mu(\alpha)c+\alpha_0\mu'(\alpha)\binom{j}{j^*}g_jc^j-\frac{\alpha_0(\omega_0 g_j)^2}{\Delta'(\alpha_0;\omega_0 i)}\left\{\frac{1}{2}\binom{j}{j^*}^2\frac{f''(\omega_0)}{(\Delta'(\alpha_0;i\omega_0))^2}\right.\\
&\qquad\left.+\sum_{m=0}^{j^*-1}(j-2m)^2\binom{j}{m}^2\frac{f((j-2m)\omega_0)}{|\Delta(\alpha_0;(j-2m)\omega_0 i)|^2}\right\}\\
&\qquad+\mathcal{O}\left(c^j\mu(\alpha)\right)+\mathcal{O}\left(c^{j+2}\mu'(\alpha)\right)+\mathcal{O}\left(c^{2j-1}|\alpha-\alpha_0|\right)+\mathcal{O}(c^{2j}).
\end{aligned}
$$

We assume $K_{2j-1}(\alpha_0)\neq 0$ and write $\mu(\alpha)=\mu_q\cdot(\alpha-\alpha_0)^q+\mathcal{O}((\alpha-\alpha_0)^{q+1})$ for some $q\geqq 2$.

If $q=2$, we scale $\alpha-\alpha_0=\beta\cdot c^{j-1}$ according to Theorem 3.1. One sees that if $0=\mu_2\beta^2+2\alpha_0\mu_2\binom{j}{j^*}g_j\beta+K_{2j-1}(\alpha_0)$ has no real solution, then no bifurcation takes place. If two distinct real roots exist, then two families of periodic orbits bifurcate from $y=0$ at $\alpha=\alpha_0$. Stabilities can be determined via Theorem 2.2.

If $q>2$ is odd, Newton's diagram [1] and the implicit function theorem imply the existence of a unique family of bifurcations defined supercritically (subcritically) if $\mu_q\cdot K_{2j-1}(\alpha_0)<0$ (>0). If q is even and $\mu_q\cdot K_{2j-1}(\alpha_0)>0$ there is no bifurcation, while for $\mu_q\cdot K_{2j-1}(\alpha_0)<0$ there is a unique family of nontrivial periodic orbits passing through $y=0$ at $\alpha=\alpha_0$. This family is defined in a full neighborhood of α_0. For $q>2$ odd or even, any such periodic solution is orbitally asymptotically stable (unstable) for $K_{2j-1}(\alpha_0)<0$ (>0).

4. Bifurcations simultaneous with a critical linearization. The algorithm of §2 and Theorem 2.2 lead to a rather direct means toward determining the stability type of the zero solution of (2.1) in the case when the linearized equation has characteristic values with zero real part. In particular, we assume (2.1) satisfies the regularity hypotheses of §2 and at $\alpha=\alpha_0\in\mathscr{A}$ the linearized equation has simple characteristic values $\pm i\omega_0$, $\omega_0>0$; all others have negative real parts. For a sufficiently small neighborhood $\mathscr{K}$ of α_0 we define the disjoint union $\mathscr{K}=\mathscr{K}_+\cup\mathscr{K}_-\cup\mathscr{K}_0$, where $\mathscr{K}_{+(-)}=\{\alpha\in\mathscr{K}\,|\,K_3(\alpha)>0\ (<0)\}$ and $\mathscr{K}_0=\{\alpha\in\mathscr{K}\,|\,K_3(\alpha)=0\}$. We may assume $\lambda(\alpha)$ as defined in §2 is defined over $\mathscr{K}$.

THEOREM 4.1. *Assume* $\lambda(\alpha)=\mu(\alpha)+i\omega(\alpha)$, *with* $\omega(\alpha)>0$ *and* $\mu(\alpha)\equiv 0$ *for all* $\alpha\in\mathscr{K}$.

(i) *If* $\alpha_0\in\mathscr{K}_-$, *then the zero solution of* (2.1) *is locally asymptotically stable.*

(ii) *If* $\alpha_0\in\mathscr{K}_+$, *then the zero solution of* (2.1) *is unstable.*

(iii) *If $\alpha_0 \in \mathscr{K}_0$ and $K_5(\alpha_0) < 0$ (>0) then for each $\alpha \in \mathscr{K}_+$ $(\mathscr{K}_-)$ near α_0 there is a unique nontrivial periodic solution $y(\alpha;\ \cdot)$ to* (2.1). *The periodic solution is orbitally asymptotically stable (unstable) and, as a function of α, is continuous, with $y(\alpha_0;\ \cdot) \equiv 0$.*

Proof. The reduced bifurcation equation reads

$$0 = K_3(\alpha)c^3 + K_5(\alpha)c^5 + \mathcal{O}(c^7) \tag{4.1}$$

for $\alpha \in \mathscr{K}$. Parts (i) and (ii) are immediate from Theorem 2.2. The assertions of (iii) are proved by dividing by c^3. For $K_5(\alpha_0) < 0$ (>0) the function $p(\rho) = K_3(\alpha) + K_5(\alpha)\rho + \mathcal{O}(\rho^2)$ is decreasing (increasing) in ρ near $\rho = 0$. The existence of a positive zero follows immediately from the given hypotheses. □

We remark that if $\mathscr{A} \subseteq \mathbb{R}$ and $K_3(\alpha_0) = 0$, $(\partial/\partial\alpha)K_3(\alpha_0) \neq 0$, then $\mathscr{K}_0$ is a submanifold of codimension 1 and both $\mathscr{K}_-$ and $\mathscr{K}_+$ are nonempty. In determining the influence of the nonlinearities on the stability of the zero solution $\dot{y} = \ell y_t + H(y_t)$, $H(0) = DH(0) = 0$ when the linearized problem has simple characteristic values $\pm i\omega_0$, $\omega_0 > 0$, we may consider $\alpha = H$ as an element of a suitable function space $\mathscr{A}$.

As an application, we consider an equation from the theory of viscoelasticity [2]. See also the discussion in [4].

Example 4.2. Consider the second order scalar equation

$$\ddot{x} + \alpha g(x) = \int_{-1}^{0} h(x(t+s))a(s)\,ds, \tag{4.2}$$

where $\alpha > 0$. We assume g, h are smooth with expansions $g(x) = x + g_2x^2 + g_3x^3 + \cdots$ and $h(x) = h_2x^2 + h_3x^3 + \cdots$, respectively. The "relaxation function" a is assumed to be C^2 and nonnegative, with $\int_{-1}^{0} a(s)\,ds = 1$.

The two-dimensional system derived from (4.1) is of the form (2.1). For $y = (y_1, y_2)^T = (x, \dot{x})^T$ we have

$$L(\alpha)y_t = \begin{bmatrix} 0 & -1 \\ -\alpha & 0 \end{bmatrix} y(t)$$

and

$$H(\alpha;\ y_t) = \begin{bmatrix} 0 \\ -\alpha[g(y_1(t)) - y_1(t)] + \int_{-1}^{0} h(y_1(t+s))a(s)\,ds \end{bmatrix}.$$

Clearly,

$$H_j(\alpha;\ [y_t]^j) = \begin{bmatrix} 0 \\ -\alpha g_j y_1^j(t) + \int_{-1}^{0} h_j y_1^j(t+s)a(s)\,ds \end{bmatrix}.$$

The only characteristic values are $\lambda(\alpha) = \pm i\sqrt{\alpha}$. The required characteristic vectors are easily determined and, after a somewhat tedious calculation, one obtains from the formulae of §2 that the bifurcation equation (2.6c) through order c^3 reads

(4.3)

$$0 = [\sqrt{\alpha}\, i - \nu i]c + \frac{1}{2i\sqrt{\alpha}} \left\{ 3\left(-\alpha g_3 + h_3 \int_{-1}^{0} e^{\nu i s} a(s)\,ds \right) \right.$$
$$\left. + 2\left(-\alpha g_2 + h_2 \int_{-1}^{0} e^{\nu i s} a(s)\,ds \right)\left(-\frac{5}{3} g_2 - \frac{h_2}{\alpha^2} \int_{-1}^{0} e^{2\nu i s} a(s)\,ds + \frac{2h_2}{\alpha} \right) \right\} c^3 + \mathcal{O}(c^5).$$

The value of $K_3(\alpha)$ is therefore given by

$$\text{(4.4)}\quad \alpha K_3(\alpha)=\frac{1}{2}\left[3h_3-\frac{10}{3}h_2g_2\right]\sqrt{\alpha}\int_{-1}^{0}\sin(\sqrt{\alpha}\,s)a(s)\,ds$$

$$+\frac{1}{6}g_2h_2\cdot 2\sqrt{\alpha}\int_{-1}^{0}\sin(2\sqrt{\alpha}\,s)a(s)\,ds+\frac{2h_2^2}{\sqrt{\alpha}}\int_{-1}^{0}\sin(\sqrt{\alpha}\,s)a(s)\,ds$$

$$-\frac{h_2^2}{3}\left[2\int_{-1}^{0}\cos(\sqrt{\alpha}\,s)a(s)\,ds\cdot\frac{1}{2\sqrt{\alpha}}\int_{-1}^{0}\sin(2\sqrt{\alpha}\,s)a(s)\,ds\right.$$

$$\left.+\frac{1}{\sqrt{\alpha}}\int_{-1}^{0}\sin(\sqrt{\alpha}\,s)a(s)\,ds\cdot\int_{-1}^{0}\cos(2\sqrt{\alpha}\,s)a(s)\,ds\right].$$

Note that $\lim_{\alpha\to 0^+}\alpha K_3(\alpha)=h_2^2\int_{-1}^{0}sa(s)\,ds$. Thus, if $h_2\neq 0$ the zero solution of (4.1) is locally asymptotically stable for all small $\alpha>0$.

Unfortunately, the expression for $K_3(\alpha)$ suggests no intuitive criterion for the stability/instability of the zero solution. However, in the special case of odd nonlinearities more specific results are possible. Here,

$$\text{(4.5)}\qquad K_3(\alpha)=\frac{3h_3}{2\sqrt{\alpha}}\int_{-1}^{0}\sin(\sqrt{\alpha}\,s)a(s)\,ds$$

and we conclude that for all $\alpha\leqq\pi^2$ the zero solution is locally asymptotically stable (unstable) for $h_3>0$ (<0).

With increasing α, the zero solution may or may not change stability depending on the choice of a. For example, if a is nondecreasing and nonlinear on $[-1,0]$, then $h_3\cdot K_3(\alpha)<0$ for all $\alpha>0$ provided $h_3\neq 0$. In general, integrating by parts we obtain

$$\text{(4.6)}\qquad K_3(\alpha)=\frac{3h_3}{2\alpha}\left[a(-1)\cos(\sqrt{\alpha})-a(0)+\int_{-1}^{0}\cos(\sqrt{\alpha}\,s)a'(s)\,ds\right].$$

If $h_3\neq 0$ and $a(0)>a(-1)$, then integration by parts shows the integral is $o(1)$ as $\alpha\to+\infty$ and, therefore, $h_3\cdot K_3(\alpha)<0$ for all sufficiently large α. There are at most finitely many values of α at which a change in stability can occur.

If $a(0)<a(-1)$, there exists an infinite number of zeros of $K_3(\alpha)$ at $\alpha=\alpha_n\to+\infty$. These values are asymptotic to the roots β_n of $a(-1)\cos(\sqrt{\beta_n})-a(0)=0$ as $\alpha_n\to+\infty$. One computes $K_3'(\alpha_n)\sim 3h_3a(-1)\sin(\sqrt{\beta_n})/(4(\sqrt{\beta_n})^3)\neq 0$ as $n\to+\infty$. Thus, for all large n there is a unique family of periodic solutions to (4.1) bifurcating from $y=0$ at $\alpha=\alpha_n\sim\beta_n$. By direct calculation from the formulae of §2, then integrating by parts, one obtains that at $\alpha=\alpha_n$,

$$K_5(\alpha)=-h_3\left[\frac{9}{2\alpha}\int_{-1}^{0}\cos(\sqrt{\alpha}\,s)a(s)\,ds+\frac{3}{16(\sqrt{\alpha})^3}\int_{-1}^{0}\sin(3\sqrt{\alpha}\,s)a(s)\,ds\right]$$

$$\cdot\left[-\alpha g_3+h_3\int_{-1}^{0}\cos(\sqrt{\alpha}\,s)a(s)\right]$$

$$\sim-\frac{9h_3g_3}{2\sqrt{\beta_n}}\sin(\sqrt{\beta_n})a(-1)+o\left(\frac{1}{\sqrt{\beta_n}}\right)$$

as $n \to +\infty$. Theorem 4.1 implies that bifurcates are subcritical (supercritical) for $g_3 > 0$ (< 0). The sign of $K_5(\alpha_n)$ determines their stability. On the other hand, if $g_3 = 0$ then as $\alpha_n \to \infty$,

$$K_5(\alpha_n) \sim \frac{9}{2}\left[h_3 a(-1)\sin\left(\sqrt{\beta}_n\right)\right]^2 / \beta_n^2 + o\left(\beta_n^{-2}\right)$$

and for all large n the bifurcations are unstable.

5. Nongeneric bifurcations in integrodifferential equations. In this section, we present a description of certain nongeneric Hopf bifurcations for the scalar equation

$$\dot{y}(t) = -\int_{-1}^{0} g(y(t+s))\,d\eta(s), \tag{5.1}$$

where $\eta \in C^*$ and $\int_{-1}^{0} d\eta(s) > 0$. The function g has the expansion $g(y) = y + h(y) = y + g_2 y^2 + g_3 y^3 + \cdots$. Our results complement those of §3 in that we do not assume the problem has "small" nonlinearities. Rather, our goal is to understand the relationships between η and h that make (5.1) generic or not. The choice of $\mathscr{A}$ depends on the variation of h, η or both, and will be clear from the context.

This section is related to [4], [5], [6], [7], and [8] in that all consider certain aspects of the asymptotic behavior of bounded solutions to equations of the type (5.1). The cosine transform $f(\eta;\ \nu) = \int_{-1}^{0} \cos(\nu s)\,d\eta(s)$ plays a central role in the work of Staffans. In particular, if $f(\eta;\ \nu) > 0$ for all ν, then $g(y(t)) \to 0$ as $t \to \infty$ [7]. If $f(\eta;\ \nu) \geqq 0$ and the zero set of $f(\eta;\ \cdot)$ is bounded on $\mathbb{R}$, then the existence of nontrivial periodic solutions to (5.1) imposes severe restrictions on g [8].

Assume that at $\eta = \eta_0$ the usual spectral conditions hold for the linearization of (5.1), and write $\lambda(\eta) = \mu(\eta) + i\omega(\eta)$ for the characteristic value near $i\omega_0$, $\omega_0 > 0$ for η near η_0. The bifurcation equation (2.6c) can again be obtained from (3.6), and considering the inductive nature of its construction, it has the form

$$0 = [\lambda(\eta) - i\nu]c - \frac{1}{\Delta'(\eta;\ \lambda(\eta))}\int_{-1}^{0} e^{\nu i s}\,d\eta(s)\left[c_3(\nu,g)c^3 + c_5(\nu,g)c^5 + \cdots\right], \tag{5.2}$$

where $c_j(\nu,g) = \binom{j}{j*}g_j - G_j^*(\nu, g_2, g_3, \cdots, g_{j-1})$ and G^* is a polynomial in $g_2, \cdots, g_{j-1}$ whose coefficients depend on ν and η. Since

$$\operatorname{Re}\left\{\frac{1}{\Delta'(\eta_0;\ i\omega_0)}\int_{-1}^{0} e^{i\omega_0 s}\,d\eta_0(s)\right\} = f'(\eta_0;\ \omega_0)/\left|\Delta'(\eta_0;\ i\omega_0)\right|^2,$$

we first consider

Case 1. $f'(\eta;\ \omega_0) \neq 0$.

Note that this implies that the characteristic value $i\omega_0$ is simple. In a sufficiently small neighborhood $\mathcal{O}$ of η_0 in C^* one can partition $\mathcal{O} = \mathcal{O}_+ \cup \mathcal{O}_0 \cup \mathcal{O}_-$, where $\mathcal{O}_0 = \{\eta \in \mathcal{O} \mid \mu(\eta) = 0\}$, $\mathcal{O}_{+(-)} = \{\eta \in \mathcal{O} \mid \mu(\eta) > 0\ (< 0)\}$. By the implicit function theorem, $\mu(\eta)$ is smooth in η, and $\mathcal{O}_0$ is easily seen to be a submanifold of C^* of codimension one. In fact, considering the real and imaginary part of the characteristic equation, there is a unique smooth function $\alpha: \mathcal{O} \to \mathbb{R}^+$ with $\alpha(\eta_0) = 1$ such that $\mathcal{O}_0 = \{\alpha(\eta)\eta \mid \eta \in \mathcal{O}\}$. For $\mathcal{O}$ sufficiently small, $f'(\tilde{\eta};\ \omega(\tilde{\eta}))$ is of constant sign for $\tilde{\eta} \in \mathcal{O}_0$.

At any $\tilde{\eta}\in\mathcal{O}_0$, the reduced bifurcation equation reads

$$0=\frac{-f'(\tilde{\eta};\,\omega(\tilde{\eta}))}{|\Delta'(\tilde{\eta};\,i\omega(\tilde{\eta}))|^2}\Bigg[3(g_3-\tilde{g}_3(\tilde{\eta};\,g_2))c^3+10(g_5-\tilde{g}_5(\tilde{\eta};\,g_2,g_3,g_4))c^5$$

$$+\binom{j}{j^*}(g_j-\tilde{g}_j(\tilde{\eta};\,g_2,\cdots,g_{j-1}))c^j+\cdots\Bigg],$$

where $\tilde{g}_j$ is a polynomial in $g_2,\cdots,g_{j-1}$ whose coefficients depend on $\tilde{\eta}$. Clearly, for each fixed g_2, there is exactly one value of $g_3=\tilde{g}_3(\tilde{\eta}p;\,g_2)$ at which (5.1) is nongeneric (i.e., $K_3(\tilde{\eta},g)=0$). If $g_3\neq\tilde{g}_3(\tilde{\eta};\,g_2)$ and $K_3(\tilde{\eta};\,g)<0$ (>0), then for each $\eta\in\mathcal{O}_+(\mathcal{O}_-)$ there is a unique continuous family of small periodic solutions $y(\eta;\,\cdot)$ of (5.1) with $y(\tilde{\eta};\,\cdot)=0$. This periodic solution is orbitally asymptotically stable (unstable).

In general, if $g_2,g_4,\cdots,g_{2k}$ are fixed, there are unique values $g_3=\tilde{g}_3$, $g_5=\tilde{g}_5,\cdots,g_{2k+1}=\tilde{g}_{2k+1}$ for which (5.1) is degenerate of order k (i.e., $K_3=K_5=\cdots=K_{2k+1}=0$). We now consider the bifurcation structure for (η,g) near a point (η,g^*), of kth order (but not $k+1$st order) degeneracy. Since $\partial K_{2k+3}/\partial g_{2k+3}$ is nonzero, the set $\{g|K_{2k+3}\}=0$ can be viewed as a submanifold of codimension 1 of the usual Banach space X of C^{2k+4} functions defined in a sufficiently small neighborhood $|y|\leqq r$ of $y=0$ and vanishing at zero. Assume $g^*\in X$ satisfies $K_3(\tilde{\eta},g^*)=K_5(\tilde{\eta},g^*)=\cdots=K_{2k+1}(\tilde{\eta},g^*)=0$ and $K_{2k+3}(\tilde{\eta},g^*)\neq 0$. (These relations define a manifold of codimension k in X.) By the mean value theorem, there can exist no more than $k+1$ periodic orbits bifurcating from $y=0$ at $\eta=\tilde{\eta}$. For any $0\leqq j\leqq k+1$, by appropriate perturbation of $\tilde{\eta}$ and $g_3,g_5,\cdots,g_{2k+1}$, the reduced bifurcation equation can be made to have j changes in sign for c near 0. By the continuity $\mu,K_3,\cdots,K_{2k+3}$ in (η,g), there are open subsets of $\mathcal{O}\times X$ containing $(\tilde{\eta},g^*)$ as a boundary point in which j small periodic solutions of (5.1) exist. (Note that since $g_2,g_4,\cdots,g_{2k}$ are arbitrary, there is no loss in generality in assuming that g^* (and its perturbations) are odd.)

The equation $\dot{y}=-\alpha g(y(t-1))$ is an example of equation that satisfies the hypotheses of this case. (Take $\alpha_0=\pi/2$ and $\omega_0=\pi/2$.) The bifurcation diagram given in [10] for this equation for g odd is representative of all first order degeneracies in case I.

Case 2. $f'(\eta_0;\,\omega_0)=0$ and $f''(\eta_0;\,\omega_0)\neq 0$.

As in Case 1, define $\mathcal{O}$ and $\mathcal{O}_+$, $\mathcal{O}_-$, $\mathcal{O}_0$. Again, $\mathcal{O}_0$ is a submanifold of C^* of codimension 1. By the implicit function theorem there is for each $\eta\in\mathcal{O}$ a value $\tilde{\omega}=\tilde{\omega}(\eta)$ for which $f'(\eta;\,\tilde{\omega}(\eta))=0$. Since $f''(\eta;\,\tilde{\omega}(\eta))$ has the same sign as $f''(\eta_0;\,\omega_0)$ for $\eta\in\mathcal{O}$, there is a unique relative extreme value of $f(\eta;\,\cdot)$ in a neighborhood of ω_0. We can write $\mathcal{O}=\mathcal{U}_0\cup\mathcal{U}_1\cup\mathcal{U}_2$, where $\mathcal{U}_1=\{\eta\in\mathcal{O}|f(\eta;\,\tilde{\omega}(\eta))=0\}$, $\mathcal{U}_{0(2)}=\{\eta\in\mathcal{O}|f(\eta;\,\tilde{\omega}(\eta))\cdot f''(\eta_0;\,\omega_0)>0\ (<0)\}$. Note that $\mathcal{U}_i$; $i=0,1,2$ are radial in the sense that if $\eta\in\mathcal{U}_i$ then so is $\alpha\eta$ for $\alpha>0$ such that $\alpha\eta\in\mathcal{O}$. Reducing $\mathcal{O}$ if necessary, one can find an open interval containing ω_0 such that $f(\eta;\,\cdot)$ has 0, 1 double, or 2 distinct simple zeros in that interval if $\eta\in\mathcal{U}_0$, $\mathcal{U}_1$ and $\mathcal{U}_2$, respectively. The manifold $\mathcal{O}_0$ can now be written as the disjoint union $\mathcal{O}_0=(\mathcal{O}_0\cap\mathcal{U}_1)\cup\mathcal{O}_0^+\cup\mathcal{O}_0^-$, where $\mathcal{O}_0^{+(-)}=\{\eta\in\mathcal{O}_0\cap\mathcal{U}_2|f'(\eta;\,\omega(\eta))>0\ (<0)\}$. By the form of the characteristic equation, note that if $\eta\in\mathcal{U}_0$ then $\alpha\eta\notin\mathcal{O}_0$ for all $\alpha>0$ such that $\alpha\eta\in\mathcal{O}$. If $\eta\in\mathcal{U}_2$, then there exist exactly two values $\alpha_1(\eta)<\alpha_2(\eta)$ near $\alpha=1$ (distinct by the simplicity of $i\omega_0$) such that $\alpha_1\eta$, $\alpha_2\eta\in\mathcal{O}_0$. Moreover, one of $\alpha_1\eta$, $\alpha_2\eta$ lies in $\mathcal{O}_0^+$, the other is in $\mathcal{O}_0^-$, and α_1, α_2 depend continuously on η in $\mathcal{U}_2$.

As in Case 1, the reduced bifurcation equation has the form

$$0=\mu(\eta)c+K_3(\eta,g)c^3+K_5(\eta,g)c^5+\mathcal{O}(c^7) \tag{5.3}$$

for $\eta \in \mathcal{O}$. Under the assumptions of this case, one computes (see Appendix)

$$K_3(\eta_0, g) = \frac{-4\omega_0^2 g_2^2 f(\eta_0;\, 2\omega_0)}{\Delta'(\eta_0;\, i\omega_0)|\Delta(\eta_0;\, 2\omega_0 i)|^2} \tag{5.4}$$

for all $g \in X$. (Recall $\Delta'(\eta_0;\, i\omega_0)$ is real.)

Case 2a. $f'(\eta_0;\, \omega_0)=0, f''(\eta_0;\, \omega_0) \neq 0$ and $f(\eta_0;\, 2\omega_0) \neq 0$.

Equation (5.1) is generic if and only if $g_2 \neq 0$. For such g and $f(\eta_0;\, 2\omega_0) \cdot \Delta'(\eta_0;\, i\omega_0) > 0$ (<0), then for every $\eta \in \mathcal{O}_+$ $(\mathcal{O}_-)$ there is a unique nontrivial periodic solution to (5.1). It is orbitally asymptotically stable (unstable).

To understand the bifurcation structure for η near η_0 for all small $|g_2|$, we compute

$$K_3(\eta, g) = -\mathrm{Re}\left\{ \int_{-1}^{0} e^{d\omega i s}\, d\eta(s) \left[3g_3 + 2g_2^2 \left(\frac{2\omega i}{\Delta(\eta, 2\omega i)} - 3 \right) \right] \Big/ \Delta'(\eta;\, \lambda(\eta)) \right\}, \tag{5.5}$$

where $\omega = \omega(\eta)$ and $\lambda(\eta) = \mu(\eta) + i\omega(\eta)$, and

$$K_5\left(\eta_0, g|_{g_2=0}\right) = -\frac{9\omega_0^2}{\Delta'(\eta_0;\, i\omega_0)} \left\{ \frac{f(\eta_0;\, 3\omega_0)}{|\Delta(\eta_0;\, 3\omega_0 i)|^2} + \frac{f''(\eta_0;\, \omega_0)}{2\left(\Delta'(\eta_0;\, i\omega_0)\right)^2} \right\} g_3^2. \tag{5.6}$$

We proceed assuming $K_5(\eta_0, g|_{g_2=0}) \neq 0$. In particular, $g_3 \neq 0$.

By direct computation, at $\alpha = 1$,

$$\frac{\partial}{\partial \alpha} K_3\left(\alpha\eta_0, g|_{g_2=0}\right) = \frac{3g_3\omega_0^2 f''(\eta_0;\, \omega_0)}{\left(\Delta'(\eta_0;\, i\omega_0)\right)^3},$$

which is nonzero. Thus, for all small $|g_2|$, $\mathcal{K}_0(g) \equiv \{\eta \in \mathcal{O} | K_3(\eta, g) = 0\}$ defines a submanifold of codimension 1 in $\mathcal{O}$. Define $\mathcal{K}_{+(-)}(g) = \{\eta \in \mathcal{O} | K_3(\eta, g) > 0\ (<0)\}$. From (5.5), $\eta \in \mathcal{O}_0$ implies $K_3(\eta, g|_{g_2=0}) = 3\omega(\eta) g_3 f'(\eta;\, \omega(\eta))/|\Delta'(\eta;\, \omega(\eta)i)|^2$. Hence $\mathcal{K}_0(g|_{g_2=0}) \cap \mathcal{O}_0 \subset \mathcal{U}_1$, $\mathcal{O}_0^+ \subset \mathcal{K}_+(g|_{g_2=0})$ and $\mathcal{O}_0^- \subset \mathcal{K}_-(g|_{g_2=0})$. Therefore, for all small $|g_2|$, $\mathcal{O}_0 \cap \mathcal{K}_+(g)$ and $\mathcal{O}_0 \cap \mathcal{K}_-(g)$ are nonempty. By the connectedness of $\mathcal{O}_0$, so is $\mathcal{O}_0 \cap \mathcal{K}_0(g) \neq \varnothing$.

If $K_5(\eta_0, g|_{g_2=0}) \neq 0$, then the implicit function theorem implies that for all (η, g) near $(\eta_0, g|_{g_2=0})$ there is a unique relative extreme value at $\tilde{\rho} = \tilde{\rho}(\eta, g)$ for

$$p(\rho) = \mu(\eta) + K_3(\eta, g)\rho + K_5(\eta, g)\rho^2 + \cdots.$$

By elementary arguments, $\tilde{\rho} = -K_3/2K_5 + \mathcal{O}(K_3^2)$ and $p(\tilde{\rho}) = \mu(\eta) - K_3^2/4K_5 + \mathcal{O}(K_3^2)$. Define $\mathcal{P}_0(g) = \{\eta \in \mathcal{O} | p(\tilde{\rho}) = 0\}$, and $\mathcal{P}_{+(-)}(g) = \{\eta \in \mathcal{O} | p(\tilde{\rho}) > 0\ (<0)\}$. Since $\mathcal{O}_0$ is a manifold of codimension 1 and $K_3(\eta_0, g|_{g_2=0}) = 0$, $\mathcal{P}_0(g)$ is for all small $|g_2|$ a submanifold of $\mathcal{O}$ of codimension 1. Reducing $\mathcal{O}$ if needed, we may write $\mathcal{O} = \mathcal{P}_+ \cup \mathcal{P}_- \cup \mathcal{P}_0$. Clearly the intersection in $\mathcal{O}$ of any two of the manifolds $\mathcal{P}_0(g)$, $\mathcal{O}_0$ and $\mathcal{K}_0(g)$ must occur at a point common to all three. Furthermore, if $K_5(\eta_0, g|_{g_2=0}) < 0$ (>0), then $\mathcal{P}_0(g) \cap (\mathcal{O}_+ \cup \mathcal{O}_-) \subset \mathcal{O}_{-(+)}$ and $\mathcal{O}_+ \subset \mathcal{P}_+(g)$ $(\mathcal{O}_- \subset \mathcal{P}_-(g))$.

To obtain more specific results, we choose $f''(\eta_0;\, \omega_0) > 0$ and $K_5(\eta_0, g|_{g_2=0}) < 0$; the other three cases are similar. Our previous discussion justifies Fig. 5.1, in which is shown the three manifolds $\mathcal{O}_0$, $\mathcal{P}_0(g)$, $\mathcal{K}_0(g)$ and their complements in $\mathcal{O}$.

Let $\tilde{\eta}\in\mathscr{K}_0(g)\cap\mathcal{O}_0$ be arbitrary. Considering the form of the reduced bifurcation equation, subcritical bifurcations (i.e., small periodic solutions existing for $y\in\mathcal{O}^-$ near $\tilde{\eta}$) exist if and only if $p(\tilde{\rho})\geqq 0$ and $\tilde{\rho}>0$. This last requirement is equivalent to $\eta\in\mathscr{K}_+(g)$. Accordingly, subcritical bifurcations exist if and only if $\eta\in\mathcal{O}\cap\mathscr{K}_+(g)\cap(\mathscr{P}_0(g)\cup\mathscr{P}_+(g))$. For such $\eta\in\mathscr{P}_0(g)$, (5.1) has a unique semistable periodic orbit (unstable from within). Otherwise, for $\eta\in\mathscr{P}_+(g)$, (5.1) has two distinct periodic orbits —the smaller of which (as measured by their corresponding roots of (5.3)) is unstable, while the larger is orbitally asymptotically stable. In contrast, for every $\eta\in\mathcal{O}^+$, there is a unique periodic orbit near $y=0$ (orbitally asymptotically stable).

Case 2b. $f'(\eta_0;\ \omega_0)=f(\eta_0;\ 2\omega_0)=0$ and $f''(\eta_0;\ \omega_0)\neq 0$. This case is remarkable in that $K_3(\eta_0;\ \cdot)\equiv 0$: Equation (5.1) is nongeneric for *all* $g\in X$. A rather tedious calculation (see Appendix) reveals that for $g\in X$,

$$K_5(\eta_0,g)=-\frac{\omega_0}{\Delta'(\eta_0;\ i\omega_0)}\left\{\left[g_3+2g_2^2\left(\frac{2\omega_0 i}{\Delta(\eta_0;\ 2\omega_0 i)}-1\right)\right]^2\frac{9\omega_0 f(\eta_0;\ 3\omega_0)}{|\Delta(\eta_0;\ 3\omega_0 i)|^2}\right.$$
$$\left.+\left[3g_3+2g_2^2\left(\frac{2\omega_0 i}{\Delta(\eta_0;\ 2\omega_0 i)}-3\right)\right]^2\frac{\omega_0 f''(\eta_0;\ \omega_0)}{2(\Delta'(\eta_0;\ \omega_0 i))^2}\right\}.$$

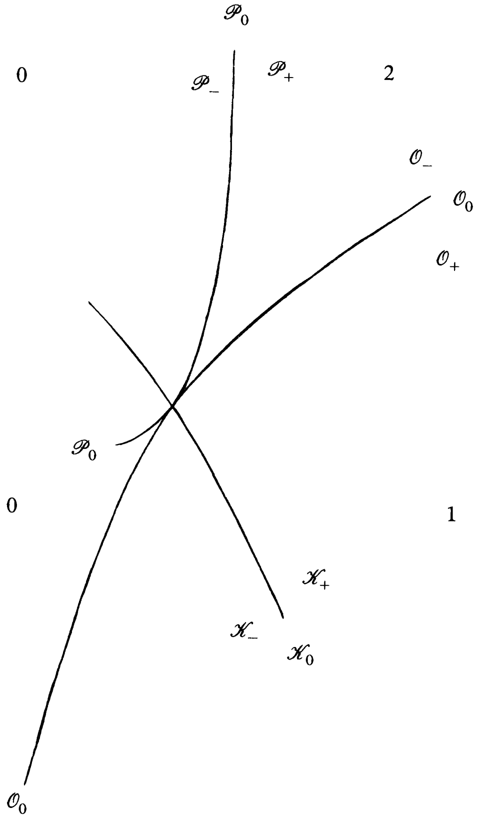


FIG. 5.1. *Intersections of the manifolds* $\mathcal{O}_0$, $\mathscr{P}_0(g)$ *and* $\mathscr{K}_0(g)$ *for* $f''(\eta;\ \omega_0)>0$, $K_5(\eta_0;\ g|_{g_2=0})<0$ *and* $|g_2|$ *small. The number of periodic solutions is indicated in each region.*

If $f(\eta_0;\ 3\omega_0)\cdot f''(\eta_0;\ \omega_0)>0$, then $\{g\in X|K_5(\eta_0,g)=0\}=\{g\in X|g_2=g_3=0\}$—a manifold of codimension two. $K_5(\eta_0,g)$ has constant sign off this manifold. If $f(\eta_0;\ 3\omega_0)=0$, then $K_5(\eta_0,g)\geqq 0$ or $\leqq 0$ on X, and the set of g with $K_5(\eta_0,g)=0$ defines a submanifold of X of codimension one. If $f(\eta_0;\ 3\omega_0)\cdot f''(\eta_0;\ \omega_0)<0$, then the (g_2,g_2) plane is subdivided into regions on which $K_5(\eta_0,g)$ takes positive and negative values. In fact, there are two distinct curves $\mathscr{C}_1$, $\mathscr{C}_2$ that intersect only at $g_2=g_3=0$. One of these curves, say $\mathscr{C}_1$, is the graph of a parabola $g_3=ag_2^2$, $a\in\mathbb{R}$. The other curve is either another parabola $g_3=bg_2^2$, $b\neq a$, or the curve $g_2=0$. Both $\mathscr{C}_1$ and $\mathscr{C}_2$ define submanifolds of X of codimension 1. Accordingly, the (g_2,g_3) plane is seen to be subdivided into 4 unbounded sets. $K_5(\eta_0,g)$ is positive on two of these and negative on the other two.

If g is such that $K_5(\eta_0,g)\neq 0$, the analysis of Case 2a applies without change. In particular, if $f''(\eta_0;\ \omega_0)>0$ and $K_5(\eta_0,g)<0$, Fig. 5.1 describes the bifurcation structure in a small neighborhood $\mathscr{O}$ of η_0 in C^*.

The equation

$$\dot{y}=-\alpha\int_{-1}^{0}(s+1)g(y(t+s))\,ds \tag{5.7}$$

satisfies at $\alpha=\alpha_0=4\pi^2$ the hypotheses of Case 2b. The associated characteristic equation has simple roots $\pm 2\pi i$ and all others have negative real parts. The cosine transform satisfies $f'(\eta_0;\ 2\pi)=f(\eta_0;\ 4\pi)=0$, where $\eta_0=2\pi^2(s+1)^2$. Moreover, $f(\eta_0;\ 6\pi)=0$ and $f''(\eta_0;\ 2\pi)=1$. Since $\Delta'(\eta_0;\ 2\pi i)=3$ and $\Delta(\eta_0;\ 4\pi i)=3\pi i$, we have $K_5(\eta_0,g)=-2\pi^2(g_3-\frac{10}{9}g_2^2)^2$. For $g_3-\frac{10}{9}g_2^2\neq 0$, Fig. 5.1 applies. Using indirect arguments, Hale [5] obtained a similar diagram for (5.7) based on certain special properties known for (5.7).

It would be misleading to suggest that the results of this section completely resolve all first order nongeneric bifurcations. In practice, arbitrary variation of η in a full neighborhood of η_0 is not allowed. For example, if $\eta=\eta(\alpha;\ \cdot)$ is a continuous function of m independent parameters $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_m)$; $\eta(0,\cdot)=\eta_0$, then the bifurcation structure for (5.1) is determined by how the range of the map $\alpha\to\eta(\alpha;\ \cdot)$ for α near 0 intersects $\mathscr{O}$. Moreover, it may be that g depends on α as well. The number and stabilities of periodic solutions for (5.1) is obtained from (5.5) with $\eta=\eta(\alpha;\ \cdot)$ and $g=g(\alpha)$. The resulting bifurcation surfaces are to be pictured in a neighborhood of $0\in\mathbb{R}^m$.

Equation (5.7) illustrates this point well. Here, $f(\eta_0;\ \nu)\geqq 0$ for all ν. Observe that for all α near $4\pi^2$, $\eta(\alpha;\ \cdot)\equiv\alpha/2(s+1)^2=\alpha\eta_0/4\pi^2\in\mathscr{P}_0\cup\mathscr{P}_-$, since if $\eta(\tilde{\alpha};\ \cdot)\in\mathscr{P}_+$, then $\eta(\tilde{\alpha};\ \cdot)+\varepsilon\delta_0\in\mathscr{P}_+$ for all small $\varepsilon>0$, where δ_0 denotes the unit mass measure at 0. Since $f(\eta(\tilde{\alpha};\ \cdot)+\varepsilon\delta_0,\ \nu)>0$ for all $\nu\in\mathbb{R}$, this contradicts the stability of Staffans mentioned earlier.

By the discussion of Hale [4, p. 112], any 1-periodic solution of

$$y''+\alpha g(y)=0, \tag{5.8}$$

$\alpha>0$, is a 1-periodic solution to (5.7). We apply (4.2) with $h=0$ in (4.1) and $\nu=2\pi$. We conclude from Theorem 2.2 that $g(\alpha;\ c)\equiv 0$; thus (5.8) has a small nontrivial 1-periodic solution if and only if there is a $c>0$ satisfying

$$0=(\sqrt{\alpha}-2\pi)+\sqrt{\alpha}\cdot\frac{3}{2}\left(g_3-\frac{10}{9}g_2^2\right)c^2+\mathscr{O}(c^4).$$

If $g_3 - \frac{10}{9}g_2^2 < 0$ (> 0), a unique root $c(\alpha) > 0$ exists for all $\alpha > 4\pi^2$($< 4\pi^2$) with $|\alpha - 4\pi^2|$ small. We conclude that for such α, $\eta(\alpha;\ \cdot) \in \mathcal{P}_0 \cap \mathcal{K}_+(\mathcal{P}_0 \cap \mathcal{K}_-)$. Thus at $\alpha = 4\pi^2$, equation (5.7) exhibits a Hopf bifurcation to a unique 1-parameter family of periodic orbits. Each is semistable.

Appendix. We briefly describe the computations for (5.3) in Case 2. Recall that $\bar{A}_{k,j} = A_{k,-j}$. The algorithm of §2 yields

$$y^{(1)}(\nu) = e^{\nu i \cdot} + e^{-\nu i \cdot}s,$$

$$y^{(2)}(\nu) = g_2\left(\frac{2\nu i}{\Delta(\eta;\, 2\nu i)} - 1\right)e^{2\nu i \cdot} - \frac{2g_2}{\Delta(\eta;\, 0)} + \cdots,$$

$$\begin{aligned} y^{(3)}(\nu) = &\left[g_3 + 2g_2^2\left(\frac{2\nu i}{\Delta(\eta;\, 2\nu i)} - 1\right)\right]\left(\frac{3\nu i}{\Delta(\eta;\, 3\nu i)} - 1\right)e^{3\nu i \cdot} \\ &+ \left[3g_3 + 2g_2^2\left(\frac{2\nu i}{\Delta(\eta;\, 2\nu i)} - 3\right)\right][\nu i - \Delta(\eta;\, \nu i)] \\ &\cdot \left[\frac{1}{\Delta(\eta;\, \nu i)} - \frac{1}{\Delta'(\eta;\, \lambda(\eta))(i\nu - \lambda(\eta))}\right]e^{\nu i \cdot} + \cdots, \end{aligned}$$

$$\begin{aligned} y^{(4)}(\nu) = \cdots &+ \left[4g_4 + 6g_3g_2\left(\frac{2\nu i}{\Delta(\eta;\, 2\nu i)} - 2\right) - 4g_2^3\left(\frac{2\nu i}{\Delta(\eta;\, 2\nu i)} - 1\right) + 2g_2(A_{3,1} + A_{3,3})\right] \\ &\cdot \left(\frac{2\nu i}{\Delta(\eta;\, 2\nu i)} - 1\right)e^{2\nu i \cdot} \\ &- \left[6g_4 + 6g_3g_2\operatorname{Re}\left(\frac{2\nu i}{\Delta(\eta;\, 2\nu i)} - 2\right) + 2g_2^3\left(\left|\frac{2\nu i}{\Delta(\eta;\, 2\nu i)} - 1\right|^2 + 2\right)\right. \\ &\left. + 4g_2\operatorname{Re}\{A_{3,1}\}\right] + \cdots. \end{aligned}$$

The term $A_{4,4}$ is not needed for our computation of (5.3) through order c^5. The terms $c_3(\nu, g)$, $c_5(\nu, g)$ in (5.2) are the coefficients of $e^{\nu i s}$ in

$$g_2[2y^{(1)}y^{(2)}] + g_3[y^{(1)}]^3$$

and

$$\begin{aligned} g_2[2y^{(1)}y^{(4)} + 2y^{(2)}y^{(3)}] &+ g_3\left[3(y^{(1)})^2y^{(3)} + 3y^{(1)}(y^{(2)})^2\right] \\ &+ g_4\left[4(y^{(1)})^3y^{(2)}\right] + g_5[y^{(1)}]^5, \end{aligned}$$

respectively. At $\nu = \omega(\eta)$, the real part of

$$M_3(\eta;\, i\nu) = \frac{-1}{\Delta'(\eta;\, \lambda(\eta))}\int_{-1}^{0} e^{\nu i s}\, d\eta(s) \cdot c_3(\nu, g)$$

is $K_3(\eta, g)$.

To compute $K_5(\eta_0, g|_{g_2=0})$, set $g_2=0$ in the above. In Case 2, $\Delta'(\eta_0; i\omega_0)$ is real, $i\int_{-1}^{0} se^{\omega_0 is}\,d\eta_0(s)$ is imaginary and $c_3(\nu, g|_{g_2=0})=3g_3$. Thus $\partial/\partial\nu M_3(\eta_0; \omega_0)$ is imaginary and $\partial/\partial\nu \operatorname{Re}\{M_3(\eta_0; \omega_0)\}=0$. Therefore $K_5(\eta_0, g|_{g_2=0})$ is the real part of

$$M_5(\eta_0; \omega_0)=\frac{\omega_0 ic_5\left(\omega_0, g|_{g_2=0}\right)}{\Delta'(\eta_0; i\omega_0)}.$$

Since $\Delta'(\eta_0; i\omega_0)$ is real, all real terms in c_5 can be ignored.

For Case 2b, we have $f(\eta_0; 2\omega_0)=0$. As above, $c_3(\omega_0, g)$ is seen to be real and $\partial/\partial\nu \operatorname{Re}\{M_3(\eta_0; \omega_0)\}=0$. Thus

$$K_5(\eta_0, g)=\frac{\omega_0}{\Delta'(\eta_0; i\omega_0)}\operatorname{Im}\{c_5(\omega_0, g)\}.$$

REFERENCES

[1] S.-N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[2] B. D. Coleman and V. J. Mizel, *On the stability of solutions of functional-differential equations*, Arch. Rational Mech. Anal., 30 (1968), pp. 173–196.

[3] J., C. F. de Oliveira and J. K. Hale, *Dynamic behavior from bifurcation equations*, Tôhoku Math. J., 32 (1980), pp. 577–592.

[4] J. K. Hale, *Functional Differential Equations*, Applied Math. Sci. Vol. 3, Springer-Verlag, New York, 1971.

[5] ______, *Generic properties of an integro-differential equation*, Amer. J. Math., to appear.

[6] J. J. Levin and J. Nohel, *On a nonlinear delay equation*, J. Math. Anal. Appl., 8 (1964), pp. 31–44.

[7] O. J. Staffans, *On the asymptotic spectra of the bounded solutions of a nonlinear Volterra equation*, J. Differential Equations, 24 (1977), pp. 365–382.

[8] ______, *On the holomorphic properties of the nonlinearity in a Volterra equation*, reprint.

[9] H. W. Stech, *Hopf bifurcation analysis in a class of scalar functional differential equations*, in Physical Mathematics and Nonlinear Partial Differential Equations, Marcel Dekker, New York, 1985.

[10] ______, *Hopf bifurcation calculations for functional differential equations*, J. Math. Anal. Appl., to appear.

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

003

STABLE EQUILIBRIA IN A SCALAR PARABOLIC EQUATION WITH VARIABLE DIFFUSION*

G. FUSCO[†] AND J. K. HALE[‡]

Abstract. A scalar parabolic equation with nonconstant diffusion and nonlinear source term is considered and some aspects of the influence of changing the diffusion on existence, stability and bifurcation properties of the equilibria are discussed.

1. Introduction. We deal with existence, stability and bifurcation properties of equilibria of the problem

$$\begin{aligned} &u_t=(cu_x)_x+f(u), \qquad x\in(-1,1),\\ &u_x(-1,t)=u_x(1,t)=0, \end{aligned} \tag{1}$$

where $c>0$ is a continuous function and f is C^1.

The initial value problem for (1) is well-posed in the Sobolev space $H^1(-1,1)$, [1], and any bounded orbit approaches an equilibrium as $t\to\infty$ [3], [5], [7]. Therefore a basic problem in understanding the dynamics of (1) is the description of the set of equilibria of (1) and of the way this set changes with the diffusion function c and with the source term f. Related important problems are the characterization of the pairs (c,f) such that (1) has stable nonconstant equilibria and to understand the role of bifurcation in the appearance of stable equilibria.

For any nonlinear function f, Chafee [6] proved that when c is constant, no stable nonconstant equilibrium exists. Chafee's result was generalized by Hale and Chipot [2] that showed that the same result holds true if $c\in C^2$ and $c_{xx}\leqq 0$. Finally, Yanagida [10] has shown that if c is written as $c=a^2$, $a>0$, a necessary and sufficient condition for the nonexistence of a function f such that (1) has a stable nonconstant equilibrium is that $a_{xx}\leqq 0$. Other results concerning the existence of stable nonconstant equilibria are due to Matano [8], [9] that has shown that, if f is a cubic polynomial as $f=u-u^3$ and $c(x)\geqq 1$ on intervals $[-1,\alpha]$, $[\beta,1]$ and $\leqq\varepsilon$ on $[\gamma,\delta]$ $\alpha<\gamma<\delta<\beta$ and ε is sufficiently small, then (1) has a stable nonconstant equilibrium. Fife and Peletier [13] have also considered equations related to (1) which have stable nonconstant equilibria.

For the n dimensional version of problem (1) in a bounded domain Ω and with constant diffusion, Casten and Holland [11] and Matano [8] have shown that, if Ω is convex, any stable equilibrium must be a constant. Matano has also shown that, assuming f of the type $f=u-u^3$, for some nonconvex domains, there exist stable nonconstant equilibria. Hale and Vegas [4] have shown the existence of stable nonconstant equilibria for a large class of nonlinearities and for domains Ω_ε that can be

*Received by the editors May 2, 1983, and in revised form January, 31, 1984.

[†] Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. Visiting from Università Degli Studi di Roma, Facoltà d'Ingegneria, Istituto di Matematica Applicata, 00161 Roma, Italy. The research of this author was supported in part by a C.N.R., NATO Senior Fellowship, (Bando n. 217.14).

[‡] Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The research of this author was supported in part by the U.S. Air Force Office of Scientific Research under contract AF-AFOSR 81-0198, in part by the National Science Foundation under contract MCS8205355 and in part by the U.S. Army Research Office under contract DAAG-29-79-C-0161.

considered as perturbations of a domain Ω_0 which is the union of two disjoint convex domains.

We assume that c is even and f is odd and such that

$$\begin{aligned} &f(0)=f(1)=0,\\ &f(u)>0 \quad \text{for } u\in(0,1),\\ &f(u)<0 \quad \text{for } u\in(1,\infty),\\ &f'(0)\neq 0, \qquad f'(1)\neq 0 \end{aligned} \tag{2}$$

(Fig. 1a)).

Under these assumptions we give an estimate of the number of equilibria of (1) in terms of c and f. We prove that, for any f of type (2), if c is sufficiently close to the step function

$$\tilde{c}=\begin{cases} 1 & \text{for } x\in[-1,-l]\cup[l,1],\\ c_0>0 & \text{for } x\in(-l,l), \quad 0<l<1 \end{cases} \tag{3}$$

(Fig. 1b)) and c_0 is sufficiently small, problem (1) has at least a pair of stable nonconstant odd monotone equilibria. Finally, we show that, if $c=c_\mu$ depends on a parameter $\mu\in[0,1]$ and u_μ is an equilibrium of (1) with exactly k zeros that bifurcates at $\mu=0$ from the zero equilibrium and becomes stable at $\mu=1$, then, as μ goes from 0 to 1, u_μ must go through at least k secondary bifurcations.

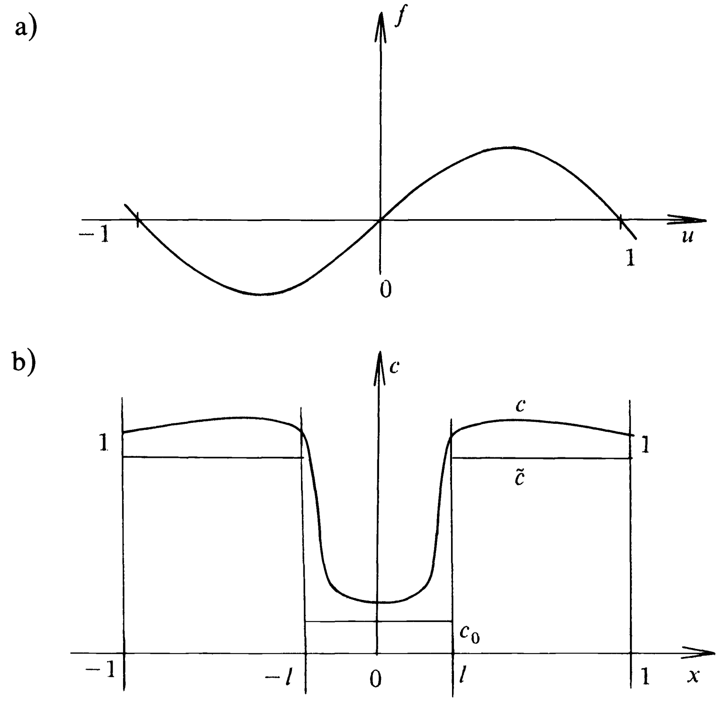


FIG. 1

2. Existence. We are interested in studying problem (1) for c in the set $\mathscr{C}$ of continuous and positive functions c: $[-1,1]\to\mathbb{R}$. Nevertheless, for the analysis that follows, in particular, for the discussion of stability where we consider functions c that are "close" to the step function $\bar{c}$ defined by (3), it is convenient to study problem (1) for a wider class of diffusion functions c that are allowed to have jump discontinuities at a finite number of fixed points in $[-1,1]$. To keep the notation simple and since we suppose c even, we consider only the case of two points of discontinuity at $x=\pm l$, $0<l<1$. Everything we say extends to the case of any number of points of discontinuity.

Let $\tilde{\mathscr{C}}$ be the set of nonnegative even functions c: $[-1,1]\to\mathbb{R}$ which have continuous restrictions to $[0,l)$ and to $[l,1]$ and possess the left limit $c(l^-)$ of $c(x)$ as $x\to l$. For any $c\in\tilde{\mathscr{C}}$, let $J_c\subset\mathbb{R}^2$ be the set $J_c\overset{\text{def}}{=}\{(x,y)|x=\pm l, y\in[c(l^-),c(l)]\}$ and $C\overset{\text{def}}{=}J_c\cup\text{graph}$ c. We suppose that $\tilde{\mathscr{C}}$ is endowed with the topology associated with the following notion of convergence that allows a sequence of continuous functions to converge to a function that has a jump at $x=\pm l$: we say that $c_n\in\tilde{\mathscr{C}}$, $n=1,\cdots$, converges to $c\in\tilde{\mathscr{C}}$ if and only if the Hausdorff distance between C_n and C approaches zero as $n\to\infty$. The Hausdorff distance $\delta(A,B)$ between two bounded subsets A, B of a metric space with metric d is defined as $\delta(A,B)=\max(p(A,B),\ p(B,A))$ where $p(A,B)$ is the distance from A to B, $p(A,B)=\sup_{x\in A}\inf_{y\in B}d(x,y)$ and $p(B,A)$ is the distance from B to A.

The class of diffusion functions that we are going to consider is the subset $\tilde{\mathscr{C}}^+\subset\tilde{\mathscr{C}}$ defined by the condition $\inf c>0$. Clearly, $\mathscr{C}$ is a dense subset of $\tilde{\mathscr{C}}^+$ and, if we assume in $\mathscr{C}$ the topology of uniform convergence in $[-1,1]$, then $\mathscr{C}$ is continuously embedded in $\tilde{\mathscr{C}}^+$. Henceforth, we allow c in problem (1) to be a generic $c\in\tilde{\mathscr{C}}^+$. This requires that (1) be complemented with the jump conditions

$$c(\pm l^+)u_x(\pm l^+)=c(\pm l^-)u_x(\pm l^-).$$

Therefore, the equilibrium problem corresponding to (1) becomes

$$\begin{aligned}&(cu_x)_x+f(u)=0,\\ &u_x(-1)=u_x(1)=0,\\ &c(\pm l^+)u_x(\pm l^+)=c(\pm l^-)u_x(\pm l^-),\end{aligned}\tag{4}$$

and reduces to the standard problem for $c\in\mathscr{C}$.

By letting $u=u$, $v=cu_x$, problem (4) transforms into the equivalent system

$$\begin{aligned}&u_x=\frac{v}{c},\\ &v_x=-f(u),\\ &v(-1)=v(1)=0.\end{aligned}\tag{5}$$

Note that the jump conditions express just continuity of v at $x=\pm l$ and, therefore, they are included in the requirement that u, v be continuous in $[-1,1]$.

The hypothesis on f and a maximum principle argument imply that solutions of (4) or (5) satisfy $-1\leqq u(x)\leqq 1$. Therefore, we can also assume that f is bounded so that the solution $u(c,a,x)$, $v(c,a,x)$ of the initial value problem

$$
\begin{aligned}
&u_x = \frac{v}{c},\\
&v_x = -f(u), \qquad\qquad (6)\\
&u(-1)=a, \qquad v(-1)=0
\end{aligned}
$$

is defined for all $(c,a,x)\in \tilde{\mathscr{C}}^{+}\times[-1,1]\times[-1,1]$.

LEMMA 1. *$u(c,a,x)$, $v(c,a,x)$ are continuous functions of (c,a,x), continuous in (c,a) uniformly in x and possess a continuous first derivative with respect to x, except possibly at $x=\pm l$.*

The proof of this lemma is a standard application of the general theory of differential equations.

To discuss the existence of space dependent equilibria of (1), i.e., the existence of nonconstant solutions of (4), we note that these solutions are in one-to-one correspondence with the $a\neq -1,0,1$ such that $v(c,a,1)=0$. If, for $a\neq 0$, we let $\delta(c,a,x)$ be the angle (positive clockwise around the x-axis in (u,v,x)-space) swept by the vector $\mathbf{u}(c,a,x')$ defined by

$$
\mathbf{u}(c,a,x')=\begin{bmatrix} u(c,a,x') \\ v(c,a,x') \\ 0 \end{bmatrix},
$$

when x' goes from -1 to x, then a necessary and sufficient condition in order that $v(c,a,1)=0$ for some $a\neq -1,0,1$ is that $\delta(c,a,1)$ be equal to πk for some integer $k\neq 0$.

The angle $\delta(c,a,x)$ can be defined also for $a=0$ so that $\delta(c,a,x)$ is continuous in (c,a,x). In fact, by performing the polar coordinate transformation $u=\rho\cos\delta$, $v=-\rho\sin\delta$, it is found that $\delta(c,a,\cdot)$ is the solution of the problem

$$
\begin{aligned}
&\delta_x = \frac{1}{c(x)}\sin^2\delta + \frac{f(u(c,a,x))}{u(c,a,x)}\cos^2\delta, \qquad (7)\\
&\delta(-1)=0.
\end{aligned}
$$

Moreover, since by Lemma 1, $u(c,a,x)$ is continuous in (c,a) uniformly in x and $u(c,0,x)=0$, it follows that, if (c',a) converges to $(c,0)$ in $\tilde{\mathscr{C}}^{+}\times[-1,1]$, then $f(u(c',a,x))/u(c',a,x)$ converges uniformly to $f'(0)$ in $[-1,1]$. This implies that, as $(c',a)\to(c,0)$, $\delta(c',a,x)$ converges uniformly to the solution $\delta(c,0,\cdot)$ of the problem

$$
\begin{aligned}
&\delta_x = \frac{1}{c(x)}\sin^2\delta + f'(0)\cos^2\delta, \qquad (8)\\
&\delta(-1)=0.
\end{aligned}
$$

From (7), (8) and Lemma 1, it also follows that $\delta(c,a,x)$ is continuously differentiable with respect to a. We also note that $\delta(c,\pm1,x)=0$ and that, for $a\in(-1,1)$, $\delta(c,a,x)$ is an increasing function of x because the right-hand sides of (7), (8) are >0.

For later use, we also introduce the angle $\sigma(c,a,x)$ which is defined as $\delta(c,a,x)$ with the vector $\mathbf{u}(c,a,x)$ replaced by its derivative $\mathbf{u}_a(c,a,x)$ with respect to a. It may be useful to note that, if Σ is the surface in the space of u, v, x defined by the solutions

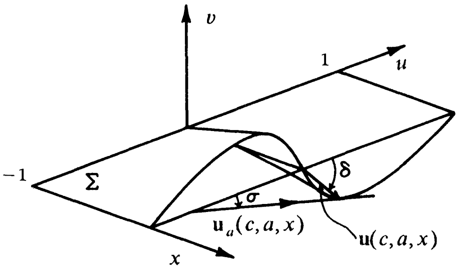


FIG. 2

of (6), then $\mathbf{u}_a(c,a,x)$ is tangent to the cross section of Σ at x at the point $(u(c,a,x), v(c,a,x),x)$ (see Fig. 2).

It is easily seen that $\sigma(c,a,\cdot)$ is the solution of the problem

$$\begin{aligned}\sigma_x&=\frac{1}{c(x)}\sin^2\sigma+f'(u(c,a,x))\cos^2\sigma,\\ \sigma(-1)&=0.\end{aligned}\tag{9}$$

Now consider an interval $[-\bar{l},\bar{l}]\subset[-1,1]$ and let $\bar{c}$ be the supremum of $c(x)$ in $[-\bar{l},\bar{l}]$. Then, for $x\in[-\bar{l},\bar{l}]$, (8) implies

$$\delta_x\geqq\frac{1}{\bar{c}}\sin^2\delta+f'(0)\cos^2\delta.\tag{10}$$

A simple computation shows that the solutions of (10) with the equality sign increase by π each time that x increases by $\pi(f'(0)/\bar{c})^{-1/2}$. From (10) and the fact that $\delta(c,a,x)$ is a nondecreasing function of x, it follows that

$$\delta(c,0,1)\geqq\pi\times\text{integer part of}\left[\frac{2\bar{l}}{\pi}\left(\frac{f'(0)}{\bar{c}}\right)^{1/2}\right].$$

This estimate, together with the continuity of $\delta(c,\cdot,1)$ and the fact that $\delta(c,\pm1,1)=0$ imply

THEOREM 1. *The number N of nonconstant equilibria of* (1) *satisfies the condition*

$$N\geqq2\times\textit{integer part of}\left[\frac{2\bar{l}}{\pi}\left(\frac{f'(0)}{\bar{c}}\right)^{1/2}\right].\tag{11}$$

Remark. In the proof of Theorem 1, no use was made of the evenness of c and oddness of f. Thus Theorem 1 holds for generic c,f. We also note that the conclusion of Theorem 1 is also true if $[-\bar{l},\bar{l}]$ is replaced by a measurable set $E\subset[-1,1]$ of measure $2\bar{l}$.

Let $s_k=\{a\,|\,\delta(c,a,1)=k\pi\}$. The set s_k can be identified with the set of equilibria of (1) that have exactly k zeros. If the right-hand side of (11) is $\geqq2k$, then s_k is nonempty and by means of equation (7), it is possible to obtain some information on the "shape" of equilibria. To this aim, let $0<\bar{u}<1$ and $a\in s_k$ be given and $\mathscr{P}\subset[-1,1]$ be the set where $|u(c,a,x)|<\bar{u}$. To get a bound for the measure of $\mathscr{P}\cap(-\bar{l},\bar{l})$, we let (x_1,x_2) be the smallest interval containing $\mathscr{P}\cap(-\bar{l},\bar{l})$ and $\bar{\eta}\overset{\text{def}}{=}\min_{|u|\leqq\bar{u}}f(u)/u$. Then, by applying to (7) the same procedure used for deriving (10) from (8), we obtain

$$\delta_x\geqq\frac{1}{\bar{c}}\sin^2\delta+\bar{\eta}\cos^2\delta,\qquad x\in\mathscr{P}\cap(-\bar{l},\bar{l})$$

and, therefore, since $\delta(c,a,x)$ is an increasing function of x,

$$\text{meas.}(\mathscr{P}\cap(-\bar{l},\bar{l}))\leqq\int_{\delta(c,a,x_1)}^{\delta(c,a,x_2)}\frac{d\delta}{(1/\bar{c})\sin^2\delta+\bar{\eta}\cos^2\delta}.$$

From this estimate, it follows that

$$\text{meas.}(\mathscr{P}\cap(-\bar{l},\bar{l}))<k\pi\left(\frac{\bar{c}}{\bar{\eta}}\right)^{1/2}\tag{12}$$

because $\delta(c,a,x_2)-\delta(c,a,x_1)<k\pi$. The estimate (12) shows that $\inf_{a\in s_k}\{|u(c,a,x)|\}$ converges in measure to 1 in $(-\bar{l},\bar{l})$ as $\bar{c}\to 0$. For $k\neq 1$, nothing can be said on the behavior of solutions in s_k outside the interval $(-\bar{l},\bar{l})$. Solutions in s_k could be almost trivial in the sense that they could be near zero outside $(-\bar{l},\bar{l})$ and oscillate in $(-\bar{l},\bar{l})$. This cannot happen when $k=1$ because solutions in s_1 are monotone and, therefore, if there is a point $\bar{x}\in(-\bar{l},\bar{l})$ where $|u(c,a,\bar{x})|$ is near 1, the same is true in $[-1,\bar{x}]$ or in $[\bar{x},1]$.

In what follows, we are interested in solutions of (4) that are odd functions of x. It is easily seen that, due to the assumption that c is even and f is odd, when on the basis of (11), it is possible to conclude that s_1 is nonempty, then it also contains at least a pair of odd solutions that transform into each other under the transformation $x\to -x$. Clearly, if $u(c,a,\cdot)$ is one of these odd solutions, and $|u(c,a,x)\mp 1|<\varepsilon$ in $[-1,\bar{x}]$, then $|u(c,a,x)\pm 1|<\varepsilon$ in $[-\bar{x},1]$. Therefore on the basis of (12), we have

Theorem 2. *For any c such that the right-hand side of* (11) *is* $\geqq 2$, *problem* (1) *has an equilibrium which is an odd and increasing function of x. If c is deformed so that $\bar{c}\to 0$, then all odd increasing equilibria of* (1) *converge to the function*

$$z=\begin{cases}-1 & \text{for } x\in[-1,0),\\ 0 & \text{for } x=0,\\ 1 & \text{for } x\in(0,1],\end{cases}$$

and the convergence is uniform in compact sets in $[-1,0)\cup(0,1]$.

In the statement of Theorem 2 and in the following, we always refer to the increasing equilibrium, with it being understood that there is also a decreasing equilibrium that transforms into the other one under the transformation $x\to -x$.

3. Stability. Let $\tilde{c}^0\in\mathscr{C}$ be the function defined by

$$\tilde{c}^0=\begin{cases}1, & x\in[-1,-l)\cup(l,1],\\ 0, & x\in(-l,l).\end{cases}$$

In this section, we prove the following

Theorem 3. *Let f be a continuously differentiable odd function that satisfies* (2). *Then there is a set $W\subset\mathscr{C}$ such that*

(i) *W is open and connected in $\mathscr{C}$,*

(ii) *c^0 belongs to the closure of W in $\mathscr{C}$,*

(iii) *for any $c\in W$, problem* (1) *has an odd increasing* (*and an odd decreasing*) *equilibrium which is stable.*

Note that Theorem 3 implies the

Corollary. *For any odd C^1-function f that satisfies* (2), *there is a $c\in\mathscr{C}$ such that problem* (1) *has a stable nonconstant equilibrium.*

To prove Theorem 3, we need a few lemmas.

LEMMA 2. *If $u(c,a,\cdot)$ is an equilibrium of* (1) *and λ is the first eigenvalue of the linear problem*

$$(13)\qquad \begin{aligned}&(cw_x)_x+f'(u(c,a,x))w=\lambda w,\\ &w_x(-1)=w_x(1)=0,\\ &c(\pm l^+)w_x(\pm l^+)=c(\pm l^-)w_x(\pm l^-),\end{aligned}$$

then $u(c,a,\cdot)$ is stable if $\lambda<0$, unstable if $\lambda>0$.

The proof of this lemma is given in [1].

LEMMA 3. *Let $\tilde{\mathscr{S}}\subset\tilde{\mathscr{C}}^+$ be the set of functions c such that* (1) *has a stable nonconstant equilibrium, the stability of which can be ascertained by the fact that the largest eigenvalue of the linear problem* (13) *is negative. Then $\tilde{\mathscr{S}}$ is open in $\tilde{\mathscr{C}}^+$.*

Proof. If $\tilde{\mathscr{S}}$ is empty, the lemma is obvious. Therefore, we assume that $\tilde{\mathscr{S}}$ is nonempty. Take any c in $\tilde{\mathscr{S}}$. Then there exist $a\in(-1,1)\setminus\{0\}$ and $k>0$ such that

$$\delta(c,a,1)=k\pi,$$

and the largest eigenvalue λ of problem (13) is negative. If one lets $w=r\cos\nu$, $cw_x=-r\sin\nu$ in (13), it is found that r, ν satisfy

$$(14)\qquad r_x=\sin\nu\cos\nu\left(f'(u(c,a,x))-\frac{1}{c(x)}-\lambda\right)r,$$

$$(15)\qquad \nu_x=\frac{1}{c(x)}\sin^2\nu+\left(f'(u(c,a,x))-\lambda\right)\cos^2\nu,$$

with the boundary conditions $\nu(-1)=0$, $\nu(1)=i\pi$ for some integer i. But i must be zero because the eigenfunction w corresponding to the largest eigenvalue never vanishes. Therefore, ν must stay in the interval $(-\pi/2,\pi/2)$. Since λ is negative, it follows from (9) and (15) that $\sigma(c,a,x)<\nu(1)=0$. On the other hand, equation (9) implies $\sigma(c,a,x)>-\pi/2$. It follows that $\sin(\delta(c,a,1)-\sigma(c,a,1))\neq0$. Since the derivative of δ with respect to a is related to σ by

$$(16)\qquad \frac{\rho^2\delta_a^2}{\rho^2\delta_a^2+\rho_a^2}=\sin^2(\delta-\sigma)$$

it results $\delta_a(c,a,1)\neq0$. The implicit function theorem implies there is a neighborhood U of c such that, for any $\tilde{c}$ in U, there is an $a(\tilde{c})$ continuous in $\tilde{c}$ such that $\delta(\tilde{c},a(\tilde{c}),1)=k\pi$. This proves existence of a solution for $\tilde{c}$ near c. The largest eigenvalue of the linear variational equation about this solution is continuous in $\tilde{c}$. This proves the lemma.

Remark. Lemma 3 is actually a special case of a general situation. In fact, the largest eigenvalue being negative for an equilibrium point u_0, implies the semigroup generated by the linear variational equation is exponentially asymptotically stable. Thus, a small perturbation in c will yield another exponentially asymptotically stable equilibrium point near u_0.

In the proof of Lemma 3, we have seen that $\sigma(c,a,1)<0$ is a necessary condition for λ to be negative. We note that this condition is also sufficient. This follows from the fact the solution of (15) depends continuously on λ, coincides with $\sigma(c,a,\cdot)$ for $\lambda=0$, and increases unboundedly as $\lambda\to-\infty$ for $x\neq-1$. Therefore, if $\sigma(c,a,1)$ is negative, there exists a unique negative λ_0 such that the solution of (15) vanishes at 1. If $r(\cdot)$ is any nonzero solution of (14), then $w(\cdot)=r(\cdot)\cos\nu(\cdot)$ is an eigenfunction of (13) that does not vanish in $[-1,1]$. Thus $\lambda_0<0$ is the largest eigenvalue of (13). Therefore, we can state

PROPOSITION 1. *A necessary and sufficient condition that the largest eigenvalue of problem* (13) *be negative is that* $\sigma(c,a,1)$ *be negative.*

LEMMA 4. *Let* $\tilde{c}\in\mathscr{C}^+$ *be a function of type* (3) *and* $(1)_{\tilde{c}}$ *be problem* (1) *with* $c=\tilde{c}$. *Then, if* c_0 *is sufficiently small, problem* $(1)_{\tilde{c}}$ *has an odd increasing equilibrium* $u(\tilde{c},a,\cdot)$ *such that the largest eigenvalue of the corresponding linear problem* $(13)_{\tilde{c}}$ *is negative.*

Proof. By Theorem 1, if $c_0<(4l^2/\pi^2)f'(0)$, there exists an $\tilde{a}\in(-1,1)\backslash\{0\}$ such that $u(\tilde{c},\tilde{a},\cdot)$ is an increasing equilibrium of $(1)_{\tilde{c}}$. The same condition together with the evenness of c and the oddness of f ensure that $\tilde{a}$ can also be chosen so that $u(\tilde{c},\tilde{a},\cdot)$ is an odd function. To prove that the largest eigenvalue of the linearized problem at $u(\tilde{c},\tilde{a},\cdot)$ is negative if c_0 is sufficiently small, we recall [12] that the eigenvalues of (13) do not decrease if $f'(u(c,a,x))$ is replaced by a function $q(x)\geqq f'(u(c,a,x))$. It follows that, if we let $\bar{q}=\max_{u\geqq u(\tilde{c},\tilde{a},l)}f'(u)$, it suffices to show that for c_0 small, the largest eigenvalue of

$$\begin{aligned}&w_{xx}+\bar{q}w=\lambda w, \qquad x\in(-1,-l)\cup(l,1),\\ &c_0w_{xx}+f'(u(\tilde{c},\tilde{a},x))w=\lambda w, \qquad x\in(-l,l),\end{aligned}\tag{17}$$

$$\begin{aligned}&w_x(-1)=w_x(1)=0,\\ &c_0w_x(-l^+)=w_x(-l^-),\\ &w_x(l^+)=c_0w_x(l^-)\end{aligned}\tag{18}$$

is negative.

From a result of Yanagida [10], it follows that the largest eigenvalue of this problem is negative if there is a strictly positive function w_0 that makes the left-hand sides of (17) equal to zero, satisfies the last two equations (18) and moreover, is such that

$$w_{0x}(-1)<0, \qquad w_{0x}(1)>0.\tag{19}$$

We look for an even such w_0 and, therefore, we may assume $w_{0x}(0)=0$ and consider only the interval $[0,1]$. Since $f'(1)<0$ and, by Theorem 2, $u(\tilde{c},\tilde{a},l)\rightarrow1$ as $c_0\rightarrow0$, $\bar{q}$ is negative for small values of c_0. Therefore, if w_0 exists, in the interval $[l,1]$, it must have the expression

$$w_0(x)=B\sinh\left[(-\bar{q})^{1/2}(x-l)\right]+A\cosh\left[(-\bar{q})^{1/2}(x-l)\right],\tag{20}$$

and the coefficients A, B must satisfy the conditions

$$A>0, \qquad \frac{B}{A}>-\tanh\left[(-\bar{q})^{1/2}(1-l)\right]\tag{21}$$

ensuring that $w_0(x)$ is positive in $[l,1]$ and $w_{0x}(1)>0$.

To compute $w_0(x)$ in the interval $[0,l]$, we must solve the problem

$$\begin{aligned}&c_0w_{0xx}+f'(u(\tilde{c},\tilde{a},x))w_0=0, \qquad x\in(0,l),\\ &w_0(0)=C, \qquad w_{0x}(0)=0,\end{aligned}\tag{22}$$

where C is a positive constant to be chosen later.

From now on, we set for simplicity $\tilde{u}=u(\tilde{c},\tilde{a},\cdot)$, $\bar{u}=\tilde{u}(l)$, $\bar{\bar{u}}=\tilde{u}(1)$. In order to solve this problem, we must overcome the difficulty lying in the fact that u is only known to be an odd increasing solution of problem $(4)_{\tilde{c}}$. To this end, we observe that, since $\tilde{u}$ is increasing, we can perform the change of variable $x=\tilde{u}^{-1}(u)\overset{\text{def}}{=}\xi(u)$. By making this change of variable in $(4)_{\tilde{c}}$ and by observing that the oddness of $\tilde{u}$ implies $\xi(0)=0$, we see that ξ satisfies

$$
\begin{aligned}
&-c_0\frac{\xi''}{\xi'^3}+f(u)=0, \qquad u\in(0,\bar{u}),\\
&-\frac{\xi''}{\xi'^3}+f(u)=0, \qquad u\in(\bar{u},\bar{\bar{u}}),
\end{aligned}
\tag{23}
$$

$$
\begin{aligned}
&\xi(0)=0, \qquad \xi(\bar{u})=l,\\
&\lim_{u\to\bar{\bar{u}}}\xi(u)=1, \qquad \lim_{u\to\bar{\bar{u}}}\xi'(u)=\infty.
\end{aligned}
\tag{24}
$$

Note that u, $\tilde{c}u$ continuous imply ξ continuous and $\xi'(\bar{u}^-)=\xi'(\bar{u}^+)c_0$. By using the fact that $d/dx=(1/\xi')d/du$, one sees that the same change of variables applied to (22) yields

$$
\begin{aligned}
&\frac{c_0}{\xi'^2}w_0''-c_0\frac{\xi''}{\xi'^3}w_0'+f'w_0=0, \qquad u\in(0,\bar{u}),\\
&w_0(0)=C, \qquad w_0'(0)=0,
\end{aligned}
\tag{25}
$$

where w_0 has been identified with the function $w_0(\xi(\cdot))$. For u in $(0,\bar{u})$, an integration of (23) from u to $\bar{u}$ yields

$$
\frac{c_0}{\xi'^2}=2\int_u^1 f(s)\,ds+K\overset{\text{def}}{=}g(u),
$$

where $K>-2\int_{\bar{u}}^1 f(s)\,ds$ is an integration constant. For u in $(\bar{u},\bar{\bar{u}})$, performing an integration in (23) from $\bar{u}$ to $\bar{\bar{u}}$, using the fact that $\xi(\bar{\bar{u}})-\xi(\bar{u})=1-l$ and requiring that $\xi'(\bar{u}^-)=\xi'(\bar{u}^+)c_0$ one observes that K, $\bar{u}$, $\bar{\bar{u}}$ must satisfy the conditions

$$
\begin{aligned}
&\int_{\bar{u}}^{\bar{\bar{u}}}\frac{du}{\left(2\int_u^{\bar{\bar{u}}}f(s)\,ds\right)^{1/2}}=1-l,\\
&c_0 g(\bar{u})=2\int_{\bar{u}}^{\bar{\bar{u}}}f(s)\,ds.
\end{aligned}
\tag{26}
$$

Since $g'=-2f$ and the first equation (23) implies that the coefficient of w_0' in equation (25) is equal to $-f$, equation (25) becomes

$$
gw_0''-fw_0'+f'w_0=\left(gw_0'+fw_0\right)'=0.
$$

Thus $gw_0'+fw_0=\text{const}=0$ because $w_0'(0)=0$ and $f(0)=0$.

It follows that, with a proper choice of the constant C appearing in (25),

$$
w_0=g^{1/2} \quad (\text{for } u\in[0,\bar{u}]).
\tag{27}
$$

If the expressions (20) (27) are patched together at $x=l$ (corresponding to $u=\bar{u}$) by imposing the conditions

$$w_0(l^+)=w_0(l^-), \qquad w_{0x}(l^+)=c_0 w_{0x}(l^-),$$

it is found that

$$A=(g(\bar{u}))^{1/2}, \qquad B=-c_0^{1/2}\frac{f(\bar{u})}{(-\bar{q})^{1/2}}.$$

Therefore, it follows that, if

$$\tag{28} \frac{c_0^{1/2}f(\bar{u})}{(-\bar{q})^{1/2}(g(\bar{u}))^{1/2}} < \tanh\left[(-\bar{q})^{1/2}(1-l)\right],$$

then w_0 satisfies all the conditions ensuring that the largest eigenvalue of problem (17), (18) is negative. We shall prove that this is the case for c_0 sufficiently small. The proof is a discussion of the asymptotic dependence of $\bar{u}$, $g(\bar{u})$ on c_0 defined by equations (26) for $c_0 \to 0$.

By the change of variables $u=\bar{u}+(\bar{\bar{u}}-\bar{u})\tau$, $s=u+(\bar{\bar{u}}-u)\sigma$, the first of equations (26) transforms as

$$\tag{29} \frac{1}{2^{1/2}}\int_0^1 (1-\tau)^{-1/2}\left(\int_0^1 \frac{f(1+\bar{\delta}(\tau,\sigma))}{\bar{\delta}(\tau,\sigma)}(\tau-\bar{\alpha}+(1-\tau)\sigma)\,d\sigma\right)^{-1/2} d\tau = 1-l,$$

where

$$\bar{\delta}(\tau,\sigma)=-\left[(1-\bar{\bar{u}})+(\bar{\bar{u}}-\bar{u})(1-\tau)(1-\sigma)\right],$$
$$\bar{\alpha}=\frac{1-\bar{u}}{\bar{\bar{u}}-\bar{u}}>1.$$

As $\bar{u}$, $\bar{\bar{u}}$ depend on c_0, so do $\bar{\delta}$ and $\bar{\alpha}$. Let $\alpha=\lim_{c_0\to 0}\sup\bar{\alpha}$. Since $\bar{u}\to 1$, $\bar{\bar{u}}\to 1$ as $c_0\to 0$, the above expression of $\bar{\delta}(\tau,\sigma)$ implies that $\bar{\delta}(\tau,\sigma)\to 0$ uniformly as $c_0\to 0$. Therefore, $f(1)=0$ implies that the ratio $f(1+\bar{\delta}(\tau,\sigma))/\bar{\delta}(\tau,\sigma)$ converges uniformly to $f'(1)\neq 0$ as $c_0\to 0$. This and equation (29) imply that $\alpha<\infty$.

For the verification of the inequality (28) for c_0 small enough, it is sufficient to show that the quantity

$$f(\bar{u})\left[\int_{\bar{u}}^{\bar{\bar{u}}} f(s)\,ds\right]^{-1/2}$$

is bounded for c_0 small since $\bar{q}\to f'(1)$ as $c_0\to 0$ and $g(\bar{u})$ satisfies (26). Using the fact that $f(1)=0$ and Taylor's theorem, we have

$$\frac{f(\bar{u})}{\left[\int_{\bar{u}}^{\bar{\bar{u}}} f(s)\,ds\right]^{1/2}} = \frac{f'(\theta(\bar{u}))(1-\bar{u})}{\left(\int_{\bar{u}}^{\bar{\bar{u}}} f(s)\,ds\right)^{1/2}} = \frac{f'(\theta(u))\bar{\alpha}}{\left((1/(\bar{\bar{u}}-\bar{u}))^2\int_{\bar{u}}^{\bar{\bar{u}}} f(s)\,ds\right)^{1/2}}$$

where $\theta(u)$ is in the interval $(\bar{u},1)$ and $\bar{\alpha}=(1-\bar{u})/(\bar{\bar{u}}-\bar{u})$. Since f' and $\bar{\alpha}$ are uniformly bounded, it remains to show that the denominator of this expression is bounded away from zero for c_0 small. Note that

$$\int_{\bar{u}}^{\bar{\bar{u}}} \frac{f(s)\,ds}{(\bar{\bar{u}}-\bar{u})^2} = \int_0^1 \frac{f(1+\bar{\delta}'(\tau))}{\delta'(\tau)}(\tau-\bar{\alpha})\,d\tau, \tag{30}$$

with $\bar{\delta}'(\tau)=(\bar{u}-1)+(\bar{\bar{u}}-\bar{u})\tau$. Since $\bar{\delta}'(\tau)\to 0$ uniformly as $c_0\to 0$, the ratio $f(1+\bar{\delta}'(\tau))/\bar{\delta}'(\tau)\to f'(1)<0$ uniformly as $c_0\to 0$. From this and the fact that $\bar{\alpha}$ is >1, it follows that

$$\int_0^1 \frac{f(1+\bar{\delta}'(\tau))}{\bar{\delta}'(\tau)}(\tau-\bar{\alpha})\,d\tau \geqq -f'(1)\int_0^1(1-\tau)\,d\tau = -\frac{1}{2}f'(1)>0.$$

This proves the lemma.

Proof of Theorem 3. By Lemma 4, there is a number $\varepsilon>0$ such that, if $\tilde{\gamma}\subset\tilde{\mathscr{C}}^+$ is the curve $\tilde{\gamma}\overset{\text{def}}{=}\{\tilde{c}\,|\,0<c_0<\varepsilon\}$ and $\tilde{c}\in\tilde{\gamma}$, then problem $(1)_{\tilde{c}}$ has an odd increasing equilibrium which is stable. Since, by Lemma 3, $\tilde{\mathscr{S}}$is open in $\tilde{\mathscr{C}}^+$, there exists an open neighborhood $\tilde{W}$ of $\tilde{\gamma}$ in $\tilde{\mathscr{C}}^+$ such that, for $c\in\tilde{W}$, problem (1) has a stable equilibrium u_c. It is easy to see that $\tilde{W}$ can be chosen so that u_c is odd and increasing for any $c\in\tilde{W}$. In fact, from the proof of Lemma 3, it follows that, for c in a neighborhood of $\tilde{c}\in\tilde{\gamma}$, u_c is the only equilibrium in a neighborhood of $u_{\tilde{c}}$. On the other hand, the evenness of c and the oddness of f imply that also v_c defined by $v_c(x)=-u_c(-x)$ is also an equilibrium of (1). Since $u_{\tilde{c}}$ is odd and $u_c\to u_{\tilde{c}}$ as $c\to\tilde{c}$, it follows that v_c converges to $u_{\tilde{c}}$ as $c\to\tilde{c}$. This contradicts uniqueness of u_c unless u_c is odd and therefore proves oddness. Since u_c is close to $u_{\tilde{c}}$, it vanishes only at $x=0$. From this and the fact that solutions of (4) with only one zero are monotone, it follows that u_c is increasing. The mapping $c_0\to\tilde{c}$ is continuous as a map from $(0,\varepsilon)$ into $\tilde{\mathscr{C}}^+$. Therefore $\tilde{\gamma}$ is locally compact as a subset of $\tilde{\mathscr{C}}^+$. Thus, by standard arguments, there is a continuous function $\phi:(0,1)\to\mathscr{C}$ such that the curve $\gamma=\{c\,|\,c=\phi(s),\ s\in(0,1)\}$ is contained in $\tilde{W}$ and $\tilde{c}^0$ is in the closure of γ in $\tilde{\mathscr{C}}$. Since $\tilde{W}$ is open and $\mathscr{C}$is continuously embedded in $\tilde{\mathscr{C}}^+$, $\tilde{W}\cap\mathscr{C}$is open in $\mathscr{C}$. From this and the continuity of ϕ, it follows that there is a subset $W\subset\tilde{W}\cap\mathscr{C}$which is open and connected in $\mathscr{C}$ and contains γ. Since $\tilde{c}^0$ is in the closure of γ in $\tilde{\mathscr{C}}$, the proof is completed.

4. Secondary bifurcation.

In this section, we consider a family $(c_\mu)_{\mu\in[-1,1]}\subset\mathscr{C}$ of diffusion functions c_μ depending continuously on a parameter μ. We let $\delta(\mu,a,x)\overset{\text{def}}{=}\delta(c_\mu,a,x)$ and assume that $c_{\mu_2}<c_{\mu_1}$ for $\mu_2>\mu_1$ and that $\delta(0,0,1)=k\pi$ for some $k>0$. Then (8) implies that $\delta(\mu,0,1)<k\pi$ for $\mu<0$ and $\delta(\mu,0,1)>k\pi$ for $\mu>0$. If we also assume that f satisfies the condition

$$f(u)<f'(0)u, \qquad u\in(0,1], \tag{31}$$

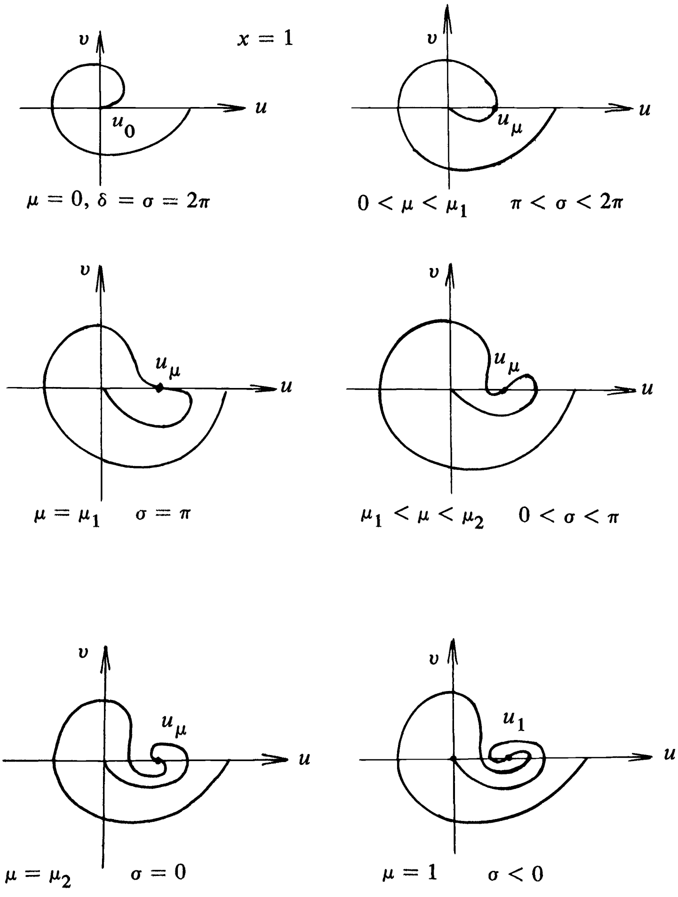
then, from (7) (8), it follows that $\delta(\mu,a,1)<\delta(\mu,0,1)$ for $a\neq 0$, $\mu\in[-1,1]$. Therefore s_k is empty for $\mu\leqq 0$, nonempty for $\mu>0$. Thus, $\mu=0$ is a bifurcation point. It is easy to see that, in this situation, for $\mu>0$ and small, s_k contains solutions that are small and converge to zero as $\mu\to 0$, i.e., solutions that bifurcate from the zero solution. These solutions are unstable for μ small because the largest eigenvalue of problem (13) with $u\equiv 0$ is $f'(0)>0$ and the eigenvalues of (13) are continuous functions of $c\in\mathscr{C}$. On the other hand, we have seen in Theorem 3 that, if c is suitably chosen, then there exist stable nonconstant equilibria of (1). Therefore, it can be expected that, if u_μ is a continuous function of $\mu\in[0,1]$ such that $u_0=0$, u_μ is a solution of (4) in s_k for $\mu\in(0,1]$, and u_1 is stable, some kind of secondary bifurcation takes place at some $\mu\in(0,1)$. This conjecture is true. We have in fact the following

THEOREM 4. *Suppose that c_μ is as before, f satisfies* (31), *u_μ is an equilibrium of* $(1)_{c_\mu}$ *which is equal to zero for $\mu=0$, has exactly k zeros for $\mu\in(0,1]$, depends continuously on μ and u_1 is stable (in the sense that the largest eigenvalue of the linearized problem at u_1 is negative), then there exist numbers $0<\mu_1<\cdots<\mu_k<1$ such that each μ_i, $i=1,\cdots,k$ is a bifurcation point.*

Proof. Let $a_\mu \overset{\text{def}}{=} u_\mu(-1)$. Then $a_0=0$ and, therefore, (7), (9) imply that $\sigma(0,a_0,1)=\delta(0,a_0,1)$. We also have $\delta(\mu,a_\mu,1)=k\pi$ for $\mu\in(0,1]$. Thus, by the continuity of δ with respect to c, a and the continuity of c_μ, u_μ with respect to μ, it follows that $\sigma(0,a_0,1)=k\pi$. On the other hand, Proposition 1 and the stability of u_1 imply $\sigma(1,a_1,1)<0$. Therefore, by continuity, there exist $0<\mu_1<\cdots<\mu_k<1$ such that

$$\sigma(\mu_i,a_{\mu_i},1)=(k-i)\pi, \qquad i=1,\cdots,k.$$

Moreover, it is obvious that μ_i, $i=1,\cdots,k$ can be chosen so that, in any neighborhood of μ_i, there exist $\bar\mu<\mu_i<\bar{\bar\mu}$ such that $\sigma(\bar\mu,a_{\bar\mu},1)\geq(k-i)\pi>\sigma(\bar{\bar\mu},a_{\bar{\bar\mu}},1)$. This, on the basis of the geometrical meaning of the angle σ, implies that μ_i is a bifurcation point.



$\mu=0,\ \delta=\sigma=2\pi$ $\quad$ $0<\mu<\mu_1 \quad \pi<\sigma<2\pi$

$\mu=\mu_1 \quad \sigma=\pi$ $\quad$ $\mu_1<\mu<\mu_2 \quad 0<\sigma<\pi$

$\mu=\mu_2 \quad \sigma=0$ $\quad$ $\mu=1 \quad \sigma<0$

FIG. 3

Theorem 4 says that going through k secondary bifurcations is a necessary condition in order that an equilibrium with k zeros that bifurcates from the zero solution becomes stable. From the proof of the theorem and Proposition 1 it follows that if, as μ goes from 0 to 1, u_μ experiences exactly k bifurcations at $0 < \mu_1 < \cdots < \mu_k < 1$, each one of which is simple in the sense that, at any μ_i, two new solutions bifurcating from u_{μ_i} appear, then u_1 is stable (see Fig. 3 for the case $k = 2$). This observation shows that in a certain sense the converse of Theorem 4 is also true.

REFERENCES

[1] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Mathematics 840, Springer-Verlag, New York, 1981.

[2] J. K. Hale and M. Chipot, *Stable equilibria with variable diffusion*, in Nonlinear Partial Differential Equations, J. Smoller, ed., Vol. 17 in Contemporary Mathematics, American Mathematical Society, Providence, RI, 1983, pp. 209–214.

[3] J. K. Hale and P. Massatt, *Asymptotic behavior of gradient-like systems*, Univ. of Florida Symposium on Dynamic Systems, II, Academic Press, New York, 1982.

[4] J. K. Hale and J. Vegas, *A nonlinear parabolic equation with varying domain*, Arch. Rat. Mech. Anal., to appear.

[5] T. J. Zelenyak, *Stabilization of solutions of boundary value problems for a second order parabolic equation with one space variable*, Differential Equations, 4 (1968), pp. 17–22; translated from Differentialniye Uravneniya.

[6] N. Chafee, *The electric ballast resistor: homogeneous and nonhomogeneous equilibria*, in Nonlinear Differential Equations: Invariance, Stability and Bifurcation, P. de Mottoni and L. Salvadori, eds., Academic Press, New York, 1981, pp. 161–173.

[7] H. Matano, *Convergence of solutions of one-dimensional semilinear parabolic equations*, J. Math. Kyoto Univ., 18 (1978), pp. 224–243.

[8] ______, *Asymptotic behavior and stability of solutions of semilinear diffusion equations*, Res. Inst. Math. Sci., Kyoto, 15 (1979), pp. 401–458.

[9] ______, *Nonincrease of lap number of a solution for a one dimensional semilinear parabolic equation*, J. Fac. Sci., Univ. Tokyo 29 (1982), pp. 401–441.

[10] E. Yanagida, *Stability of stationary distributions in space-dependent population growth process*, J. Math. Biol., 15 (1982), pp. 37–50.

[11] R. G. Casten and C. J. Holland, *Instability results for a reaction diffusion equation with Neumann boundary conditions*, J. Differential Equations, 27 (1978), pp. 266–273.

[12] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. I Interscience, New York.

[13] P. C. Fife and L. A. Peletier, *Clines induced by variable selection and migration*, Proc. Roy. Soc. London B, 214 (1981), pp. 99–123.

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

004

THE INITIAL BOUNDARY PROBLEM FOR THE MAXWELL EQUATIONS IN THE PRESENCE OF A MOVING BODY*

JEFFERY COOPER[†] AND WALTER STRAUSS[‡]

Abstract. Existence and uniqueness of finite energy solutions of the Maxwell equations is proved in the presence of a moving body which may be either a perfect conductor or a dielectric. For the perfect conductor, it is assumed that the speed of the body is less than the speed of propagation in a vacuum, while for the dielectric it is assumed that the speed of the body is less than the speed of propagation within the body when at rest. The proof involves localization of the problem to a neighborhood of the moving boundary and a change of coordinates using the techniques of general relativity. The Neumann problem for a moving body and the scalar wave equation is also treated.

1. Introduction. We consider the problem of proving the existence and uniqueness of a solution of Maxwell's equations with given initial conditions at $t=0$ in the presence of a moving body. We consider both a perfect conductor and a dielectric. It is perhaps surprising that this problem has never before been treated in the mathematical literature. Of course, stationary bodies have been treated many times. The only case of which we are aware for a moving body is that of the scalar wave equation $\Box u=0$ with Dirichlet boundary conditions, which was treated by Cooper and Bardos [1] and by Inoue [2]. In the engineering literature, Van Bladel [10] has constructed approximate solutions for specific geometries.

More precisely, the problem considered here is the well-posedness in the energy norm. Is the energy finite if it is finite initially? As for the body, we assume:

(i) it is compact, its boundary is smooth and its motion is smooth, and

(ii) it moves slower than the wave speed (the speed of light).

Our notation is as follows. The space-time region exterior to the body is denoted by Q. It is an open set in spacetime $\mathbb{R}\times\mathbb{R}^3$. The region occupied by the body at time t is

$$\mathcal{O}(t)=\{x\in\mathbb{R}^3\mid(t,x)\notin Q\}.$$

Assumption (i) means that $\mathcal{O}(t)$ is compact for all t and $\Sigma=\partial Q$ is a C^∞ hypersurface in $\mathbb{R}\times\mathbb{R}^3$. We denote by $\nu=(\nu_t,\nu_x)$ the unit space-time normal to Σ pointing into Q. Assumption (ii) means that $|\nu_t|<|\nu_x|$ at each point of Σ. It also means that Σ is timelike. (We have taken units so that the speed of light in a vacuum is unity.)

The reader should note that whenever $\nu_t=0$, the boundary Σ is characteristic for the hyperbolic part of the Maxwell's equations, that is, the Maxwell's equations with the divergence equations removed. A standard local change of variables to a coordinate system in which the body is at rest would still leave a boundary which is characteristic at some points, but not others. To avoid this difficulty we make a covariant change of variable involving the full Maxwell system which leads to a uniformly characteristic boundary. We are then able to study the existence question for the hyperbolic part of

*Received by the editors March 15, 1984.

[†] Department of Mathematics, University of Maryland, College Park, Maryland 20742. The research of this author was supported by the National Science Foundation under grant MCS-82-00496.

[‡] Department of Mathematics, Brown University, Providence, Rhode Island 02912. The research of this author was supported by the National Science Foundation under grant MCS-81-21487.

the transformed system, with the added complication that the constitutive relations now have coefficients depending on time. In §§2 and 3 we discuss the coordinate changes and constitutive relations. We prove the well-posedness in §4 for the perfect conductor and §5 for the dielectric body. In both cases the problem is reduced to an abstract evolution equation in Hilbert space with a time-dependent generator and a general theorem of Kato [1] is applied. Explicit methods of partial differential equations could also have been used instead of the abstract approach (see [4]).

Finally in §6 we return to the scalar wave equation but we consider a general class of boundary conditions of the Neumann type. However, the reader should be warned that neither the classical Neumann condition $\partial u/\partial \nu_x = 0$ nor the condition $\partial u/\partial \nu = 0$ is well-posed in the energy norm! What is well-posed is the condition $\partial u/\partial \nu^* = 0$ where $\nu^* = (-\nu_t, +\nu_x)$ is the conormal derivative. Also allowed in the boundary condition is a dissipative term and a Robin-type term. Our approach is much simpler that that of [1] or [2], as well as more general.

2. Constitutive relations. Due to the nature of our problem, we are forced to write the time variable explicitly. Therefore we write the Maxwell equations in their classical form

$$\partial_t D - \nabla \times H = -J, \qquad \nabla \cdot D = q, \tag{2.1a}$$

$$\partial_t B + \nabla \times E = 0, \qquad \nabla \cdot B = 0, \tag{2.1b}$$

where the electric field E, the displacement D, the magnetic field H, the magnetic induction B, the charge density q and the current density J may be distributions. It is convenient to introduce the field strength tensors

$$F = \begin{bmatrix} 0 & -E_1 & -E_2 & -E_3 \\ E_1 & 0 & B_3 & -B_2 \\ E_2 & -B_3 & 0 & B_1 \\ E_3 & B_2 & -B_1 & 0 \end{bmatrix}, \qquad G = \begin{bmatrix} 0 & D_1 & D_2 & D_3 \\ -D_1 & 0 & H_3 & -H_2 \\ -D_2 & -H_3 & 0 & H_1 \\ -D_3 & H_2 & -H_1 & 0 \end{bmatrix}. \tag{2.2}$$

The constitutive relations, which link the fields F and G, can be written in the form

$$F = gGg, \tag{2.3}$$

where $g = g(t,x)$ is a 4×4 symmetric matrix function such that

$$g_{00} < 0 \tag{2.4a}$$

and

$$\tilde{g} = [g_{ij}]_{i,j=1,2,3} \quad \text{is positive definite.} \tag{2.4b}$$

For brevity, we shall call such a matrix function g a *good metric*. (More precisely, we should call it the matrix of a given Lorentz metric in a "good" coordinate system.) For instance, a homogeneous isotropic medium has $g = \sqrt{\mu}\ \mathrm{diag}(-1/\varepsilon\mu, 1, 1, 1)$ and $D = \varepsilon E$, $B = \mu H$.

It is easy to see that the constitutive relations (2.3) always permit D and B to be expressed in terms of E and H. This is the concern of the first proposition.

PROPOSITION 2.1. *If* (2.3) *holds for a good metric g, then there is a unique positive definite, real symmetric* 6×6 *matrix* $\mathscr{M}=\mathscr{M}(t,x)$ *such that*

$$\begin{bmatrix} D \\ B \end{bmatrix}=\mathscr{M}\begin{bmatrix} E \\ H \end{bmatrix}. \tag{2.5}$$

The first step in the proof is

LEMMA 2.2. *The relation* (2.3) *implies that there is a real symmetric* 6×6 *matrix* $\mathscr{N}$ *such that*

$$\begin{bmatrix} -E \\ B \end{bmatrix}=\mathscr{N}\begin{bmatrix} D \\ H \end{bmatrix}. \tag{2.6}$$

Proof. The mapping $(D_1,D_2,D_3,H_1,H_2,H_3)\rightarrow G$ is an isomorphism σ: $\mathbb{R}^6\rightarrow\mathscr{S}$ where $\mathscr{S}$ is the space of 4×4 real skew matrices. The operation $S\rightarrow gSg$ takes $\mathscr{S}$ into itself because g is symmetric, and thus defines a transformation $\mathscr{N}$ on $\mathbb{R}^6$ given by the diagram

$$\begin{tikzcd}
\mathbb{R}^6 \arrow[r, "\sigma"] \arrow[d, "\mathscr{N}"'] & \mathscr{S} \arrow[d] & S \arrow[d] \\
\mathbb{R}^6 \arrow[r, "\sigma"] & \mathscr{S} & gSg
\end{tikzcd}$$

Thus $\mathscr{N}a=\sigma^{-1}g\sigma(a)g$ for $a\in\mathbb{R}^6$ and $\mathscr{N}$ satisfies (2.6). We let $\langle\cdot,\cdot\rangle_6$ denote the scalar product on $\mathbb{R}^6$. Because

$$\langle a,b\rangle_6=-\tfrac{1}{2}\mathrm{Tr}(\sigma(a)\sigma(b)),$$

we have

$$\begin{aligned}\langle\mathscr{N}a,b\rangle_6&=-\tfrac{1}{2}\mathrm{Tr}(g\sigma(a)g\sigma(b))\\ &=-\tfrac{1}{2}\mathrm{Tr}(\sigma(a)g\sigma(b)g)=\langle a,\mathscr{N}b\rangle_6\end{aligned}$$

where we have used the fact that $\mathrm{Tr}(AB)=\mathrm{Tr}(BA)$ for square matrices A and B. This proves the lemma.

Proof of Proposition 2.1. We claim that the matrix for $\mathscr{N}$ is

$$\begin{bmatrix} -R & Q \\ Q^T & S \end{bmatrix} \tag{2.7}$$

where R and S are positive definite 3×3 symmetric matrics. To determine R, we compute $\mathscr{N}a$ on vectors $a=(a_1,a_2,a_3,0,0,0)$. The first three components of $\mathscr{N}a$ can be read off from the first row of

$$g\sigma(a)g=g\begin{bmatrix} 0 & a_1 & a_2 & a_3 \\ -a_1 & 0 & 0 & 0 \\ -a_2 & 0 & 0 & 0 \\ -a_3 & 0 & 0 & 0 \end{bmatrix}g.$$

This yields the matrix $R=-g_{00}\tilde{g}+[g_{i0}g_{j0}]$ $i,j=1,2,3$. But $g_{00}<0$ and $\tilde{g}$ is positive definite by assumption, while $[g_{i0}g_{j0}]$ is always nonnegative.

To show that $S>0$, we consider vectors $a\in\mathbb{R}^6$, $a\neq 0$, of the form $(0,0,0,a_4,a_5,a_6)$. If we calculate $g\sigma(a)g$ for these a, we see that it involves only $\tilde{g}$. Thus

$$\begin{aligned}\langle Sa,a\rangle_6=\langle \mathcal{N}a,a\rangle_6&=-\tfrac{1}{2}\mathrm{Tr}(g\sigma(a)g\sigma(a))\\ &=-\tfrac{1}{2}\mathrm{Tr}(\tilde{g}\sigma(a)\tilde{g}\sigma(a))\\ &=-\tfrac{1}{2}\mathrm{Tr}\big([\tilde{g}^{1/2}\sigma(a)\tilde{g}^{1/2}]^2\big)>0.\end{aligned}$$

By (2.6) and (2.7) we have

$$-E=-RD+QH \quad\text{and}\quad B=Q^{T}D+SH.$$

Therefore

$$D=R^{-1}E+R^{-1}QH \quad\text{and}\quad B=Q^{T}R^{-1}E+Q^{T}R^{-1}QH+SH$$

so that

$$\mathcal{M}=\begin{bmatrix} R^{-1} & R^{-1}Q\\ Q^{T}R^{-1} & Q^{T}R^{-1}Q+S\end{bmatrix}.$$

Finally to show $\mathcal{M}$ is positive definite, we will show $\langle\mathcal{M}a,a\rangle_6>0$ if $a=(E,H)\neq 0$. If $H=0$, then

$$\langle\mathcal{M}a,a\rangle_6=\langle R^{-1}E,E\rangle_3>0.$$

If $H\neq 0$, then

$$\langle\mathcal{M}a,a\rangle_6=\langle R^{-1}E,E\rangle_3+\langle R^{-1}QH,QH\rangle_3+\langle SH,H\rangle_3+2\langle R^{-1}QH,E\rangle_3.$$

Since $R^{-1}>0$, $2|\langle R^{-1}QH,E\rangle|\leqq\langle R^{-1}QH,QH\rangle+\langle R^{-1}E,E\rangle$ so that

$$\langle\mathcal{M}a,a\rangle_6\geqq\langle SH,H\rangle_3>0.$$

Proposition 2.1 is thus proved.

3. Coordinate transformations. We will call a coordinate transformation *proper* if it is given by an equation

$$x=\psi(t,x')\quad\text{where}\quad\left|\frac{\partial\psi}{\partial t}\right|<1 \tag{3.1}$$

and ψ is assumed to be smooth with nonvanishing Jacobian $\partial\psi_j/\partial x'_k$. Writing $x=(x_1,x_2,x_3)$ and $x'=(x'_1,x'_2,x'_3)$, this means $x_j=\psi_j(t,x'_1,x'_2,x'_3)$ for $j=1,2,3$ and t remains unchanged.

Remark. If (3.1) is written in terms of the inverse mapping ϕ, it takes a more complicated form. Indeed, write the equation as $x'=\phi(t,x)$. Then $x'=\phi(t,\psi(t,x'))$. Let us write ϕ_t for the vector $\partial\phi/\partial t$ and ϕ_x for the Jacobian matrix $(\partial\phi_i/\partial x_j)$. Then $\psi_t=\phi_x^{-1}\phi_t$. To be proper means that this vector has length less than one. Simple linear algebra shows that this is equivalent to the condition that the 3×3 matrix

$$\left[\sum_{k=1}^{3}\frac{\partial\phi_i}{\partial x_k}\frac{\partial\phi_j}{\partial x_k}-\frac{\partial\phi_i}{\partial t}\frac{\partial\phi_j}{\partial t}\right]\text{ is positive definite.} \tag{3.2}$$

This was the assumption of Inoue [2] who did not use the simpler statement (3.1).

Our assumption can be interpreted very simply by saying that if x' is fixed and x depends on t according to (3.1), then its speed $|dx/dt|$ is less than the speed of wave propagation which we have taken to be unity.

PROPOSITION 3.1. *Let Σ be a (smooth) hypersurface in space-time $\mathbb{R}^4$. Then Σ is timelike if and only if there exists a proper transformation (defined in a neighborhood of any point on Σ) $x \to x'$, $x = \psi(t, x')$, such that Σ goes into a stationary hypersurface Σ'.*

Proof. By a stationary hypersurface we mean a cylinder parallel to the t-axis. Suppose first that ψ is a proper transformation which takes Σ into Σ' where Σ' is stationary. Let us write $\nu = (\nu_t, \nu_x)$ as a normal vector for Σ, split into t and x components. Then $\nu' = (\nu_t', \nu_x')$ is a normal vector for Σ', split into t and x' components, where

$$\nu_t' = \nu_t + \frac{\partial\psi}{\partial t}\cdot\nu_x \quad\text{and}\quad \nu_x' = \left(\frac{\partial\psi}{\partial x'}\right)\nu_x.$$

Since Σ' is stationary, $\nu_t' = 0$. Since ψ is proper, $|\partial\psi/\partial t| < 1$. Hence $|\nu_t| \leqq |\partial\psi/\partial t||\nu_x| < |\nu_x|$. This means that ν is a spacelike vector and Σ is a timelike surface.

Conversely, let Σ be timelike. Then we may assume Σ is given locally by an equation $x_3 = l(t, x_1, x_2)$. This equation holds in a neighborhood of a point (t^0, x^0), at which point we may assume $\partial l/\partial x_1 = \partial l/\partial x_2 = 0$. Since Σ is timelike and a normal vector is $(\partial_t l, \partial_1 l, \partial_2 l, -1)$, we have

$$(\partial_t l)^2 < (\partial_1 l)^2 + (\partial_2 l)^2 + 1.$$

Therefore in a smaller neighborhood of the point (t^0, x^0) we have $(\partial_t l)^2 < 1$. In that neighborhood we define the transformation

$$\begin{aligned} x_1 &= \psi_1(t, x') = x_1', \\ x_2 &= \psi_2(t, x') = x_2', \\ x_3 &= \psi_3(t, x') = x_3' + l(t, x_1', x_2'). \end{aligned}$$

Then $\partial_t\psi = (0, 0, \partial_t l)$ so that $|\partial_t\psi| < 1$ and the transformation is proper. Furthermore, in the new coordinates the surface Σ' is given (locally) by the equation $x_3' = 0$, which is clearly stationary.

Next we investigate how Maxwell's equations are affected by a change of coordinates. The field tensor F and the metric g transform as cotensors, while G transforms as a contratensor (see Møller [6]). Thus F, G, and g in the (t, x) coordinates become F', G', and g' in the (t', x') coordinates given by

$$\begin{aligned} g' &= |\det\mathcal{T}|^{-1/2}\mathcal{T}^{\tau}g\mathcal{T}, \\ F' &= \mathcal{T}^{\tau}F\mathcal{T}, \\ G' &= |\det\mathcal{T}|\mathcal{T}^{-1}G(\mathcal{T}^{\tau})^{-1} \end{aligned} \tag{3.3}$$

where $\mathcal{T}$ is the Jacobian matrix of the mapping $(t', x') \to (t, x)$. For the mapping $(t, x') \to (t, \psi(t, x'))$ we have

$$\mathcal{T} = \begin{bmatrix} 1 & 0 \\ v & T \end{bmatrix}$$

where T is the 3×3 matrix $(\partial\psi_i/\partial x_j')$ and $v = \partial\psi/\partial t$ is a column 3 vector.

Remark 3.2. If g is the flat metric, $g=\operatorname{diag}(-1,1,1,1)$, and ψ is proper, then g' is a good metric (in the sense of (2.4)). Indeed, writing 4×4 matrices in blocks,

$$\begin{aligned} g' &= |\det T|^{-1/2}\begin{pmatrix} 1 & 0 \\ v & T\end{pmatrix}\begin{pmatrix} -1 & 0 \\ 0 & I\end{pmatrix}\begin{pmatrix} 1 & v^{\tau} \\ 0 & T^{\tau}\end{pmatrix} \\ &= |\det T|^{-1/2}\begin{pmatrix} -1+v^{\tau}v & v^{\tau}T \\ T^{\tau}v & T^{\tau}T\end{pmatrix} \end{aligned}$$

so that $g'_{00}=-1+|v|^2<0$ and $\tilde{g}=T^{\tau}T$ is positive definite. More generally, if $g=\sqrt{\mu}\,\operatorname{diag}(-(\varepsilon\mu)^{-1},1,1,1)$, and $|\partial_t\psi|^2<(\varepsilon\mu)^{-1}$, then g' is a good metric.

In the new coordinates the constitutive relations take the same form:

$$g'G'g'=\mathscr{T}^{\tau}g\mathscr{T}\mathscr{T}^{-1}G(\mathscr{T}^{\tau})^{-1}\mathscr{T}^{\tau}g\mathscr{T}=\mathscr{T}^{\tau}gGg\mathscr{T}=F'.$$

Furthermore, F', G' satisfy the Maxwell equations (2.1) with $(q',J')=|\det\mathscr{T}|^{-1}\mathscr{T}^{-1}(q,J)$ transformed as a 4-vector (see [6]).

4. The perfect conductor. The conducting body occupies the region $\mathscr{O}(t)$ at time t while the exterior region is a vacuum with $\varepsilon=\mu=1$. In a perfect conductor all the fields are assumed to vanish so that the current and charge are distributions supported by the boundary surface. Thus Maxwell's equations (2.1) become:

$$(4.1)\qquad \partial_t E-\nabla\times H=0,\qquad \nabla\cdot E=0\quad \text{in } Q,$$

$$(4.2)\qquad \partial_t H+\nabla\times E=0,\qquad \nabla\cdot H=0\quad \text{in } \mathbb{R}^4$$

with $E=H=0$ in $\mathscr{O}(t)$. These equations are to hold in the sense of distributions. In particular, (4.2) is to hold across Σ. Formally this implies the boundary conditions

$$(4.3)\qquad \nu_t H+\nu_x\times E=0,\qquad \nu_x\cdot H=0\quad \text{on } \Sigma.$$

(Note that when the body is stationary, $\nu_t=0$ and the boundary conditions reduce to the usual conditions for a perfect conductor that E be normal to the boundary surface and that H be tangential.)

We shall be considering vector fields $f=(f_1,f_2,f_3)$ on $\mathbb{R}^3$ with each component square integrable. We write the norm in $L^2(\mathbb{R}^3)^3$ as

$$\|f\|=\left[\int|f(x)|^2dx\right]^{1/2}$$

where $|f(x)|^2=|f_1(x)|^2+|f_2(x)|^2+|f_3(x)|^2$ is the pointwise Euclidean norm. For vector fields E and H we shall write

$$\|E,H\|=\left[\|E\|^2+\|H\|^2\right]^{1/2}.$$

We let $\Omega(t)$ be the exterior region at time t: $\Omega(t)=\mathbb{R}^3\backslash\mathscr{O}(t)$.

THEOREM 4.1 (well-posedness). *Let $e, h\in L^2(\mathbb{R}^3)^3$ with $e=h=0$ in $\mathscr{O}(0)$, $\nabla\cdot e=0$ in $\Omega(0)$ and $\nabla\cdot h=0$ in $\mathbb{R}^3$. Then there exists a unqiue solution $E(t,x)$, $H(t,x)$ to* (4.1), (4.2) *with the initial condition $E(0,x)=e(x)$, $H(0,x)=h(x)$, and such that $t\to(E,H)$ is continuous with values in $L^2(\mathbb{R}^3)^6$ with* $\sup_{0\leqq t\leqq T}\|E(t),H(t)\|\leqq c_T\|e,h\|$.

THEOREM 4.2 (causality). *Let Λ be the backward cone $\Lambda=\{(t,x): |x-y|\leqq s-t\leqq s\}$ with vertex at (s,y). Let (E,H) be a solution of* (4.1), (4.2) *in Λ which is continuous with values in $L^2(\mathbb{R}^3)^6$. If (E,H) vanishes on $\Lambda\cap\{t=0\}$, then $(E,H)=0$ in Λ.*

Theorems 4.1 and 4.2 are a consequence of the following result on local existence and uniqueness.

THEOREM 4.3. *Let e,h be given as in Theorem* 4.1. *Let* $(0,x^0)$ *be a point in the boundary* $\{0\}\times\partial\Omega(0)$. *Then there exists a unique solution to Maxwell's equations* (4.1), (4.2) *in some space-time neighborhood V of* $(0,x^0)$ *with* (i) $E(0)=e$, $H(0)=h$ *in* $V\cap\{t=0\}$ *and* (ii) $t\to(E(t),H(t))$ *continuous with values in* $L^2(\omega)^6$ *for t in a neighborhood of zero. Here ω is an open neighborhood of* x^0, $\omega\subset\{x\colon(0,x)\in V\}$.

First we show how Theorems 4.1 and 4.2 follow from Theorem 4.3.

Proof of Theorem 4.1. Define K as the set of points $(t,x)\in\mathbb{R}^4$ such that the ball $\{|y-x|\leqq|t|\}\subset\Omega(0)$. $K\subset Q$ by the timelike condition assumed for Σ. Of course, in K the desired solution will be equal to the usual free solution of Maxwell's equation with the same initial data. At each point $(0,x^0)$ of $\{0\}\times\partial\Omega(0)$ Theorem 4.3 guarantees the existence of a solution of (4.1), (4.2) in some neighborhood V of $(0,x^0)$. The compact set $\{0\}\times\partial\Omega(0)$ may be covered by a finite number of such neighborhoods. Where these neighborhoods overlap, the local solutions will agree by the uniqueness part of Theorem 4.3. In this way a unique solution is shown to exist in $Q\cap\{|t|\leqq\varepsilon\}$ for some $\varepsilon>0$. The $\varepsilon>0$ can be determined in a uniform fashion as we march ahead in time with steps of length ε.

Proof of Theorem 4.2. The result is clearly true if $\Lambda\cap\{t=0\}\subset Q$ because then $\Lambda\subset Q$ by the timelike property of Σ and we may use the well-known causality properties of the free Maxwell equations. Thus we assume Λ meets Σ. It suffices to show that $(E,H)=0$ in $\Lambda_\delta\cap Q$ for each $\delta>0$ where Λ_δ is the slightly smaller cone with vertex at $(s-\delta,y)$. By the local Theorem 4.3, $(E,H)=0$ in a neighborhood of $\Lambda_\delta\cap\{(0)\}$. It follows that $(E,H)=0$ in $\Lambda_\delta\cap Q\cap\{t\leqq\varepsilon\}$ for some $\varepsilon>0$. Due to the compactness of $\Lambda_\delta\cap\Sigma$, we may repeat this argument a finite number of times to show that $(E,H)=0$ in $\Lambda_\delta\cap Q$.

Now we turn to the proof of Theorem 4.3. We apply Proposition 3.1 to the hypersurface $\Sigma=\partial Q$ in a neighborhood of the point $(0,x^0)$. Thus there exists a proper transformation $(t,x)\to(t,x')$ which takes $(0,x^0)$ into $(0,0)$ and Σ into $\Sigma'=\{x_3'=0\}$. That is, Σ' is stationary. The flat metric g goes into a good metric g' and the Maxwell equations are satisfied for the fields E', D', H', B' in a neighborhood V' of $(0,0)$ (see Møller [6]).

$$\begin{aligned}&\partial_t D'-\nabla\times H'=0, &&\nabla\cdot D'=0 &&\text{in } V'\cap Q',\\ &\partial_t B'+\nabla\times E'=0, &&\nabla\cdot B'=0 &&\text{in } V'\end{aligned}\tag{4.4}$$

with initial conditions

$$D'(0,x')=d'(x),\qquad B'(0,x')=b'(x)$$

where $\nabla\cdot d'=0$ in $V'\cap Q'\cap\{t=0\}$ and $\nabla\cdot b'=0$ in $V'\cap\{t=0\}$. The fields still vanish on the body, that is, on $V'\backslash\overline{Q}'$. Since Σ' is stationary, the boundary condition has become (formally) $n\times E'=0$ and $n\cdot B'=0$ on Σ' where $n=(0,0,1)$.

Let Ω' be the hemisphere $\{x'\colon|x'|\leqq\delta\}\cap Q'\cap\{t=0\}$, and choose T, $\delta>0$ so small that $(-T,T)\times\Omega'\subset V'\cap Q'$. By Proposition 2.1, D' and B' are expressible in terms of E' and H'. Therefore the evolution equations of (4.4) may be written

$$\partial_t\left(\mathscr{M}\begin{bmatrix}E'\\H'\end{bmatrix}\right)+\begin{bmatrix}-\nabla\times H'\\\nabla\times E'\end{bmatrix}=0\quad\text{in }(-T,T)\times\Omega'\tag{4.5}$$

with initial conditions

$$\begin{bmatrix}E'\\H'\end{bmatrix}(0)=\begin{bmatrix}e'\\h'\end{bmatrix}=\mathscr{M}^{-1}\begin{bmatrix}d'\\b'\end{bmatrix}\quad\text{in }\Omega'.$$

We shall in addition impose the boundary condition $n \times E' = 0$ on $\partial\Omega'$ where n is the unit normal to $\partial\Omega'$.

The initial-boundary value problem (4.5) has a unique solution, continuous with values in $L^2(\Omega')^6$ by virtue of Proposition 4.4 to be proved shortly. Assuming the existence and uniqueness of solutions of (4.5), it remains to check the auxiliary conditions $\nabla \cdot D' = 0$ and $\nabla \cdot B' = 0$ of (4.4). But this follows immediately from the evolution equations (4.4) and the fact that $\nabla \cdot d' = 0$ and $\nabla \cdot b' = 0$. The solution thus constructed on $(-T, T) \times \Omega'$ may be then transformed back into a solution of the original problem (4.1), (4.2) in some neighborhood V of $(0, x^0)$.

Remark. If Σ is a C^∞ hypersurface and the initial data are C^∞, so is the unique solution. We omit the proof, which is fairly standard.

Proposition 4.4. *Let e', h' be given in $L^2(\Omega')^3$. Then the problem* (4.5) *has a unique solution (E', H') continuous on $|t| \leqq T$ with values in $L^2(\Omega')^6$.*

Proof. We shall write (4.5) as an equation in the Hilbert space $X = L^2(\Omega')^6$. Let $u = (E', H')$ and define the operator

$$A_0 = \begin{bmatrix} 0 & \nabla \times \\ -\nabla \times & 0 \end{bmatrix}$$

with domain $D(A_0) = \{u = (E', H') \colon u \in C^\infty(\bar{\Omega}')^6,\ n \times E' = 0 \text{ on } \partial\Omega'\}$. Define A as the closure of A_0. It is not difficult to show that A is skew-adjoint in X: $A^* = -A$ (see Schmidt [8, p. 313]). Furthermore let $M(t)$ be the operator on X of multiplication by $\mathscr{M}(t, x')$. Then $M(t)$ is a bounded self-adjoint operator on X which is (strictly) positive and depends smoothly on t. It suffices to show that the initial value problem

$$\frac{d}{dt}[M(t)u(t)] = Au(t), \qquad u(0) = u_0 \in X \tag{4.6}$$

can be solved uniquely in X. This is true using only the abstract properties of A and $M(t)$ mentioned above.

To do so, we apply the theory of Kato [3]. We rewrite (4.6) in the form

$$\frac{du}{dt} = [M(t)]^{-1} Au(t) - [M(t)]^{-1}\left[\frac{dM}{dt}\right]u(t). \tag{4.7}$$

Let $A(t) = [M(t)]^{-1}[A - M'(t)]$ with the constant domain $D(A(t)) = D(A)$. Here the prime denotes the time-derivative. We claim that the family of operators $\{A(t)\}$ is stable in the sense of Kato.

Indeed, we apply [3, Prop. 3.4] with $(u, v)_t = (M(t)u, v)_X$. Then (for $0 \leqq t, s \leqq T$)

$$\begin{aligned} \left| \|u\|_t^2 - \|u\|_s^2 \right| &= |(M(t) - M(s)u, u)| \\ &\leqq k|t-s| \, \|u\|_X^2 \leqq k'|t-s| \, \|u\|_s^2, \end{aligned}$$

or

$$\left| \|u\|_t^2 \|u\|_s^{-2} - 1 \right| \leqq k'|t-s|,$$

whence

$$\exp(-c|t-s|) \leqq \|u\|_t \|u\|_s^{-1} \leqq \exp(c|t-s|)$$

for some constant c. Furthermore

$$\left([M(t)]^{-1}Au,v\right)_t=(Au,v)=0$$

so that $[M(t)]^{-1}A$ generates a contraction semigroup in $X_t(=X$ with $\| \|_t)$. By Propositions 3.4 and 3.5, $\{A(t)\}$ is stable.

Furthermore we define $Y=D(A)$ so that $t\rightarrow A(t)$ is a C^1 map from $\mathbb{R}$ into $\mathcal{L}(Y,X)$. Now apply [3, Remark 6.2]. Thus all the conclusions of Kato's Theorem 6.1 are valid. Therefore there exists a unique family of bounded operators $U(t,s)$ on X $0\leqq s\leqq t\leqq T$ such that:

(a) $U(t,s)$ is strongly continuous in t,s with values in $\mathcal{L}(X,X)$, $U(t,t)=I$ and

$$\|U(t,s)\|_{\mathcal{L}(X,X)}\leqq K\exp(\beta(t-s));$$

(b) $U(t,r)=U(t,s)U(s,r)$;

(d) $(d/ds)[U(t,s)f]=U(t,s)A(s)f$ for all $f\in Y$;

(e) $U(t,s)Y\subset Y$ with norm $\leqq\tilde{K}\exp(\tilde{\beta}(t-s))$;

(f′) $U(t,s)$ is strongly continuous with values in $\mathcal{L}(Y,Y)$;

(h) $(d/dt)[U(t,s)f]=-A(t)U(t,s)f$ for all $f\in Y$.

Since A is skew-adjoint, we may reverse time in (4.7) and apply the same results to obtain the existence of $U(t,s)$ for $-T\leqq s\leqq t\leqq T$.

5. The dielectric. In this section we assume that the body is a dielectric with constants ε and μ which moves at a speed v which is less than the speed of light $(\varepsilon\mu)^{-1/2}$ in the body:

$$|v|<\frac{1}{\sqrt{\varepsilon\mu}}<1. \tag{5.1}$$

The second inequality says that light travels at a slower speed in the body than in a vacuum (c has been normalized to 1). As in the introduction, let $\mathcal{O}(t)$ be the region occupied by the body at time t. Let $\mathcal{O}'$ be a fixed compact set in $\mathbb{R}^3$ with smooth boundary. We assume that the motion is described by a function $\psi(t,x')$ defined on a neighborhood of $\mathbb{R}\times\mathcal{O}'$. For each reference point $x'\in\mathcal{O}'$, the point $\psi(t,x')$ is the position at time t of the corresponding material point. Thus $x'\rightarrow\psi(t,x')$ carries $\mathcal{O}'$ onto $\mathcal{O}(t)$. A particular point $x=\psi(t,x')$ moves with the velocity $v=\partial\psi/\partial t$. We assume that ψ is smooth, has nonvanishing Jacobian $(\partial\psi_j/\partial x_k')$ and satisfies (5.1). In particular it is a proper transformation in the sense of §3 so that $\Sigma=\partial Q$ is timelike.

Since the moving body is a dielectric, the Maxwell equations (2.1) and (2.2) are valid in $\mathbb{R}^4$ with $J=q=0$. However, the four fields have jumps across the boundary. Denote by $[D]$ the jump of D across Σ, etc. Then (2.1), (2.2) imply the jump conditions

$$\begin{aligned}\nu_t[D]-\nu_x\times[H]&=0, & \nu_x\cdot[D]&=0,\\ \nu_t[B]+\nu_x\times[E]&=0, & \nu_x\cdot[B]&=0\end{aligned}$$

on Σ, at least if the solutions are piecewise smooth. (When the body is stationary, these reduce to the usual conditions that the tangential components of H and E and the normal components of D and B be continuous across Σ.)

In the vacuum, occupying the region Q, the constitutive relations are $D=E$ and $B=E$. When the body is at rest, the fields in the body satisfy $D=\varepsilon E$ and $B=\mu H$ where ε and μ are the assumed dielectric constant and permeability. However when the body is in motion, the constitutive relations must be modified. Their exact nature is a matter of some controversy [7] but their most widely accepted formulation, called the Minkowski formulation or the "instantaneous rest frame hypothesis" [10], is as follows. The constitutive relations at a point (t,x) in the body are taken to be those of a body moving with constant velocity equal to the instantaneous velocity of the body at that point, ignoring acceleration. Thus they are given in [9] as

$$\begin{aligned} D+v\times H&=\varepsilon[E+v\times B],\\ B-v\times E&=\mu[H-v\times D] \end{aligned} \tag{5.2}$$

where $v=v(t,x)=\partial\psi/\partial t$ is the velocity of the body at that point.

Relations (5.2) can be written in the standard form

$$F=gGg \tag{5.3}$$

if we use the notation of §2. Indeed, fix the point (t,x) and let L denote the Lorentz transformation corresponding to velocity v. That is,

$$L=\begin{bmatrix} \gamma & -\gamma\nu^{T} \\ -\gamma\nu & I+(\gamma-1)vv^{T}|v|^{-2} \end{bmatrix},$$

a 4×4 matrix with the lower right corner a 3×3 block where $\gamma^2=(1-|v|^2)^{-1}=(1-\nu^{T}v)^{-1}$. (This matrix transforms a point moving with velocity v into a stationary point.) Let

$$g=L\sqrt{\mu}\begin{bmatrix} -(\varepsilon\mu)^{-1} & 0 \\ 0 & I \end{bmatrix}L.$$

Then an easy calculation shows that inside the body,

$$g=\sqrt{\mu}\,\gamma^2\begin{bmatrix} -(\varepsilon\mu)^{-1}+|v|^2 & \left[(\varepsilon\mu)^{-1}-1\right]v^{T} \\ \left((\varepsilon\mu)^{-1}-1\right)v & \Gamma \end{bmatrix}$$

where Γ is the 3×3 matrix

$$\Gamma=\frac{1}{\gamma^2}\left(I-\frac{vv^{T}}{|v|^2}\right)+\left(1-\frac{|v|^2}{\varepsilon\mu}\right)\frac{vv^{T}}{|v|^2}.$$

Because of (5.1), $g_{00}<0$ and Γ is positive definite. So g is a good metric in the sense of §2. By Proposition 2.1 we can write

$$\begin{bmatrix} D \\ B \end{bmatrix}=\mathscr{R}\begin{bmatrix} E \\ H \end{bmatrix} \tag{5.4}$$

where $\mathscr{R}=\mathscr{R}(t,x)$ is a positive definite symmetric 6×6 matrix. Of course for this very simple transformation it is easy to write (5.4) explicitly. The result is (see [9])

$$\begin{aligned} &D_{\|}=\varepsilon E_{\|}, \qquad B_{\|}=\mu H_{\|},\\ &\left(1-\varepsilon\mu|v|^2\right)D_{\perp}=\varepsilon\left(1-|v|^2\right)E_{\perp}+(\varepsilon\mu-1)(v\times H_{\perp}),\\ &\left(1-\varepsilon\mu|v|^2\right)B_{\perp}=\mu\left(1-|v|^2\right)H_{\perp}-(\varepsilon\mu-1)(v\times E_{\perp}) \end{aligned} \tag{5.5}$$

where the fields are resolved into their components parallel and normal to the velocity.

We set $\mathscr{R}(x,t)=I$ in Q. Then Maxwell's equations and the constitutive relations (5.4) may be combined in the set of equations

$$\partial_t \mathscr{R}(t)\begin{bmatrix} E \\ H \end{bmatrix} = \begin{bmatrix} \nabla \times E \\ -\nabla \times H \end{bmatrix}, \tag{5.6}$$

$$\nabla \cdot B = \nabla \cdot D = 0 \quad \text{where} \begin{bmatrix} D \\ B \end{bmatrix} = \mathscr{R}\begin{bmatrix} E \\ H \end{bmatrix}. \tag{5.7}$$

Both sets of equations are to hold in the sense of distributions on $\mathbb{R}^4$.

THEOREM 5.1. *Assume* (5.2). *Let* e, $h \in L^2(\mathbb{R}^3)^6$ *such that* $[d,b]^\tau = \mathscr{R}(0)[e,h]^\tau$ *satisfy* $\nabla \cdot d = \nabla \cdot b = 0$ *in* $\mathbb{R}^3$. *Then there exists a unique solution* E, H *of* (5.6), (5.7) *such that* $t \to (E(t), H(t))$ *is continuous with values in* $L^2(\mathbb{R}^3)^6$ *with* $E(0,x) = e(x)$ *and* $H(0,x) = h(x)$. *For any* $T > 0$ *we have*

$$\sup_{0 \leqq t \leqq T} \|E(t), H(t)\| \leqq C_T \|e,h\|.$$

Furthermore the causality principle holds as in Theorem 4.2.

Proof. As in §4, Theorem 5.1 can be shown to be a consequence of a local existence theorem. Thus it suffices to construct a solution in a space-time neighborhood of $\{0\} \times \mathcal{O}(0)$. We transform to new spatial coordinates x' by the equation $x = \psi(t,x')$, thereby mapping Σ into a stationary surface $\Sigma' = \mathbb{R} \times \partial\mathcal{O}'$. The metric g given by (5.3) transforms to a new metric g'. We claim that g' is again a good metric. Indeed let us use the previous notation $v = \partial\psi/\partial t$ and $T = (\partial\psi_i/\partial x_j')$. Then in $\mathcal{O}'$,

$$g' = |T|^{-1/2}\mathscr{T}^\tau g \mathscr{T} = \begin{bmatrix} g_{00}' & w \\ w^\tau & \tilde{g}' \end{bmatrix}$$

where $\mathscr{T}$ is as in §3 and g is given by (5.3). A direct calculation shows

$$g_{00}' = -(\mu/|T|)^{1/2}(\varepsilon\mu)^{-1}(1-|v|^2) < 0$$

and

$$\tilde{g}' = -(\mu/|T|)^{1/2}\gamma^2 T^\tau \Gamma T$$

is positive definite since $\Gamma > 0$. Therefore g' satisfies (2.4) inside $\mathcal{O}'$. Outside of $\mathcal{O}'$, g' is given by (3.3), and is also a good metric. Thus by Proposition 2.1 there is a positive definite symmetric 6×6 matrix $\mathscr{M} = \mathscr{M}(t,x')$ such that $[D',B']^\tau = \mathscr{M}[E',H']^\tau$, and the Maxwell equations in the (t,x') coordinates may be written in the form (4.5) in a neighborhood of $\mathcal{O}'$. Note that $x' \to \mathscr{M}(t,x')$ has a jump discontinuity across $\partial\mathcal{O}'$. With an appropriate use of cut-off functions we can consider the equation to hold in all of $\mathbb{R}^4$, with the addition of an inhomogeneous term. We can again apply Kato's theorem as in Proposition 4.4 to prove existence and uniqueness of solutions for the evolution equations. The conditions $\nabla \cdot D' = \nabla \cdot B' = 0$ then follow in the standard manner from the initial conditions.

6. The scalar wave equation. For a function $u(t,x)$, $x \in \mathbb{R}^n$, we consider the problem

$$u_{tt} - \Delta u = 0 \quad \text{in } Q, \tag{6.1}$$

$$\frac{\partial u}{\partial \nu^*} + \alpha\frac{\partial u}{\partial \zeta} + \beta u = 0 \quad \text{on } \Sigma \tag{6.2}$$

where $\alpha \geqq 0$ and β are smooth real functions on Σ. Here as before $\nu = (\nu_t, \nu_x)$ is the space-time normal to Σ, $\nu^* = (-\nu_t, \nu_x)$ the conormal, and $\zeta = (\zeta_t, \zeta_x)$ is a fixed tangential vector field on Σ with $\zeta_t > 0$ and $|\zeta_x| < \zeta_t$.

Let $\Omega(t)$ denote the exterior region in $\mathbb{R}^n$: $\Omega(t) = \mathbb{R}^n \setminus \mathcal{O}(t)$.

We define $L(t)$ to be the closure of $C_0^\infty(\bar{\Omega}(t)) \times C_0^\infty(\bar{\Omega}(t))$ in the energy norm

$$\|f\|^2 = \frac{1}{2} \int_{\Omega(t)} \left(|\nabla f_1|^2 + |f_2|^2 \right) dx$$

for a pair of functions $f = [f_1, f_2]$. For a solution of (6.1), the energy norm is

$$\|u(t)\|^2 = \frac{1}{2} \int_{\Omega(t)} \left\{ |\nabla u(t,x)|^2 + |u_t(t,x)|^2 \right\} dx.$$

THEOREM 6.1. *Let $f \in L(0)$. Then there is a unique solution $u(t,x)$ of* (6.1) *and* (6.2) *such that*

(i) *$u(t, \cdot) \in L(t)$ for each t;*

(ii) *when extended by zero inside $\mathcal{O}(t)$, $t \to u_t(t, \cdot)$ and $t \to \nabla u(t, \cdot)$ are continuous with values in $L^2(\mathbb{R}^3)$;*

(iii) *$u(0,x) = f_1(x)$ and $u_t(0,x) = f_2(x)$.*

For each $T > 0$

$$\sup_{0 \leqq t \leqq T} \|u(t)\| \leqq c_T \|f\| \tag{6.3}$$

where c_T does not depend on f.

Before proving Theorem 6.1, we make a brief digression to show that the boundary condition (6.2) is the "natural" one which yields an energy estimate (6.3).

PROPOSITION 6.2. *Let $u(t,x)$ be a smooth solution of* (6.1), (6.2) *with $u(0,x)$, $u_t(0,x)$ having compact support. Then* (6.3) *holds.*

Proof. We may assume the vector field $\zeta = (1, h)$ where h is a smooth vector field on Σ with values in $\mathbb{R}^n$ with $|h| < 1$. We extend h to a neighborhood of Σ and then smoothly cut off so that we have h defined on $\bar{Q}$ with $|h| < 1$. It suffices to consider only real solutions. We multiply (6.1) by $u_t + h \cdot \nabla u$. There results

$$\begin{aligned} 0 &= (u_{tt} - \Delta u)(u_t + h \cdot \nabla u) \\ &= \partial_t \left\{ \frac{1}{2}\left(u_t^2 + |\nabla u|^2\right) + u_t h \cdot \nabla u \right\} \\ &\quad - \nabla \cdot \left\{ \frac{1}{2} h \left(u_t^2 - |\nabla u|^2\right) + \nabla u (u_t + h \cdot \nabla u) \right\} - q \end{aligned} \tag{6.4}$$

where q is quadratic in u_t and ∇u. If we integrate the identity (6.4) over $Q \cap \{0 \leqq t \leqq T\}$, we find

$$\begin{aligned} &\frac{1}{2} \int_{\Omega(T)} \left(u_t^2 + |\nabla u|^2 + 2u_t h \cdot \nabla u \right) dx + \int_{\Sigma \cap \{0 \leqq t \leqq T\}} B\, ds \\ &= \frac{1}{2} \int_{\Omega(0)} \left(u_t^2 + |\nabla u|^2 + 2u_t h \cdot \nabla u \right) dx + \int_0^T \int_{\Omega(t)} q\, dx\, dt. \end{aligned} \tag{6.5}$$

Since $|h|<1$, the energy form

$$\frac{1}{2}\left(u_t^2+|\nabla u|^2\right)+u_t h\cdot\nabla u \geqq \frac{1}{2}(1-|h|)\left(u_t^2+|\nabla u|^2\right)$$

is positive definite, and, of course, also bounded above by a constant times the energy norm of u. It remains to show that the boundary integral $\int_\Sigma B\,dS$ is bounded below. Now

$$B=\tfrac{1}{2}\nu_t\left\{u_t^2+|\nabla u|^2+2u_t h\cdot\nabla u\right\}$$
$$-\tfrac{1}{2}(\nu_x\cdot h)\left\{u_t^2-|\nabla u|^2\right\}-(\nu_x\cdot\nabla u)\{u_t+h\cdot\nabla u\}.$$

Since $(1,h)$ is tangent to Σ, $\nu_t+\nu_x\cdot h=0$. Therefore

$$B=\nu_t\left(u_t^2+u_t h\cdot\nabla u\right)-(\nu_x\cdot\nabla u)(u_t+h\cdot\nabla u)$$
$$=-\left(\frac{\partial u}{\partial\nu^*}\right)\left(\frac{\partial u}{\partial\zeta}\right)=\alpha\left(\frac{\partial u}{\partial\zeta}\right)^2+\frac{1}{2}\beta\frac{\partial}{\partial\zeta}(u^2).$$

The first term is nonnegative since $\alpha\geqq 0$. The second term can be integrated to get a lower bound. Then an application of Gronwall's lemma to (6.5) shows that the energy is bounded over the interval $[0,T]$.

To prove Theorem 6.1, we first localize the problem and then transform Σ into a fixed boundary, thereby introducing variable coefficients into the equation as we did for Maxwell's equations. Without loss of generality we assume that in a neighborhood of a point $(0,x^0)$, $x^0\in\partial\Omega(0)$, Σ is given by

$$x_n=l(t,x_1,x_2,\cdots,x_{n-1})$$

where $|l_t|<1$ and $l_j=\partial l/\partial x_j=0$ at $(0,x^0)$. We change variables

$$y_j=x_j \quad \text{for } j=1,2,\cdots,n-1,$$
$$z=x_n-l(t,x_1,\cdots,x_{n-1}).$$

Thereby the wave equation (6.1) transforms into

$$(6.6)\qquad u_{tt}-2au_{zt}-b^2u_{zz}-\Delta_y u+2c\cdot\nabla_y u_z+u_z=0$$

for $z>0$ where Δ_y and ∇_y are taken with respect to $y_1,\cdots,y_{n-1}$. Keeping in mind that ν is the normal exterior to Q, we see that the boundary condition (6.2) becomes

$$(6.7)\qquad b^2u_z-c\cdot\nabla_y u+au_t-\alpha\left(u_t+h'\cdot\nabla_y u\right)-\beta u=0 \quad \text{on } z=0.$$

In terms of the function l, we have

$$a=l_t,\qquad c=(l_1,\cdots,l_{n-1}),$$
$$b^2=1-a^2+|c|^2,\qquad d=\frac{\partial^2 l}{\partial t^2}-\sum_{j=1}^{n-1}\frac{\partial^2 l}{\partial x_j^2}.$$

The tangent vector $\zeta=(1,h)$ goes into $(1,h',0)$ where $h'=(h_1,\cdots,h_{n-1})$. This kind of problem has been studied by Miyatake [5]. We shall verify Miyatake's condition for the well-posedness in energy norm of (6.6), (6.7). By the localization procedure we have used earlier, this will prove Theorem 6.1.

We write τ,ξ,η for the dual variables to t,z,y where $\tau=\sigma-i\gamma\in\mathbb{C}$, $\xi\in\mathbb{R}$ and $\eta\in\mathbb{R}^{n-1}$. The principal symbol of the operator (6.6) is $P(\tau,\xi,\eta)$ where

$$-b^2P(\tau,\xi,\eta)=\tau^2-2a\xi\tau-b^2\xi^2-|\eta|^2+2\xi c\cdot\eta.$$

The principal symbol of the boundary operator is

$$B(\tau,\xi,\eta)=b^2\xi-c\cdot\eta+a\tau-\alpha(\tau+h'\cdot\eta).$$

The root $\xi_+(\tau,\eta)$ of $P(\tau,\xi,\eta)=0$ with positive imaginary part is

$$\xi_+(\tau,\eta)=b^{-2}\left[c\cdot\eta-a\tau+\sqrt{D}\right]$$

where $\sqrt{D}$ is the root with positive imaginary part of

$$D=(c\cdot\eta)^2-b^2|\eta|^2-2a(c\cdot\eta)\tau+(a^2+b^2)\tau^2.$$

To verify Miyatake's condition, we only have to prove

LEMMA 6.3. *$|B(\tau,\xi_+(\tau,\eta),\eta\,|\geqq\text{const}\,\gamma^{1/2}$ for (y,t) in a neighborhood of* (0,0) *and for $\gamma^2+\sigma^2+|\eta|^2=1$, $\gamma>0$.*

Proof. We see that

$$B(\tau,\xi_+(\tau,\eta),\eta)=\sqrt{D}-\alpha(\tau+h'\cdot\eta).$$

Here $\tau=\sigma-i\gamma$ is complex, but a,b,c,h', and α are real. Writing $D=D_1+iD_2$, we have

$$D_1=\operatorname{Re}D=(c\cdot\eta)^2-b^2|\eta|^2-2\sigma ac\cdot\eta+\left(1+|c|^2\right)\left(-1+2\sigma^2+|\eta|\right),$$

and

$$D_2=\operatorname{Im}D=2\gamma\left(ac\cdot\eta-\sigma\left(1+|c|^2\right)\right).$$

First suppose $\alpha>0$ and $\sigma^2+|\eta|^2+\gamma^2=1$, $\gamma\geqq0$. We claim that in this case $B\neq0$ at $(y,t)=(0,0)$. At $(y,t)=(0,0)$, we have $c=0$, and $a^2+b^2=1$, which yields

$$D_1=(a^2-1)|\eta|^2+\sigma^2-\gamma^2\quad\text{and}\quad D_2=-2\gamma\sigma.$$

In case $\eta=0$, $D=\tau^2$ and $\sqrt{D}=-\tau$ so that $|B|=|\sqrt{D}-\alpha\tau|=1+\alpha\neq0$. Thus we may assume $\eta\neq0$. Since D_2 has the opposite sign to σ, so does $\operatorname{Re}\sqrt{D}$. Therefore

$$|\operatorname{Re}B|\geqq|\operatorname{Re}\sqrt{D}-\alpha\sigma|-\alpha|h'||\eta|\geqq\alpha\left(|\sigma|-|h'||\eta|\right).$$

Suppose now that $\operatorname{Re}B=0$ at $(y,t)=(0,0)$. Then $|\sigma|\leqq|h'||\eta|$. Hence

$$D_1\leqq\left(a^2-1+|h'|^2\right)|\eta|^2-\gamma^2<-\gamma^2$$

because $a=l_t=l_t-c\cdot h'=h_n$ at (0,0) and $a^2-1+|h'|^2=|h|^2-1<0$. Since $D_1<-\gamma^2$, we must have $\operatorname{Im}\sqrt{D}>\gamma$. Therefore

$$\operatorname{Im}B=\operatorname{Im}\sqrt{D}+\alpha\gamma>(1+\alpha)\gamma\geqq0$$

whence $\operatorname{Im}B\neq0$. In any case $B\neq0$ at $(y,t)=(0,0)$, which establishes the claim. But this implies that B is bounded away from zero for (y,t) in a neighborhood of (0,0) so that the lemma is proved when $\alpha>0$.

Next we consider the case $\alpha=0$, but do not restrict our attention just to the point $(0,0)$. We note that we can write

$$D_1=(c\cdot\eta)^2+a^2|\eta|^2-1-|c|^2-\sigma\gamma^{-1}D_2\leqq a^2-1-\sigma\gamma^{-1}D_2$$

because $|\eta|\leqq 1$. This inequality implies that when $\sigma>0$, D lies below the line with negative slope $D_2=-\gamma\sigma^{-1}(D_1+1-a^2)$. When $\sigma<0$, D lies above the line with positive slope $D_2=\gamma|\sigma^{-1}|(D_1+1-a^2)$. Both lines pass through $(a^2-1,0)$ where $a^2-1<0$ because $|a|=|l_t|<1$. When $\sigma=0$, $D_1\geqq a^2-1<0$. In any case, $|D|\geqq \text{const}\,\gamma$, so that when $\alpha=0$,

$$|B(\tau,\xi_+(\tau,\eta),\eta)|=|\sqrt{D}\,|\geqq \text{const}\,\gamma^{1/2}$$

for $\sigma^2+\gamma^2+|\eta|^2=1$, $\gamma>0$. The lemma is proved.

Acknowledgments. We thank J. Ralston for some spirited discussions of the Neumann problem for the wave equation, R. Showalter for suggesting the possible utility of Kato's theory, and especially T. Kato for helping us apply his elegant paper [3] in our special situation.

REFERENCES

[1] J. Cooper and C. Bardos, *A non linear wave equation in a time dependent domain*, J. Math. Anal. Appl., 42 (1973), pp. 29–60.

[2] A. Inoue, *Sur* $\Box u+u^3=f$ *dans un domaine noncylindrique*, J. Math. Anal. Appl., 46 (1973), pp. 777–819.

[3] T. Kato, *Linear evolution equations of "hyperbolic" type*, J. Fac. Sci. Univ. Tokyo, 17 (1970), pp. 241–258.

[4] P. Lax and R. Phillips, *Local boundary conditions for dissipative symmetric linear differential operators*, Comm. Pure Appl. Math., 13 (1960), pp. 427–455.

[5] S. Miyatake, *Mixed problem for hyperbolic equation of second order*, J. Math. Kyoto Univ., 13 (1973), pp. 435–487.

[6] C. Møller, *The Theory of Relativity*, 2nd ed. Oxford Univ. Pres, London, 1972.

[7] P. Penfield and H. Haus, *Electrodynamics of Moving Media*, Research Monograph No. 40, MIT Press, Cambridge, MA, 1967.

[8] G. Schmidt, *Spectral and scattering theory for Maxwell's equations in an exterior domain*, Arch. Rat. Mech. Anal., 28 (1968), pp. 284–322.

[9] A. Sommerfeld, *Electrodynamics*, Lectures on Theoretical Physics, Vol. III, Academic Press, New York, 1952.

[10] J. Van Bladel, *Electromagnetic fields in the presence of rotating bodies*, Proc. IEEE, 64 (1976), pp. 301–318.

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

A NONLINEAR INTEGRAL OPERATOR ARISING FROM A MODEL IN POPULATION GENETICS III. HETEROZYGOTE INFERIOR CASE*

ROGER LUI[†]

Abstract. We study asymptotic behavior of the solutions to the recursion $u_{n+1}=Q[u_n]$ for $n \geqq 0$. Here $Q[u]=K*(g \circ u)$ acts on functions bounded between 0 and 1, K is a probability density, $g \in C^2[0,1]$ is increasing, $g(0)=0$, $g'(0)<1$, $g(1)=1$, $g'(1)<1$ and there exists $\alpha \in (0,1)$ such that $g(u)<u$ in $(0,\alpha)$, $g(u)>u$ in $(\alpha, 1)$. It is known that a nonincreasing travelling wave w facing right with speed c exists for one value of $c=c_+^*$ and a nondecreasing travelling wave $\bar{w}$ facing left exists for $c=c_-^*$. We prove here that if $c_-^*<c_+^*$, $u_0(\pm\infty)<\alpha$ and u_0 is superthreshold, then given w, $\bar{w}$, u_n is trapped by suitable translations of the function $w(x-nc_+^*)+\bar{w}(x-nc_-^*)-1$ as $n \rightarrow \infty$. If in addition, K is the normal density and u_0 has compact support, then u_n converges exponentially to the above function for some w and $\bar{w}$.

1. Introduction. This is the fourth in a series of papers (see [23], [24], [25]) concerning the long-term behavior of a discrete time population genetics model in which the individuals are assumed to be living in a homogeneous one-dimensional habitat. The model, proposed by Weinberger [36] to describe the spread of an advantageous gene, is actually an improvement of a similar model proposed by R. A. Fisher in his classic paper [11].

In Fisher's model, the fraction $u(x,t)$ of the advantageous genes in the population, at time t and at point x, is governed by a partial differential equation of the form

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2}+f(u), \tag{1.1}$$

where $f(u)$ satisfies the conditions $f \in C^1[0,1]$, $f(0)=0$, $f(1)=0$.

We are interested in how the advantageous genes spread through the population in the long run. Mathematically, this corresponds to describing the limiting behavior of the solution $u(x,t)$ of (1.1) with the initial condition $u(x,0)=u_0(x)$.

Many far reaching results have been obtained over the years on this problem, [1], [3], [4], [6]–[8], [12], [17], [18], [21], [26], [29], [30], [33]. Extensions have also been made. For example, the case when the habitat is no longer homogeneous and f depends on both u and x has been considered in [9], [10], [12], [28]. Partial results for the fully nonlinear case, $u_t=f(u_{xx}, u_x, u)$, where $(\partial/\partial\alpha) f(\alpha,\beta,\gamma) \geqq 1$, are obtained in [13] and [14] and for the quarter-space problem in [34], [35]. However, if we allow more interactions between the species in the population, then we have to consider a system rather than a single equation, or if we assume that the habitat is multi-dimensional, then $\partial^2 u/\partial x^2$ in (1.1) must be replaced by the Laplace operator Δu. These problems are substantially more difficult than (1.1) and so far very little is known, [2], [16], [20].

The purpose of this series of papers is to show that most of the results obtained for Fisher's equations are also valid for Weinberger's model. We begin by briefly describing the model itself. Further details may be found in [36] and [38].

* Received by the editors June 23, 1983, and in revised form July 15, 1984. This work was done while the author was at San Diego State University.

[†] Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, Massachusetts 01609.

Consider a diploid population whose members carry a certain type of gene that occurs in two variant forms labeled as A and a. There are then three genotypes: the homozygotes AA and aa and the heterozygote Aa. Individuals from this population are classified according to genotype. We shall make simplifying assumptions in order to work with a single equation involving the gene fraction instead of a system of equations relating the frequencies of the genotypes. The gene fraction $u(x)$ here is defined as the ratio of the number of alleles of type A to the total number of alleles of type A and a at the point x.

We assume the habitat is the entire $\mathbb{R}^N(N=1, 2 \text{ or } 3)$ and that time is divided into discrete nonoverlapping generations. The life cycle of a new (say nth) generation begins when the parent generation randomly mate without regard to genotype, produce offspring and die. These offspring undergo various hazards for a period of time before they are mature and migrate. The ability of an individual to survive these hazards depends only on its genotypes. Let the fitnesses of the three genotypes AA, Aa and aa be in the constant ratios $1+s:1:1+\sigma$ and let $u_n(x)$ denote the gene fraction of the nth generation right after birth. Then, assuming the Hardy-Weinberg Law holds, the gene fraction of the population just before migration is given by $g(u_n(x))$ where

$$g(u)=\frac{su^2+u}{1+su^2+\sigma(1-u)^2}. \tag{1.2}$$

We assume the total number of individuals that survive to migrate is of a constant carrying capacity and that migration occurs randomly, independently of time or genotype. Since the habitat is homogeneous, the fraction of the population that migrates from the point y to the point x depends only on $x-y$ and is given by $K(x-y)\,dy$, where $K(x)\geqq 0$. Since every individual must go somewhere, $\int K(x)\,dx=1$, i.e., K is a probability density.

After the migration, the species mate randomly, produce offspring and die, thus completing one life-cycle. Under these assumptions, the gene fraction of the $(n+1)$st generation immediately after birth is equal to the gene fraction of the nth generation after migration. We have therefore arrived at the following formula,

$$u_{n+1}(x)=\int K(x-y)g(u_n(y))\,dy. \tag{1.3}$$

Equation (1.3) is an example of a recursion of the kind

$$u_{n+1}=Q[u_n], \tag{1.4}$$

where

$$Q[u](x)=\int K(x-y)g(u(y))\,dy. \tag{1.5}$$

It is clear from (1.2) that $g(u)$ increases from 0 to 1 on the interval $[0,1]$. Since K is a probability density, we see from (1.5) that if $0\leqq u_0\leqq 1$, then $0\leqq u_n\leqq 1$ for all n.

Without loss of generality, we may always assume that A is the advantageous gene so that $\sigma\leqq s$. From (1.2), there are then three cases to consider.

(i) $\sigma<0<s$. This is the heterozygote intermediate case meaning that AA is the most fit to survive and aa is the least fit to survive. We note that $g(u)>u$ if $0<u<1$.

(ii) $\sigma\leqq s<0$. This is the heterozygote superior case with Aa most fit to survive. We note that $g(u)>u$ if $0<u<\sigma/(s+\sigma)$ and $g(u)<u$ if $\sigma/(s+\sigma)<u<1$.

(iii) $0<\sigma\leqq s$. This is the heterozygote inferior case with Aa least fit to survive. We note that $g(u)<u$ if $0<u<\sigma/(s+\sigma)$ and $g(u)>u$ if $\sigma/(s+\sigma)<u<1$.

The model described above is of course very simple, and situations when one or more of the assumptions are not satisfied are of much biological interest. Some of these situations are discussed in [27] and [37].

We now turn to the question of examining how the advantageous gene A advances through the population after many generations. Mathematically, this is equivalent to determining the limiting behavior of the function $u_n(x)$ as $n\to\infty$. The most important concept involved here is the wave speed, $c^*(\xi)$, defined for every unit vector $\xi\in R^N$. It is, in an asymptotic sense, the speed with which initial disturbances are propagated in the direction of the vector ξ. The precise definition of $c^*(\xi)$ is given in [38, §5].

Consider for the moment, the one-dimensional heterozygote intermediate case. Let u_0 have compact support in $\mathbb{R}$ and define u_n recursively by (1.4). Then under appropriate conditions on K and g, we have

$$\lim_{n\to\infty}\ \max_{x\notin[nc_1,nc_2]} u_n(x)=0 \quad \text{for every } c_1<-c^*(-1)<c^*(1)<c_2$$

and

$$\lim_{n\to\infty}\ \min_{x\in[nc_1',nc_2']} u_n(x)=1 \quad \text{for every} -c^*(-1)<c_1'<c_2'<c^*(1)$$

(see [38]). This indicates that $c^*(1)$ and $c^*(-1)$ are the asymptotic speeds of propagation for initial data with compact support in the positive and negative directions respectively. However, on the intervals $[nc_1,nc_1']$ and $[nc_2',nc_2]$, we have no information about the function u_n. It is proved in [24] (with more assumptions on K and g), that $u_n(x)$ actually develops uniformly in x, as $n\to\infty$, into a pair of diverging waves, with speed $c^*(1)$ and $c^*(-1)$, facing opposite directions.

A nonconstant solution of the recursion (1.4) which is of the form $u_n(x)=w(x\cdot\xi-nc)$, where ξ is a fixed unit vector in $\mathbb{R}^N$, is called a travelling wave of speed c. For the heterozygote intermediate case, nonincreasing travelling waves of speed c are known to exist if and only if $c\geqq c^*(\xi)$, [38]. For the heterozygote inferior case, monotone travelling waves exist if and only if $c=c^*(\pm1)$, [25]. We shall return to this in §2.

Consider now the behavior of u_n in the one-dimensional heterozygote inferior case. As in the intermediate case, if we assume that $-c^*(-1)<c^*(1)$ and that $u_0(x)>\sigma/(s+\sigma)$ in a sufficiently large interval, then

$$(1.6)\qquad \lim_{n\to\infty}\ \max_{x\notin[nc_1,nc_2]} u_n(x)=0 \quad \text{for any } c_1<-c^*(-1)<c^*(1)<c_2$$

and

$$(1.7)\qquad \lim_{n\to\infty}\ \min_{x\in[nc_1',nc_2']} u_n(x)=1 \quad \text{for any } -c^*(-1)<c_1'<c_2'<c^*(1).$$

The main purpose of this paper is to describe the behavior of u_n in the rest of $\mathbb{R}$. We are able to show that under very general conditions on K and g, if $u_0(x)>\sigma/(s+\sigma)$ on a sufficiently large interval and if $u_0(x)<\sigma/(s+\sigma)$ for x near $\pm\infty$, then u_n is trapped in between two translations of the function $w(x-nc_+^*)+\bar{w}(x-nc_-^*)-1$ as $n\to\infty$. Here $c_+^*=c^*(1)$, $c_-^*=-c^*(-1)$, w is a nonincreasing travelling wave of speed c_+^*, and $\bar{w}$ is a nondecreasing travelling wave of speed c_-^*. Furthermore, if K is the normal density and if $c_+^*>0$, then u_n converges exponentially to a pair of diverging waves $w(x-nc_+^*-x_0)+\bar{w}(x-nc_-^*-x_1)-1$ as $n\to\infty$.

The results we obtained here are quite similar to those obtained in [23], [24] except that there, K only needed to be in PF_3 (see §2) while here, K must be the normal density. On the other hand, results here are true for a wider class of u_0, and the rate of convergence is exponential.

The rest of this section contains the hypotheses on K and g. Section 2 contains the mathematical preliminaries. Section 3 contains the statement of the results while the proofs are presented in §§4, 5, and 6.

The following assumptions on K are identical to those listed in (1.3) of [25]. They are assumed to hold throughout the rest of this paper.

(i) $K(x)\geqq 0$. If $B_1=\inf\{x:K(x)>0\}$, $B_2=\sup\{x:K(x)>0\}$, then $K(x)>0$ in (B_1,B_2). We allow $B_1=-\infty$ or $B_2=\infty$ so that K need not have compact support.

(ii) $K(x)$ is continuous in $\mathbb{R}$ except possibly at B_1, B_2, where $\lim_{x\downarrow B_1}K(x)=p_1$, $\lim_{x\uparrow B_2}K(x)=p_2$. Also K may be written in the form

$$K(x)=K_a(x)-p_1\chi_{(-\infty,B_1]}-p_2\chi_{[B_2,\infty)},$$

(1.8) where K_a is absolutely continuous and χ_s is the indicator function of the set S.

(iii) $\int K(x)\,dx=1$,

(iv) $\int K(x)e^{\mu x}dx$ is finite for every real μ,

(v) $\int_x^\infty K(y)\,dy\leqq$ const. $K(x)$ for large x and $\int_{-\infty}^x K(y)\,dy$ $\leqq$ const. $K(x)$ for small x.

Remark 1.1. If K is the normal density, then all of the above assumptions hold. The constant in (v) may be replaced by $1/x$. See [5, Chap. 7]. Instead of assuming that g has the form (1.2) in the heterozygote inferior case, we assume the following about g throughout the entire paper:

(vi) $g\in C^2[0,1]$.

(vii) $g(0)=0, g(1)=1$.

(viii) There exists a constant $\alpha\in(0,1)$ such that $g(u)<u$ in $(0,\alpha)$ and $g(u)>u$ in $(\alpha,1)$.

(1.9) (ix) $g'(u)>0$ in $[0,1]$.

(x) $g'(0)<1, g'(1)<1$.

(xi) $g(u)\geqq g'(\alpha)(u-\alpha)+\alpha$ in $[0,\alpha]$ and $g(u)\leqq g'(\alpha)(u-\alpha)+\alpha$ in $[\alpha,1]$.

(xii) $g'(0)u\leqq g(u)\leqq g'(1)(u-1)+1$ in $[0,1]$.

Remark 1.2. Conditions (vi), (vii), (viii), (ix), and (x) are obviously satisfied when g has the form (1.2) with $\alpha=\sigma/(s+\sigma)$, $g'(0)=1/(1+\sigma)$ and $g'(1)=1/(1+s)$ (s, σ are positive). Also condition (xii) is valid. The right-hand inequality reduces to showing that the polynomial $\phi(u)=(s+\sigma)u^3+(\sigma s-2\sigma-s)u^2+(\sigma-2\sigma s-s)u+(s+\sigma s)$ is nonnegative in the interval $[0,1]$. This is accomplished by observing that $\phi(0)>0$ and $\phi(1)=0$ is a local minimum.

Remark 1.3. It is easy to prove that (xi) implies

$$\max_{[0,1]} \frac{g(u)}{u} < g'(\alpha) \quad \text{and} \quad \max_{[0,1]} \frac{1-g(1-u)}{u} < g'(\alpha),$$

which is sufficient for the results in [25] to hold. In [25, (1.4)] it was assumed that $g'(u) \leqq g'(\alpha)$, which holds if and only if $s=\sigma$.

2. Mathematical preliminaries. Let $0<\gamma<1$ and define $m_n^+(\gamma)=\sup\{x: u_n(x) \geqq \gamma\}$, $m_n^-(\gamma)=\inf\{x: u_n(x) \geqq \gamma\}$ whenever possible. The following theorem, which is a special case of [38, Thms. 6.1 and 6.2], implies that $m_n^+(\gamma)$ and $m_n^-(\gamma)$ will be defined on every compact subset of $(0,1)$ for sufficiently large n.

THEOREM 2.1 (asymptotic speed of propagation). *Let u_0 have compact support and $c_-^* < c_+^*$. Then for any $\eta>0$, there exists an $L>0$ such that if $u_0(x)>\alpha+\eta$ on an interval of length greater than L, then* (1.6) *and* (1.7) *hold. In particular, for any* $0<\gamma<1$,

$$\text{(2.1a)} \qquad \lim_{n\to\infty} \frac{m_n^+(\gamma)}{n} = c_+^*$$

and

$$\text{(2.1b)} \qquad \lim_{n\to\infty} \frac{m_n^-(\gamma)}{n} = c_-^*.$$

A similar statement holds if u_0 vanishes for sufficiently large x and $\liminf_{x\to-\infty} u_0(x)>\alpha$.

Remark 2.1. If u_0 satisfies the assumptions in Theorem 2.1, then u_n propagates as $n\to\infty$. In the future, we shall say u_0 is superthreshold.

THEOREM 2.2 (existence of travelling waves). *There exists a nonincreasing function $w(x)$, $w(-\infty)=1$, $w(\infty)=0$ and a nondecreasing function $\bar{w}(x)$, $\bar{w}(\infty)=1$, $\bar{w}(-\infty)=0$, such that if $u_n(x)=w(x-nc_+^*)$ or $u_n(x)=\bar{w}(x-nc_-^*)$, then u_n satisfies the recursion* (1.4).

Proof. [25, Thm. 5].

Remark 2.2. w, $\bar{w}$ are of course determined only up to translations. From now on, w, $\bar{w}$ will denote travelling waves with speeds c_+^* and c_-^* respectively.

THEOREM 2.3. *Let u_0 satisfy the conditions* $\liminf_{x\to-\infty} u_0(x)>\alpha$ *and* $\limsup_{x\to\infty} u_0(x)<\alpha$. *Then given a travelling wave w, there exist constants x_1, x_2, q_0 and μ, the last two positive, such that*

$$\text{(2.2)} \qquad w(x-x_1)-q_0e^{-\mu n} \leqq u_n(x+nc_+^*) \leqq w(x-x_2)+q_0e^{-\mu n} \quad \textit{for all } n.$$

A similar statement holds if $\liminf_{x\to\infty} u_0(x)>\alpha$ *and* $\limsup_{x\to-\infty} u_0(x)<\alpha$ *when w and c_+^* are replaced by $\bar{w}$ and c_-^*, respectively.*

Proof. [25, Thm. 4].

LEMMA 2.4. *Suppose u_0 satisfies the hypotheses of Theorem* 2.3, *then given $\varepsilon>0$, there exists a $\delta>0$ such that if $|u_0(x)-w(x)|<\delta$ for some travelling wave w, then $|u_n(x+nc_+^*)-w(x)|<\varepsilon$ for all n.*

Proof. Theorem 2.3 is proved by first choosing z_0, q_0 such that $w(x-z_0)-q_0 \leqq u_0(x)$ and $\alpha<1-q_0<\liminf_{x\to-\infty} u_0(x)$. Then by defining $z_{n+1}=-kq_0e^{-\mu n}+z_n$ recursively, where k, $\mu>0$ depend only on w and g, we can show that $w(x-z_n)-q_0e^{-\mu n} \leqq u_n(x+nc_+^*)$ for all n. Since $z_n=z_0-kq_0((1-(e^{-\mu})^n)/(1-e^{-\mu}))$ decreases to the limit $x_1=z_0-k'q_0$ as $n\to\infty$ and since w is nonincreasing, we may replace z_n in the above inequality by x_1 to obtain the left side of (2.2). Having recalled all these, we note that our hypotheses imply that $w(x)-\delta<u_0(x)$ for all x. Letting $z_0=0$ and $q_0=\delta$ in

the above argument, we have from (2.2)

$$w(x)-u_n(x+nc_+^*)\leqq w(x)-w(x-x_1)+q_0e^{-\mu n}\leqq (Mk'+e^{-\mu n})\delta,$$

where according to (ii) of (1.8), we may take $M=\|K'\|_1+p_1+p_2\geqq\|w'\|_\infty$. A similar argument will prove the opposite inequality.

LEMMA 2.5 (uniqueness of travelling waves). *Let w_1 satisfy the conditions $\liminf_{x\to-\infty} w_1(x)>\alpha$, $\limsup_{x\to\infty} w_1(x)<\alpha$ and let $u_n(x)=w_1(x-nc)$ be a solution to the recursion* (1.4). *Then $c=c_+^*$. Furthermore, if K is the normal density, then $w_1(x)=w(x-\tau)$ for some constant τ.*

Proof. Letting $u_0=w_1$ in Theorem 2.3, we have $w(x-x_1)-q_0e^{-\mu n}\leqq w_1(x+n(c_+^*-c))\leqq w(x-x_2)+q_0e^{-\mu n}$ for all n. If $c\neq c_+^*$, then by letting $n\to\infty$ we arrive at a contradiction. Thus $c=c_+^*$ and $w(x-x_1)\leqq w_1(x)\leqq w(x-x_2)$. Let $x^*=\inf\{x_2: w_1(x)\leqq w(x-x_2)$ on $\mathbb{R}\}$ and define $w_2(x)=w(x-x^*)$. If $w_1(x_0)=w_2(x_0)$, then

$$0=w_1(x_0)-w_2(x_0)=\int K(x_0+c_+^*-y)[g(w_1(y))-g(w_2(y))]\,dy.$$

Since g is increasing and $w_1(x)\leqq w_2(x)$, the above inequality implies that $w_1(x)=w_2(x)$ on an interval containing x_0 (note that $B_1<c_+^*<B_2$). Thus the set where $w_1(x)=w_2(x)$ is open and is obviously closed. To show that it is nonempty, we need the techniques in §6. The case when K is the normal density is shown in the appendix. The theorem is true under (1.8) but the proof will be published elsewhere.

LEMMA 2.6. *Let $0<\beta<1$ and $\Phi(\mu,\beta)=(1/\mu)\ln\{\beta\int K(x)e^{\mu x}dx\}$. Let μ^* be the unique positive root of $\Phi(\mu,g'(0))=c_+^*$ and $-\bar{\mu}^*$ be the unique negative root of $\Phi(\mu,g'(1))=c_+^*$. Then*

$$w(x)\sim e^{-\mu^*x}\quad as\ x\to\infty, \tag{2.3a}$$

$$1-w(x)\sim e^{\bar{\mu}^*x}\quad as\ x\to-\infty. \tag{2.3b}$$

Similarly, let $-\bar{\mu}_*$ be the unique negative root of $\Phi(\mu,g'(0))=c_-^*$ and μ_* the unique positive root of $\Phi(\mu,g'(1))=c_-^*$, then

$$\bar{w}(x)\sim e^{\bar{\mu}_*x}\quad as\ x\to-\infty, \tag{2.4a}$$

$$1-\bar{w}(x)\sim e^{-\mu_*x}\quad as\ x\to\infty. \tag{2.4b}$$

Proof. [25, Prop. 5].

In this paper, $f(x)\sim g(x)$ as $x\to\pm\infty$ means that $f(x)/g(x)$ converges to a positive constant as $x\to\pm\infty$.

LEMMA 2.7. *Suppose that K_a' changes signs a finite number of times; then $w'(x)\sim-\mu^*e^{-\mu^*x}$ as $x\to\infty$ and $w'(x)\sim-\bar{\mu}^*e^{\bar{\mu}^*x}$ as $x\to-\infty$. Similar results hold for $\bar{w}$.*

Proof. The proof of [23, Lemma 7] is valid here to show that $w'(x)\sim-\mu^*w(x)$ as $x\to\infty$. We simply replace w_c, μ_c, β in the proof by w, μ^* and $g'(0)$ respectively and then make use of (2.3a) and $\Phi(\mu^*,g'(0))=c_+^*$. This result and (2.3a) obviously imply our lemma.

Remark 2.3. Using the facts that $w'(x)=\int K(x+c_+^*-y)g'(w(y))w'(y)dy$ and $g'>0$, we see that $w'(x)<0$ in $\mathbb{R}$. In particular, on every compact subset of $\mathbb{R}$, $w'<-\varepsilon<0$ for some $\varepsilon>0$.

Remark 2.4. If $K(x)=K(-x)$, then $\bar{w}(-x)$ is a nonincreasing travelling wave with speed $-c_-^*$. From Lemma 2.5., $c_+^*=-c_-^*$. Thus $c_+^*>0$ if and only if $c_+^*>c_-^*$. It is conjectured that if K is even, then $c_+^*>0$ if and only if $\int_0^1 g(x)dx>\frac{1}{2}$.

Remark 2.5. Let $K(x)=(2\pi)^{-1/2}\exp\{-x^2/2\}$ so that $\int K(x)e^{\mu x}dx=\exp\{\mu^2/2\}$. From Lemma 2.6 and the facts that $g'(0)<1$, $g'(1)<1$, we have

$$2c_+^*-\mu^*<0, \qquad -2c_+^*-\bar{\mu}^*<0, \tag{2.5a}$$

$$-2c_-^*-\bar{\mu}_*<0, \qquad 2c_-^*-\mu_*<0. \tag{2.5b}$$

LEMMA 2.8 (comparison principle). *Let $0\leqq v_n\leqq 1$ and $0\leqq w_n\leqq 1$ be two sequences of functions such that $v_{n+1}\geqq Q[v_n]$ and $w_{n+1}\leqq Q[w_n]$ for all n. Suppose further that $v_0\geqq w_0$, then $v_n\geqq w_n$ for all n.*

Proof. This follows from an inductive argument.

We close this section by introducing the class of PF_r functions. They will be needed later on for a result like Lemma 2.4 but with u_0 having compact support.

A PF_1 function is just a nonnegative function in $\mathbb{R}$. A function $f\in PF_r$ if $f\in PF_{r-1}$ and if for every ϕ with no more than $r-1$ number of sign changes in $\mathbb{R}$, $f*\phi$ also has no more than $r-1$ number of sign changes in $\mathbb{R}$. If $f\in PF_r$ for every $r>0$, then $f\in PF_\infty$. The normal density belongs to PF_∞.

We remark that the hypothesis in Lemma 2.7 is satisfied if $K\in PF_2$. Furthermore, if $K\in PF_3$, then $p_1=p_2=0$ in (ii) of (1.8) so that K is absolutely continuous.

LEMMA 2.9. *Let $K\in PF_3$ and $c_+^*>0$. Suppose u_0 has compact support and is superthreshold. Suppose also that for some $\varepsilon>0$, the line $u=l$ crosses u_0 exactly twice for every $\alpha-\varepsilon\leqq l\leqq\alpha+\varepsilon$. Then the line $u=l$ crosses u_n exactly twice for every $g^n(\alpha-\varepsilon)<l<g^n(\alpha+\varepsilon)$, $n\geqq 0$.*

Proof. The argument is the same as [24, Lemma 2]. Note that $g^n(\alpha-\varepsilon)\downarrow 0$ and $g^n(\alpha+\varepsilon)\uparrow 1$ as $n\to\infty$.

LEMMA 2.10. *Suppose $K\in PF_2$, $u_0(x)=0$ for $x\geqq A$. Then there exist w and L such that $u_n(x+nc_+^*)\leqq w(x-L)$ for $x\geqq L$.*

Proof. Let $f_0(x)=1$ for $x\leqq 0$ and $f_0(x)=0$ for $x>0$. Define f_n recursively by $f_{n+1}=Q[f_n]$. Then for any $0<\gamma<1$, $f_n(x+f_n^{-1}(\gamma))$ increases uniformly to the travelling wave $w(x+w^{-1}(\gamma))$ for $x\geqq 0$ and decreases uniformly to $w(x+w^{-1}(\gamma))$ for $x\leqq 0$. This assertion is proved in [23, Lemma 13] assuming that K has compact support and g is in the heterozygote intermediate case. But the proof only makes use of the fact that $K\in PF_2$ and g is nondecreasing; hence it is also valid here.

By our hypothesis, we may assume without loss of generality that $A=0$. Thus $u_0\leqq f_0$ and from Lemma 2.8, $u_n(x+nc_+^*)\leqq f_n(x+nc_+^*)$. From (2.2) with $u_n=f_n$, and $x=f_n^{-1}(\gamma)-nc_+^*$, we see that $|f_n^{-1}(\gamma)-nc_+^*|\leqq L$ for all n. Thus $f_n(x+nc_+^*)\leqq f_n(x+f_n^{-1}(\gamma)-L)\leqq w(x+w^{-1}(\gamma)-L)$ whenever $x\geqq L$. We may choose γ small so that $w^{-1}(\gamma)\geqq 0$; this and the previous inequality imply our lemma.

3. Statement of the results.

For the convenience of the reader, we summarize our theorems in this section. Their proofs will be presented in the subsequent sections.

THEOREM 3.1. *Let $c_-^*<c_+^*$, and let u_0 be superthreshold and satisfy the condition $\limsup_{|x|\to\infty}u_0(x)<\alpha$. Then given w and $\bar{w}$, there exist constants x_1, $\bar{x}_1$, x_2, $\bar{x}_2$, q_0 and μ, the last two positive, such that*

$$w(x-nc_+^*-x_1)+\bar{w}(x-nc_-^*-\bar{x}_1)-1-q_0e^{-\mu n}$$
$$\leqq u_n(x)\leqq w(x-nc_+^*-x_2)+\bar{w}(x-nc_-^*-\bar{x}_2)-1+q_0e^{-\mu n}$$

for all n.

THEOREM 3.2. *Let $K_1(x)=(2\pi)^{-1/2}\exp\{-x^2/2\}$ and $K(x)=\frac{1}{\sigma}K_1((x-\tau)/\sigma)$ for some τ and $\sigma>0$. Suppose $u_0(x)=0$ for large x, $\liminf_{x\to-\infty}u_0(x)>\alpha$ if $c_+^*\geqq\tau$ and $u_0(x)=1$ for small x, $\limsup_{x\to\infty}u_0(x)<\alpha$ if $c_+^*\leqq\tau$, then there exists a travelling wave w*

and constants C, $\varepsilon > 0$ such that $|u(x+nc_+^*)-w(x)| \leqq Ce^{-\varepsilon n}$ for all n. Similar results hold for the negative direction with w and c_+^* replaced by $\bar{w}$ and c_-^* respectively.

THEOREM 3.3. *Let K satisfy the same hypothesis as in Theorem 3.2. Suppose $c_-^* < \tau < c_+^*$, and that u_0 has compact support and is superthreshold; then there exist travelling waves w and $\bar{w}$ and constants C, $\varepsilon > 0$ such that $|u_n(x)-w(x-nc_+^*)-\bar{w}(x-nc_-^*)+1| \leqq Ce^{-\varepsilon n}$ for all n.*

Remark 3.1. Theorem 3.3 is probably still true if we replace the condition that u_0 have compact support by the condition $\limsup_{|x|\to\infty} u_0(x) < \alpha$. Theorem 3.2 is probably true if u_0 satisfies the conditions $\limsup_{x\to\infty} u_0(x) < \alpha$ and $\liminf_{x\to-\infty} u_0(x) > \alpha$ regardless of the sign of $c_+^* - \tau$.

Remark 3.2. Theorem 3.1 is like Theorem 2.3 and so is its proof. Observe that if $c_-^* < 0 < c_+^*$ and $x \geqq 0$, then for large n, $\bar{w}(x-nc_-^*-\bar{x}_1)$ and $\bar{w}(x-nc_-^*-\bar{x}_2$ are close to 1 so that the inequalities in Theorem 3.1 look exactly like (2.2). As mentioned in Theorem 2.3, (2.2) has a counterpart involving $\bar{w}$, c_-^* and u_n facing left which looks like the inequalities in Theorem 3.1 for large n and $x \leqq 0$.

Remark 3.3. Theorems 3.1 and 2.3 are stronger than Theorem 2.1. For example, the former two theorems would imply that for any $\gamma_1, \gamma_2 \in (0,1)$, $m_n^+(\gamma_1)-m_n^+(\gamma_2)$ is bounded as $n \to \infty$. However, (2.1a) could hold even if $m_n^+(\gamma_1)-m_n^+(\gamma_2) = O(\log n)$.

4. Proof of Theorem 3.1.

LEMMA 4.1. *Let the hypotheses of Theorem 3.1 be satisfied. Then there exist two sequences of real numbers $\{a_n\}$, $\{b_n\}$, both decreasing and bounded, and positive constants q_0, μ, such that if $v_n(x) = w(x-nc_+^*-a_n)+\bar{w}(x-nc_-^*+b_n)-1-q_0e^{-\mu n}$, then $v_{n+1} \leqq Q[v_n]$ for all n.*

Proof. The proof is rather long and will be given in several steps. We first assume that $c_-^* < 0 < c_+^*$.

Step 1. Choose q_0' such that $0 < q_0' < 1-\alpha$; then there exist constants $\delta > 0$, $\theta_1 \in (0,1)$ such that

$$g(w-q)-g(w) \geqq -\theta_1 q \tag{4.1}$$

for all $0 \leqq w \leqq \delta$ or $1-\delta \leqq w \leqq 1$ and $0 \leqq q \leqq q_0'$. To see this, let

$$\phi(w,q) = \begin{cases} \dfrac{g(w)-g(w-q)}{q} & \text{if } q \neq 0, \\ g'(w) & \text{if } q = 0 \end{cases}$$

on the interval $0 \leqq q \leqq q_0'$, $\alpha \leqq w \leqq 1$. We shall define $g(u) = 0$ if $u < 0$. From the assumptions on g, we see that $\phi(w,q)$ is uniformly continuous, and there exists θ_1 such that $\phi(w,q) < \theta_1 < 1$ for $0 \leqq q \leqq q_0'$ and $w = 1$. Therefore the same inequality also holds if w is in a left neighborhood of 1. This implies (4.1) when $1-\delta \leqq w \leqq 1$. If w is near 0 and $w-q \geqq 0$, then (4.1) follows from the mean value theorem provided; if necessary, we increase θ_1 so that $g'(0) < \theta_1 < 1$. If $w-q < 0$, then since $\lim_{w\downarrow 0}(g(w)/w) = g'(0) < \theta_1$, we have $g(w)/q \leqq g(w)/w \leqq \theta_1$ for sufficiently small w. This completes the proof of (4.1).

We remark here that once (4.1) is established it continues to hold if we decrease $q_0' > 0$ with no change in θ_1 and δ.

Step 2. After having chosen θ_1, δ and q_0', we define all other constants necessary for the rest of the proof in this step. The order of their dependence is important.

Let $\max_{[0,1]} g'(u) = M > 1$. According to (2.3) and (2.4), there exist A, $\nu > 0$ such that

$$1 - w(x) \leqq Ae^{\nu x} \quad \text{and} \quad 1 - \overline{w}(x) \leqq Ae^{-\nu x} \quad \text{for all } x. \tag{4.2}$$

Let $c^* = \max(c_-^*, -c_+^*)$ and $\mu > 0$ be sufficiently small that

$$\nu c^* + \mu < 0 \tag{4.3}$$

and

$$\theta_1 < e^{-\mu}. \tag{4.4}$$

From (4.4) we choose $\varepsilon > 0$ such that

$$\varepsilon M + \theta_1 < e^{-\mu}. \tag{4.5}$$

Choose l sufficiently large that

$$\int_{-\infty}^{-l} K(x)\,dx \leqq \frac{\varepsilon}{2} \quad \text{and} \quad \int_{l}^{\infty} K(x)\,dx \leqq \frac{\varepsilon}{2}. \tag{4.6}$$

For any $0 < \gamma < 1$, we introduce the notation $E_\gamma = w^{-1}(\gamma)$ and $\overline{E}_\gamma = \overline{w}^{-1}(\gamma)$. Then since $w' < 0$ in $\mathbb{R}$, there exist θ_2, $\theta_3 > 0$ such that

$$\begin{aligned} w(x) - w(y) &\leqq -\theta_2(x-y) \quad \text{if } E_{1-\delta} - c_+^* - 2l \leqq y < x \leqq E_\delta - c_+^* + 2l, \\ \overline{w}(y) - \overline{w}(x) &\leqq -\theta_3(x-y) \quad \text{if } \overline{E}_\delta - c_-^* - 2l \leqq y < x \leqq \overline{E}_{1-\delta} - c_-^* + 2l. \end{aligned} \tag{4.7}$$

Choose $q_0 < q_0'$ such that, for sufficiently large m,

$$-(\varepsilon M + \theta_1 - M)(Ae^{-\nu m} + q_0) \leqq l \min(\theta_2, \theta_3). \tag{4.8}$$

Fix m large so that (4.8) holds together with the following:

$$(\varepsilon M + \theta_1)\left[\frac{Ae^{-\nu m} + q_0}{q_0}\right] < e^{-\mu} \tag{4.9}$$

and

$$Ae^{-\nu m} + q_0 < q_0'. \tag{4.10}$$

Finally we define the sequences $\{a_n\}$, $\{b_n\}$ recursively by

$$a_{n+1} = a_n + \theta_2^{-1}(\varepsilon M + \theta_1 - M)(Ae^{-\nu m} + q_0)e^{-\mu n}, \tag{4.11}$$

$$b_{n+1} = b_n + \theta_3^{-1}(\varepsilon M + \theta_1 - M)(Ae^{-\nu m} + q_0)e^{-\mu n}. \tag{4.12}$$

Then

$$a_n = a_0 + \sum_{k=0}^{n-1} (a_{k+1} - a_k) = a_0 + C\sum_{k=0}^{n-1} (e^{-\mu})^k = a_0 + C\left[\frac{1 - (e^{-\mu})^n}{1 - e^{-\mu}}\right], \tag{4.13}$$

where $C = \theta_2^{-1}(\varepsilon M + \theta_1 - M)(Ae^{-\nu m} + q_0)$ is negative. Similarly

$$b_n = b_0 + C'\left[\frac{1 - (e^{-\mu})^n}{1 - e^{-\mu}}\right], \tag{4.14}$$

where $C'=\theta_3^{-1}(\varepsilon M+\theta_1-M)(Ae^{-\nu m}+q_0)$ is also negative. It is clear from (4.11) and (4.12) that a_n, b_n are decreasing. If we choose a_0, b_0 sufficiently large such that

$$a_0 \geqq \frac{-C}{1-e^{-\mu}}+m \quad \text{and} \quad b_0 \geqq \frac{-C'}{1-e^{-\mu}}+m,$$

then from (4.13) and (4.14), we have

$$(4.15) \qquad a_n \geqq m, \qquad b_n \geqq m \quad \text{for every } n.$$

Step 3. For every n, define the intervals

$$\begin{aligned}
\Gamma_n &= \left[E_{1-\delta}+nc_+^*+a_n,\ E_\delta+nc_+^*+a_n\right],\\
\Gamma_n' &= \left[E_{1-\delta}+nc_+^*+a_n-l,\ E_\delta+nc_+^*+a_n+l\right],\\
\bar\Gamma_n &= \left[\bar E_\delta+nc_-^*-b_n,\ \bar E_{1-\delta}+nc_-^*-b_n\right],\\
\bar\Gamma_n' &= \left[\bar E_\delta+nc_-^*-b_n-l,\ \bar E_{1-\delta}+nc_-^*-b_n+l\right].
\end{aligned}$$

The relation $v_{n+1}\leqq Q[v_n]$ is equivalent to showing that the following expression is bounded above by $q_{n+1}=q_0e^{-\mu(n+1)}$:

(4.16)

$$\begin{aligned}
&w\left(x-(n+1)c_+^*-a_{n+1}\right)-w\left(x-(n+1)c_+^*-a_n\right)\\
&\quad+\bar w\left(x-(n+1)c_-^*+b_{n+1}\right)-\bar w\left(x-(n+1)c_-^*+b_n\right)\\
&\quad-\int_{\Gamma_n\cap\mathbb{R}^+}K(x-y)\left\{g\left[w\left(y-nc_+^*-a_n\right)-\left(1-\bar w\left(y-nc_-^*+b_n\right)+q_0e^{-\mu n}\right)\right]\right.\\
&\qquad\qquad\left.-g\left(w\left(y-nc_+^*-a_n\right)\right)\right\}dy\\
&\quad-\int_{\Gamma_n^c\cap\mathbb{R}^+}K(x-y)\left\{g\left[w\left(y-nc_+^*-a_n\right)-\left(1-\bar w\left(y-nc_-^*+b_n\right)+q_0e^{-\mu n}\right)\right]\right.\\
&\qquad\qquad\left.-g\left(w\left(y-nc_+^*-a_n\right)\right)\right\}dy\\
&\quad-\int_{\bar\Gamma_n\cap\mathbb{R}^-}K(x-y)\left\{g\left[\bar w\left(y-nc_-^*+b_n\right)-\left(1-w\left(y-nc_+^*-a_n\right)+q_0e^{-\mu n}\right)\right]\right.\\
&\qquad\qquad\left.-g\left(\bar w\left(y-nc_-^*+b_n\right)\right)\right\}dy\\
&\quad-\int_{\bar\Gamma_n^c\cap\mathbb{R}^-}K(x-y)\left\{g\left[\bar w\left(y-nc_-^*+b_n\right)-\left(1-w\left(y-nc_+^*-a_n\right)+q_0e^{-\mu n}\right)\right]\right.\\
&\qquad\qquad\left.-g\left(\bar w\left(y-nc_-^*+b_n\right)\right)\right\}dy\\
&\quad+\int_{\mathbb{R}^-}K(x-y)\left[g\left(w\left(y-nc_+^*-a_n\right)\right)-1\right]dy\\
&\quad+\int_{\mathbb{R}^+}K(x-y)\left[g\left(\bar w\left(y-nc_-^*+b_n\right)\right)-1\right]dy.
\end{aligned}$$

We shall call the first four integrals I_1, I_2, I_3 and I_4 respectively. Since the last two integrals are negative, it suffices to show that all but the last two integrals are bounded above by q_{n+1}. We do this in two steps.

Step 4. Here $x \notin \Gamma_n' \cup \bar{\Gamma}_n'$. We have according to (4.2), (4.6), (4.3) and (4.15) that

$$\begin{aligned} I_1 &= \int_{\Gamma_n \cap \mathbb{R}^+} K(x-y) g'(\xi)\left[1-\bar{w}\left(y-nc_-^*+b_n\right)+q_0 e^{-\mu n}\right] dy \\ &\leqq M \int_{\Gamma_n \cap \mathbb{R}^+} K(x-y)\left[A e^{-\nu(y-nc_-^*+b_n)}+q_0 e^{-\mu n}\right] dy \\ &\leqq M\left[A e^{\nu c_-^* n-\nu b_n}+q_0 e^{-\mu n}\right] \int_{\Gamma_n \cap \mathbb{R}^+} K(x-y)\, dy \\ &\leqq \frac{\varepsilon}{2} M\left[A e^{(\nu c_-^*+\mu) n-\nu b_n}+q_0\right] e^{-\mu n} \leqq \frac{\varepsilon}{2} M\left[A e^{-\nu m}+q_0\right] e^{-\mu n}. \end{aligned}$$

For I_2, if $y \in \Gamma_n^c$, then $y \leqq E_{1-\delta}+nc_+^*+a_n$ or $y \geqq E_\delta+nc_+^*+a_n$, in both cases $w(y-nc_+^*-a_n) \in [0, \delta] \cup [1-\delta, 1]$. Also, since $y \geqq 0$, we have from (4.2), (4.3), (4.15) and (4.10),

$$\begin{aligned} &1-\bar{w}\left(y-nc_-^*+b_n\right)+q_0 e^{-\mu n} \\ &\quad \leqq A e^{-\nu(y-nc_-^*+b_n)}+q_0 e^{-\mu n} \leqq \left[A e^{(\nu c_-^*+\mu) n-\nu b_n}+q_0\right] e^{-\mu n} \leqq \left[A e^{-\nu m}+q_0\right] e^{-\mu n} \leqq q_0'. \end{aligned}$$

Therefore according to step 1,

$$\begin{aligned} I_2 &\leqq \theta_1 \int_{\Gamma_n^c \cap \mathbb{R}^+} K(x-y)\left[1-\bar{w}\left(y-nc_-^*+b_n\right)+q_0 e^{-\mu n}\right] dy \\ &\leqq \theta_1\left[A e^{-\nu m}+q_0\right] e^{-\mu n} \int_{\mathbb{R}^+} K(x-y)\, dy. \end{aligned}$$

Similarly,

$$\begin{aligned} I_3 &= \int_{\bar{\Gamma}_n \cap \mathbb{R}^-} K(x-y) g'(\xi)\left[1-w\left(y-nc_+^*-a_n\right)+q_0 e^{-\mu n}\right] dy \\ &\leqq M \int_{\bar{\Gamma}_n \cap \mathbb{R}^-} K(x-y)\left[A e^{\nu(y-nc_+^*-a_n)}+q_0 e^{-\mu n}\right] dy \\ &\leqq \frac{\varepsilon}{2} M\left[A e^{-\nu m}+q_0\right] e^{-\mu n}. \end{aligned}$$

For I_4, if $y \in \bar{\Gamma}_n^c \cap \mathbb{R}^-$, then $\bar{w}(y-nc_-^*+b_n) \in [0, \delta] \cup [1-\delta, 1]$ and

$$\begin{aligned} 1-w\left(y-nc_+^*-a_n\right)+q_0 e^{-\mu n} &\leqq \left[A e^{(-\nu c_+^*+\mu) n-\nu a_n}+q_0\right] e^{-\mu n} \\ &\leqq \left[A e^{-\nu m}+q_0\right] e^{-\mu n} \leqq q_0'. \end{aligned}$$

Thus

$$I_4 \leqq \theta_1\left[A e^{-\nu m}+q_0\right] e^{-\mu n} \int_{\mathbb{R}^-} K(x-y)\, dy.$$

Since a_n, b_n are decreasing, the sum of the first four terms in (4.16) are nonpositive so that altogether, if $x \notin \Gamma_n' \cup \bar{\Gamma}_n'$ (4.16) is bounded above by $(\varepsilon M+\theta_1)(A e^{-\nu m}+q_0) e^{-\mu n}$ which according to (4.9) is less than q_{n+1}.

Step 5. To treat the case when $x \in \Gamma_{\mathrm{n}}'$ or $x \in \bar{\Gamma}_{\mathrm{n}}'$, we combine I_1, I_2 and I_3, I_4. As in step 4,

$$I_1 + I_2 = \int_{\mathbb{R}^+} K(x-y) g'(\xi)\left[1 - \bar{w}\left(y - nc_-^* + b_n\right) + q_0 e^{-\mu n}\right] dy$$

$$\leqq M\left[Ae^{-\nu m} + q_0\right] e^{-\mu n} \int_{\mathbb{R}^+} K(x-y)\, dy,$$

and

$$I_3 + I_4 = \int_{\mathbb{R}^-} K(x-y) g'(\xi)\left[1 - w\left(y - nc_+^* - a_n\right) + q_0 e^{-\mu n}\right] dy$$

$$\leqq M\left[Ae^{-\nu m} + q_0\right] e^{-\mu n} \int_{\mathbb{R}^-} K(x-y)\, dy.$$

Hence $I_1 + I_2 + I_3 + I_4 \leqq M[Ae^{-\nu m} + q_0]e^{-\mu n}$.

Now if $x \in \Gamma_n'$, then

$$E_{1-\delta} - c_+^* - l \leqq x - (n+1)c_+^* - a_n \leqq E_\delta - c_+^* + l.$$

Since a_n is decreasing, we have from (4.8) and (4.11) that $x - (n+1)c_+^* - a_{n+1} \leqq E_\delta - c_+^* + 2l$. Hence according to (4.7) $w(x - (n+1)c_+^* - a_{n+1}) - w(x - (n+1)c_+^* - a_n) \leqq -\theta_2(a_n - a_{n+1})$. Therefore, if $x \in \Gamma_n'$, (4.16) is bounded above, according to (4.11) and (4.9), by

$$-\theta_2\left(a_n - a_{n+1}\right) + M\left[Ae^{-\nu m} + q_0\right] e^{-\mu n} = \left(\varepsilon M + \theta_1\right)\left[Ae^{-\nu m} + q_0\right] e^{-\mu n} \leqq q_0 e^{-\mu(n+1)}.$$

Similarly, if $x \in \bar{\Gamma}_n'$, then

$$\bar{E}_\delta - c_-^* - l \leqq x - (n+1)c_-^* + b_n \leqq \bar{E}_{1-\delta} - c_-^* + l.$$

Since b_n is decreasing we have, from (4.8) and (4.12), $x - (n+1)c_-^* + b_{n+1} \geqq \bar{E}_\delta - c_-^* - 2l$. Hence

$$\bar{w}\left(x - (n+1)c_-^* + b_{n+1}\right) - \bar{w}\left(x - (n+1)c_-^* + b_n\right) \leqq -\theta_3\left(b_n - b_{n+1}\right).$$

Therefore, if $x \in \bar{\Gamma}_n'$, (4.16) is bounded above, according to (4.12) and (4.9), by

$$-\theta_3\left(b_n - b_{n+1}\right) + M\left[Ae^{-\nu m} + q_0\right] e^{-\mu n} = \left(\varepsilon M + \theta_1\right)\left[Ae^{-\nu m} + q_0\right] e^{-\mu n} \leqq q_0 e^{-\mu(n+1)}.$$

This completes the proof of the lemma in the case $c_-^* < 0 < c_+^*$. In general, since $c_-^* < c_+^*$, there exists τ such that $\tilde{c}_-^* < 0 < \tilde{c}_+^*$ where $\tilde{c}_\pm^* = c_\pm^* + \tau$. Let $K_1(X) = K(x - \tau)$ and $Q_1[u] = K_1 * g(u)$, then w, $\bar{w}$ are travelling waves for the operator Q_1 with speeds $\tilde{c}_+^*$ and $\tilde{c}_-^*$ respectively. What we have proved for this case plus a simple change of variables will prove the lemma in the general case.

To prove Theorem 3.1, we let v_n be as defined in Lemma 4.1. Recall q_0, q_0' are only required to satisfy the conditions $0 < q_0 < q_0'$, q_0 small and $0 < q_0' < 1 - \alpha$ in Lemma 4.1. It is therefore possible to choose them so that

$$\alpha < 1 - q_0' < 1 - q_0 < \alpha + \eta \tag{4.17}$$

for some $\eta > 0$.

According to our hypotheses, u_n propagates and hence there exists k such that $u_k \geqq \alpha + \eta$ on the interval $[\bar{w}^{-1}(q_0) - b_0, w^{-1}(q_0) + a_0]$. On this interval $v_0(x) = w(x - a_0) + \bar{w}(x + b_0) - 1 - q_0 \leqq u_k(x)$ by (4.17). On the complement of this interval,

$v_0 \leqq 0$, so that $v_0(x) \leqq u_k(x)$ for all x. According to Lemma 4.1 and Lemma 2.8 $v_n \leqq u_{n+k}$ for all n. Let $n' = n+k$, then $w(x - n'c_+^* + kc_+^* - a_n) + \overline{w}(x - n'c_-^* + kc_-^* + b_n) - 1 - q_0 e^{\mu k} e^{-\mu n'} \leqq u_{n'}(x)$ for $n' \geqq k$. It is also clear from the monotonicity of w, $\overline{w}$, a_n and b_n that we may replace a_n, b_n by their limits in the above inequality. Doing so and renaming the constants, we obtain the inequality,

$$w\left(x - nc_+^* - x_1\right) + \overline{w}\left(x - nc_-^* - \bar{x}_1\right) - 1 - \bar{q}_0 e^{-\mu n} \leqq u_n(x)$$

for $n \geqq k$. Finally we increase $\bar{q}_0$ so that $1 - \bar{q}_0 e^{-\mu k} \leqq 0$. This completes the proof for the left inequality in Theorem 3.1.

For the right inequality, choose $\bar{u}_0$ that satisfies the hypotheses of Theorem 2.3 with $u_0 \leqq \bar{u}_0$. Then from Lemma 2.8 and Theorem 2.3, $u_n(x) \leqq w(x - nc_+^* - x_2) + q_1 e^{-\mu_1 n}$ for all x. Similarly, $u_n(x) \leqq \overline{w}(x - nc_-^* - \bar{x}_2) + q_2 e^{-\mu_2 n}$ for all x. If $x \geqq 0$, then $1 - \overline{w}(x - nc_-^* - \bar{x}_2) \leqq Ae^{-\nu(x - nc_-^* - \bar{x}_2)} \leqq C_1 e^{\nu c_-^* n}$ so that

$$\begin{aligned} u_n(x) &\leqq w\left(x - nc_+^* - x_2\right) + \overline{w}\left(x - nc_-^* - \bar{x}_2\right) - 1 + C_1 e^{\nu c_-^* n} + q_1 e^{-\mu_1 n} \\ &\leqq w\left(x - nc_+^* - x_2\right) + \overline{w}\left(x - nc_-^* - \bar{x}_2\right) - 1 + \bar{q}_1 e^{-\bar{\mu} n}, \end{aligned}$$

where $\bar{\mu} = \min(-\nu c_-^*, \mu_1)$, $\bar{q}_1 = C_1 + q_1$. A similar argument holds for $x \leqq 0$. The proof of Theorem 3.1 is now complete.

LEMMA 4.2. *Let u_0 satisfy the hypotheses of Theorem* 3.2. *Suppose that some subsequence $u_{n_j}(x + n_j c_+^*)$ converges to a travelling wave w uniformly in $\mathbb{R}$, then $\lim_{n \to \infty} u_n(x + nc_+^*) = w(x)$ uniformly in $\mathbb{R}$.*

Proof. For any $\varepsilon > 0$, there exists j such that $u_0(x) = u_{n_j}(x + n_j c_+^*)$ satisfies the δ condition of Lemma 2.4 which implies Lemma 4.2.

LEMMA 4.3. *Let $K \in PF_3$, $c_-^* < 0 < c_+^*$, w, $\overline{w}$ be given and u_0 satisfy the hypotheses of Theorem* 3.3. *Suppose that every subsequence $u_{n_j}(x + n_j c_+^*)$ contains a further subsequence $u_{n'_j}(x + n'_j c_+^*)$ converging to a travelling wave $w(x - x_0)$ uniformly for $x \geqq -n'_j c_+^*$ (x_0 may depend on the final subsequence). Suppose also that the same is true for $u_n(x + nc_-^*)$ on the interval $x \leqq -nc_-^*$. Then $\lim_{n \to \infty} u_n(x + nc_+^*) = w(x - x_0)$ uniformly for $x \geqq -nc_+^*$ and $\lim_{n \to \infty} u_n(x + nc_-^*) = \overline{w}(x - x_1)$ uniformly for $x \leqq -nc_-^*$.*

Proof. *Step* 1. By our hypotheses, there exists a subsequence n_j such that $u_{n_j - 1}(x + (n_j - 1)c_+^*)$ converges to $w(x - x_0)$ uniformly for $x \geqq -(n_j - 1)c_+^*$. From (2.1), u_n converges to 1 uniformly on $[n\bar{c}, 0]$ for every $c_-^* < \bar{c} < 0$. Thus $u_{n_j - 1}(x + (n_j - 1)c_+^*)$ converges to $w(x - x_0)$ uniformly for $x \geqq (n_j - 1)(\bar{c} - c_+^*)$. From (1.3),

$$\begin{aligned} &(4.18)\quad u_{n_j}\left(x + n_j c_+^*\right) - w\left(x - x_0\right) \\ &\qquad = \int K\left(x + c_+^* - y\right)\left[g\left(u_{n_j - 1}\left(y + \left(n_j - 1\right)c_+^*\right)\right) - g\left(w\left(y - x_0\right)\right)\right] dy. \end{aligned}$$

It is clear that the integral over the region $y \geqq (n_j - 1)(\bar{c} - c_+^*)$ goes to 0 uniformly in x. On the rest of $\mathbb{R}$, if $x \geqq -n_j c_+^*$, we have

$$\int_{-\infty}^{(n_j - 1)(\bar{c} - c_+^*)} K\left(x + c_+^* - y\right) dy \leqq \int_{-(n_j - 1)\bar{c}}^{\infty} K(z)\, dz \to 0 \quad \text{as}\, j \to \infty.$$

Therefore $u_{n_j}(x + n_j c_+^*) \to w(x - x_0)$ uniformly for $x \geqq -n_j c_+^*$. Differentiating (4.18) and using the fact that K' is integrable, we see by the same argument that $u'_{n_j}(x + n_j c_+^*) \to w'(x - x_0)$ uniformly for $x \geqq -n_j c_+^*$. Now apply the second part of the hypotheses to the sequence $u_{n_j - 1}(x + (n_j - 1)c_-^*)$ and as above select a subsequence $j \equiv n'_j$ of n_j such that $u_j(x+jc_-^*)\to\bar{w}(x-x_1)$, $u_j'(x+jc_-^*)\to\bar{w}'(x-x_1)$ uniformly for $x\leqq -jc_-^*$. Since w', $\bar{w}'$ do not vanish in ℝ, the above assertions imply that there exists an integer N and a $\varepsilon>0$ such that every line $u=l$, where $\alpha-\varepsilon\leqq l\leqq\alpha+\varepsilon$, crosses $u_N(s)$ twice in ℝ. From Lemma 2.9, u_n is nondecreasing in the interval $[m_n^-(g^n(\alpha-\varepsilon)),\ m_n^-(g^n(\alpha+\varepsilon))]$ and nonincreasing in the interval $[m_n^+(g^n(\alpha+\varepsilon)),\ m_n^+(g^n(\alpha-\varepsilon))]$ for all $n\geqq N$.

Step 2. Let $\delta>0$ be arbitrarily small and let $u_{n_j}(x+n_jc_+^*)$ converge to $w(x)$ uniformly for $x\geqq -n_jc_+^*$. We may assume $x_0=0$ as long as we work with one subsequence. There exists an integer J such that $n_J\geqq N$ and $u_{n_j}(x)\geqq w(x-n_jc_+^*)-\delta$ for $x\geqq 0$ and $j\geqq J$.

Now in analogy to Lemma 4.1, there exists a decreasing sequence $\{z_n\}$ and positive constants $\bar{q}_0$, μ_1 such that $v_n(x)=w(x-nc_+^*-z_n)-\bar{q}_0e^{-\mu_1 n}$ satisfies $v_{n+1}\leqq Q[v_n]$ for all n. It is clear that this inequality still holds if we replace $w(x)$ by $w(x-\tau)$. From the way the z_n's are defined (see the proof of Lemma 2.4), the inequality holds independently of the exact choice of z_0. Let $z^*=\lim_{n\to\infty}z_n=-k'\bar{q}_0+z_0$. Finally, $\bar{q}_0$, μ_1 only need to be sufficiently small, like the constants q_0 and μ in (4.5), (4.8)–(4.10).

Choose $1-\delta<\bar{\alpha}<1$ and $j\geqq J$ large enough that

$$\alpha<w\big(-n_jc_+^*-z^*\big)-\delta<\bar{\alpha}\quad\text{and}\quad g^{n_j}(\alpha+\varepsilon)>\bar{\alpha}. \tag{4.19}$$

From Theorem 3.1, there exist constants x_1, $\bar{x}_1$, q_0, μ such that $w(x-nc_+^*-x_1)+\bar{w}(x-nc_-^*-\bar{x}_1)-1-q_0e^{-\mu n}\leqq u_n(x)$ for all n. From (4.2), we have

$$1-w\big(-(n+n_j+1)c_+^*-x_1\big)\leqq\frac{g'(1)\delta}{5}e^{-\nu c_+^*n}, \tag{4.20a}$$

$$1-\bar{w}\big(-(n+n_j+1)c_-^*-\bar{x}_1\big)\leqq\frac{g'(1)\delta}{5}e^{\nu c_-^*n}, \tag{4.20b}$$

and

$$q_0e^{-\mu(n_j+1)}\leqq\frac{g'(1)\delta}{2} \tag{4.21}$$

for sufficiently large j. We then fix such a $j(\geqq J)$ and define $\bar{n}=n_j$.

We replace $w(x)$ by $w(x-\bar{n}c_+^*)\equiv w_1(x)$ in the above definition of $v_n(x)$ and note that $v_{n+1}\leqq Q[v_n]$ for all n. As remarked earlier, we may let $z_0=0$, $\bar{q}_0=\delta$ and choose $\mu_1>0$ small enough that

$$0<\mu_1<\min\{\mu,\ -\nu c_-^*,\nu c_+^*\}. \tag{4.22}$$

Step 3. We establish some properties of the sequence u_n. First,

$$v_n\leqq g^n(\bar{\alpha})\quad\text{for all } n. \tag{4.23}$$

This is obvious when $n=0$. Assuming that it is true up to n, then

$$v_{n+1}(x)\leqq Q[v_n](x)=\int K(x-y)g(v_n(y))\,dy\leqq g^{n+1}(\bar{\alpha}).$$

Thus (4.23) is proved. Next, we claim that

$$Q[v_n](0)<u_{\bar{n}+n+1}(0)\quad\text{for all } n. \tag{4.24}$$

We have from condition (xii) of (1.9),

$$Q[v_n](0)=\int K(-y)g\left[w_1\left(y-nc_+^*-z_n\right)-\delta e^{-\mu_1 n}\right]dy$$
$$\leqq g\left(1-\delta e^{-\mu_1 n}\right)\leqq 1-g'(1)\delta e^{-\mu_1 n}.$$

On the other hand, according to (4.20),

$$u_{\bar n+n+1}(0)\geqq w\left(-(\bar n+n+1)c_+^*-x_1\right)+\bar w\left(-(\bar n+n+1)c_-^*-\bar x_1\right)-1-q_0e^{-\mu(\bar n+n+1)}$$
$$\geqq -\frac{g'(1)\delta}{5}e^{-\nu c_+^* n}-\frac{g'(1)\delta}{5}e^{\nu c_-^* n}+1-q_0e^{-\mu(\bar n+n+1)}.$$

From (4.21) and (4.22),

$$g'(1)\delta e^{-\mu_1 n}-q_0e^{-\mu(\bar n+n+1)}\geqq g'(1)\delta e^{-\mu_1 n}-\frac{g'(1)\delta}{2}e^{-\mu n}$$
$$\geqq\frac{g'(1)}{2}\delta e^{-\mu_1 n}>\frac{g'(1)\delta}{5}e^{-\nu c_+^* n}+\frac{g'(1)\delta}{5}e^{\nu c_-^* n}.$$

Combining this with the previous two inequalities we obtain (4.24).

Finally, we claim that $v_0(x)=w_1(x)-\delta$ crosses $u_{\bar n}(x)$ once in $\mathbb{R}^-$. We know from the beginning of step 2 that $u_{\bar n}(x)\geqq w(x-\bar nc_+^*)-\delta$ for $x\geqq 0$. On $\mathbb{R}^-$, however, (4.19) and (4.23) imply that $\alpha\leqq v_0(0)\leqq v_0(x)<\bar\alpha$. Thus v_0 and $u_{\bar n}$ can be equal only in the interval $[m_{\bar n}^-(\alpha),m_{\bar n}^-(\bar\alpha)]$. This interval is part of $[m_{\bar n}^-(\alpha),m_{\bar n}^-(g^{\bar n}(\alpha+\varepsilon))]$ by (4.19) and on the larger interval, $u_{\bar n}$ is nondecreasing according to step 1. Thus v_0 and $u_{\bar n}$ must cross exactly once in $\mathbb{R}^-$.

Step 4. We now prove that v_n and $u_{\bar n+n}$ cross once in $\mathbb{R}^-$ for all n. The proof is by induction and the case $n=0$ has just been proved. Assuming that this is true up to n, then since K is PF_2 and g is increasing, we see that $Q[v_n]$ crosses $Q[u_{\bar n+n}]=u_{\bar n+n+1}$ once. From (4.24), this crossing must occur in $\mathbb{R}^-$. Hence $v_{n+1}\leqq Q[v_n]\leqq u_{\bar n+n+1}$ in $\mathbb{R}^+$. If $x^*<0$ and $v_{n+1}(x^*)=u_{\bar n+n+1}(x^*)$, then from (4.19) and (4.23), $\alpha<v_{n+1}(0)\leqq v_{n+1}(x^*)\leqq g^{n+1}(\bar\alpha)$ and hence x^* lies in the interval $[m_{\bar n+n+1}^-(\alpha),m_{\bar n+n+1}^-(g^{n+1}(\bar\alpha))]$. This interval is contained in $[m_{\bar n+n+1}^-(\alpha),m_{\bar n+n+1}^-(g^{\bar n+n+1}(\alpha+\varepsilon))]$ by (4.19). According to step 1, $u_{\bar n+n+1}$ is nondecreasing there and hence v_{n+1} crosses $u_{\bar n+n+1}$ exactly once in $\mathbb{R}^-$. This completes our induction.

Step 5. Since $v_{n+1}(0)\leqq Q[v_n](0)<u_{\bar n+n+1}(0)$ for all n, we have $v_n(x)\leqq u_{\bar n+n}(x)$ for all n and $x\geqq 0$, or what is the same,

$$w\left(x-(\bar n+n)c_+^*-z_n\right)-\delta e^{-\mu_1 n}\leqq u_{\bar n+n}(x)\quad\text{for } x\geqq 0,\ n\geqq 0.$$

Since $z_n\to -k'\delta$ as $n\to\infty$, and since k', as in the proof of Lemma 2.4, is independent of δ, we have $0\leqq\liminf_{n\to\infty}[u_n(x)-w(x-nc_+^*)]$ uniformly for $x\geqq 0$. The opposite inequality is proved similarly and the proof of the first half of Lemma 4.3 is complete. The proof of the second half is the same.

Remark 4.1. Steps 3, 4, and 5 of the above proof are similar to comparing two functions u and v which satisfy a parabolic equation in the positive quadrant. If $u(x,0)\geqq v(x,0)$ for $x\geqq 0$, and $u(0,t)\geqq v(0,t)$ for $t\geqq 0$, then the maximum principle implies that $u(x,t)\geqq v(x,t)$ for $x\geqq 0$, $t\geqq 0$.

5. Proof of Theorems 3.2 and 3.3 (uniform convergence). We may assume in the statements of Theorems 3.2 and 3.3. that $\tau=0$ and $\sigma=1$. For let $\bar u_0(x)=u_0(\sigma x)$ and define recursively $\bar u_{n+1}=Q_1[\bar u_n]$ for all n where $Q_1[\bar u](x)=\int K_1(x-y)g(\bar u(y))dy$. Then a simple inductive argument shows that $\bar{u}_n(x)=u_n(\sigma x+n\tau)$. If w and $\bar{w}$ are the travelling waves of the operator Q with speed c_+^* and c_-^* respectively, then $w_1(x)=w(\sigma x)$, $\bar{w}_1(x)=\bar{w}(\sigma x)$ are the travelling waves of the operator Q_1 with speed $\rho_+^*=(c_+^*-\tau)/\sigma$ and $\rho_-^*=(c_-^*-\tau)/\sigma$. Furthermore, $\rho_-^*<0<\rho_+^*$. If we can prove Theorem 3.3 for the case $\tau=0$, $\sigma=1$, then $|\bar{u}_n(x)-w_1(x-n\rho_+^*)-\bar{w}_1(x-n\rho_-^*)+1|\leqq Ce^{-\varepsilon n}$ for all n and x. Changing back to the original variables, Theorem 3.3 follows. A similar remark holds for Theorem 3.2.

From now on, we assume that $K=K_1$. We give the details of the proof only for Theorem 3.3 and indicate at the end how to modify the proof to suit Theorem 3.2. The following notation will be used for the rest of this paper: $\mathscr{H}=L^2(\mathbb{R})$, $\beta^2=\exp\{-(c_+^*)^2/2\}$ and if a function $G(x,y)$ defines a linear operator $\int G(x,y)\phi(y)\,dy$ on a subspace of $\mathscr{H}$, then the operator will also be denoted by G. Recall the operator norm $\|G\|=\sup_{\|\phi\|_2\leqq 1}\|G\phi\|_2$ for G defined on $\mathscr{H}$. G^* denotes the adjoint of G and we shall use the inequality $\|f*\phi\|_2\leqq\|f\|_1\|\phi\|_2$ frequently without drawing attention to it. Finally, we let $v_n(x)=u_n(x+nc_+^*)$ satisfy the recursion

$$v_{n+1}(x)=\int K(x+c_+^*-y)g(v_n(y))\,dy.$$

To begin the proof, we rearrange the above so that

$$e^{c_+^*x}v_{n+1}(x)=\beta^2\int K(x-y)e^{c_+^*y}g(v_n(y))\,dy. \tag{5.1}$$

Let $K_2(x)=\beta\pi^{-1/2}\exp\{-x^2\}$. It is easily verified that $\beta^2K=K_2*K_2$. Since K_2 is even and integrable, it defines a self-adjoint bounded operator from $\mathscr{H}$ into $\mathscr{H}$. Also, the Fourier transform of K_2 is nonvanishing so that the operator K_2 is one-to-one.

Write (5.1) as

$$e^{c_+^*x}v_{n+1}(x)=K_2*K_2^*\left[e^{c_+^*y}g(v_n)\right](x). \tag{5.2}$$

From the fact that $c_+^*>0$, Lemma 2.10, (2.3a) and (2.5a), we see that $e^{c_+^*x}g(v_n)\in\mathscr{H}$ for every n. We write (5.2) in the form

$$e^{c_+^*x}(K_2^*)^{-1}K_2^{-1}\left[e^{c_+^*y}v_{n+1}\right](x)=e^{2c_+^*x}g(v_n(x)). \tag{5.3}$$

Let $A:S\to\mathscr{H}$ be the linear operator $A\phi=K_2^{-1}[e^{c_+^*x}\phi]$ defined on the subspace $S=\{\phi:e^{c_+^*x}\phi=K_2\psi$ for some $\psi\in\mathscr{H}\}$. Then (5.3) becomes

$$A^*A[v_{n+1}]=e^{2c_+^*x}g(v_n), \tag{5.4}$$

where $A^*=e^{c_+^*x}[K_2^*]^{-1}$. A direct calculation using (5.2) will show that

$$\int(Av_{n+1}-Av_n)Av_{n+1}\,dx=\int(v_{n+1}-v_n)A^*Av_{n+1}\,dx.$$

We define the functional

$$I[\phi]=\int\left\{\frac{(A\phi)^2}{2}-e^{2c_+^*x}\int_0^{\phi(x)}g(s)\,ds\right\}dx \tag{5.5}$$

on the set S. The second term in (5.5) is finite beause $g(s)\leqq Cs$ for some constant C if $s\geqq 0$, and $g(s)=0$ if $s<0$. Also $\phi\in S$ implies that $e^{c_+^*x}\phi\in\mathscr{H}$ and $A\phi\in\mathscr{H}$.

From (5.2), $v_n \in S$ for all $n \geqq 1$. Therefore,

$$
\begin{aligned}
&I[v_{n+1}]-I[v_n]\\
&\quad = \int \left\{ \frac{(Av_{n+1})^2-(Av_n)^2}{2} - e^{2c_+^*x}\int_{v_n(x)}^{v_{n+1}(x)} g(s)\,ds \right\} dx\\
&\quad = \int \left\{ \frac{-(Av_{n+1}-Av_n)(Av_{n+1}-Av_n-2Av_{n+1})}{2} - e^{2c_+^*x}\int_{v_n}^{v_{n+1}} g(s)\,ds \right\} dx\\
&\quad = -\int \frac{(Av_{n+1}-Av_n)^2}{2}dx + \int \left\{ (v_{n+1}-v_n)A^*Av_{n+1} - e^{2c_+^*x}\int_{v_n}^{v_{n+1}} g(s)\,ds \right\} dx\\
&\quad = -\int \frac{(Av_{n+1}-Av_n)^2}{2}dx + \int e^{2c_+^*x}\left[(v_{n+1}-v_n)g(v_n) - \int_{v_n}^{v_{n+1}} g(s)\,ds \right] dx\\
&\quad \leqq -\int \frac{(Av_{n+1}-Av_n)^2}{2}dx.
\end{aligned}
$$

The last equality follows from (5.4) and the last inequality follows from the fact that the term inside the square bracket is nonpositive.

Summing this inequality from 1 to $n-1$, we have

$$I[v_n]-I[v_1] \leqq -\frac{1}{2}\sum_{k=1}^{n-1}\int (Av_{k+1}-Av_k)^2 dx. \tag{5.6}$$

Suppose we can show that $I[v_n] > -M$ for some $M>0$; then the series in (5.6) converges and in particular $\lim_{n\to\infty}\int(Av_{n+1}-Av_n)^2dx = 0$. Since K_2 is bounded on $\mathscr{H}$, we have from the definition of A, $\|e^{c_+^*x}[v_{n+1}-v_n]\|_2 \to 0$ as $n\to\infty$. This implies that the hypotheses of Lemma 4.3 are satisfied. To see this, let $\{v_{n_j}\}$ be given. Then since $0 \leqq v_{n_j} \leqq 1$ and $|(v_{n_j})'| \leqq \|K'\|_1$, $\{v_{n_j}\}$ is equicontinuous and there exists a subsequence $\{v_{n_j'}\}$ such that $v_{n_j'}$ converges uniformly on compact sets to a continuous function ϕ. The same is also true for the functions $v_{n_j'+1}$ which converge uniformly on compact sets to $Q[\phi](x+c_+^*)$. Since $\lim_{j\to\infty}\int_{\mathscr{J}} e^{2c_+^*x}[v_{n_j'+1}(x)-v_{n_j'}(x)]^2dx = 0$ on any bounded interval $\mathscr{J}$, we see that $Q[\phi](x+c_+^*) = \phi(x)$. From Theorem 3.1, $\phi(-\infty) > \alpha$, $\phi(\infty) < \alpha$ and from Lemma 2.5 $\phi(x) = w(x-x_0)$, for some constant x_0. If we substitute $x+nc_+^*$ for x in Theorem 3.1, we see that $u_n(x+nc_+^*)$ is near 0 for x sufficiently large independent of n, and is near 1 for $-nc_+^*+n\bar{c} \leqq x \leqq L$, where $c_-^* < \bar{c} \leqq 0$ and L depends only on the constant x_1 in Theorem 3.1. Thus $\lim_{j\to\infty}u_{n_j'}(x+n_j'c_+^*) = w(x-x_0)$ uniformly for $x \geqq -n_j'c_+^* + n_j'\bar{c}$ and Lemma 4.3 implies that $\lim_{n\to\infty}u_n(x+nc_+^*) = w(x-x_0)$ uniformly for $x \geqq -nc_+^*$. A similar argument applies to $u_n(x+nc_-^*)$. Finally, since $\lim_{n\to\infty}[1-w(x-nc_+^*-x_0)] = 0$ uniformly for $x \leqq 0$, and $\lim_{n\to\infty}[1-\bar{w}(x-nc_-^*-x_1)] = 0$, uniformly for $x \geqq 0$, the uniform convergence part of Theorem 3.3 follows.

It remains to show that $I[v_n] > -M$ for all n. Since the normal density is in PF_2, we have from Lemma 2.10, $u_n(x+nc_+^*) \leqq w(x-L)$ for $x \geqq L$. From (2.3a), (2.5a) and the fact that $c_+^*>0$, we see that the L^1-norm of the function $e^{2c_+^*x}\int_0^{v_n(x)}g(s)ds \leqq e^{2c_+^*x}v_n(x)$ is bounded by some constant $M>0$. Thus $-M \leqq I[v_n] \leqq I[v_1]$ for all $n \geqq 1$. The proof for the functions $u_n(x+nc_-^*)$ in the negative direction, assuming that $c_-^*<0$ and $u_0(x)=0$ for x small, is similar. This completes the proof of that part of Theorem 3.3, on the uniform converence to a pair of diverging waves.

The above argument is equally valid for Theorem 3.2 provided $c_+^* > 0$. The difference is we need to prove that the hypotheses of Lemma 4.2 (instead of Lemma 4.3) are satisfied and after we have shown that $v_{n'_j}$ converges uniformly on compact subsets of $\mathbb{R}$, we deduce the uniform convergence in $\mathbb{R}$ from Theorem 2.3.

If $c_+^* < 0$, then the easiest way to obtain Theorem 3.2 is by using the dual wave speed introduced in [25, §2]. If we let $\tilde{g}(u) = 1 - g(1-u)$ and $\tilde{Q}[u] = K * \tilde{g}(u)$, then $\tilde{g}$ satisfies a list of hypotheses similar to (1.9) with α replaced by $1-\alpha$. The wave speed in the positive and negative directions from $\tilde{Q}$ are c_-^* and c_+^* respectively and the corresponding travelling waves are $1-w$ and $1-\bar{w}$. Let $z_0 = 1 - u_0$, and define z_n recursively by $z_{n+1} = \tilde{Q}[z_n]$. From an inductive argument, $z_n = 1 - u_n$ for all n. Since $z_0 = 0$ for x sufficiently small and $c_+^* < 0$, the analogue of the above proof in the negative direction will imply that $z_n(x + nc_+^*)$ converges to $1 - w(x - x_0)$ uniformly in $\mathbb{R}$. This proves our result when $c_+^* < 0$.

If $c_+^* = 0$, then (5.4) becomes $A^*A[v_{n+1}] = g(v_n)$ and $\beta = 1$. We replace (5.5) by the functional

$$\tilde{I}[\phi] = \int \left\{ \frac{(A\phi)^2}{2} - \int_0^{\phi} g(s)\,ds + a\chi_{\{x<0\}} \right\} dx, \tag{5.7}$$

where $a = \int_0^1 g(s)\,ds - 1/2$. From Lemma 2.10, the integral in (5.7) over $\mathbb{R}^+$ with $\phi = v_n$ is uniformly bounded. On the other hand,

$$\int K_2 = 1 \quad \text{and} \quad \frac{1}{2} - \frac{(Av_n)^2}{2} = \int_{K_2 * g(v_{n-1})}^{1} s\,ds \leqq 1 - K_2 * g(v_{n-1}) = K_2 * [1 - g(v_{n-1})].$$

Since $u_0(x) = 1$ for small x, a result like Lemma 2.10 implies that $v_n(x) \geqq w(x - L')$ if $x \leqq L'$. From (viii) of (1.9), (2.3b) and the above inequality, we see that $1/2 - (Av_n)^2/2$ is dominated for all n by a function which is integrable near $-\infty$. A similar argument shows that the same is true for the function $\int_{v_n}^1 g(s)ds$. Thus $\tilde{I}[v_n]$ is bounded below for all n. The rest of the proof parallels the case $c_+^* > 0$. We note that $v_{n+1} - v_n$ is square integrable even though $v_n \notin \mathscr{H}$. This completes the proof of the uniform convergence part of Theorem 3.2.

6. Proof of Theorems 3.2 and 3.3 (exponential rate of convergence). In §5, we showed that u_n converges to a pair of diverging waves facing opposite directions as $n \to \infty$ uniformly in $\mathbb{R}$. We now show that the rate of convergence is actually exponential. The proof relies on spectral properties of positive linear operators. We begin by summarizing the ideas of the proof.

Recall that K is the standard normal density and w is a travelling wave with speed c_+^* facing right. Let $G^*(x,y) = K(x-y)h^2(y)$, where $h^2(x) = \beta^2 g'(w(x))$. The linear integral operator G^* is positive (order-preserving) and has 1 as an eigenvalue with the nonpositive eigenfunction $e^{c_+^* x} w'$.

Suppose we can show that $\|G^*\| = 1$, 1 is a simple eigenvalue and in the orthogonal complement of $e^{c_+^* x} w'$ the norm of G^* is less than 1. Then linearizing Q about w, we obtain by letting $v_n(x) = u_n(x + nc_+^*)$,

$$e^{c_+^* x}[v_{n+1}(x) - w(x)] = \int G^*(x,y) e^{c_+^* y}[v_n(y) - w(y)]\,dy + O\left(\|v_n - w\|_2^2\right).$$

However, $e^{c_+^* x}[v_n - w]$ is not orthogonal to $e^{c_+^* x} w'$. To achieve this, we must replace $w(x)$ by $w_n(x) = w(x - z_n)$ in the above formula where $|z_n| \to 0$ as $n \to \infty$. Then we obtain (essentially) the estimate,

$$\left\| e^{c_+^* x}[v_{n+1} - w_{n+1}] \right\|_2 \leqq \gamma \left\| e^{c_+^* x}[v_n - w_n] \right\|_2,$$

where $0 < \gamma < 1$. This proves the exponential decay of $v_n - w_n$ in a weighted L^2-norm. The final result follows from showing that $|z_n|$ converges to 0 exponentially and estimating the supremum norm of $v_n - w$ by its L^2-norm.

We mention in passing that there is a lot of interest in the topic of wave stability for nonlinear parabolic equations, especially on when and how stability may be deduced from the stability of the linearized equations. See [31] and [32] in relation to (1.1) and also [15] for further references.

We shall only present the proof for Theorem 3.3, it being the more difficult. In fact, we only show that $u_n(x + nc_+^*)$ converges to $w(x - x_0)$ exponentially for $x \geqq -nc_+^*$. The proof for the negative direction is the same. Theorem 3.3 then follows from these and (4.2).

Let $u_n(x + nc_+^*) \to w(x)$ uniformly for $x \geqq -nc_+^*$ and let $\bar{u}_n$ be defined for the rest of this section by

$$\bar{u}_n(x) = \begin{cases} w(x - nc_+^*), & x < 0, \\ u_n(x), & x \geqq 0. \end{cases}$$

Let $\bar{v}_n(x) = \bar{u}_n(x + nc_+^*)$; then $\bar{v}_n(x) \to w(x)$ uniformly in $\mathbb{R}$.

LEMMA 6.1. *There exists a sequence* $\{z_n : n \geqq N\}$ *such that* $|z_n| \to 0$ *and for* $n \geqq N$,

$$\text{(6.1a)} \quad \int e^{2c_+^* x}[g(w(x - z_n)) - g(w(x - z_{n+1}))][\bar{v}_{n+1}(x) - w(x - z_{n+1})]\,dx = 0$$

if $z_n \neq z_{n+1}$, *and*

$$\text{(6.1b)} \quad \int e^{2c_+^* x} g'(w(x - z_n)) w'(x - z_n)[\bar{v}_{n+1}(x) - w(x - z_n)]\,dx = 0$$

if $z_n = z_{n+1}$.

Proof. Let

$$\varepsilon_2 = e^{-2c_+^*} \int e^{2c_+^* x} g'(w(x))[w'(x)]^2 dx > 0,$$

$$M = \frac{1}{2} \|[g \circ w]''\|_\infty e^{2c_+^*} \int e^{2c_+^* x} |w'(x)|\,dx$$

and $0 < \varepsilon_1 < \varepsilon_2$. From Lemma 2.10, $u_n(x + nc_+^*) \leqq w(x - L)$ for x near ∞ independently of n. Thus, there exist $\varepsilon > 0$ and $N_1(\varepsilon)$ such that if $|\xi| < \varepsilon < 1$ and $n \geqq N_1$,

$$\text{(6.2)} \quad \frac{3}{2} \|[g \circ w]''\|_\infty \int e^{2c_+^* x} |\bar{v}_{n+1}(x) - w(x - \xi)|\,dx \leqq \varepsilon_1.$$

We choose ε small enough so that

$$\text{(6.3)} \quad \varepsilon^* = -\varepsilon_1 - 2\varepsilon M + \varepsilon_2 > 0.$$

Consider the function

$$(6.4)\quad \phi_n(\xi)=\frac{\int e^{2c_+^*x}\big[g(w(x-z_n))-g(w(x-\xi))\big]\big[\bar{v}_{n+1}(x)-w(x-\xi)\big]dx}{\xi-z_n}$$

if $\xi\neq z_n$ and define

$$\phi_n(z_n)=\int e^{2c_+^*x}g'(w(x-z_n))w'(x-z_n)\big[\bar{v}_{n+1}(x)-w(x-z_n)\big]\,dx.$$

By the mean value theorem, $|\phi_n(0)|\to 0$ as $n\to\infty$ independently of how we define z_n. Let $|\phi_n(0)|<\varepsilon\varepsilon^*$ if $n\geqq N_2$ and define $N=\max(N_1,N_2)$. We now choose $\{z_n: n\geqq N\}$ inductively.

Let $z_N\in(-\varepsilon,\varepsilon)$ and assume that $z_n\in(-\varepsilon,\varepsilon)$, $n>N$, has been chosen so that (6.1) holds with n replaced by $n-1$. Define $g(x,\xi)=g(w(x-z_n))-g(w(x-\xi))$ and $v(x,\xi)=\bar{v}_{n+1}(x)-w(x-\xi)$ on $|\xi|<\varepsilon$. From (6.4),

$$\phi_n'(\xi)=\frac{\int e^{2c_+^*x}\big[(\xi-z_n)g_\xi(x,\xi)-g(x,\xi)\big]v(x,\xi)\,dx}{(\xi-z_n)^2}+\frac{\int e^{2c_+^*x}g(x,\xi)w'(x-\xi)\,dx}{(\xi-z_n)}$$

$$\equiv I_1+I_2.$$

From the mean value theorem,

$$g(x,\xi)=\big[g_\xi(x,\xi)-g_{\xi\xi}(x,\theta)(\xi-z_n)\big](\xi-z_n)+\frac{g_{\xi\xi}(x,\bar{\theta})}{2}(\xi-z_n)^2.$$

Inequality (6.2) then implies that $|I_1|\leqq\varepsilon_1$ if $|\xi|<\varepsilon<1$. To estimate I_2, we use $g(x,\xi)=g'(w(x-\xi))w'(x-\xi)(\xi-z_n)+\frac{1}{2}[g\circ w]''(\theta)(\xi-z_n)^2$. Therefore,

$$I_2=\int e^{2c_+^*x}g'(w(x-\xi))\big[w'(x-\xi)\big]^2dx+\frac{(\xi-z_n)}{2}\int e^{2c_+^*x}[g\circ w]''(\theta)w'(x-\xi)\,dx.$$

The second term above is bounded in absolute value by $M|\xi-z_n|\leqq 2\varepsilon M$ since $|z_n|<\varepsilon$ and $|\xi|<\varepsilon<1$. From (6.3) we have $\phi_n'(\xi)\geqq-\varepsilon_1-2\varepsilon M+\varepsilon_2=\varepsilon^*>0$ if $|\xi|<\varepsilon<1$. From this and the fact that $|\phi_n(0)|<\varepsilon\varepsilon^*$, $\phi_n(\xi)$ must have a unique zero in the interval $(-\varepsilon,\varepsilon)$ which we shall call z_{n+1}. This completes our induction step. Since $|\phi_n(0)|\to 0$ as $n\to\infty$, the same is true for $|z_n|$. The proof of the lemma is complete.

Remark 6.1. The above proof does not prevent some of the z_n's from being equal.

From the definitions of $\bar{u}_n$ and $\bar{v}_n$, it is clear that

$$\bar{v}_n(x)=v_n(x)+H(-x-nc_+^*)\big[w(x)-v_n(x)\big],$$

where $v_n(x)=u_n(x+nc_+^*)$ and $H(x)$ is the Heaviside function. Looking at

$$v_{n+1}(x)=\int K(x+c_+^*-y)g(v_n(y))\,dy$$

$$=\int K(x+c_+^*-y)\big\{g(\bar{v}_n(y))+g'(\theta)\big[v_n(y)-\bar{v}_n(y)\big]\big\}\,dy,$$

we have

$$\text{(6.5)}\qquad \begin{aligned}\bar{v}_{n+1}(x)=&\int K(x+c_+^*-y)g(\bar{v}_n(y))\,dy\\&+H(-x-(n+1)c_+^*)[w(x)-v_{n+1}(x)]+A_n(x),\end{aligned}$$

where

$$\text{(6.6)}\qquad A_n(x)=\int K(x+c_+^*-y)g'(\theta)[v_n(y)-\bar{v}_n(y)]\,dy.$$

Define $w_n(x)=w(x-z_n)$ and

$E_n(x)=H(-x-(n+1)c_+^*)[w(x)-v_{n+1}(x)]+A_n(x)+[w_n(x)-w_{n+1}(x)]$ for $n\geqq N$.

From (6.5),

$$\text{(6.7)}\quad \bar{v}_{n+1}(x)-w_{n+1}(x)=\int K(x+c_+^*-y)[g(\bar{v}_n(y))-g(w_n(y))]\,dy+E_n(x).$$

Define the functions,

$$h_n^2(x)=\begin{cases}\beta^2[g(w_{n-1}(x))-g(w_n(x))]/[w_{n-1}(x)-w_n(x)] & \text{if } z_{n-1}\neq z_n,\\ \beta^2 g'(w_n(x)) & \text{if } z_{n-1}=z_n,\end{cases}$$

and

$$k_n(x)=\begin{cases}\beta^2\left[\dfrac{g(\bar{v}_n(x))-g(w_n(x))}{\bar{v}_n(x)-w_n(x)}-\dfrac{g(w_{n-1}(x))-g(w_n(x))}{w_{n-1}(x)-w_n(x)}\right] & \text{if } z_{n-1}\neq z_n,\\ \beta^2\left[\dfrac{g(\bar{v}_n(x))-g(w_n(x))}{\bar{v}_n(x)-w_n(x)}-g'(w_n(x))\right] & \text{if } z_{n-1}=z_n.\end{cases}$$

We then rewrite (6.7) as

$$\text{(6.8)}\quad \begin{aligned}e^{c_+^*x}[\bar{v}_{n+1}(x)-w_{n+1}(x)]=&\int K(x-y)h_n^2(y)e^{c_+^*y}[\bar{v}_n(y)-w_n(y)]\,dy\\&+\int K(x-y)k_n(y)e^{c_+^*y}[\bar{v}_n(y)-w_n(y)]\,dy+e^{c_+^*x}E_n(x).\end{aligned}$$

Again define the functions $p_n(x)=h_n(x)e^{c_+^*x}[\bar{v}_n(x)-w_n(x)]$, $r_n(x)=k_n(x)/h_n(x)$ and $G_n(x,y)=h_n(x)K(x-y)h_n(y)$ and rewrite (6.8) as

$$\text{(6.9)}\quad \begin{aligned}p_{n+1}(x)=\frac{h_{n+1}(x)}{h_n(x)}\int G_n(x,y)p_n(y)\,dy+h_{n+1}(x)\int K(x-y)r_n(y)p_n(y)\,dy\\+h_{n+1}(x)e^{c_+^*x}E_n(x).\end{aligned}$$

We now prove several lemmas concerning the terms on the right side of (6.9).

LEMMA 6.2. *$\|h_{n+1}[K*(r_np_n)]\|_2\leqq C_1\|r_n\|_\infty\|p_n\|_2$ for large n and $\lim_{n\to\infty}\|r_n\|_\infty=0$.*

Proof. The inequality is clear because $\|h_n\|_\infty$ is uniformly bounded and $\int K=1$. To prove the rest, it suffices to show that $\|k_n\|_\infty\to 0$ as $n\to\infty$. This is clear from the definition and the fact that w_n and $\bar{v}_n$ converge to w uniformly in $\mathbb{R}$ as $n\to\infty$.

LEMMA 6.3. *There exists* $\delta_1 > 0$ *such that*

$$\left\| e^{c_+^* x} H\left(-x-(n+1)c_+^*\right) h_{n+1}[w-v_{n+1}] \right\|_2 \leqq C_2 e^{-\delta_1 n}$$

for large n.

Proof. This follows from the calculation

$$\int_{-\infty}^{-(n+1)c_+^*} e^{2c_+^* x}\left[w(x)-v_{n+1}(x)\right]^2 dx \leqq C_2' e^{-2\delta_1 n},$$

where $\delta_1 = (c_+^*)^2$.

LEMMA 6.4. *There exists* $\delta_2 > 0$ *such that* $\|h_{n+1} e^{c_+^* x} A_n\|_2 \leqq C_3 e^{-\delta_2 n}$ *for large* n.

Proof. From (6.6) and the fact that $v_n(x) = \bar{v}_n(x)$ if $x \geqq -nc_+^*$, we have $|A_n(x)| \leqq C_3' \int_{x+(n+1)c_+^*}^{\infty} K(z)\,dz \leqq C_3'$. Choose $\bar{c} < 0$ such that $c_+^* + \bar{c} > 0$. Using C_3' as a generic constant, we have

$$(6.10) \qquad \int e^{2c_+^* x} |A_n(x)|^2 dx \leqq C_3' \int_{-\infty}^{(n+1)\bar{c}} e^{2c_+^* x} dx + \int_{(n+1)\bar{c}}^{\infty} e^{2c_+^* x} |A_n(x)|^2 dx.$$

From condition (iv) of (1.8) there exists $\delta > c_+^*$ such that $K(x) \leqq C_3' e^{-\delta x}$ for $x \geqq 0$. If $x \geqq (n+1)\bar{c}$, we have

$$\int_{x+(n+1)c_+^*}^{\infty} K(z)\,dz \leqq C_3' e^{-\delta[x+(n+1)c_+^*]}$$

so that

$$\begin{aligned}\int_{(n+1)\bar{c}}^{\infty} e^{2c_+^* x} |A_n(x)|^2 dx &\leqq C_3' \int_{(n+1)\bar{c}}^{\infty} e^{2c_+^* x} e^{-2\delta[x+(n+1)c_+^*]} dx \\ &= C_3' e^{-2\delta c_+^*(n+1)} \int_{(n+1)\bar{c}}^{\infty} e^{-2(\delta - c_+^*)x} dx \\ &= C_3' e^{-2\delta c_+^*(n+1)} e^{-2(\delta - c_+^*)\bar{c}(n+1)} \\ &= C_3' e^{-2\delta(c_+^* + \bar{c})(n+1)} e^{2c_+^* \bar{c}(n+1)}.\end{aligned}$$

Since $\bar{c} < 0$, the first integral in (6.10) decays exponentially and since $c_+^* + \bar{c} > 0$, the above estimate implies the same is true of the second integral. This completes the proof of our lemma.

LEMMA 6.5. *For sufficiently large* n,

(i) $\left(e^{c_+^* x} h_{n+1}[w_n - w_{n+1}], p_{n+1}\right) = 0$ *if* $z_n \neq z_{n+1}$, *and*

(ii) $\left(e^{c_+^* x} h_{n+1} w_n', p_{n+1}\right) = 0$ *if* $z_n = z_{n+1}$.

Proof. This lemma follows from Lemma 6.1.

We now take the L^2-inner product of (6.9) with the function p_{n+1}. From Lemmas 6.2 to 6.5, we obtain for large n, with the obvious change of notations, the estimate

$$(6.11) \qquad \|p_{n+1}\|_2^2 \leqq \left| \left(\frac{h_{n+1}}{h_n} G_n p_n, p_{n+1} \right) \right| + C_1 \|r_n\|_\infty \|p_n\|_2 \|p_{n+1}\|_2 + C_2 \|p_{n+1}\|_2 \gamma^n,$$

where $0<\gamma<1$ is some constant. At this point, in order not to interrupt the flow of ideas, we state a lemma and defer its proof until the end of this section.

LEMMA 6.6. *For large n, there exist constants $0<\gamma_n<1$ such that $\|G_n p_n\|_2 \leqq \gamma_n \|p_n\|_2$ and $\limsup_{n\to\infty}\gamma_n=\gamma^*<1$.*

Let $\theta>0$ be such that for large n, $\gamma_n \leqq (1-\sqrt{2\theta})$ and $\|h_{n+1}/h_n\|_\infty \leqq (1+\sqrt{2\theta})$. Therefore, from (6.11), Lemmas 6.2 and 6.6,

$$\|p_{n+1}\|_2 \leqq (1-\theta)\|p_n\|_2 + C_2\gamma^n \quad \text{for large } n. \tag{6.12}$$

LEMMA 6.7. *Let $\{a_n : n \geqq N\}$ be a sequence of positive numbers such that $a_{n+1} \leqq (1-\theta)a_n + C_2\gamma^n$ for some C_2, $\theta>0$ and $0<\gamma<1$. Then there exist C, $\varepsilon>0$ such that $a_n \leqq Ce^{-\varepsilon n}$ for $n \geqq N$.*

Proof. Since the inequality is satisfied if we decrease θ, we may assume that $\bar{\gamma}=\gamma e^{\theta}\in(0,1)$. For $n\geqq N$, we have

$$\begin{aligned} e^{\theta(n+1)}a_{n+1}-e^{\theta n}a_n &\leqq (1-\theta)e^{\theta(n+1)}a_n + C_2\gamma^n e^{\theta(n+1)} - e^{\theta n}a_n \\ &= (e^{\theta}-\theta e^{\theta}-1)e^{\theta n}a_n + C_2'\bar{\gamma}^n. \end{aligned}$$

For small $\theta>0$, $e^{\theta}\leqq 1+\theta e^{\theta}$ and hence $e^{\theta(n+1)}a_{n+1}-e^{\theta n}a_n \leqq C_2'\bar{\gamma}^n$. Summing the left side from N to $n-1$ and on the right from N to ∞, we obtain $e^{\theta n}a_n - e^{\theta N}a_n \leqq C_2'\bar{\gamma}^N(1-\bar{\gamma})^{-1}$ for $n \geqq N$. This last inequality implies our lemma.

Summarizing our results so far, we have from Lemma 6.7, (ix) of (1.9) and (6.12),

$$\left\| e^{c_+^* x}[\bar{v}_n - w_n] \right\|_2 \leqq Ce^{-\varepsilon n} \quad \text{for large } n. \tag{6.13}$$

We now show that (6.13) implies that $v_n(x)=u_n(x+nc_+^*)$ converges to a travelling wave facing right exponentially for $x \geqq -nc_+^*$ and then give the proof of Lemma 6.6.

LEMMA 6.8. *Let $f(x)$ be continuously differentiable on $[L,\infty)$ and let $f_0=\sup_{[L,\infty)}|f(x)|$, $f_1=\sup_{[L,\infty)}|f'(x)|$ be finite. Then*

$$f_0 \leqq (3)^{1/3}\left[\int_L^\infty f^2(x)\,dx\right]^{1/3} f_1^{1/3}.$$

Proof. Given $0<\delta<f_0$, there exists x_0 such that $|f(x_0)|>f_0-\delta$. Without loss of generality, we assume that $f(x_0)>0$. For $x>L$,

$$f(x)=f(x_0)+\int_{x_0}^{x} f'(y)\,dy \geqq f_0-\delta-|x-x_0|f_1.$$

Let $l=(f_0-\delta)/f_1$; then

$$\int_L^\infty f^2(x)\,dx \geqq \int_{x_0}^{x_0+l} f^2(x)\,dx \geqq \int_{x_0}^{x_0+l}\left(f_0-\delta-|x-x_0|f_1\right)^2 dx = \frac{1}{3}\frac{(f_0-\delta)^3}{f_1}.$$

Letting $\delta \downarrow 0$, we obtain our lemma.

Consider for a fixed and sufficiently large n, the function $\bar{v}_n - w$ on the interval $x \geqq -n\delta$. Here δ is chosen so that $0<\delta<c_+^*$ and $c_+^*\delta-\varepsilon<0$. The constant ε is taken from (6.13). Then from (6.13),

$$\int_{-n\delta}^{\infty}[v_n(x)-w_n(x)]^2 dx \leqq e^{2c_+^*\delta n}\int_{-n\delta}^{\infty} e^{2c_+^* x}[v_n(x)-w_n(x)]^2 dx \leqq Ce^{-\varepsilon n}$$

for large n and a different $\varepsilon>0$. Since v_n' and w_n' are bounded independently of n, Lemma 6.8 implies, again with a different C and ε, that $|v_n(x)-w_n(x)| \leqq Ce^{-\varepsilon n}$ for large

n and $x \geqq -n\delta$. On $[-nc_+^*, -n\delta]$, the same inequality holds because of Theorem 3.1, (2.3b) and (2.4b). The first finite number of terms can be handled by increasing the constant C. Thus the above inequality holds for all n and $x \geqq -nc_+^*$.

It remains to show that $|z_n|$ converges to 0 exponentially. To this end, we rearrange (6.9), when $z_n \neq z_{n+1}$, leaving only the term $h_{n+1}e^{c_+^* x}[w_n - w_{n+1}]$ on one side and then take the L^2-norm of both sides. From Lemmas 6.2 to 6.4, Lemma 6.6 and (6.13), we have $\|e^{c_+^* x}[w_n - w_{n+1}]\|_2 \leqq Ce^{-\varepsilon n}$ for large n and some $\varepsilon > 0$. Since $[w_n(x) - w_{n+1}(x)]/(z_{n+1} - z_n) \to w'(x)$ as $n \to \infty$, we have from Fatou's lemma, $\liminf_{n\to\infty} \|e^{c_+^* x}[(w_n - w_{n+1})/(z_{n+1} - z_n)]\|_2 \geqq \|e^{c_+^* x} w'\|_2 > 0$. Thus for large n, $|z_{n+1} - z_n| \leqq Ce^{-\varepsilon n}$. This implies that the $|z_n|$ decay exponentially as $n \to \infty$.

In the last part of this section, we are concerned with the spectral properties of the operator G_n and in particular the proof of Lemma 6.6. G_n as a bounded linear operator from $\mathscr{H}$ into $\mathscr{H}$ is positive, in the sense that $\phi \geqq 0$ implies that $G\phi \geqq 0$. A lot has been done on the spectral properties of positive operators, beginning with the fundamental paper of Krein and Rutman [22]. After verifying Lemma 6.9, Lemma 6.10 may be proved by using some powerful results on irreducible operators (consult the work by H. H. Schaefer).[1] However, a simple proof can be given here. We assume that n is sufficiently large in the following lemmas and that $z_n \neq z_{n+1}$. The case $z_n = z_{n+1}$ may be dealt with similarly.

LEMMA 6.9. *Let* $e_n = e^{c_+^* x} h_n [w_{n-1} - w_n]$. *Then*

(i) $G_n : \mathscr{H} \to \mathscr{H}$ *is a self-adjoint positive linear operator.*

(ii) $G_n e_n = e_n$ *and* e_n *is of one sign.*

(iii) *Let* $T_n\phi(x) = \int K(x-y)h_n^2(y)\phi(y)\,dy$; *then* $T_n : \mathscr{H} \to \mathscr{H}$ *is positive and quasi-compact, i.e., there exists a compact operator* V *and an integer* $m \geqq 0$ *such that* $\|T_n^m - V\| < 1$.

(iv) *Let* $r(T_n)$ *be the spectral radius of* T_n; *then* $r(T_n)$ *is an eigenvalue of* T_n *with a positive eigenfunction.*

(v) λ *is an eigenvalue of* T_n *if and only if* λ *is an eigenvalue of* G_n.

(vi) $(\lambda I - T_n)$ *is onto if and only if* $(\lambda I - G_n)$ *is onto.*

(vii) $r(T_n) = r(G_n) = 1$.

Proof. G_n is self-adjoint because $G_n(x,y) = G_n(y,x)$ and is positive because $G_n(x,y) > 0$. (ii) follows from a direct calculation and the fact that w is nonincreasing. To show that T_n is quasi-compact, we choose a bounded interval $\mathscr{J}$ such that on $\mathscr{J}^c$, $|h_n^2(y)| \leqq \beta^2 < 1$. This is possible because $g'(0) < 1$, $g'(1) < 1$ and $\beta^2 < 1$. Let $g_1(y) = h_n^2(y)\chi_{\mathscr{J}}(y)$, $g_2(y) = h_n^2(y) - g_1(y)$ and $V : \mathscr{H} \to \mathscr{H}$ be the operator $V\phi(x) = \int K(x-y)g_1(y)\phi(y)\,dy$. Then V is a compact operator since $\int\int K^2(x-y)g_1^2(y)\,dy\,dx < \infty$. Also $\|T_n - V\| < 1$ since $\|(T_n - V)\phi\|_2 = \|\int K(x-y)g_2(y)\phi(y)\,dy\|_2 \leqq \beta^2 \|K\|_1 \|\phi\|_2 < 1$ if $\|\phi\|_2 \leqq 1$. Thus from [19, Cor. 1 of Thm. 5], (iv) is valid. Note that 1 is an eigenvalue of T_n with eigenfunction $e^{c_+^* x}[w_{n-1} - w_n]$. (v) follows from the obvious fact that $T_n\phi = \lambda\phi$ if and only if $G_n[h_n\phi] = \lambda[h_n\phi]$. To prove (vi), let $f \in \mathscr{H}$ be given and suppose that $\lambda I - T_n$ is onto; then there exists $u \in \mathscr{H}$ such that $(\lambda I - T_n)u = f/h_n$ and hence $(\lambda I - G_n)[h_n u] = f$. Conversely, let f be given and $(\lambda I - G_n)$ be onto; then there exists $u \in \mathscr{H}$ such that $(\lambda I - G_n)u = h_n f$, and hence $(\lambda I - T_n)[u/h_n] = f$. Finally, since 1 is an eigenvalue of T_n, $r(T_n) \geqq 1$. If $r(T_n) > 1$ is an eigenvalue of T_n with positive eigenfunction e^*, then $r(T_n)$ is an eigenvalue of G_n with eigenfunction $h_n e^* \geqq 0$. Since G_n is self-adjoint, $h_n e^*$ is orthogonal to e_n which is impossible because $e_n > 0$. Thus $r(T_n) = 1$. Since G_n is a positive operator, [19, Thm. 4] implies that $r(G_n)$ is in the spectrum of G_n. Also from

[1] The author is grateful to the referee for pointing out this reference.

(ii), $r(G_n)\geqq 1$ and from (v) and (vi), $r(G_n)$ is in the spectrum of T_n so that $r(G_n)\leqq r(T_n)=1$. Therefore $r(G_n)=1$ and this completes the proof of Lemma 6.9.

Let H_n be the one-dimensional subspace spanned by e_n and let $\mathscr{H}=H_n+H_n'$ where $H_n'=\{\phi\in\mathscr{H}\colon (\phi,e_n)=0\}$. Define $\gamma_n=\sup\{\|G_n\phi\|_2\colon \phi\in H_n',\ \|\phi\|_2\leqq 1\}$.

LEMMA 6.10. $\gamma_n<1$.

Proof. Suppose $\gamma_n=1$; then $G_n\colon H_n'\to H_n'$ since G_n is self-adjoint and e_n is an eigenfunction of G_n corresponding to the eigenvalue 1. We claim that $(I-G_n)\colon H_n'\to H_n'$ is onto so that 1 is an eigenvalue for $G_n\colon H_n'\to H_n'$. To this end, let $f\in H_n'$ and consider $(I-T_n)\colon\mathscr{H}\to\mathscr{H}$. Since T_n is quasicompact, the result in the last section of [39] implies that f/h_n belongs to the range of $I-T_n$ if and only if f/h_n is orthogonal to all members of the null-space of $I-T_n^*$. Let $(I-T_n^*)\psi=0$; then $\psi-h_n^2[K*\psi]=0$ so that $(I-G_n)[\psi/h_n]=0$. We shall show later on that this implies $\psi/h_n\in H_n$. Assuming this for the moment, we then have $(f,\psi/h_n)=(f/h_n,\psi)=0$ and f/h_n belongs to the range of $(I-T_n)$. Let $(I-T_n)[\phi]=f/h_n$ and ϕ^* be the projection of $h_n\phi$ into H_n'; then $(I-G_n)[h_n\phi]=f$, $\phi^*\in H_n'$ is nontrivial and $(I-G_n)[\phi^*]=f$. Thus $I-G_n\colon H_n'\to H_n'$ is onto.

Finally, if 1 is an eigenvalue of $G_n\colon H_n'\to H_n'$, let $e^*\in H_n'$ be such that $(I-G_n)[e^*]=0$. Since $(e^*,e_n)=0$, the set $\{e^*>0\}$ and $\{e^*<0\}$ must be of positive measure and $|e^*|=|G_ne^*|<G_n|e^*|$. Therefore, $(e_n,|e^*|)<(e_n,G_n|e^*|)=(G_ne_n,|e^*|)=(e_n,|e^*|)$ which is a contradiction.[2] The only way out is to accept that $\gamma_n<1$. This last argument may be used to show that $\psi/h_n\in H_n$. For let $\psi/h_n=\lambda e_n+\theta$ where $\theta\in H_n'$; then $(I-G_n)[\theta]=0$ and $(\theta,e_n)=0$ which we have seen to be impossible. This completes the proof of Lemma 6.10.

LEMMA 6.11. *Let* $h^2(x)=\beta^2g'(w(x))$, $e^*=e^{c_+^*x}hw'$, $G(x,y)=h(x)K(x-y)h(y)$ *and* $\gamma^*=\sup\{\|G\phi\|_2\colon (\phi,e^*)=0,\ \|\phi\|_2\leqq 1\}$. *Then*:

(i) $\lim_{n\to\infty}h_n(x)=h(x)$ *uniformly in* $\mathbb{R}$.

(ii) $\|G_n-G\|\to 0$ *as* $n\to\infty$.

(iii) G *has* 1 *as an eigenvalue with the negative eigenfunction* e^*.

(iv) *If* $\bar{e}_n=e_n/(z_n-z_{n-1})$, *then* $\lim_{n\to\infty}\|\bar{e}_n-e^*\|_2=0$.

(v) $\lim_{n\to\infty}\gamma_n=\gamma^*<1$.

Proof. (i) follows from the mean value theorem and the fact that w_n converges to w uniformly in $\mathbb{R}$. (ii) is a consequence of (i) and (iii) follows from a direct calculation. To prove (iv), we first observe that

$$\frac{w(x-z_{n-1})-w(x-z_n)}{z_n-z_{n-1}}-w'(x)=w'(x-z_n)-w'(x)+\frac{w''(\theta)}{2}(z_n-z_{n-1}).$$

Since $|w''|$ is bounded in $\mathbb{R}$, we have $(w(x-z_{n-1})-w(x-z_n))/(z_n-z_{n-1})$ converges to $w'(x)$ uniformly in $\mathbb{R}$. This together with (i), Lemma 2.7 and the dominated convergence theorem imply (iv). Finally the methods of proof in Lemmas 6.9 and 6.10 are also applicable to the operator G so that $\gamma^*<1$. Let P_n and P be the projection operator from $\mathscr{H}$into the orthogonal complement of $\bar{e}_n$ and e^*; then $\gamma_n=\|G_nP_n\|$, $\gamma^*=\|GP\|$ and $|\|G_nP_n\|-\|GP\||\leqq\|G_nP_n-GP\|\leqq\|P_n-P\|+\|G_n-G\|\ \|P\|$. From (ii), it suffices to show that $\|P_n-P\|\to 0$ and this follows from (iv). Thus (v) is proved and so is Lemma 6.11.

Lemma 6.6 follows from Lemmas 6.5, 6.10 and 6.11 (v). If $z_{n-1}=z_n$, we replace e_n by $e^{c_+^*x}h_nw_n'$; then Lemmas 6.9, 6.10 and 6.11 are still valid. In Lemma 6.11, we may take $\bar{e}_n=e_n$. The proof of Theorems 3.2 and 3.3 is finally complete.

[2] The author is grateful to Professor John W. Lee for showing him this argument.

Appendix (proof of Lemma 2.5, conclusion). Without loss of generality, we may assume that the following situation has occurred and derive from it a contradiction. Let $e^* = e^{c_+^* x}[w_2 - w_1] > 0$ and $\phi_\varepsilon(x) = e^{c_+^* x}[w_2(x+\varepsilon) - w_1(x)]$ becomes negative infinitely often near ∞ for every $\varepsilon > 0$. Let $h_\varepsilon^2(x) = \beta^2(g(w_2(x+\varepsilon)) - g(w_1(x)))/(w_2(x+\varepsilon) - w_1(x))$, G_ε be defined like G_n right before (6.9) and T_ε be defined like T_n in Lemma 6.9. Then $T_\varepsilon = K_\varepsilon + V_\varepsilon$, where K_ε is compact, $\|V_\varepsilon\| < 1$ and both K_ε, V_ε converge. Also G_ε is self-adjoint and $T_\varepsilon \phi = \lambda \phi$ if and only if $G_\varepsilon[h_\varepsilon \phi] = \lambda[h_\varepsilon \phi]$. Since $T_\varepsilon \phi_\varepsilon = \phi_\varepsilon$ and ϕ_ε changes sign, we must have $r(T_\varepsilon) = r(G_\varepsilon) = \lambda_\varepsilon > 1(\lambda_\varepsilon \to 1)$ and there exists $e_\varepsilon \geqq 0$, $\|e_\varepsilon\|_2 = 1$ such that $T_\varepsilon e_\varepsilon = \lambda_\varepsilon e_\varepsilon$. Thus $(h_\varepsilon \phi_\varepsilon, h_\varepsilon e_\varepsilon) = 0$ for $\varepsilon > 0$. On the other hand, $h_\varepsilon \phi_\varepsilon$ converges in L^2 to a positive function. It remains to show that e_ε has a strong limit to arrive at a contradiction. This is obvious from writing $e_\varepsilon = (\lambda_\varepsilon I - V_\varepsilon)^{-1} K_\varepsilon e_\varepsilon$. The proof of Lemma 2.5 is complete.

Acknowledgment. Many of the ideas in this paper grew out of reading [7]. The author would like to acknowledge his enjoyment in reading that paper.

REFERENCES

[1] D. G. ARONSON AND H. F. WEINBERGER, *Nonlinear diffusion in population genetics, combustion, and nerve propagation*, in Partial Differential Equations and Related Topics, J. Goldstein, ed., Lecture Notes in Mathematics 446, Springer, New York, 1975, pp. 5–49.

[2] ———, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. Math., 30 (1978), pp. 33–76.

[3] M. BRAMSON, *Maximal displacement of branching Brownian motion*, Comm. Pure Appl. Math., 31 (1978), pp. 531–581.

[4] ———, *The convergence of solutions of the Kolmogorov nonlinear diffusion equation to travelling waves*, AMS Memoirs, No. 285, July 1983.

[5] W. FELLER, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed., John Wiley, New York, 1973.

[6] P. C. FIFE, *Long time behavior of solutions of bistable nonlinear diffusion equations*, Arch. Rat. Mech. Anal., 70 (1979), pp. 31–46.

[7] P. C. FIFE AND J. B. MCLEOD, *The approach of solutions of nonlinear diffusion equations to travelling wave solutions*, A.M.S. Bull., 81 (1975), pp. 1076–1078; Arch. Rat. Mech. Anal., 65 (1977), pp. 335–361.

[8] ———, *A phase plane discussion of convergence to travelling fronts for nonlinear diffusion*, Arch. Rat. Mech. Anal., 75 (1981), pp. 281–314.

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

[10] ———, *Clines induced by variable selection and migration*, Proc. Royal Soc. London B, 214 (1981), pp. 99–123.

[11] R. A. FISHER, *The advance of advantageous genes*, Ann. Eugenics, 7 (1937), pp. 355–369.

[12] K. P. HADELER AND F. ROTHE, *Travelling fronts in nonlinear diffusion equations*, J. Math. Biology, 2 (1975), pp. 251–263.

[13] P. S. HAGAN, *The instability of nonmonotonic wave solutions of parabolic equations*, Studies in Applied Math., 64 (1981), pp. 57–88.

[14] ———, *Traveling wave and multiple traveling wave solutions of parabolic equations*, this Journal, 13 (1982), pp. 717–738.

[15] D. HENRY, *Geometric theory of semilinear parabolic equations*, Lecture Notes in Mathematics 840, Springer, New York, 1981.

[16] C. K. R. T. JONES, *Spherically symmetric solutions of a reaction-diffusion equation*, J. Differential Equations, 49 (1983), pp. 142–169.

[17] JA. I. KANEL', *Stabilization of solutions of the Cauchy problem for equations encountered in combustion theory*, Mat. Sbornik (N.S.), 59,(101) (1962), supplement, pp. 245–288.

[18] ———, *On the stability of solutions of the equations of combustion theory for finite initial functions*, Mat. Sbornik (N.S.) 65 (107) (1964), pp. 398–413.

[19] S. KARLIN, *Positive operators*, J. Math. Mech., 8 (1959), pp. 907–937.

[20] G. A. KLAASEN AND W. C. TROY, *The stability of traveling wave front solutions of a reaction-diffusion system*, SIAM J. Appl. Math., 41 (1981), pp. 145–167.

[21] A. Kolmogoroff, I. Petrovsky and N. Piscounoff, *Étude de l'équations de la diffusion avec croissance de la quantité de matière et son application a un problème biologique*, Bull. Univ. Moscow. Ser. Internat., Sec. A, 1 (1937), 6, pp.1–25.

[22] M. G. Krein and M. A. Rutman, *Linear operators leaving invariant a cone in Banach space*, Amer. Math. Soc., Transl. No. 26, 1950.

[23] R. Lui, *A nonlinear integral operator arising from a model in population genetics* I, *monotone initial data*, this Journal, 13 (1982), pp. 913–937.

[24] ———, *A nonlinear integral operator arising from a model in population genetics* II, *initial data with compact support*, this Journal, 13 (1982) pp. 938–953.

[25] ———, *Existence and stability of travelling wave solutions of a nonlinear integral operator*, J. Math. Biol., 16 (1983), pp. 199–220.

[26] H. P. McKean, *Application of Brownian motion to the equation of Kolmogorov-Petrovskii-Piskunov*, Comm. Pure Appl. Math., 28 (1975), pp. 323–331.

[27] T. Nagylaki, *The geographical structure of populations*, Studies in Mathematics, Vol. 16: Studies in Mathematical Biology, Part II, S. A. Levin, ed., Mathematical Association of America, Washington, 1978, pp. 588–623.

[28] J. P. Pauwelussen and L. A. Peletier, *Clines in the presence of asymmetric migration*, J. Math. Biol. 11 (1981), pp. 207–233.

[29] F. Rothe, *Convergence to travelling fronts in semilinear parabolic equations*, Proc. Roy. Soc. Edinburgh, 80A (1978), pp. 213–234.

[30] ———, *Convergence to pushed fronts*, Rocky Mountain J. Math., 11 (1981), pp. 617–633.

[31] D. H. Sattinger, *On the stability of waves of nonlinear parabolic systems*, Adv. Math., 22 (1976), pp. 312–355.

[32] ———, *Weighted norms for the stability of travelling waves*, J. Differential Equations, 25 (1977), pp. 130–144.

[33] K. Uchiyama, *The behavior of solutions of some nonlinear diffusion equations for large time*, J. Math. Kyoto U., 18 (1978), pp. 453–508.

[34] E. J. M. Veiling, *Travelling waves in an initial-boundary value problem*, Proc. Royal Soc. Edinburgh, 90A (1981), pp. 41–61.

[35] ———, *Pushed travelling waves in an initial-boundary value problem for Fisher type equations*, Nonlinear Anal. TMA., 6 (1982), pp. 1271–1286.

[36] H. F. Weinberger, *Asymptotic behavior of a model in population genetics*, in Nonlinear Partial Differential Equations and Applications, J. Chadam, ed., Lecture Notes in Mathematics 648, Springer, New York, 1978, pp. 47–98.

[37] ———, *Some deterministic models for the spread of genetic and other alterations*, Proc. Conference on Models of Biological Growth and Spread–Mathematical Theories and Applications, P. Tautu, ed., Lecture Notes in Biomathematics 38, Springer, New York, 1981.

[38] ———, *Long-time behavior of a class of biological models*, this Journal, 13 (1982), pp. 353–396.

[39] K. Yosida, *Quasi-completely-continuous linear functional operators*, Japan J. Math., 15 (1939), pp. 297–301.

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

006

TRAVELING WAVE SOLUTIONS TO COMBUSTION MODELS AND THEIR SINGULAR LIMITS*

HENRI BERESTYCKI[†], BASIL NICOLAENKO[‡] AND BRUNO SCHEURER[§]

Abstract. We consider the deflagration wave problem for a compressible reacting gas, with species involved in a single step chemical reaction. In the limit of small Mach numbers, the one-dimensional traveling wave problem reduces to a system of reaction-diffusion equations. Existence is proved by first considering the problem in a bounded domain, and taking an infinite domain limit. In the singular limit of high activation energy within the Arrhenius exponential reaction term, we prove strong convergence to a limiting free boundary problem; the latter is characterized by a jump of the derivatives, which we determine.

Introduction. We consider the deflagration wave problem for a compressible reacting gas, with one reactant involved in a single step chemical reaction. In the limit of small Mach numbers, the one-dimensional traveling wave problem reduces to a system of two reaction-diffusion equations (cf. §1 for review of the basic flame equations). We assume both heat conductivity and diffusion coefficients are temperature-dependent. For the sake of clarity, we will first develop our methods on a simpler scalar case (corresponding to a Lewis number equal to one). The renormalized model is:

$$\begin{aligned} &-(k(u)u')'+cu'=g(u) \quad \text{on } \mathbb{R}, \\ &u(-\infty)=0, \qquad u(+\infty)=1 \end{aligned} \tag{0.1}$$

where u is the renormalized temperature, $0 \leqq u \leqq 1$; $k(u)$ is a C^1 function of u, which is strictly positive; $g(u)$ is a renormalized reaction term such that $g(u) \equiv 0$ on $[0, \theta)$ and $g(u) > 0$ on $(\theta, 1)$ where θ is some ignition temperature $(0 < \theta < 1)$. Moreover, $g(1) = 0$. c is the unknown mass flux of the wave. Next, we investigate the system

$$\begin{aligned} &-u''+cu'=f(u)v \quad \text{on } \mathbb{R}, \\ &-\Lambda v''+cv'=-f(u)v \quad \text{on } \mathbb{R}, \\ &u(-\infty)=0, \qquad u(+\infty)=1, \\ &v(-\infty)=1, \qquad v(+\infty)=0, \end{aligned} \tag{0.2}$$

where u is again the renormalized temperature $(0 \leqq u \leqq 1)$, v is a renormalized reactant concentration $(0 \leqq v \leqq 1)$, and $f(u)$ has the same properties of $g(u)$, except that now $f(1) > 0$. In (0.2), Λ is taken to be constant for simplicity.

The above equations (0.1), (0.2) are nonlinear eigenvalue problems for c. Existence for the systems is proved by first considering the problem on a bounded domain. This allows the reduction of the corresponding problem to a fixed point formulation w.r.t.

*Received by the editors March 28, 1984. This research was supported by the Theoretical Division and the Center for Nonlinear Studies, Los Alamos National Laboratory. Work was also performed under the auspices of the U. S. Department of Energy under contracts W-7405-ENG-36 and KC-04-02-01, Office of Basic and Engineering Sciences, Division of Applied Mathematical Sciences.

[†] Université Paris-Nord, Avenue J. B. Clement, 93430 Villetaneuse, France. Part of this research was performed while this author visited the Center for Nonlinear Studies at Los Alamos National Laboratory.

[‡] Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

[§] Centre d'Etudes de Limeil, B. P. 27, 91490 Villeneuve St. Georges, France. Part of this research was performed while this author visited the Center for Nonlinear Studies at Los Alamos National Laboratory.

the triplet (u, v, c). Then the usual Leray–Schauder degree gives the existence in a *bounded domain*. Taking an infinite domain limit completes the proof. Obtaining strictly positive lower and upper bounds for c, independent of the size of the domain, is essential for the above.

A considerable amount of work has been performed on the above equations in the asymptotic limit of infinite activation energy in the Arrhenius reaction term (see [4] and the bibliography there). Equivalently, f in (0.2) is now allowed to depend on ε (proportional to the reciprocal of the activation energy); $f_\varepsilon(u)v$, in (0.2), formally behaves as $\delta(u-1)$ when $\varepsilon \to 0$, where δ is the Dirac distribution centered at zero. Such formal asymptotic limits have not been rigorously established from a mathematical point of view. In this paper, we prove strong convergence of the traveling waves, both for (0.1) and (0.2), to singular limit free boundary solutions, with discontinuous derivatives. Again, obtaining strictly positive lower and upper bounds for c independent of ε, is crucial for this analysis. The plan of the paper is as follows.

The main theorems of the paper are summarized in §2.

1. The basic equations of the premixed one-dimensional laminar flame.

1.1. Reactive flow equations in one dimension. We summarize the equations of a chemically reacting mixture; this is essentially a compressible, heat-conducting viscous fluid with the added complexity of species diffusion and source terms representing the chemical reaction [4], [7]. Let ρ be the total mass density, T the temperature, p the hydrostatic pressure, and $\mathbf{v}$ the mass-average velocity. The reacting mixture is considered to be made of N fluids whose separate densities are ρY_i $(i=1,2,\cdots,N)$; here the Y_i are mass fractions of species i, with molecular mass m_i:

$$\sum_{i=1}^{N} Y_i = 1. \tag{1.1}$$

In what follows, we will investigate in detail the case of a single reactant A which yields a global product P, in a one-dimensional geometry:

$$A \to P. \tag{1.2}$$

The total mass density satisfies the equation

$$\frac{D\rho}{Dt} + \rho\frac{\partial v}{\partial x} = 0, \tag{1.3}$$

where $D/Dt \equiv \partial/\partial t + v\partial/\partial x$ is the convective derivative. The balance law for momentum is the same as for nonreactive flows, that is,

$$\rho\frac{Dv}{Dt}+\frac{\partial p}{\partial x}=\frac{4}{3}\frac{\partial}{\partial x}\left(\kappa\frac{\partial v}{\partial x}\right), \tag{1.4}$$

where κ is the dynamic viscosity coefficient and external forces on the mixture have been neglected. Energy conservation is expressed by

$$\rho c_p\frac{DT}{Dt}-\frac{\partial}{\partial x}\left(\lambda\frac{\partial T}{\partial x}\right)=Q\omega+\frac{Dp}{Dt}+\frac{4}{3}\kappa\left(\frac{\partial v}{\partial x}\right)^2, \tag{1.5}$$

where c_p is the specific heat at constant pressure, $\lambda(T)$ the coefficient of thermal conductivity, Q the chemical heat release of the reaction (1.2), ω measures the rate at which reaction (1.2) is proceeding:

$$\omega=B(T)\frac{\rho Y}{m}\exp\left(-\frac{E}{RT}\right), \tag{1.6}$$

where E is the activation energy of the reaction (a constant); in some sense, E/R is the temperature below which the reaction is negligible; R is the perfect gas constant, and $B(T)$ has a weak dependence on T. Equation (1.6) encompasses both the law of mass action and Arrhenius kinetics [4], [17].

The continuity equation for the mass fraction Y of reactant A with molecular mass m is

$$\rho\frac{DY}{Dt}-\frac{\partial}{\partial x}\left(\mu\frac{\partial Y}{\partial x}\right)=-m\omega, \tag{1.7}$$

where $\mu(T,Y)$ is the diffusion coefficient. From (1.1), the mass concentration of the combustion product P is $1-Y$, which enables its elimination. Finally, the equation of state for a perfect gas yields a supplementary constitutive law:

$$p=R\rho T. \tag{1.8}$$

1.2. Classical approximation of combustion. A flame is a low-speed wave whose Mach number M_0 is $\ll 1$. As a consequence (see [4], [19] for details) the variations in pressure are also small, i.e.,

$$p=p_c+\delta p, \qquad \delta p=O(M_0^2),$$

so that we may set

$$p=p_c=\text{const}$$

everywhere except in the momentum equation (1.4). The momentum equation now only controls small "flow-induced" variations δp in pressure [4] and uncouples from the remaining equations. Similarly, in the energy equation (1.5), one can neglect $Dp/Dt \equiv D\delta p/Dt$ and $(\partial v/\partial x)^2=O(M_0^2)$. Finally, in the combustion approximation, (1.3)–(1.7) reduce to

$$
\begin{aligned}
&\frac{D\rho}{Dt}+\rho\frac{\partial v}{\partial x}=0,\\
(1.9)\qquad &\rho c_p\frac{DT}{Dt}-\frac{\partial}{\partial x}\left(\lambda\frac{\partial T}{\partial x}\right)=+Q\omega,\\
&\rho\frac{DY}{DT}-\frac{\partial}{\partial x}\left(\mu\frac{\partial}{\partial x}Y\right)=-m\omega,
\end{aligned}
$$

together with the equation of state (1.8). Note that the combustion approximation does not imply a constant density approximation in one dimension.

1.3. The flame front equations for a single reactant. To study a one-dimensional flame moving with constant velocity $V_0>0$ to the left, it is appropriate to write (1.9) in the frame of reference of an observer moving at this same speed. Let $\xi=x+V_0t$ be the observer's space variable, and "prime" denote the differentiation with respect to ξ, then (1.9) becomes:

$$(1.10a)\qquad V_0\rho'+(\rho v)'=0,$$

$$(1.10b)\qquad c_p(\rho V_0+\rho v)T'-(\lambda T')'=Q\omega,$$

$$(1.10c)\qquad (\rho V_0+\rho v)Y'-(\mu Y')'=-m\omega.$$

Equation (1.10a) integrates to yield

$$(1.11a)\qquad \rho(V_0+v)=\text{const}\equiv c,$$

where c is the *mass flux*; then the energy and species equations uncouple from the rest of the system:

$$(1.11b)\qquad cc_pT'-(\lambda T')'=+Q\omega,$$

$$(1.11c)\qquad cY'-(\mu Y')'=m\omega.$$

The following boundary conditions apply to (1.11): at $\xi=-\infty$, the mixture is cold and unburned:

$$(1.12a)\qquad T=T(-\infty),\qquad Y=Y(-\infty);$$

at $\xi=+\infty$ the reactant is burned out and

$$(1.12b)\qquad Y(+\infty)=0,$$

$T(+\infty)$ is determined through Rankine–Hugoniot-like conditions obtained by integrating (1.11) from $-\infty$ to $+\infty$ and using (1.2a–b):

$$(1.12c)\qquad T(+\infty)=T(-\infty)+\frac{Q}{c_pm}Y(-\infty),$$

where we have assumed c_p *independent from* T.

Starting in §7 of the paper, we shall consider a renormalized version of (1.11–1.12):

$$(1.13)\qquad cu'-u''=vf(u),\qquad cv'-\Lambda v''=-vf(u)$$

where u (resp. v) is the renormalized temperature (resp. concentration of reactant):

$$u=\frac{T-T(-\infty)}{T(+\infty)-T(-\infty)},\qquad v=\frac{Y}{Y(-\infty)},$$

$f(u)$ is the renormalized formulation of ω/Y. For simplification, λ,μ and c_p are assumed constant. The renormalized boundary conditions are

$$\begin{aligned} u(-\infty)&=0, & v(-\infty)&=1,\\ u(+\infty)&=1, & v(+\infty)&=0. \end{aligned}$$

1.4. The case of Lewis number equal to 1. Consider again the (unnormalized) single reactant system (1.11). If the Lewis number

$$\text{(1.14)}\qquad \mathrm{Le}=\frac{\lambda}{\mu c_p}$$

is such that $\mathrm{Le}\equiv 1$, $\forall T$, we can eliminate the concentration Y, through the use of the "Shvab–Zeldovich variable" [4]

$$\text{(1.15)}\qquad Z=Q\frac{Y}{m}+c_pT$$

Z satisfies a purely convective-diffusive equation

$$\text{(1.16)}\qquad cZ'-(\mu Z')'=0,$$

whose only bounded solution is

$$\text{(1.17a)}\qquad Z=\text{const}=Z(+\infty);$$

since the reactant Y is depleted at $+\infty$, $Y(+\infty)=0$ and

$$\text{(1.17b)}\qquad Z\equiv c_pT(+\infty),$$

where $T(+\infty)$ is the adiabatic flame temperature

$$\text{(1.18)}\qquad T(+\infty)\equiv\frac{Q}{c_p}\frac{Y(-\infty)}{m}+T(-\infty).$$

Injecting the identity

$$\text{(1.19)}\qquad Y\equiv\frac{mc_p}{Q}(T(+\infty)-T)$$

into (1.11b), we obtain the reduced scalar equation

$$\text{(1.20)}\qquad cT'-\left(\frac{\lambda}{c_p}T'\right)'=\rho B(T)(T(+\infty)-T)\exp\left(-\frac{E}{RT}\right).$$

In §§3–6 we will investigate the normalized version of (1.20)

$$\text{(1.21)}\qquad cu'-(k(u)u')=g(u),$$

where $g(u)\equiv(1-u)f(u)$, and $f(u)$ is defined in (1.13).

1.5. Considerations of the ignition temperature. Consider the flame equations (1.11). If we fix boundary conditions at $\xi = -\infty$, i.e., $T(-\infty)$ and $Y(-\infty)$, then the problem is improperly posed, since

$$\exp\left(\frac{-E}{RT(-\infty)}\right) \neq 0$$

and there exist no bounded solutions to (1.11). The origin of the difficulty is clear: this formulation requires the mixture to react all the way in from $\xi = -\infty$, so that by the time finite ξ is reached, the combustion would be complete [4]; this is the "cold-boundary" difficulty, on which a considerable amount of ingenuity has been spent [10], [17], [19]. To resolve it, we classically modify the reaction term ω through the introduction of an *ignition temperature* θ^c, such that $\omega \equiv 0$, $\forall T < \theta^c$; this is equivalent to replacing ω by $H(T-\theta^c)\omega$, where H is the Heaviside function. It has been proven [9] that if one takes a sequence θ_i^c such that $\theta_i^c \to T(-\infty)$, c_i converges to some limit c_∞ from below. However, this lack of universal significance for c disappears as the activation energy E becomes large; the high activation energy asymptotics investigated in §§5 and 9 completely circumvent this difficulty by yielding unique limiting formulas for c independent of θ^c [4].

Moreover, the results of [10] have recently been (see [14]) well defined and extended to the full system (1.13), without ignition cut-off temperature. There, the limit c_∞ from [10] is shown to be the universal limit of sequences c_a obtained by considering (1.13) on *finite* truncated domains $[-a, +a]$; in some sense c_∞ is the unique numerically stable limit.

1.6. The high activation energy limit. A remarkable limit in (0.1)–(0.2), (1.5)–(1.7) is the asymptotic limit of infinite activation energy in the Arrhenius term (1.6). Specifically, one defines a small parameter as the reciprocal of the Zeldovich number [19]:

$$\frac{1}{\varepsilon} = \frac{E}{RT(+\infty)} \frac{T(+\infty) - T(-\infty)}{T(+\infty)} \gg 1. \tag{1.22}$$

Then, in terms of ε and the renormalized temperature and concentration u and v, we can transform the exponential term in (1.6), as:

$$\begin{aligned} \exp - \frac{E}{RT} &= \exp\left(-\frac{E}{RT(+\infty)}\right) \cdot \exp\left(\frac{E}{R}\left(\frac{1}{T(+\infty)} - \frac{1}{T}\right)\right) \\ &= \exp\left(-\frac{E}{RT(+\infty)}\right) \cdot \exp\left(\left(\frac{u-1}{\varepsilon} \frac{1}{1+\varepsilon\alpha(u-1)/\varepsilon}\right)\right) \end{aligned} \tag{1.23}$$

where $\alpha \equiv (T(+\infty) - T(-\infty))/T(+\infty)$ is the thermal expansion coefficient. The expression (1.23) is the starting point for the abstract setting leading to the rigorous asymptotic analysis of §§5 and 9. As $\varepsilon \to 0$, the exponential is very small, except for u such that $u - 1 = O(\varepsilon)$.

2. Main results and summary. First, we study the scalar equation (0.1), where c is an unknown nonlinear eigenvalue. Many earlier works have assessed the question of existence for problems of the type (0.1). A good survey of the question with a discussion of the literature can be found in the articles of P. C. Fife [5], [6] and the monograph by J. Smoller [15]. In the context of population genetics, the paper of D. G. Aronson and H. F. Weinberger [1] is also relevant. The works of Ya. I Kanel' [12], [13] and Ya. Zeldovich [18], [19] are specifically devoted to combustion with an ignition cut-off hypothesis. All these papers use a "phase plane" approach to solve (0.1). Here we use a more general analytical technique, namely a shooting method, together with more general hypothesis on the reaction term $g(u)$. In particular, we do not assume any differentiability on $g(u)$, but only piecewise Lipschitz continuity.

The main existence and uniqueness result in Theorem 3.1 of §3, which also contains the full shooting argument. Uniqueness is established up to a translation of the origin (§4) via hodograph transformation. In §5, we address the question of asymptotic analysis as ε goes to 0 (see (1.22)); under abstract conditions (5.1)–(5.3) on $g_\varepsilon(u)$, we prove the main Theorem 5.2. The main step in the proof consists in using energy estimates; this approach is very flexible and extends to more general situations, especially to higher order systems of complex chemical reactions. We show that the limit problem associated with (0.1) is the free boundary problem

$$\begin{gathered} -(k(u)u')' + c_0 u' = c_0 \delta_{x=\bar{x}}, \\ u(-\infty)=0, \quad u(+\infty)=1, \quad u(0)=\theta \end{gathered}$$

where $\delta_{x=\bar{x}}$ is the Dirac measure at the point $\bar{x}$ and $\bar{x}$ is uniquely determined by the condition $u(0)=\theta$ (the latter removes the translational invariance). In §6, we conclude the scalar case study with some remarks on approximating (0.1) by a problem on a finite truncated domain $[-a, +a]$. Second, we study similar questions for the system (0.2). To our knowledge, the only work in this direction is that of Ya. I. Kanel' [13], who uses formal phase plane arguments. Here, in §7, following the point of view of §6, we approximate the problem through a truncated domain $[-a, +a]$ reducing translational invariance by fixing $u(0)=\theta$. This closely follows actual numerical steady state schemes currently developed for the computation of traveling waves in systems with complex chemistry [16]. The general energy estimates, which we develop, enable us to demonstrate the existence Theorem 8.11 of §8; the latter is, in fact, a numerical scheme convergence result. We combine the energy estimates with a general comparison result (Proposition 8.4) which bounds the concentration v from above and below in terms of $1-u$. The above tools are also essential in studying the asymptotic limit $\varepsilon \to 0$ for system (0.2) in §9. Indeed, they enable us to show that the sequence c_ε is bounded from above and below: this singular perturbation problem is definitely not of a classical type. We prove the equivalent of Theorem 5.2, that is, Theorem 9.4. Now, the limit free boundary problem is

$$\begin{aligned} &-u'' + c_0 u' = c_0 \delta_{x=\bar{x}}, \\ &-\Lambda u'' + c_0 u' = -c_0 \delta_{x=\bar{x}}, \\ &u(-\infty)=0, \quad u(+\infty)=1, \quad u(0)=\theta \\ &v(-\infty)=1, \quad v(+\infty)=0, \end{aligned}$$

where $\bar{x}$ is defined as before.

In §10, we extend our results to the more general case of a single step, nth order reaction; there $vf(u)$ in (0.2) is replaced by $v^n f(u)$, $n>0$. In §11, we give (Theorem 11.1) the exact formula for the unique limiting c_0 in the system case. In §5, the limiting c_0 for the scalar case was automatically obtained through sharp a priori estimates. For the system case, the latter gives only a permissible band for c_0. Then, we need to establish rigorously the existence of a local Shvab–Zeldovich variable $u + \Lambda u - 1$ inside the reaction zone (see §1.4).

We hope to use the techniques developed in this paper to investigate deflagration waves with complex chemical networks [5]–[9].

3. Existence of a solution in the scalar case. In this first part, up to §7, we study the following problem:

$$\text{(3.1)}\qquad \begin{array}{l}\text{To find a function } u\colon \mathbb{R}\to[0,1] \text{ and } c\in\mathbb{R} \text{ satisfying}\\ -(k(u)u')'+cu'=g(u) \text{ in } \mathbb{R},\quad u(-\infty)=0,\quad u(+\infty)=1.\end{array}$$

We recall that this equation corresponds to the model of premixed laminar flames with a single reaction $A\to B$, in the particular case that the Lewis number $\Lambda=1$. Since there is no physical ground to make such an assumption, it should be kept in mind that we study (3.1) as a model for the more complex systems to be investigated in the forthcoming sections of this paper. Let us emphasize that the number c is an *unknown* of the problem.

The function g verifies the following hypotheses:

(3.2) $g\colon [0,1]\to\mathbb{R}_+$, (i.e., $g\geqq 0$) and $g(1)=0$;

(3.3) there is some $\theta\in(0,1)$ such that $g\equiv 0$ on $[0,\theta)$ and $g>0$ on $(\theta,1)$;

(3.4) g is Lipschitz continuous on $[\theta,1]$.

We recall that whenever g is strictly positive on $(\theta,1)$ θ represents a (reduced) *ignition temperature*. $k(u)$ is a (possibly) nonlinear diffusion coefficient. It will be assumed that it verifies:

$$\text{(3.5)}\qquad k\colon [0,1]\to\mathbb{R} \text{ is a } C^1\text{-function},\qquad k(s)\geqq\alpha>0\quad \forall s\in[0,1].$$

The smoothness assumption on k could be weakened but we impose it here as we are looking for classical solutions. Note, however, that since g is allowed to be *discontinuous* at $u=\theta$, u is not necessarily of class C^2 at the points where $u=\theta$.

In this and the next section we will prove the following

THEOREM 3.1. *Under conditions* (3.2)–(3.5), *there exists a solution* $u\colon\mathbb{R}\to[0,1]$ *and* $c>0$ *of problem* (3.1). *u is of class C^1, and of class C^2 on $\mathbb{R}-\{x_0\}$ for some x_0; u is monotone increasing on $\mathbb{R}$. Furthermore, u and c are uniquely determined from* (3.1) (*up to a translation of the origin*). *There are positive constants A and δ such that $0<u(x)\leqq Ae^{\delta x}$, $\forall x<0$. Lastly, if $g'(1)$ exists and $g'(1)<0$, then $0<1-u(x)\leqq Ae^{-\delta x}$, $\forall x>0$ for some positive constants A,δ.*

Remark 3.2. Existence results for problems of the type (3.1) have been obtained in many earlier works (see the references and the discussion of the literature in §2 above). Nevertheless, the hypotheses of Theorem 3.1 seem to be more general than those in all the works we know of. Moreover, the proof of existence and uniqueness which we present here are quite simple. To start with, let us extend g and k to be defined on $\mathbb{R}$ by setting

$$\text{(3.6)}\qquad g(s)=0\quad \forall s\leqq 0,\qquad g(s)=g(1)=0\quad \forall s\geqq 1,$$

$$\text{(3.7)}\qquad k(s)=k(0)\quad \forall s\leqq 0,\qquad k(s)=k(1)\quad \forall s\geqq 1.$$

Indeed, we wish to consider functions which do not necessarily verify $0\leqq u\leqq 1$. Observe that k verifies

$$\text{(3.8)}\qquad 0<\alpha\leqq k(s)\leqq\beta<\infty\quad \forall s\in\mathbb{R}.$$

In the sequel g and k will always be required to verify (3.2)–(3.8). By a solution to (3.1) we mean a pair (u,c) such that u is a C^1 function.

LEMMA 3.3. *Suppose u,c is a solution of* (3.1). *Then,* $c>0$, $0<u<1$ *and u is monotone increasing,* $u'>0$, *and* $\lim_{x\to\pm\infty}u'(x)=0$.

Proof of Lemma 3.3. Let $a<b$ and let us integrate the equation on $[a,b]$ to obtain

$$(3.9)\qquad -k(u(b))u'(b)+k(u(a))u'(a)+c(u(b)-u(a))=\int_a^b g(u(s))\,ds.$$

We take $a=0$ and let $b\to+\infty$ in (3.9). Since $g\geqq 0$, the integral $\int_0^{+\infty}g(u(s))\,ds$ converges (possibly to $+\infty$). Since $u(b)\to 1$ as $b\to+\infty$, (3.9) shows (using (3.8)) that $u'(b)$ has a limit and, hence, this limit must be 0 to ensure $u(+\infty)=1$. Similarly, by taking $b\to-\infty$, we show that

$$(3.10)\qquad \lim_{x\to\pm\infty}u'(x)=0.$$

Now, letting $b\to+\infty$ and $a\to-\infty$ in (3.9), we have

$$(3.11)\qquad c=\int_{-\infty}^{+\infty}g(u(s))\,ds.$$

Hence, c is positive. Let us now show that u is monotone increasing. We argue by contradiction. In view of the limits of u as $x\to\pm\infty$, we may assume, if u is not monotone increasing, that there exist $a<b$ such that $u(a)>u(b)$ with $u'(a)=u'(b)=0$. From (3.9), we then derive

$$(3.12)\qquad 0>c(u(b)-u(a))=\int_a^b g(u(s))\,ds\geqq 0,$$

a contradiction.

It remains to show that $u'>0$ on $\mathbb{R}$. Suppose that $x_0\in\mathbb{R}$ with $u'(x_0)=0$. Let us first assume $u(x_0)\neq\theta$. In this case from the equation we have $-k(u(x_0))u''(x_0)=g(u(x_0))$. Now, if $g(u(x_0))=0$, then by the uniqueness in the initial value problem, we would have $u(x)=u(x_0)$, $\forall x\in\mathbb{R}$, which is impossible. On the other hand, if $g(u(x_0))>0$, then $u''(x_0)<0$ which is also impossible for u is monotone. Suppose now that $u(x_0)=\theta$. This case is singled out as g may be discontinuous at θ. If $g(\theta)=0$, then g is Lipschitz continuous on $[0,1]$ and the above argument shows $u(x)=\theta$, $\forall x\in\mathbb{R}$. Therefore, we assume $g(\theta)>0$. In this case, since $u(x)>\theta$, $\forall x>x_0$, we know by letting $x\downarrow x_0$ in (3.1) that $u''(x_0+0)<0$, which is again a contradiction. □

COROLLARY 3.4. *Let u,c be a solution of* (3.1). *Then after a shift of the origin* (*if need be*) *u,c verify*

$$(3.13)\qquad -(k(u)u')'+cu'=g(u)\quad on\ \mathbb{R}_+,\qquad u(0)=\theta,\qquad u'(0)=c\theta k(\theta)^{-1}$$

Conversely, if u,c is a solution of (3.13), *then u can be extended on R in such a way that u,c is a solution to problem* (3.1).

Proof. If u is a solution of (3.1), then $u(x+a)$ is also a solution with the same c ((3.1) is translation invariant). Therefore, after translating the origin if need be, there is no loss in generality to assume that $u(0)=\theta$. We then know that $u(x)<\theta$ for $x<0$ while $u(x)>\theta$ for $x>0$. Hence, $-k(u)u'+cu$ is constant $\forall x\leqq 0$ which shows $u'(0)=c\theta k(\theta)^{-1}$. Conversely, let u,c be a solution of (3.13). Let us consider the backward initial value problem:

$$(3.14)\qquad \begin{aligned}&-k(v)v'+cv=0\quad \text{for } x\leqq 0,\\ &v(0)=\theta.\end{aligned}$$

It is straightforward to show that (3.14) has a unique solution defined on $\mathbb{R}_-$ satisfying $0<v<\theta$, $v'>0$ on $\mathbb{R}_-$ and $\lim_{x\to\infty} v(x)=0$. (It suffices to observe that if $x_0<0$ with $v(x_0)=0$, then $v'(x_0)=0$ and $v\equiv 0$. Hence, $0<v<\theta$ and $v'>0$. Since $\lim_{x\to-\infty} v'(x)=\lim_{v\to-\infty} v/k(v)$, we have $\lim_{x\to-\infty} v(x)=0$).

Now extending u by setting $u(x)=v(x)$ $\forall x\leqq 0$ we have a solution of (3.1). □

Remark 3.5. From (3.14) it is clear that

$$v(x)\leqq \theta e^{cx/\beta} \quad \forall x\leqq 0.$$

If $g'(1)$ exists and $g'(1)<0$ then it is classical that $1-u$ has exponential decay as $x\to\infty$. This is shown by using the "linearized equation" in $w=1-u$ as $x\to+\infty$, namely:

$$-k(1)w''+cw'-g'(1)w=0.$$

The proof will be omitted here. Note also that since $u=\theta$ only at $x=0$, u is of class C^2 on $\mathbb{R}-\{0\}$. □

In view of the preceding results it remains to show existence and uniqueness of a solution for (3.13). The proof of uniqueness will be delayed to §4. Existence of a solution is derived here by a shooting argument in the same spirit as the one used in H. Berestycki, P. L. Lions and L. A. Peletier [2].

We consider the initial value problem associated with (3.13):

$$\begin{aligned} &-(k(u)u')'+cu'=g(u), \qquad x\geqq 0,\\ &u(0)=\theta, \\ &u'(0)=c\theta k(\theta)^{-1}. \end{aligned} \tag{3.15}$$

Since g is bounded and by (3.2)–(3.8) it is clear that for any $c>0$, (3.15) has a unique solution defined on all of $\mathbb{R}_+$.

LEMMA 3.6. *Let $c>0$ and u be the solution of* (3.15). *Then*

i) *If $x_1>0$ with $u(x_1)=1$, then $u'(x_1)>0$ for $x>x_1$* (*whence $u>1$*).

ii) *If $x_1>0$ with $u(x_1)=\theta$, then $u'(x_1)<0$ and $u'<0$ for for $x>x_1$* (*whence $u<\theta$*).

Proof of i). If $u'(x_1)=0$, then since $g(1)=0$, from uniqueness for the initial value problem (IVP) we would have $u\equiv 1$, which is impossible. So let us assume that $u'(x_1)<0$. In this case, there exists $\bar{x}$, $0<\bar{x}<x_1$ with $u(\bar{x})=1$, $u'(\bar{x})>0$ and $u>1$ on $(\bar{x},x_1)$. In (3.9) we set $a=\bar{x}$, $b=x_1$ to get

$$-k(1)\left[u'(x_1)-u(\bar{x})\right]=0,$$

which is obviously impossible. Thus, $u'(x_1)>0$. The same type of argument using (3.9) shows that u' cannot vanish for $x\geqq x_1$.

Proof of ii). If $x_1>0$ with $u(x_1)=\theta$, there must exist $\bar{x}$, $0<\bar{x}<x_1$, with $u(\bar{x})>\theta$ and $u'(\bar{x})=0$. Set $a=\bar{x}$, $b=x_1$ in (3.9) to obtain

$$-k(\theta)u'(x_1)+c(\theta-u(x_1))\geqq 0, \tag{3.16}$$

which shows $u'(x_1)<0$. Should there exist $x_2>x_1$ with $u(x_2)<u(x_1)=\theta$ and $u'(x_2)=0$, we may assume $u(x)<\theta$, $\forall x\in(x_1,x_2]$, and we would then obtain from (3.9) (using (3.3) and (3.6)) that

$$k(\theta)u'(x_1)+c(u(x_2)-\theta)=0,$$

which is impossible. Hence $u'<0$, $\forall x\geqq x_1$. □

Remark 3.7. It is straightforward to show that in case i) $\lim_{x\to+\infty} u(x) = +\infty$ and in case ii) $\lim_{x\to+\infty} u(x) = -\infty$.

COROLLARY 3.7. *Let $c>0$ and u be a corresponding solution of* (3.15). *Suppose that there exists $x_0>0$ with $u'(x_0)=0$. Then $\theta<u(x_0)<1$ and there exists $x_1>x_0$ such that $u(x_1)=\theta$.*

Proof. From Lemma 3.6 we know that if $u'(x_0)=0$ then $\theta<u(x_0)<1$. Using (3.16) it is clear that $u'(x)<0$, $\forall x>x_0$.

Hence, if $u>\theta$ for all $x>x_0$, $l=\lim_{x\to+\infty} u(x)$ verifies $\theta\leqq l<u(x_0)$ and from (3.9) (where $a=x_0$ and $b\to+\infty$) we find that $\lim_{x\to+\infty} u'(x)=0$ and

$$c(l-u(x_0))=\int_{x_0}^{+\infty} g(u(s))\,ds\geqq 0,$$

which is impossible. □

For $c>0$, let us denote by u_c the solution to the IVP (3.15). We consider two subsets of $\mathbb{R}_+^*=\{c>0\}$:

$$\Gamma_+=\{c>0;\ x_1>0 \text{ with } u_c(x_1)=1\},$$
$$\Gamma_-=\{c>0;\ x_1>0 \text{ with } u_c(x_1)=\theta\}.$$

LEMMA 3.8. *Γ_+ and Γ_- are open disjoint subsets of $\mathbb{R}_+$.*

Proof. That $\Gamma_+\cap\Gamma_-=\varnothing$ follows from Lemma 3.6. The fact that Γ_+ and Γ_- are open sets is a consequence of the continuous dependence of u_c with respect to c (at least for the values of x such that $u(x)\geqq\theta$ since g might have a discontinuity at θ). Indeed, let $c\in\Gamma_+$. Then, $\exists x_1>0$ with $u_c(x_1)=1$. By Lemma 3.6 we know that $u_c'(x_1)>0$, hence $u_c(x_1+\varepsilon)>1$ for some $\varepsilon>0$. Therefore, for c' close to c we also have $u_{c'}(x_1+\varepsilon)>1$ which shows that $c'\in\Gamma_+$. The proof is similar for Γ_-. □

LEMMA 3.9. *Γ_+ is nonempty. More precisely, $[c_+,+\infty)\subset\Gamma_+$, where $c_+=(1/\theta)\{2\int_0^1 k(s)g(s)\,ds\}^{1/2}$.*

Proof. Let us set $K(z)=\int_0^z k(s)g(s)\,ds$. Multiply the equation by $k(u)u'$ and integrate by parts between $0\leqq a<b$. This yields

$$-\frac{1}{2}k(u(x))^2u'(x)^2\Big|_{x=a}^{x=b}+c\int_a^b k(u(s))u'(s)^2\,ds=K(u(x))\Big|_{x=a}^{x=b}. \tag{3.17}$$

Now let $c\geqq c_+$. Suppose $c\notin\Gamma_+$, that is $u_0(x)<1$, $\forall x>0$. Then, either $x_0>0$ with $u'(x_0)=0$ or $u'>0$, $\forall x>0$ and $l=\lim_{x\to+\infty} u(x)$ verifies $\theta<l\leqq 1$. In the first case, for $a=0$, $b=x_0$, (3.17) reads:

$$\frac{1}{2}c^2\theta^2+c\int_0^{x_0} k(u)u'^2\,dx=K(u(x_0))-K(\theta)\leqq K(1)$$

for $K(\theta)=0$ and K is nondecreasing. Therefore, we have $c^2\theta^2<2K(1)$, i.e., $c<c_+$. In the second case, when $u'>0$, $\forall x>0$, it is seen from (3.9) that $\lim_{x\to+\infty} u'(x)=0$. Set $a=0$ and $b\to+\infty$ in (3.17). We obtain

$$\frac{1}{2}c^2\theta^2<K(l)\leqq K(1),$$

that is, again, $c<c_+$.

Thus, for all $c\geqq c_+$, c must be in Γ_+. □

LEMMA 3.10. *Γ_- is nonempty. More precisely, $[0,c_-]\subset\Gamma_-$, where $c_-=\{2K(1)\}^{1/2}$.*

Proof. Let $c \leqq c_-$; if $c \notin \Gamma_-$, then by Lemma 3.6 and Corollary 3.7, we know that $u' > 0$ for all $x > 0$. There are two possibilities: either $\bar{x} > 0$ with $u(\bar{x}) = 1$ or $u < 1$, $\forall x > 0$. In the latter case, $\lim_{x \to +\infty} u(x) = l$ verifies $\theta < l \leqq -1$. In view of (3.9) we have $\lim_{x \to +\infty} u'(x) = 0$ and from the equation we derive

$$\lim_{x \to +\infty} u''(x) = -\frac{g(l)}{k(l)}.$$

Hence for u to remain bounded, we must have $g(l) = 0$, whence $l = 1$. In this situation we set $\bar{x} = +\infty$. Thus, if $c \notin \Gamma_-$ one has for $\bar{x}$, $0 < \bar{x} \leqq +\infty$:

$$(3.18) \qquad u' > 0 \quad \forall x \in [0, \bar{x}], \qquad u(\bar{x}) = 1, \qquad u'(\bar{x}) \geqq 0.$$

Now multiply the equation by u and integrate by parts on $[a, b]$:

$$(3.19) \qquad -k(u)uu'\Big|_{x=a}^{x=b} + \int_a^b k(u)u'^2\,dx + \frac{c}{2}u(x)^2\Big|_{x=a}^{x=b} = \int_a^b g(u)u\,dx.$$

Let us evaluate (3.9), (3.19), and (3.17) for $a = 0$ and $b = \bar{x}$. We obtain, respectively,

$$(3.20) \qquad -k(1)u'(\bar{x}) + c = \int_0^{\bar{x}} g(u)\,dx,$$

$$(3.21) \qquad -k(1)u'(\bar{x}) + \frac{c}{2}(1 + \theta^2) + \int_0^{\bar{x}} k(u)u'^2\,dx = \int_0^{\bar{x}} g(u)u\,dx$$

$$(3.22) \qquad -\frac{1}{2}(1)^2 u'(\bar{x})^2 + \frac{1}{2}c^2\theta^2 + c\int_0^{\bar{x}} k(u)u'^2\,dx = K(1).$$

Now we subtract (3.20) from (3.21) and we use that $u < 1$ on $(0, \bar{x})$ to obtain,

$$(3.23) \qquad \int_0^{\bar{x}} k(u)u'^2\,dx < \frac{c}{2}(1 - \theta^2).$$

Thus, combining (3.22) and (3.23) we derive

$$(3.24) \qquad K(1) < \frac{c^2}{2},$$

that is, $c > c_-$. Therefore, any $c \leqq c_-$ must belong to Γ_-. □

Conclusion. We are now ready to prove the existence of a solution to (3.13), whence by Corollary 3.4, the existence of a solution to (3.1). Since Γ_+ and Γ_- are nonempty disjoint open subsets of $(0, +\infty)$, there must exist a $c > 0$ such that

$$(3.25) \qquad c \notin \Gamma_+ \cup \Gamma_-.$$

Observe that such a c verifies the bounds

$$(3.26) \qquad c_- < c < c_+.$$

Now if $u = u_c$ denotes the corresponding solution of (3.15), we know that $u' > 0$ and $\theta < u < 1$, $\forall x > 0$. We have already seen that this implies $\lim_{x \uparrow +\infty} u(x) = 1$. Clearly u is a solution of (3.13), hence of (3.1).

Therefore, apart from the uniqueness, the proof of Theorem 3.1 is complete. □

4. Uniqueness of the solution in the scalar case. We will now show that the solution u, c of problem (3.1) is unique. In view of Corollary 3.4, it suffices to show that the solution to problem (3.13) is unique. Notice that for this range of values of u, that is $[\theta, 1)$, g is locally Lipschitz, whence there is a uniqueness for the IVP.

Let u, c be a solution of (3.13). Since u is increasing, we may define a function $x(s)=u^{-1}(s)$ $(u(x(s))=s)$. Then, x: $[\theta,1)\to\mathbb{R}_+$ is continuous, increasing and verifies $x(\theta)=0$, $x(1)=+\infty$.

Set $dx(s)/ds=z(s)$. The function z: $[\theta,1)\to\mathbb{R}_+$ is continuous, and it verifies

$$\frac{1}{z(\theta)}=u'(0)=\frac{c\theta}{k(\theta)},\quad z(1)=+\infty,\qquad z(s)>0,\quad \forall s\in[\theta,1). \tag{4.1}$$

The equation (3.1) translates into

$$-\frac{d}{ds}\left(\frac{k(s)}{z(s)}\right)+c=g(s)z(s). \tag{4.2}$$

Now, suppose there are two solutions (c_1,u_1), (c_2,u_2) of problem (3.13). We denote by $x_i(s)=u_i^{-1}(s)$ and $z_i(s)=dx_i(s)/ds$, the corresponding functions defined above.

By the uniqueness to the IVP it suffices to show that $c_1=c_2$. So let us argue by contradiction and assume that $c_1<c_2$. Then, by (4.1), one has $z_1(\theta)>z_2(\theta)$, whence by continuity for some $\sigma>\theta$:

$$z_1(s)>z_2(s)\quad \forall s\in[\theta,\sigma). \tag{4.3}$$

We claim that $\sigma=1$. Indeed, if not, $\theta<\sigma<1$ and

$$z_1(\sigma)=z_2(\sigma). \tag{4.4}$$

Then, using the difference of the equations (4.2) for z_1 and z_2, respectively, we obtain

$$\frac{d}{ds}\left[\frac{k}{z_1}-\frac{k}{z_2}\right]\Bigg|_{s=\sigma}=c_2-c_1>0. \tag{4.5}$$

That is, letting $w=k/z_1-k/z_2$ gives $w'(\sigma)<0$. This, however, is impossible since by (4.3), (4.4), w verifies $w<0$ for $s<\sigma$ and $w(\sigma)=0$. Therefore, one has $\sigma=1$.

Let us now integrate the equations (4.2) for z_1 and z_2, respectively, over $[\theta,1)$. This yields

$$c_1-c_2=\int_\theta^1 g(s)(z_1(s)-z_2(s))\,ds. \tag{4.6}$$

This is an obvious contradiction since $c_1<c_2$ and $z_1>z_2$, and the condition (3.3) on g.

The proof of uniqueness is thereby complete, which concludes the proof of Theorem 3.1. □

Remark 4.1. The preceding uniqueness theorem extends a result of Kanel' [12] by allowing more general hypotheses on g and a variable diffusion coefficient. Furthermore, the proof above is much simpler than the one in [12], [18], [19].

5. Asymptotic analysis for large activation energy. In this section, we assume that g depends on a parameter $\varepsilon>0$ and we let $g=g_\varepsilon$. The corresponding (unique) solution of (3.1) will be denoted by u_ε, c_ε. We set

$$K_\varepsilon(z)=\int_0^z k(s)g_\varepsilon(s)\,ds.$$

The following conditions will be assumed here on the family of functions g_ε:

(5.1) For each $\varepsilon>0$, g_ε verifies conditions (3.2)–(3.4) with a fixed $\theta\in(0,1)$;

(5.2) $\exists\theta_\varepsilon, \theta\leqq\theta_\varepsilon<1$ such that $\lim\limits_{\varepsilon\downarrow 0}\uparrow\theta_\varepsilon=1$ and $\lim\limits_{\varepsilon\downarrow 0}\max\limits_{u\in[\theta,\theta_\varepsilon]} g_\varepsilon(u)=0$;

(5.3) $\lim\limits_{\varepsilon\downarrow 0} K_\varepsilon(1)=m>0, m<+\infty$.

Remark 5.1. The typical example in the applications is the function $g_\varepsilon(u)=f_\varepsilon(u)(1-u)$ with $f_\varepsilon(u)=0$ for $0\leqq u<\theta$ and $f_\varepsilon(u)=(1/\varepsilon^2)\phi((u-1)/\varepsilon,\varepsilon)$ for $\theta\leqq u\leqq 1$, where ϕ is some fixed positive function satisfying

$$\lim_{\varepsilon\downarrow 0}\int_{-\infty}^{0} -\sigma\phi(\sigma,\varepsilon)\,d\sigma<\infty$$

and, for instance, $\lim_{y\to-\infty}|y|^p\phi(y,0)=0$, with $p>2$ and ϕ is nondecreasing near $-\infty$. Notice that (5.1)–(5.3) are satisfied in this case. The prototype of such a function ϕ is the exponential $\phi(\sigma)=e^\sigma$ (cf. §1.6). This gives rise to the Arrhenius term (3.1). In scaled variables, the parameter ε repesents the inverse of an activation energy E. In this section we are deriving in a rigorous fashion the asymptotic limits as E becomes infinite.

Our main result is the following:

THEOREM 5.2. *Under conditions* (5.1)–(5.3), *the unique solution* u_ε, c_ε *to* (3.1) *such that* $u_\varepsilon(0)=\theta$ *has the following behavior when* $\varepsilon\downarrow 0$:

$$\lim_{\varepsilon\downarrow 0} c_\varepsilon=\sqrt{2m}\equiv c_0,$$

and u_ε *converges to* u_0 *in the sense that*

$$\max_{x\in\mathbb{R}}|u_\varepsilon(x)-u_0(x)|\to 0 \quad \text{as } \varepsilon\to 0,$$

$$\|u_\varepsilon-u_0\|_{H^1(\mathbb{R})}\to 0 \quad \text{as } \varepsilon\to 0,$$

where u_0 *is uniquely determined by* $u_0(x)=1$, $\forall x\geqq\bar{x}$,

$$-k(u_0)u_0'+c_0u_0=0 \quad \forall x\leqq\bar{x},$$

and $\bar{x}$ *is uniquely determined by the condition* $u_0(0)=\theta$.

Proof. First, from §3, by (3.26), we know that c_ε is bounded from above and from below (away from zero) independently of ε:

$$0<\{2K_\varepsilon(1)\}^{1/2}<c_\varepsilon<\frac{1}{\theta}\{2K_\varepsilon(1)\}^{1/2} \quad \text{and} \quad K_\varepsilon(1)\to m>0 \quad \text{as } \varepsilon\downarrow 0. \tag{5.4}$$

From the relations (3.9), (3.19) and (3.17) respectively, for each $\varepsilon>0$, for $a=0$, $b=x$, we obtain

$$-k(u_\varepsilon(x))u_\varepsilon'(x)+c_\varepsilon u_\varepsilon(x)=\int_0^x g_\varepsilon(u_\varepsilon)\,dx, \tag{5.5}$$

$$-k(u_\varepsilon(x))u_\varepsilon(x)u_\varepsilon'(x)+\frac{c_\varepsilon}{2}\left(u_\varepsilon(x)^2+\theta^2\right)+\int_0^x k(u_\varepsilon)u_\varepsilon'^2\,dx=\int_0^x g_\varepsilon(u_\varepsilon)u_\varepsilon\,dx, \tag{5.6}$$

$$-\frac{1}{2}k(u_\varepsilon(x))^2u_\varepsilon'(x)^2+\frac{1}{2}c_\varepsilon^2\theta^2+c_\varepsilon\int_0^x k(u_\varepsilon)u_\varepsilon'^2\,dx=K_\varepsilon(u_\varepsilon(x)). \tag{5.7}$$

For each $\varepsilon>0$, we let $x_\varepsilon>0$ be defined by $u_\varepsilon(x_\varepsilon)=\theta_\varepsilon$. From (5.7) we get, for all x, $0<x\leqq x_\varepsilon$,

$$\frac{\beta^2}{2}u_\varepsilon'(x)^2+\frac{1}{2}c_\varepsilon^2\theta^2\leqq K_\varepsilon(\theta_\varepsilon) \tag{5.8}$$

for K_ε is nondecreasing. Using the lower bound on c_ε in (5.4), we see from (5.8) that

$$K_\varepsilon(1)\theta^2-K_\varepsilon(\theta_\varepsilon)\leqq\frac{\beta^2}{2}u_\varepsilon'(x)^2. \tag{5.9}$$

The hypotheses on g_ε, (5.2) and (5.3), imply that $K_\varepsilon(\theta_\varepsilon)\to 0$ and $K_\varepsilon(1)\to m>0$ as $\varepsilon\to 0$. Hence, for ε small enough, say $0<\varepsilon\leqq\varepsilon_0$, (5.9) shows that there is a constant $\alpha>0$ with

$$u_\varepsilon'(x)\geqq\alpha>0\quad\forall x\in[0,x_\varepsilon]. \tag{5.10}$$

Therefore, for all ε, $0<\varepsilon\leqq\varepsilon_0$,

$$1-\theta\geqq u_\varepsilon(x_\varepsilon)-u_\varepsilon(0)=\int_0^{x_\varepsilon}u_\varepsilon'(s)\,ds\geqq\alpha x_\varepsilon. \tag{5.11}$$

This shows that x_ε is bounded from above by $(1-\theta)\alpha^{-1}$ independently of ε ($\varepsilon\leqq\varepsilon_0$). Hence, by (5.2):

$$0<\int_0^{x_\varepsilon}g_\varepsilon(u_\varepsilon(s))\,ds\leqq(1-\theta)\alpha^{-1}\max_{u\in[0,\theta_\varepsilon]}g_\varepsilon(u)\to 0\quad\text{as }\varepsilon\to 0. \tag{5.12}$$

Then (5.5), read at the point $x=x_\varepsilon$, implies

$$\lim_{\varepsilon\downarrow 0}\{k(\theta_\varepsilon)u_\varepsilon'(x_\varepsilon)-c_\varepsilon\theta_\varepsilon\}=0. \tag{5.13}$$

From (3.17) written at $a=x_\varepsilon$ and $b\to+\infty$ we know that

$$\frac{1}{2}k(\theta_\varepsilon)^2u_\varepsilon'(x_\varepsilon)^2\leqq K_\varepsilon(1)-K_\varepsilon(\theta_\varepsilon). \tag{5.14}$$

Thus, combining (5.13) and (5.14) yields

$$\overline{\lim_{\varepsilon\downarrow 0}}\,c_\varepsilon\leqq\sqrt{2m}\,. \tag{5.15}$$

On the other hand, (5.4) shows that

$$\underline{\lim_{\varepsilon\downarrow 0}}\,c_\varepsilon\geqq\sqrt{2m}\,, \tag{5.16}$$

and therefore

$$\lim_{\varepsilon\downarrow 0}c_\varepsilon=\sqrt{2m}\equiv c_0. \tag{5.17}$$

It is straightforward to show from the equation (see (3.14)) that on $\mathbb{R}_-$, u_ε converge in the $C^2(\mathbb{R}_-)$ and $H^2(\mathbb{R}_-)$ topologies to the unique solution u_0 of the equation

$$\begin{gathered}-(k(u_0)u_0')+c_0u_0'=0\quad\text{on }\mathbb{R}_-,\\ u_0(0)=\theta,\qquad u_0'(0)=c_0\theta[k(\theta)]^{-1}.\end{gathered} \tag{5.18}$$

We also know that $x_\varepsilon \leqq x_0 = (1-\theta)\alpha^{-1}$, $\forall \varepsilon \leqq \varepsilon$. Hence, for all $\varepsilon \leqq \varepsilon_0$, and for all $x \geqq x_0$, one has $\theta_\varepsilon \leqq u_\varepsilon(x) < 1$ which shows that $u_\varepsilon \to u_0$ in the sup-norm on $[x_0, +\infty)$ where $u_0(x)=1$, $\forall x \geqq x_0$. It remains to study the limit of u_ε on the bounded interval $[0, x_0]$.

Letting $x \to +\infty$ in (5.7) shows that

$$c_\varepsilon \int_0^{+\infty} k(u_\varepsilon) u_\varepsilon'^2 \, dx = K_\varepsilon(1) - \frac{1}{2} c_\varepsilon^2 \theta.$$

Hence, $\int_0^{+\infty} u_\varepsilon'(x)^2 dx$ is bounded independently of ε. In particular, u_ε remains bounded in $H^1([0,x_0])$. Using the bound on x_ε and the compact embedding $H^1([0,x_0]) \subset C^0[0,x_0]$, we may extract a subsequence $\varepsilon_j \downarrow 0$ such that

$$(5.19) \qquad x_{\varepsilon_j} \to \bar{x} \leqq x_0,$$

$$(5.20) \qquad u_{\varepsilon_j} \to u_0 \quad \text{in } C^0[0,x_0].$$

Since $\theta_\varepsilon \uparrow 1$, we know that $u_0(x)=1$, $\forall x \geqq \bar{x}$. On the other hand, by (5.3), $g_{\varepsilon_j}(u_{\varepsilon_j}(x))$ converges uniformly to 0 on any interval $[0, \bar{x}-\delta]$ with $\delta > 0$. Therefore, the equation shows that u_0 is of class C^2 on $\mathbb{R} - \{\bar{x}\}$ and

$$(5.21) \qquad -\left(k(u_0)u_0'\right)' + c_0 u_0' = 0 \quad \forall x \leqq \bar{x}.$$

We claim that this determines $\bar{x}$ uniquely. Indeed, since $u_0(0)=\theta$ and $u_0'(0) = c_0\theta[k(\theta)]^{-1}$, it is easily checked by the uniqueness in the IVP (5.21) that $\bar{x}$ is uniquely determined. One could use, for instance, the argument in §4 to show that the solution of (5.21) together with the initial condition at $x=0$ is monotonous (pointwise) with respect to c_0.

Now, since $\bar{x}$ is unique, (5.19) holds for *any* subsequence $\varepsilon_j \downarrow 0$, which shows that

$$(5.22) \qquad \lim_{\varepsilon \downarrow 0} x_\varepsilon = \bar{x}.$$

Consequently, since u_0 is also unique, (5.20) holds for any subsequence and therefore

$$(5.23) \qquad \lim_{\varepsilon \downarrow 0} \max_{x \in \mathbb{R}} |u_\varepsilon(x) - u_0(x)| = 0.$$

It is then straightforward to show that

$$(5.24) \qquad \lim_{\varepsilon \downarrow 0} \|u_\varepsilon - u_0\|_{H^1(\mathbb{R})} = 0.$$

The proof of Theorem 5.2 is thereby complete. □

Remark 5.3. Theorem 5.2 and its proof give a rigorous justification for the heuristic "internal layer analysis" by which this type of asymptotic limit was usually obtained [4]. It is of interest to observe that the limiting solution u_0 verifies the equation

$$(5.25) \qquad -\left(k(u_0)u_0'\right)' + c_0 u_0' = c_0 \delta_{x=\bar{x}}$$

where $\delta_{x=\bar{x}}$ is the Dirac measure at $x=\bar{x}$ and $c_0 = \sqrt{2m}$. It is an open question whether one can pass to the limit in the equation in order to derive (5.25) in some more direct fashion.

6. Some remarks related to the numerical approximation of the scalar equation. From the viewpoint of the numerical approximation, it is desirable to study the analogous problem to (3.1) on a bounded interval and how a solution of the latter converges to a solution of (3.1). For $a>0$ we let $I_a=[-a,a]$ and we consider the problem:

To find $u\in C^2(I_a,[0,1])$ and $c>0$ satisfying

$$(6.1)\qquad \begin{aligned} &-(k(u)u')'+cu'=g(u) \quad \text{on } I_a,\\ &-k(u(-a))u'(-a)+cu(-a)=0, \quad u(+a)=1, \quad u(0)=\theta. \end{aligned}$$

The purpose of this section is to show under conditions (3.2)–(3.8) the following:

PROPOSITION 6.1. *For any $a>0$, there exists a unique solution u_a, c_a to problem* (6.1) *with the following properties*: $c_a>0$, $0<u_a<1$ *and u_a is increasing. As a function of a, c_a is decreasing. When $a\to+\infty$, c_a converges to c, and u_a converges to u in the sense that*

$$\lim_{a\to+\infty}\max_{x\in\mathbb{R}}|u_a(x)-u(x)|=0, \qquad \lim_{a\to+\infty}\|u_a-u\|_{H^1(\mathbb{R})}=0$$

where u_a is extended by $u_a(x)=1$, $\forall s\geqq a$ and $u_a(x)=u_a(-a)$, $\forall x\leqq -a$. Further (u,c) is the solution to problem (3.1).

Proof. The existence of u_a and c_a are actually consequences of the more general Theorem 7.6 concerning systems in the next section. Indeed, it will be seen that (6.1) is obtained in system (7.1), when the Lewis number Λ is made equal to 1. From the next section, we also know (cf. Proposition 8.1) that for any solution u,c of (6.1) one has u increasing and $c>0$.

Let us now prove the uniqueness of this solution. For a solution u, since u is increasing, it verifies $u<\theta$ on $(-a,0)$ whence $-k(u)u' +cu=0$ on$(-a,0)$. Therefore, it suffices to prove uniqueness of the solution c,u to the IVP:

$$(6.2)\qquad \begin{aligned} &-(k(u)u')'+cu'=g(u) \quad \text{on } (0,a),\\ &u(0)=\theta, \qquad u'(0)=c\theta k(\theta)^{-1} \end{aligned}$$

together with the boundary condition

$$(6.3)\qquad u(+a)=1.$$

Let u be a solution to the IVP (6.2). Let $x(s)=u^{-1}(s)$ and $z(s)=dx(s)/ds$. Then z satisfies equation (4.2) on $s\in[\theta,1]$. Now let $0<c_1<c_2$ and denote by u_1, u_2 the corresponding solutions of (6.2), $x_i(s)=u_i^{-1}(s)$, $z_i(s)=dx_i(s)/ds$, $i=1,2$, $s\in[\theta,1]$. Following the same argument as in §4 we obtain that

$$(6.4)\qquad z_1(s)>z_2(s) \quad \forall s\in[\theta,1],$$

whence it follows that

$$(6.5)\qquad x_1(s)>x_2(s) \quad \forall s\in(\theta,1].$$

In particular,

$$(6.6)\qquad x_1(1)>x_2(1).$$

This shows the solution to (6.2)–(6.3) to be unique. Indeed if $u(+a)=1$ and if $\bar{c}<c$ (resp., $\bar{c}>c$) then the solution $\bar{u}$ of the IVP (6.2) corresponding to $\bar{c}$ verifies $\bar{u}(+a)<1$ (resp., $\bar{u}(+a)>1$). This, of course, also proves that c_a is a *decreasing* function of $a>0$.

Therefore, c_a decreases toward a limit c. That u_a, c_a converge to u,c (in the sense of the proposition) and that u,c is *the* solution of (3.1) follow from the more general results of §8 (again corresponding to the particular case of a Lewis number equal to 1). □

Remark 6.2. Prescribing $u(0)=\theta$ is essential in Proposition 6.1. Indeed, c is a priori an unknown of the problem and is determined by imposing a constraint. The condition $u(0)=\theta$ is adequate since the limiting problem is translation invariant; indeed, if one wants to avoid weak convergences to $u=0$, $c\to+\infty$ of the solution of (6.1) as $a\to+\infty$, it is necessary to fix u at some point. Current numerical steady states (on a truncated domain) use the same centering condition [16].

7. The system case: existence of a solution on a bounded domain $[-a,+a]$. In this section, we construct a problem similar to the system introduced in (0.3), (1.13), but posed on a bounded domain $[-a,+a]$. On a bounded domain, this problem is then equivalent to a fixed point equation which we can solve with Leray–Schauder degree theory. The existence result obtained will be used in the next section. We should also point out its interest in numerical applications.

7.1. The boundary value problem and the hypothesis on the reaction term. In what follows a will be a fixed real number, strictly positive. We now consider the following boundary value problem posed on $[-a,+a]$: Find the triplet (u,v,c) satisfying

$$\begin{aligned}&-u''+cu'=f(u)v \quad \text{on } (-a,+a),\\&-\Lambda v''+cv'=-f(u)v \quad \text{on } (-a,+a),\end{aligned}\tag{7.1}$$

and the boundary conditions

$$\begin{aligned}&-u'(-a)+cu(-a)=0, &&u(a)=1,\\&-\Lambda v'(-a)+cv(-a)=c, &&v(a)=0.\end{aligned}\tag{7.2}$$

In what follows, θ is a given real number such that $0<\theta<1$. We are assuming on the nonlinear term f the following:

(7.3) $f\colon[0,1]\to\mathbb{R}^+$ is continuous, locally Lipschitz on $[\theta,1]$, possibly discontinuous at $u=\theta$;

(7.4) $f(u)=0$ on $[0,\theta[$, $\quad f(u)>0$ on $]\theta,1]$.

Remark 7.1. The real number θ will refer to the ignition temperature and the hypothesis (7.4) is made in order to circumvent the so-called "cold boundary difficulty." The hypothesis (7.3) is nearly optimal and corresponds to all known reaction terms.

Remark 7.2. Thanks to the hypothesis (7.4), the solution u, for $x<0$, is explicit: $u(x)=\theta e^{cx}$ and moreover $-\lambda v'(0)+cv(0)=c$. (We are anticipating the result $u'>0$ proved in §8.) The boundary conditions (7.2) are therefore equivalent to

$$\begin{aligned}&u(0)=\theta \quad u'(0)=\theta c, &&u(a)=1,\\&-\Lambda v'(0)+cv(0)=c, &&v(a)=0.\end{aligned}$$

We point out the analogy with the boundary conditions used in steady state numerical schemes [16]. Loosely speaking, when $a\to+\infty$, $u'(-a)$ and $v'(-a)$ are going to 0; taking the limit $a=+\infty$ in (7.2), we, therefore, recover (0.2) and (1.13). □

7.2. Equivalence with a fixed point problem. The transformation of a nonlinear equation into a fixed point problem is by now classical. The interest of the transformation performed here lies in the fact that we are solving a nonlinear eigenvalue problem. Indeed, not only u,v but also the parameter c (representing the mass flux, see §1.3) must be found to satisfy (7.1) and (7.2). We will denote $I_a=(-a,+a)$ and set by definition

$$X=C^1(\bar{I}_a)\times C^1(\bar{I}_a)\times\mathbb{R}. \tag{7.5}$$

This is a Banach space equipped with the $\|(u,v,c)\|_X\equiv\max(\|u\|_{C^1(\bar{I}_a)},\|v\|_{C^1(\bar{I}_a)},|c|)$. Let us consider the mapping sending each element (u,v,c) of X to the unique solution (U,V) of the following linear system, indexed by the parameter τ with $0\leqq\tau\leqq 1$:

$$\begin{aligned} &-U''+cU'=\tau f(u)v \quad \text{on } I_a,\\ &-\Lambda V''+cV'=-\tau f(u)v \quad \text{on } I_a; \end{aligned} \tag{7.6}$$

$$\begin{aligned} &-U'(-a)+cU(-a)=0, \qquad U(a)=1,\\ &-\Lambda V'(-a)+cV(-a)=c, \qquad V(a)=0. \end{aligned} \tag{7.7}$$

We can easily check that U and V are in the Sobolev space $H^2(I_a)$ and even in $W^{2,\infty}(I_a)$, if u and v are given in $C^1(\bar{I}_a)$. If the space dimension is one, the space $H^2(I_a)$ is embedded into $C^1(\bar{I}_a)$, the embedding being compact. Therefore, we can define a compact mapping K_τ,[1] indexed by $\tau\in[0,1]$, from X into X:

$$K_\tau\colon X\to X\colon (u,v,c)\to(U,V,c-u(0)+\theta). \tag{7.8}$$

Similarly, the mapping K: $X\times[0,1]\to X$ defined by $(u,v,c,\tau)\to K_\tau(u,v,c)$ is compact and uniformly continuous with respect to τ. Let us then notice that every solution of (7.1), (7.2) is a fixed point of K_τ and conversely: setting $F_\tau\equiv I-K_\tau$ (I is the identity mapping in X), solving (7.1), (7.2) is therefore, equivalent to proving existence of (u,v,c) such that $F_\tau(u,v,c)=0$. Thanks to the property of K_τ, it suffices then to compute the degree of F_τ at 0. It will remain to prove that the degree is indeed well defined and nonzero. Let us introduce the open bounded set $\Omega\subset X$:

$$\Omega=\left\{(u,v,c)\in X|\ \|u\|_{C^1(\bar{I}_a)}<M,\|v\|_{C^1(\bar{I}_a)}<M,\ \underline{c}<c<\bar{c}\right\}, \tag{7.9}$$

where $0<M<+\infty$, $0<\underline{c}<\bar{c}<+\infty$. The degree of F_τ in $\bar{\Omega}$ at 0 will be defined if there exist $M,\underline{c},\bar{c}$ such that $F_\tau(\partial\Omega)\neq 0$. This is demonstrated in the next section, where we will also compute the degree of F_τ in Ω at 0, using both the invariance by homotopy and the multiplicative property of the degree.

7.3. Justification of the degree and existence of one solution.

Let us state:

PROPOSITION 7.3. *Let K_τ and Ω be defined by* (7.8) *and* (7.9). *Then there exist finite constants $M,\underline{c},\bar{c}$, independent of a, such that*

$$F_\tau(\partial\Omega)=(I-K_\tau)(\partial\Omega)\neq 0, \qquad 0\leqq\tau\leqq 1. \tag{7.10}$$

Remark 7.4. It is not necesary to have $M,\underline{c},\bar{c}$ independent of a in order to prove (7.10). The result is indeed a consequence of stronger estimates, proved in §8, allowing it to pass to the limit $a=+\infty$.

[1] Actually, if $f(s)$ has a discontinuity at the point $s=\theta$ (i.e. $f(\theta)>0$), the operator K_τ will not be continuous. To overcome this problem, one more technical step is required: To approximate f by a continuous $f_\varepsilon=f\chi_\varepsilon$ with $\chi_\varepsilon(s)=0$ if $s\leqq\theta$, 1 if $s\geqq\theta+\varepsilon$, and $(s-\theta)/\varepsilon$ if $\theta\leqq s\leqq\theta+\varepsilon$. Now, the solution u we find is montonically increasing. Then, in view of the a priori estimates derived in §7.3. below, the limiting procedure as $\varepsilon\to 0$ is fairly straightforward. Details will be omitted here. This yields the existence result for a discontinuous f as well. The authors are indebted to V. Giovangigli for pointing out to them that this point was overlooked in a first draft.

Proof. It is sufficient to prove the existence of $M, \underline{c}, \bar{c}$, independent of a, such that

$$\{F_\tau(u,v,c)=0, \forall\tau\in[0,1]\} \Rightarrow \{(u,v,c)\in\Omega\}.$$

The existence of $\underline{c}$ is given by Proposition 8.7. The existence of $\bar{c}$ is given by Proposition 8.10. From Proposition 8.1 we deduce $\sup_{I_a}|u(x)|\leqq 1$, $\sup_{I_a}|v(x)\leqq 1$ and $\sup_{I_a}|u'(x)\leqq c$, $\sup_{I_a}|v'(x)|\leqq c/\Lambda$; therefore, $\|u\|_{C^1(\bar{I}_a)}\leqq \max(1,c)$, $\|v\|_{C^1(\bar{I}_a)}\leqq \max(1,c/\Lambda)$ the existence of M is then again a consequence of Proposition 8.10. □

The value of deg $(F_\tau, \Omega, 0)$ is then given by

PROPOSITION 7.5. *Let K_τ and Ω be defined by* (7.8) *and* (7.9). *Then the mapping $F_\tau = I - K_\tau$ satisfies*

$$\deg(F_\tau,\Omega,0)=\deg(F_0,\Omega,0)=-1, \qquad 0\leqq\tau\leqq 1. \tag{7.11}$$

Proof. The homotopy invariance property of the degree proves the first equality in (7.11). When $\tau=0$, a straightforward computation gives U and V solutions of (7.6), (7.7):

$$U_0(c)\colon x\to e^{c(x-a)}, \qquad V_0(c)\colon x\to 1+e^{c(x-a)/\Lambda}.$$

Therefore, F_0 is known explicitly:

$$F_0\colon X\to X\colon (u,v,c)\to(u-U_0(c), v-V_0(c), u(0)-\theta)$$

and this mapping is homotopic to

$$\Phi\colon X\to X\colon (u,v,c)\to(u-U_0(c), v-V_0(c), e^{-ca}-\theta).$$

Using the multiplicative property of the degree, we find that the degree of Φ is -1 (note that the function $c\to e^{-ca}-\theta$ is decreasing.) □

The main result of this section is a direct consequence of Propositions 7.3 and 7.5.

THEOREM 7.6. *Under assumptions* (7.3) *and* (7.4), *the problem* (7.1), (7.2) *does have at least one solution* (u,v,c).

Remark 7.7. Uniqueness for the problem (7.1), (7.2) is an open question for $\Lambda>1$ [19].

8. Existence of a solution on $\mathbb{R}$ for the system. Existence of one solution (u,v,c) for the problem

$$\begin{aligned} &-u''+cu'=uf(u) \quad \text{on } \mathbb{R},\\ &-\Lambda v''+cv'=-vf(u) \quad \text{on } \mathbb{R},\\ &u(-\infty)=0, \qquad u(+\infty)=1,\\ &v(-\infty)=1, \qquad v(+\infty)=0,\\ &u(0)=\theta \end{aligned} \tag{8.1}$$

is a consequence of Theorem 7.6. More precisely, for each $a>0$, the problem (7.2) does have at least one solution (u_a, v_a, c_a). In this section we show that, for a going to $+\infty$, the sequence (u_a, v_a, c_a) (or an extracted subsequence) converges to one solution of (8.1). The proof is based mainly on qualitative properties of a solution (u_a, v_a, c_a) (defined on I_a) and a priori estimates, independent of a, of c_a, then of u_a, v_a.

8.1. Qualitative properties of (u_a, v_a, c_a). For the sake of simplicity, we will drop the explicit dependency on a. We consider the problem

$$\begin{array}{ll} & -u''+cu'=f(u)v \quad \text{on } (-a,+a),\\ & -\Lambda v''+cv'=-f(u)v \quad \text{on } (-a,+a),\\ (8.2) & +u'(-a)+cu(-a)=0, \qquad u(a)=1,\\ & -\Lambda v'(-a)+cv(-a)=c, \qquad v(a)=0,\\ & u(0)=\theta, \end{array}$$

where $\Lambda>0$.

PROPOSITION 8.1. *For every solution* (u,v,c) *of the problem* (8.2) *with* $c\geqq 0$, *the following holds*:

$$\begin{array}{lll} (8.3) & c>0, & \\ (8.4) & 0<u\leqq 1, & 0\leqq v<1 \quad \textit{on } [-a,+a],\\ (8.5) & 0<u'\leqq c, & -c/\Lambda\leqq v'<0 \quad \textit{on } [-a,+a]. \end{array}$$

Remark 8.2. In particular, $u>\theta$ on $(0, +a]$ and $u<\theta$ on $[-a,0)$. As f satisfies for hypothesis (7.4), for $x\in[-a,0]$, one has:

$$u(x)=\theta e^{cx}, \qquad v(x)=1-\alpha e^{cx/\Lambda},$$

where α is a positive (unknown) constant.

Proof. A. To prove (8.3), we show that $c=0$ is impossible. If $c=0$, since $u'(-a)=v'(-a)=0$ and $u(a)=1$, $v(a)=0$, every solution (u,v) satisfies $u+\Lambda v-1\equiv 0$ on $(-a,a)$. Therefore, u is a solution of the problem

$$(8.6) \qquad \begin{array}{l} -u''-\dfrac{1}{\Lambda}f(u)(1-u)=0,\\ u'(-a)=0, \quad u(0)=\theta, \quad u(a)=1. \end{array}$$

Necessarily, $u\leqq 1$ on $[-a,+a]$; if not, some x_0 would exist with $-a\leqq x_0<a$ such that $u'(x_0)=0$ and $u(x_0)=\max_{-a\leqq x\leqq a}u(x)>1$, $u''(x_0)>0$ and that is impossible in view of (8.6). Let us now define $w\equiv u'$; as $f(u)(1-u)\geqq 0$ on $[-a,+a]$ and $w(-a)=0$, we deduce from (8.6)

$$w(x)=-\frac{1}{\Lambda}\int_{-a}^{x}f(u(y)(1-u(y))\,dy\leqq 0$$

for $x\in[-a,+a]$. The function u is, therefore, decreasing on $[-a,+a]$. In particular, $\theta=u(0)\geqq u(a)=1$, which is in contradiction with the hypothesis $\theta<1$. Thus, $c=0$ is impossible.

B. Let us show that $v\geqq 0$ on $[-a,+a]$, again arguing by contradiction. Suppose $\min_{-a\leqq x\leqq a}v(x)<0$; then either there exists x_1 such that $v(x_1)<0$ and $v'(x_1)=0$, or $v\leqq 0$ and $v'>0$ on $[-a,+a]$. In the first case, there exist α,β such that $-a\leqq\alpha<\beta\leqq+a$ and

$$v'(\alpha)=0, \qquad v(\beta)=0, \qquad v'\geqq 0 \quad \text{on } [\alpha,\beta].$$

In particular $v\leqq 0$ on $[\alpha,\beta]$, and as v is solution of $+\Lambda v''-cv'=f(u)v$, v satisfies $(v'\exp-cx/\Lambda)'=(1/\Lambda)\exp(-c/\Lambda)\mathrm{f}(u)v\leqq 0$ on $[\alpha,\beta]$. Integrating this inequality from α to x, $\alpha\leqq x\leqq\beta$ we obtain $v'\leqq 0$ on $[\alpha,\beta]$, for $v'(\alpha)=0$. Thus, $v'\equiv 0$ on $[\alpha,\beta]$ and

$v\equiv 0$ on $[\alpha,\beta]$ since $v(\beta)=0$. Therefore, $v\equiv 0$ on $[-a,+a]$; but $-\Lambda v'(-a)+cv(-a)=c$ and so $c=0$, which contradicts the conclusion of A.

In the second case, as $-\Lambda v'(-a)=c(1-v(-a))$, the hypothesis $c\geqq 0$ and $v\leqq 0$ show that $v'(-a)\leqq 0$ which contradicts $v'>0$. In conclusion, $\min_{-a\leqq x\leqq +a} v<0$ is impossible, therefore $v\geqq 0$ on $[-a,+a]$.

C. Let us remark that the system (8.2) is equivalent to a first order system. This remark is crucial in order to prove (8.4) knowing that $v\geqq 0$. Let us introduce a new dependent variable w (mass flux fraction) defined by

$$w\equiv -\Lambda v'+cv. \tag{8.7}$$

Then, for (u,v,c) a solution of (8.2), $w'=-vf(u)$ and $v'=cv/\Lambda-w/\Lambda$; moreover, $-(u+\Lambda v)''+c(u+v)'=0$, after integration between $-a$ and x, yields $u'=c(u-1)+w$, thanks to (8.7) and $w(-a)=c$. The system (8.2) is therefore equivalent to

$$\begin{aligned}
&u'=c(u-1)+w,\\
&v'=\frac{c}{\Lambda}v-\frac{1}{\Lambda}w,\\
&w'=-f(u)v,\\
&-u'(-a)+cu(-a)=0, \qquad u(a)=1,\\
&w(-a)=c, \qquad v(a)=0,\\
&u(0)=\theta.
\end{aligned}\tag{8.8}$$

Since $f(u)\geqq 0$, $v\geqq 0$, the function w is decreasing from c to $w(a)=-\lambda v'(a)>0$ (necessarily $v'(a)<0$ if $v\geqq 0$): $0\leqq w\leqq c$. But $u'-cu=w-c$; we thus obtain $-ce^{-cx}<(e^{-cx}u)'\leqq 0$. By integration of these two inequalities from x to $+a$ ($u(a)=1$), we get $u(x)\geqq e^{-c(a-x)}>0$ and $e^{-cx}(u(x)-1)\leqq 0$ and so $0<u\leqq 1$. Similarly, from $v'-cv/\Lambda=-w/\Lambda$, we get $(-c/\Lambda)\exp(-cx/\Lambda)\leqq(v\exp(-cx/\Lambda))'\leqq 0$ and by integration from x to $+a$ ($v(a)=0$), we obtain $0\leqq v<1$.

D. The function u satisfies $-u''+cu'=f(u)v\geqq 0$ and, therefore, $-(e^{-cx}u')'\geqq 0$. By integration from x to $+a$, we have $e^{-ca}u'(a)\leqq e^{-cx}u'(x)$. But, integrating the relation $-(u+\Lambda v)''+c(u+v)'=0$ from $-a$ to $+a$, we get $u'(a)+\Lambda v'(a)=0$ and thus $u'(a)>0$ since $v'(a)<0$. Therefore, $u'>0$. In the same way, as v satisfies $-\Lambda v''+cv'=-f(u)v\leqq 0$, the inequality $(v'\exp-cx/\Lambda)'\geqq 0$ provides, after integration from x to $+a$, $v'<0$. In particular, $0<v<1$ on $]-a,+a[$. Similarly, as $u'>0$, $0<u<1$ on $]-a,+a[$. Finally, from the first equation in (8.8), as $0<u\leqq 1$ and $0<w\leqq c$, we immediately obtain $u'(x)\leqq c$ on $[-a,+a]$. The second equation in (8.8) also gives $-c/\Lambda\leqq v'(x)$ on $[-a,+a]$. □

Remark 8.3. The former qualitative properties have been obtained under the sole assumption of positivity of f. The function f is possibly discontinuous.

8.2. A priori estimates on (u,v,c). A priori estimates from above and below for c will be obtained (using the conservation laws satisfied by u) by comparison of v to $1-u$. So we will prove the simple, but very useful for the sequel, following:

PROPOSITION 8.4. *Every solution* (u,v,c) *of the problem* (8.2), *with* $c\geqq 0$, *satisfies on* $[-a,+a]$:

$$\begin{aligned}
&|u(x)+v(x)-1|\leqq|\Lambda-1|v(x),\\
&|u(x)+v(x)-1|\leqq\left|\frac{\Lambda-1}{\Lambda}\right|(1-u(x));
\end{aligned}\tag{8.9.1}$$

$$(8.9.2)\qquad \begin{aligned} &|u(x)+\Lambda v(x)-1| \leqq |\Lambda-1| v(x), \\ &|u(x)+\Lambda v(x)-1| \leqq |\Lambda-1|(1-u(x)). \end{aligned}$$

In particular,

$$(8.10.1)\qquad \frac{1}{\Lambda}(1-u(x)) \leqq v(x) \leqq (1-u(x)) \quad \text{if } \Lambda>1,$$

$$(8.10.2)\qquad (1-u(x)) \leqq v(x) \leqq \frac{1}{\Lambda}(1-u(x)) \quad \text{if } 0<\Lambda<1.$$

Remark 8.5. $u+v-1$ and $u+\Lambda v-1$ are the only linear relations possible between u and v. (The constant -1 comes from the boundary condition $u(a)=1$.) The physical meaning of $u+v-1$ is the enthalpy.

Remark 8.6. If $\Lambda=1$, then $u+v-1\equiv 0$ and we verify the conservation of enthalpy [4]. In particular, $u(x_0)=0$ if and only if $v(x_0)=0$.

Proof. It is a simple differential inequality. Setting $z\equiv u+v-1$, where u and v are solutions of (8.2), the function z satisfies $-z''+cz'=(\Lambda-1)v''$. By integration between $-a$ and x for $-a\leqq x\leqq +a$, we get:

$$-z'(x)+cz(x)-(\Lambda-1)v'(x)=-z'(-a)+cz(-a)-(\Lambda-1)v'(-a).$$

The right-hand side of this equality is 0 because $-u'(-a)+cu(-a)=-\Lambda v'(-a)+cv(-a)=0$. Therefore, $-z'(x)+cz(x)=(\Lambda-1)v'(x)$ and integrating again, now between x and $+a$ $(z(a)=0)$:

$$e^{-cx}z(x)=(\Lambda-1)\int_x^a e^{-cs}v'(s)\,ds.$$

From Proposition 8.1, we know that $v'<0$ and thus:

$$e^{-cx}|z(x)| \leqq |\Lambda-1| e^{-cx}\int_x^a (-v'(s))\,ds.$$

The first inequality (8.9.1) follows as $v(+a)=0$. To prove the second, we remark that $z=u+v-1$ does also satisfy $-\Lambda z''+cz'=(1-\Lambda)u''$. Integrating between $-a$ and x, using the left boundary condition, we obtain

$$-\Lambda z'(x)+cz(x)=(1-\Lambda)u'(x).$$

A new integration between x and $+a(z(a)=0)$ gives

$$e^{-cx/\Lambda}z(x)=\frac{1-\Lambda}{\Lambda}\int_x^a e^{-cs/\Lambda}u'(s)\,ds$$

and as (Proposition 8.1) $u'>0$ and $u(a)=1$, we get $|z(x)|\leqq|((\Lambda-1)/\Lambda)|(1-u(x))$. The two inequalities (8.9.2) are obtained in a similar way introducing $y\equiv u+\Lambda v-1$, that satisfy $-y''+cy'=(\Lambda-1)cv'$ and $-\Lambda y''+cy'=(1-\Lambda)cu'$. The two inequalities (8.10) follow directly from (8.9.1) and (8.9.2). □

Now we state the main result of this section.

PROPOSITION 8.7. *Let* $g(s)\equiv(1-s)f(s)$ *and assume that* $G(1)\equiv\int_\theta^1 g(s)\,ds<+\infty$. *Then, if* (u,v,c) *is any solution of* (8.2), *c satisfies*:

$$\frac{1}{\Lambda}\leqq\frac{c^2}{2G(1)}\leqq\frac{1}{\theta^2}\left(1+\frac{|u'(a)|^2}{2G(1)}\right)\quad \textit{if}\ \Lambda>1, \tag{8.11}$$

$$1\leqq\frac{c^2}{2G(1)}\leqq\frac{1}{\theta^2}\left(\frac{1}{\Lambda}+\frac{|v'(a)|^2}{2G(1)}\right)\quad \textit{if}\ \Lambda<1. \tag{8.12}$$

Remark 8.8. Let us recall that the function $s\to f(s)$ is identically 0 for $s<\theta(0<\theta<1)$; therefore $G(1)=\int_0^1 g(s)\,ds$.

The proposition will result from Proposition 8.4 and the following one, which make precise L^2 estimates for u' and v'.

PROPOSITION 8.8. *For any solution* (u,v,c) *of* (8.2) *with* $c\geqq 0$:

$$\int_0^a |u'(x)|^2\,dx\leqq\frac{c}{2}(1-\theta^2), \tag{8.13}$$

$$\frac{c}{2}|u(-a)|^2+\int_{-a}^a |u'(x)|^2\,dx\leqq\frac{c}{2}, \tag{8.14}$$

$$\frac{c}{2}|v(-a)|^2+\Lambda\int_{-a}^a |v'(x)|^2\,dx\leqq c. \tag{8.15}$$

Proof. By integration, between 0 and $+a$, of the equation $-u''+cu'=vf(u)$ multiplied by u, we get as $u(0)=\theta$:

$$-u'(a)+\frac{c}{2}(1+\theta^2)+\int_0^a |u'(x)|^2\,dx=\int_0^a f(u(x))v(x)u(x)\,dx.$$

The right-hand side is bounded from above ($f(u)vu\geqq 0$ and $u\leqq 1$) by $\int_0^a f(u(x))v(x)\,dx$; but this integral is equal to $c-u'(a)$, as easily seen by integrating $-u''+cu'=f(u)v$ between 0 and $+a$. As a result we get (8.13). Now, we integrate $-u''+cu'=f(u)v$, between $-a$ and $+a$, after multiplication by 1 and u. We get two identities:

$$c-u'(a)=\int_{-a}^{+a} f(u(x))v(x)\,dx,$$

$$c-u'(a)-\frac{c}{2}\left(|u(a)|^2-|u(-a)|^2\right)+\int_{-a}^{+a} |u'(x)|^2\,dx=\int_{-a}^{+a} f(u(x))v(x)u(x)\,dx.$$

The right-hand side of the second identity is bounded from above by $c-u'(a)$. From the first identity and $|u(a)|^2<1$ we get (8.14). Finally, we integrate between $-a$ and $+a$, the equation $-\Lambda v''+cv'=-f(u)v$ multiplied by v and obtain ($f(u)\geqq 0$):

$$+\int_{-a}^a (-\Lambda v''(x)+cv'(x))v(x)\,dx=-\int_{-a}^a f(u(x))v(x)^2\,dx\leqq 0.$$

The inequality (8.15) follows, integrating by parts since $v(-a)\leqq 1$. □

Now we can give the

Proof of Proposition 8.7. Let us prove (8.11), the proof of (8.12) being identical. We start with the identity obtained by intgration of $-u''+cu'=f(u)v$ between 0 and $+a$, after multiplication by u',

$$-\frac{1}{2}|u'(a)|^2+\frac{1}{2}\theta^2c^2+c\int_0^a |u'(x)|^2\,ds=\int_0^a f(u(x))v(x)u'(x)\,dx \tag{8.16}$$

($u'(0)=\theta c$, see Remark 8.2). As $f(u)\geqq 0$ and $u'\geqq 0$ (Proposition 8.1), we obtain from (8.10.1)

$$(8.17)\quad \frac{1}{2}\theta^2c^2+c\int_0^a |u'(x)|^2\,dx \geqq \frac{1}{\Lambda}\int_0^a f(u(x))(1-u(x))u'(x)\,dx \equiv \frac{1}{\Lambda}G(1).$$

From (8.13) (Proposition 8.8), the left-hand side of this inequality is bounded above by

$$\frac{\theta^2c^2}{2}+(1-\theta^2)\frac{c^2}{2}=\frac{c^2}{2}$$

and so we get the first part of (8.11). Now from (8.16), we can also deduce:

$$(8.18)\qquad \frac{1}{2}\theta^2c^2 \leqq \int_0^a f(u(x))v(x)u'(x)\,dx+\frac{|u'(a)|^2}{2}$$

(note that $|u'(a)|\leqq c$, using (8.5), Proposition 8:1). Using again (8.10.1), we get from (8.18) and the definition of $G(1)$:

$$\frac{1}{2}\theta^2c^2 \leqq G(1)+\frac{|u'(a)|^2}{2},$$

which completes the proof of (8.11). □

Proposition 8.7 will play a key role to bound c when passing to the limit $a=+\infty$. For a finite a, the necessary upper bound for c (independent of a) is obtained by constructing an appropriate upper solution.

PROPOSITION 8.10. *Let* $M=\sup_{\theta\leqq s\leqq 1}f(s)$ *and* $a_0>0$ *is fixed. Then, for each* $a\geqq a_0$, *if* (u,v,c) *is any solution of* (8.2), *c satisfies*

$$(8.19)\qquad c\leqq \max\left(-\frac{\log\theta}{a_0},\max\left(2M,\frac{\sqrt{2M}}{\theta}\right)\right).$$

Proof. Let us define the function $\bar{u}$ as the unique solution of:

$$(8.20)\qquad \begin{aligned} &-\bar{u}''+c\bar{u}'=MH \quad \text{on } (-a,+a),\\ &-\bar{u}'(-a)+c\bar{u}(-a)=0, \qquad \bar{u}(a)=1, \end{aligned}$$

where H is the Heaviside function at $x=0$. From Proposition 8.1, we know that $u(x)\leqq\theta$ for $x\leqq 0$, $u\leqq 1$ and $0\leqq v\leqq 1$. Thus a simple comparison principle shows that $u(x)\leqq\bar{u}(x)$ on $(-a,+a)$. In particular, $\theta=u(0)\leqq\bar{u}(0)$; but an explicit computation gives $\bar{u}(0)$:

$$(8.21)\qquad \theta=u(0)\leqq\bar{u}(0)=e^{-ca}\left(1-\frac{M}{c}\right)+\frac{M}{c^2}(1-e^{-ca}).$$

Set $c_0=\max(2M,\sqrt{2M}/\theta)$; then either $c\leqq c_0$ and the proof is complete or $c\geqq c_0$. In this latter case, $\theta\leqq\frac{1}{2}e^{-ca}+\theta//2$, due to (8.21), and then $c\leqq -(\log\theta)/a\leqq -(\log\theta)/a_0$ since $0<\theta<1$, $a\geqq a_0$ which gives an upper bound for c. The proof is complete. □

8.3. The passage to the limit $a=+\infty$. By passing to the limit $a=+\infty$, we shall obtain, using the a priori estimates of former sections, existence of one solution for the problem (8.1). More precisely:

THEOREM 8.11. *Let* θ, a_0 *such that* $0<\theta<1$, $a_0>0$. *Assuming the conditions* (7.3), (7.4) *on F, there exists an increasing sequence* $\{a_n\}_{n\in N}$ *with* $a_n>a_0, \lim_{n\to\infty}a_n=+\infty$

such that $(u_{a_n}, v_{a_n}, c_{a_n})$ *solution of* (8.2) *on* $(-a_n, +a_n)$ *converges, for the topology of* $C^1_{\text{loc}}(\mathbb{R}) \times C^1_{\text{loc}}(\mathbb{R}) \times \mathbb{R}$, *to one solution of* (u, v, c) *of* (8.1). *Moreover*:

$$0 \leqq u, v \leqq 1 \quad \textit{on } \mathbb{R}, \tag{8.22}$$

$$0 \leqq u' \leqq c, \ -\frac{c}{\Lambda} \leqq v' \leqq 0 \quad \textit{on } \mathbb{R}, \tag{8.23}$$

$$u, v \in W^{2,\infty}(\mathbb{R}) \tag{8.24}$$

$$0 < \underline{c} \leqq c \leqq \bar{c} < +\infty. \tag{8.25}$$

Remark 8.9. It can be proved that the problem (8.1) has in fact only one solution if $\Lambda < 1$. This uniqueness result will be made more precise §11.

Proof. Let us take a solution (u_a, v_a, c_a) of (8.2). Using Propositions 8.7 and 8.10, there exist two constants $\underline{c}, \bar{c}$, independent of a, with $0 < \underline{c} < \bar{c} + \infty$ and such that $\underline{c} \leqq c_a \leqq \bar{c}$. Using (8.4) and (8.5), we have $0 < u_a \leqq 1$, $0 < u'_a \leqq c_a < \bar{c}$ and $-\bar{c}/\Lambda \leqq -c_a/\Lambda \leqq v' < 0$. As $0 \leqq f(s) \leqq M$ for $0 \leqq s \leqq 1$, we then deduce that $u''_a = c_a u'_a - f(u_a) v_a$ and $v''_a = c_a v'_a / \Lambda + f(u_a) v_a / \Lambda$ are bounded independently of a. Therefore, u_a, v_a are bounded independently of a in $W^{2,\infty}(-a, +a)$. As a consequence, we obtain the convergence, in the topology of $C^1_{\text{loc}}(\mathbb{R}) \times C^1_{\text{loc}}(\mathbb{R}) \times \mathbb{R}$, of the sequence (u_{a_n}, v_{a_n}, c_n) to (u, v, c) satisfying:

$$-u'' + cu' = f(u)v \quad \text{on } \mathbb{R}, \tag{8.26}$$

$$-\Lambda v'' + cv' = -f(u)v \quad \text{on } \mathbb{R}. \tag{8.27}$$

Properties (8.23)–(8.25) are clearly satisfied. From Remark 8.2, we get $u(-\infty) = 0$ and $v(-\infty) = 1$. Moreover, $u(0) = \theta$, since $u_a(0) = \theta$. Let us find $l \equiv \lim_{x \to +\infty} u(x)$ and $l' \equiv \lim_{x \to +\infty} v(x)$. Since v' and v are bounded on R, $v'(+\infty)$ is finite and $v'(+\infty) = 0$; similarly $v''(+\infty) = 0$. But v does satisfy (8.27) and therefore $f(u(+\infty))v(+\infty) = f(l)l' = 0$. Since $l > \theta$ (u is strictly increasing), with the assumption (7.3) we deduce that $f(l) \neq 0$ and thus $l' = 0$. Proposition 8.4 and Remark 8.6 give finally $l = 1$. This proves $u(+\infty) = 1$, $v(+\infty) = 0$ and (8.22). □

9. High activation energy values: asymptotic analysis.

9.1. Setting of the problem. Section 5 has shown the interest of the asymptotic analysis for high activation energy values. Until now, the only assumptions on f have been hypotheses (7.3) and (7.4). We will now make precise the singular behavior of $f(u)v$ (reaction rate) with respect to the "small" parameter $\varepsilon > 0$ (inverse of the reduced activation energy). We will always write f_ε instead of f in order to emphasize the dependency on ε. In practical cases (see Remark 5.1), f_ε often takes the form:

$$f_\varepsilon(u) \equiv \frac{1}{\varepsilon^2} \phi\left(\frac{u-1}{\varepsilon}, \varepsilon\right), \qquad u > \theta \tag{9.1}$$

($f_\varepsilon(u) \equiv 0$ for $u < \theta$), where ϕ satisfies

$$\lim_{\varepsilon \to 0} -\frac{1}{\varepsilon} \frac{s-1}{\varepsilon} \phi\left(\frac{s-1}{\varepsilon}, \varepsilon\right) = 0, \qquad 0 < s < 1, \tag{9.2}$$

$$m \equiv \lim_{\varepsilon \to 0} \int_{-\infty}^{0} -\sigma \phi(\sigma, \varepsilon)\, d\sigma < +\infty. \tag{9.3}$$

Remark 9.1. Hypotheses (7.3) and (7.4) being still satisfied, ϕ is a function defined on $\mathbb{R}^-$ with values into $\mathbb{R}^+$.

Remark 9.2. One typical example of function ϕ is the exponential $\sigma \to \exp \sigma$ that corresponds to the Arrhenius Law (see §1.6).

The main result of this section will be stated under hypotheses that generalize (9.1)–(9.3). More precisely, $\{f_\varepsilon\}_{\varepsilon>0}$ satisfy (7.3), (7.4). Moreover, if we set by definition,

$$G_\varepsilon(u)\equiv\int_\theta^u f_\varepsilon(s)(1-s)\,ds,$$

there exists θ_ε, $\theta\leqq\theta_\varepsilon<1$, such that

$$\lim_{\varepsilon\to0}\theta_\varepsilon=1 \text{ and } \lim_{\varepsilon\to0}\max_{\theta\leqq s\leqq\theta_\varepsilon} f_\varepsilon(s)(1-s)=0, \tag{9.4}$$

$$\lim_{\varepsilon\to0}\int_\theta^1 f_\varepsilon(s)(1-s)\,ds=\lim_{\varepsilon\to0}G_\varepsilon(1)\equiv m<+\infty. \tag{9.5}$$

Remark 9.3. The function $s\to f_\varepsilon(s)(1-s)$ will play the role of the function g of Part I.

9.2. The main result and its proof. Let us consider problem (8.1). For $\varepsilon>0$ fixed, there exists one solution $(u_\varepsilon,v_\varepsilon,c_\varepsilon)$ (in fact, unique if $\Lambda<1$, see §11) and we will study its behavior as ε goes to 0. The main result is:

THEOREM 9.4. *Under the hypotheses* (7.3), (7.4) *and* (9.4), (9.5) *for* f_ε, *there exists one sequence* $\{\varepsilon_n\}_{n\in N}$, *decreasing to* 0, *such that* $(u_{\varepsilon_n},v_{\varepsilon_n},c_{\varepsilon_n})$, *a solution of* (8.1), *converge* (*strongly*) *for the topology* $H^1(\mathbb{R})\times H^1(\mathbb{R})\times\mathbb{R}$ *to* (u,v,c). *Moreover*, (u,v) *is a solution of the problem*

$$-u''+cv'=c\delta_{x=\bar{x}}, \tag{9.6}$$

$$-\Lambda v''+cv'=-c\delta_{x=\bar{x}}, \tag{9.7}$$

$$u(-\infty)=0,\quad u(0)=\theta,\quad u(+\infty)=1 \tag{9.8}$$

$$v(-\infty)=1,\qquad v(+\infty)=0 \tag{9.9}$$

for $\bar{x}=-\log\theta/\varepsilon$ ($\delta_{x=\bar{x}}$ *is the Dirac function at* $\bar{x}$).

Remark 9.5. The condition $u(0)=\theta$ fixes the value of $\bar{x}$.

Remark 9.6. The theorem justifies the so called "model of Dirac" commonly used in combustion [4].

Remark 9.7. The precise value of c will be given in §11.

Proof. First of all, let us remark that inequalities (8.10) of Proposition 8.4 are still satisfied on $\mathbb{R}$:

$$\min(1,\Lambda^{-1})(1-u_\varepsilon(x))\leqq v_\varepsilon(x)\leqq\max(1,\Lambda^{-1})(1-u_\varepsilon(x))\quad\text{on }\mathbb{R}. \tag{9.10}$$

Also, inequalities (8.11) and (8.12) are satisfied on $\mathbb{R}$; since $u'(+\infty)=\lim_{a\to+\infty}u'(a)=0$, we get a more precise result:

$$\min(1,\Lambda^{-1})\leqq\frac{c_\varepsilon^2}{2G_\varepsilon(1)}\leqq\frac{1}{\theta^2}\max(1,\Lambda^{-1}). \tag{9.11}$$

These two inequalities allow us to generalize the proof of Theorem 5.2 to the case of this system. As before, let us define x_ε, such that $u_\varepsilon(x_\varepsilon)=\theta_\varepsilon$ ($u_\varepsilon'>0$ and $\theta_\varepsilon<1$). We start with the three familiar identities obtained in integration, between 0 and x, of the equation $-u''_\varepsilon+c_\varepsilon u'_\varepsilon=f_\varepsilon(u_\varepsilon)v_\varepsilon$ successively multiplied by 1, u_ε, u'_ε:

$$-u'_\varepsilon()+c_\varepsilon u_\varepsilon(x)=\int_0^x f_\varepsilon(u_\varepsilon(y))v_\varepsilon(y)\,dy, \tag{9.12}$$

$$\begin{aligned}-u_\varepsilon(x)u'_\varepsilon(x)+\frac{c_\varepsilon}{2}\left(|u_\varepsilon(x)|^2+\theta^2\right)+\int_0^x |u'_\varepsilon(y)|^2\,dy\\ =\int_0^x f_\varepsilon(u_\varepsilon(y))v_\varepsilon(y)u_\varepsilon(y)\,dy,\end{aligned} \tag{9.13}$$

$$\begin{aligned}-\frac{1}{2}|u'_\varepsilon(x)|^2+\frac{1}{2}c_\varepsilon^2\theta^2+c_\varepsilon\int_0^x |u'_\varepsilon(y)|^2\,dy\\ =\int_0^x f_\varepsilon(u_\varepsilon(y))v_\varepsilon(y)u'_\varepsilon(y)\,dy.\end{aligned} \tag{9.14}$$

Since $f_\varepsilon(u_\varepsilon)u'_\varepsilon>0$, $f_\varepsilon(u_\varepsilon)(1-u_\varepsilon)u'_\varepsilon\geqq 0$, we get from (9.14) and (9.10):

$$-\frac{1}{2}|u'_\varepsilon(x)|^2+\frac{1}{2}c_\varepsilon^2\theta^2\leqq\max(1,\Lambda^{-1})G_\varepsilon(\theta_\varepsilon),\qquad 0<x\leqq x_\varepsilon.$$

Using (9.11), we obtain for $0<x\leqq x_\varepsilon$:

$$\min(1,\Lambda^{-1})G_\varepsilon(1)\theta^2-\max(1,\Lambda^{-1})G_\varepsilon(\theta_\varepsilon)\leqq\frac{1}{2}|u'_\varepsilon(x)|^2.$$

The assumptions (9.4), (9.5) allow us to find a constant $\alpha>0$, independent of ε, such that for each $\varepsilon\leqq\varepsilon_0$:

$$u'_\varepsilon(x)\geqq\alpha,\qquad 0<x\leqq x_\varepsilon.$$

Integration of this inequality between 0 and x_ε gives (note that $\theta\leqq\theta_\varepsilon$):

$$0<x_\varepsilon<\frac{1-\theta}{\alpha}\equiv x_0. \tag{9.15}$$

From (9.12), written at $x=x_\varepsilon$, and (9.10) we deduce

$$0<-u'_\varepsilon(x_\varepsilon)+c_\varepsilon\theta_\varepsilon\leqq\max(1,\Lambda^{-1})\int_0^{x_\varepsilon} f_\varepsilon(u_\varepsilon(y))(1-u_\varepsilon(y))\,dy,$$

and, therefore, with (9.15):

$$0<-u'_\varepsilon(x_\varepsilon)+c_\varepsilon\theta_\varepsilon\leqq\max(1,\Lambda^{-1})\frac{1-\theta}{\alpha}\max_{\theta\leqq s\leqq\theta_\varepsilon} f_\varepsilon(s)(1-s).$$

With assumption (9.4), we then obtain

$$\lim_{\varepsilon\to 0}\left(-u'_\varepsilon(x_\varepsilon)+c_\varepsilon\theta_\varepsilon\right)=0. \tag{9.16}$$

Integration from x_ε to $+\infty$ of $-u''_\varepsilon+c_\varepsilon u'_\varepsilon=f_\varepsilon(u_\varepsilon)$ multiplied by u'_ε and use of (9.10) gives:

$$\frac{1}{2}|u'_\varepsilon(x_\varepsilon)|^2\leqq\max(1,\Lambda^{-1})(G_\varepsilon(1)-G_\varepsilon(\theta_\varepsilon)). \tag{9.17}$$

From (9.17), (9.4), and (9.16), we deduce

$$\overline{\lim_{\varepsilon\to 0}}\, c_\varepsilon\leqq\sqrt{2\max(1,\Lambda^{-1})m} \tag{9.18}$$

and from (9.11)

$$\underline{\lim}_{\varepsilon\to 0} c_\varepsilon \geqq \sqrt{2\min(1,\Lambda^{-1})m}\,. \tag{9.19}$$

We have proven that there exists a sequence ε_n, decreasing such that

$$c_{\varepsilon_n}\to c. \tag{9.20}$$

Let us now look at the convergence of u_{ε_n}, v_{ε_n}; it will be convenient to distinguish three cases $x\leqq 0$, $0\leqq x\leqq x_0$, $x\geqq x_0$, where $x_0=(1-\theta)/\alpha$. As u_ε is strictly increasing (Theorem 8.11), $u_\varepsilon(x)\leqq\theta$ for $x\leqq 0$ and u_ε, v_ε satisfy $-u_\varepsilon''+c_\varepsilon u_\varepsilon'=-\Lambda v_\varepsilon''+c_\varepsilon v_\varepsilon'=0$ for $x\leqq 0$ and $u_\varepsilon(-\infty)=0$, $u_\varepsilon(+\infty)=1$, $u_\varepsilon(0)=\theta$, $u_\varepsilon'(0)=\theta c_\varepsilon$. Therefore, thanks to (9.20), u_{ε_n} and v_{ε_n} converges, for the H^1-topology, to u and v satisfying $-u''+cu'=-\Lambda v''+cv'=0$ for $x\leqq 0$ and $u(-\infty)=0$, $v(-\infty)=1$, $v(0)=\theta$, $v'(0)=\theta c$.

If $0\leqq x\leqq x_0$ we may apply Proposition 8.8, still true if $a=+\infty$, to prove with (9.18) that u_ε' and v_ε' are bounded in $L^2((0,x_0))$ independently of ε. Truly, we also have $0\leqq u_\varepsilon$, $v_\varepsilon\leqq 1$. Therefore, u_ε and v_ε are bounded in $H^1((0,x_0))$, independently of ε. By compactness, eventually taking a new sequence ε_n, we deduce with (9.75)

$$x_{\varepsilon_n}\to\bar{x}\leqq x_0, \tag{9.21}$$

$$u_{\varepsilon_n}, v_{\varepsilon_n}\to u, v \quad \text{in } C^0([0,x_0]). \tag{9.22}$$

Since $\theta_{\varepsilon_n}\leqq u_{\varepsilon_n}(x)<1$ for $\bar{x}\leqq x$, then $u(x)=1$ and ((9.10)) $v(x)=0$. On the other hand, from (9.10), we get that the function $x\to f_{\varepsilon_n}(u_{\varepsilon_n}(x))v_{\varepsilon_n}(x)$ is bounded from above from $\max(1,\Lambda^{-1})f_{\varepsilon_n}(u_{\varepsilon_n}(x))(1-u_{\varepsilon_n}(x))$, and, using (9.4), converge to 0 uniformly over any compact set of $[0,\bar{x}]$. Therefore $-u''+cu'=-\Lambda v''+cv'=0$ for $x\leqq\bar{x}$. Finally, for $x\geqq x_0$ we have $\theta_{\varepsilon_n}\leqq u_{\varepsilon_n}(x)<1$ and using (9.10) we obtain

$$u_{\varepsilon_n}\to 1 \quad \text{in } C^0([x_0,\infty)), \tag{9.23}$$

$$v_{\varepsilon_n}\to 0 \quad \text{in } C^0([x_0,+\infty)). \tag{9.24}$$

To summarize, the functions u,v are of class C^2 except at the point $\bar{x}$: if $x\leqq\bar{x}$, they satisfy $-u''+cu'=-\Lambda v''+cv'=0$ and $u(x)=e^{c(x-\bar{x})}$, $v(x)=1-e^{c(x-\bar{x})/\Lambda}$; if $x\geqq\bar{x}$ $u(x)=1$, $v(x)=0$. The proof is complete. □

We now address the questions of uniqueness. The relevant result is as follows (see [13], [19], and [14] for a rigorous proof):

THEOREM 9.8. *Let the assumptions* (7.3), (7.4) *be satisfied. If, in addition,* $0<\Lambda<1$, *then the solution* (u,v,c) *of* (8.1) *is unique. As a consequence, we may precise the result of Theorem* 9.4.

COROLLARY 9.9. *Let the assumptions of Theorem* 9.4 *be satisfied; moreover, assume* $0<\Lambda<1$. *Then, for any sequence* $\{\varepsilon_n\}_{n\in N}$, *decreasing to* $(u_{\varepsilon_n},v_{\varepsilon_n},c_{\varepsilon_n})$ *converge (strongly) for the topology* $H^1(\mathbb{R})\times H^1(\mathbb{R})\times\mathbb{R}$ *to the unique* (u,v,c) *solution of* (9.6)–(9.9).

The proof of Theorem 9.8 has been sketched first by Kanel' [13]. A precise form of it can be found in the book of Zeldovich [19] and in Marion [14].

10. Remarks on the case of *n*th order reaction. In this section we mention the extension of the previous results to the case of a single step reaction of order n, where n is an integer greater than 1. Precisely, we are looking for (u,v,c) solution to the following problem:

$$\begin{aligned} &-u''+cu'=f(u)v^n \quad \text{on } \mathbb{R}, \\ &-\Lambda v''+cv'=-f(u)v^n \quad \text{on } \mathbb{R}; \end{aligned} \tag{10.1}$$

$$\begin{aligned} &u(-\infty)=0, \qquad u(+\infty)=1, \\ &v(-\infty)=1, \qquad v(+\infty)=0 \\ &u(0)=\theta. \end{aligned} \tag{10.2}$$

The function f always satisfies the assumptions (7.3), (7.4). Indeed, it depends also on ε and we address the question of the asymptotic behavior of (u,v,c) as in §9. The relevant assumptions that generalize (9.4), (9.5) are now:

there exists θ_ε, $\theta \leqq \theta_\varepsilon < 1$, with $\lim_{\varepsilon\to 0}\theta_\varepsilon = 1$, such that

$$\lim_{\varepsilon\to 0}\max_{\theta\leqq s\leqq\theta_\varepsilon} f_\varepsilon(s)(1-s)^n=0 \tag{10.3}$$

$$\lim_{\varepsilon\to 0}\int_0^1 f_\varepsilon(s)(1-s)^n ds=\lim_{\varepsilon\to 0}G_\varepsilon(1)=m<+\infty. \tag{10.4}$$

We can now state:

THEOREM 10.1. *Under assumptions* (7.3), (7.4), *there exists one solution of problem* (10.1), (10.2) *such that*

$$\begin{aligned} &0\leqq u,v\leqq 1 \quad \textit{on } \mathbb{R} \\ &0<v'\leqq c, \qquad -\frac{c}{\Lambda}\leqq v'<0 \quad \textit{on } \mathbb{R} \\ &u,v\in W^{2,\infty}(R). \end{aligned}$$

Moreover, there exist two constants $\underline{c},\bar{c}$, *independent of* ε, *for which* $0<\underline{c}\leqq c\leqq \bar{c}<+\infty$.

THEOREM 10.2. *Under assumptions* (7.3), (7.4) *and* (10.3), (10.4), *the conclusion of Theorem* 10.1 *still holds for problem* (10.1), (10.2).

The proofs follow arguments similar to those found in the preceding sections. The only new argument is the distinction between n odd or even in proving Proposition 8.1.

11. The precise value of $c=\lim_{\varepsilon\to 0}c_\varepsilon$; rigorous internal layer analysis. Here we address the question left open in §9, that is, the value of $c=\lim_{\varepsilon\to 0}c_\varepsilon$. For simplicity, we will consider the case where

$$f_\varepsilon(u)=\frac{1}{\varepsilon^2}\phi\left(\frac{u-1}{\varepsilon}\right), \qquad u>\theta, \tag{11.1}$$

for ϕ satisfying (9.2), (9.3). We have seen (Remark 9.1) that this choice of f_ε is the most important in practice.

THEOREM 11.1. *Set* $f_\varepsilon(u)=(1/\varepsilon^2)\phi(u-1/\varepsilon)$ *if* $u>\theta$ *and* $f_\varepsilon(u)\equiv 0$ *if* $u<\theta$. *Assume* (9.2), (9.3) *and local Lipschitz continuity as previously discussed. Moreover, assume that on some bounded interval* $[-L,0]$, $\phi(\sigma)$ *possesses a finite number of extrema*; L *is some positive constant defined in* (A.5) *of the Appendix. Then, the conclusion of Theorem* 9.4 *holds, and specifically*

$$c=\lim_{\varepsilon_n\to 0}c_{\varepsilon_n}=\sqrt{2m/\Lambda}, \tag{11.2}$$

where (*see* (9.3))

$$m \equiv \int_{-\infty}^{0} -\sigma\phi(\sigma)\,d\sigma.$$

For the proof, we will need the following lemma, proved in the Appendix.

LEMMA 11.2. *Let* θ_ε, $\theta \leqq \theta_\varepsilon < 1$ *such that*

$$\lim_{\varepsilon\to 0} \theta_\varepsilon = 1, \qquad \frac{1-\theta_\varepsilon}{\varepsilon} = O(\varepsilon^{-\gamma}),$$

$0<\gamma<\frac{1}{2}$. *Then there exists* $0<\delta<1-2\gamma$, *such that as* $\varepsilon\to 0$:

$$(11.3) \qquad \int_{x_\varepsilon}^{+\infty} (1-u_\varepsilon(x))\,dx = O(\varepsilon^{1+\delta}),$$

$$(11.3)\text{ bis} \qquad \int_{x_\varepsilon}^{+\infty} v_\varepsilon(x)\,dx = O(\varepsilon^{1+\delta}).$$

Moreover, the local Shvab–Zeldovich variable is bounded as

$$(11.4) \qquad |u_\varepsilon(x)+\Lambda v_\varepsilon(x)-1| = O(\varepsilon^{1+\delta}),$$

uniformly for $x \geqq x_\varepsilon$.

Commentary. The spirit of the proof can be better understood if we switch to the usual internal layer rescaled variables

$$\hat{u}_\varepsilon = \frac{u_\varepsilon - 1}{\varepsilon}, \qquad \hat{v}_\varepsilon = \frac{v_\varepsilon}{\varepsilon}.$$

Then the choice of θ_ε corresponds to a $\hat{\theta}_\varepsilon$ going to $-\infty$ as $\varepsilon^{-\gamma}$. To center the internal layer, we choose a point η_ε such that $u_\varepsilon(y_\varepsilon) \equiv \eta_\varepsilon$, with $\lim_{\varepsilon\to 0}((1-\eta_\varepsilon)/\varepsilon) = L$. Equivalently, $\lim_{\varepsilon\to 0}\hat{\eta}_\varepsilon = -L$, where $L>0$ is defined in (A.5). This leads to the following specific stretching:

$$\xi = \frac{x-\bar{y}}{\varepsilon},$$

where $\bar{y}$ is the finite limit of the monotone sequence $\{y_\varepsilon\}_{\varepsilon>0}$. Then, the proof relies heavily on the uniform boundedness for $\xi \geqq (x_\varepsilon - \bar{y})/\varepsilon$ of the local Shvab–Zeldovich variable:

$$|(\hat{u}_\varepsilon + \Lambda\hat{v}_\varepsilon)(\xi)| = O(\varepsilon^\delta).$$

This type of estimate has been taken for granted in formal internal layer analysis [4]. The demonstration of Lemma 11.2 corresponds to a nonclassical singular perturbation analysis; indeed, in terms of rescaled variables and coordinates the system read as:

$$-(\hat{u}_\varepsilon)_{\xi\xi} + \varepsilon c_\varepsilon (u_\xi)_\xi = v_\varepsilon \phi(u_\varepsilon),$$
$$-\Lambda(\hat{v}_\varepsilon)_{\xi\xi} + \varepsilon c_\varepsilon(\hat{v}_\varepsilon)_\xi = -\hat{v}_\varepsilon\phi(\hat{u}_\varepsilon).$$

Interestingly, to obtain the sup norm estimates for the local Shvab–Zeldovich variable, we must distinguish between $\xi \leqq -L$ and $-L \leqq \xi \leqq +\infty$. In his thesis, Joulin [11] made some formal remarks along the same lines. □

Proof of Theorem 11.1. We start with the identity

$$\frac{1}{2}\left|u'_\varepsilon(x_\varepsilon)\right|^2+c_\varepsilon\int_{x_\varepsilon}^{-\infty}\left|u'_\varepsilon(x)\right|^2dx \tag{11.5}$$
$$=\frac{1}{\Lambda}\int_{x_\varepsilon}^{+\infty}f_\varepsilon(u_\varepsilon(x))(1-u_\varepsilon(x))u'_\varepsilon(x)\,dx$$
$$+\frac{1}{\Lambda}\int_{x_\varepsilon}^{+\infty}f_\varepsilon(u_\varepsilon(x))(u_\varepsilon(x)+\Lambda v_\varepsilon(x)-1)u'_\varepsilon(x)\,dx,$$

which is obtained by integration between x_ε and $+\infty$ of $-u''_\varepsilon+c_\varepsilon u'_\varepsilon=f_\varepsilon(u_\varepsilon)v_\varepsilon$ multiplied by u'_ε. Using (9.16), we have by definition of c:

$$\lim_{\varepsilon\to0}\frac{1}{2}\left|u'_\varepsilon(x_\varepsilon)\right|^2=\frac{1}{2}c^2. \tag{11.6}$$

Taking u_ε as independent variables, we set $p_\varepsilon(u'_\varepsilon)=u'_\varepsilon$, and therefore

$$c_\varepsilon\int_{x_\varepsilon}^{+\infty}\left|u'_\varepsilon(x)\right|^2dx=c_\varepsilon\int_{\theta_\varepsilon}^{1}p_\varepsilon(u)\,du\leqq|c_\varepsilon|(1-\theta_\varepsilon)\sup_{\theta_\varepsilon<u\leqq1}|p_\varepsilon(u)|.$$

But c_ε and $p_\varepsilon(u)=u'_\varepsilon$ and bounded independently of ε (see (9.11) and (8.5)). Hence:

$$\lim_{\varepsilon\to0}c_\varepsilon\int_{x_\varepsilon}^{+\infty}\left|u'_\varepsilon(x)\right|^2dx\leqq C\lim_{\varepsilon\to0}(1-\theta_\varepsilon)=0. \tag{11.7}$$

Using u_ε again as an independent variable, we have:

$$\frac{1}{\Lambda}\int_{x_\varepsilon}^{+\infty}f_\varepsilon(u_\varepsilon(x))(1-u_\varepsilon(x))u'_\varepsilon(x)\,dx=\frac{1}{\Lambda}\int_{(\theta_\varepsilon-1)/\varepsilon}^{0}-\sigma\phi(\sigma)\,d\sigma.$$

Hence, with the assumption on θ_ε and the definition of m

$$\lim_{\varepsilon\to0}\frac{1}{\Lambda}\int_{x_\varepsilon}^{+\infty}f_\varepsilon(u_\varepsilon(x))(1-u_\varepsilon(x))u'_\varepsilon(x)\,dx=\frac{m}{\Lambda}. \tag{11.8}$$

It remains to handle the last term of (11.5). Clearly, if u_ε, v_ε are solutions of (8.1), then $u'_\varepsilon(x)+\Lambda v'_\varepsilon(x)=c_\varepsilon(u_\varepsilon+v_\varepsilon(x)-1)$ and integrating this relation we get:

$$u_\varepsilon(x)+\Lambda v_\varepsilon(x)-1=-c_\varepsilon\int_x^{+\infty}(u_\varepsilon(y)+v_\varepsilon(y)-1)\,dy.$$

From the former identity and (8.9), we deduce:

$$|u_\varepsilon(x)+\Lambda v_\varepsilon(x)-1|\leqq\frac{|\Lambda-1|}{\Lambda}c_\varepsilon\int_x^{+\infty}(1-u_\varepsilon(y))\,dy.$$

Consequently, since $f(u_\varepsilon(x))u'_\varepsilon(x)>0$:

$$|I_\varepsilon|\equiv\left|\frac{1}{\Lambda}\int_{x_\varepsilon}^{+\infty}f_\varepsilon(u_\varepsilon(x))(u_\varepsilon(x)+\Lambda v_\varepsilon(x)-1)u'_\varepsilon(x)\,dx\right|$$
$$\leqq\frac{|\Lambda-1|}{\Lambda^2}|c_\varepsilon|\int_{x_\varepsilon}^{+\infty}f_\varepsilon(u_\varepsilon(x))u'_\varepsilon(x)\left(\int_x^{+\infty}(1-u_\varepsilon(y))\,dy\right)dx.$$

But c_ε is bounded using (9.18), (9.19) and $\int_x^{+\infty}(1-u_\varepsilon(y))\,dy$ is decreasing, so there exists C, independent of ε such that:

$$|I_\varepsilon| \leqq C\int_{x_\varepsilon}^{+\infty} (1-u_\varepsilon(y))\,dy \cdot \int_{x_\varepsilon}^{+\infty} f_\varepsilon(u_\varepsilon(x))u_\varepsilon'(x)\,dx.$$

And so finally, with (11.1):

$$(11.9) \qquad |I_\varepsilon| \leqq \frac{C}{\varepsilon}\int_{(\theta_\varepsilon-1)/\varepsilon}^{0} \phi(\sigma)\,d\sigma \cdot \int_{x_\varepsilon}^{+\infty} (1-u_\varepsilon(y))\,dy.$$

One easily verified by using (9.2), (9,3), that

$$\int_{(\theta_\varepsilon-1)/\varepsilon}^{0} \phi(\sigma)\,d\sigma \leqq C'\int_{(\theta_\varepsilon-1)/\varepsilon}^{0} -\sigma\phi(\sigma)\,d\sigma,$$

where the constant C' depends only on ϕ. Hence, as $\lim_{\varepsilon\to 0}((\theta_\varepsilon-1)/\varepsilon) = -\infty$, and with (9.8):

$$(11.10) \qquad \lim_{\varepsilon\to 0}|I_\varepsilon| \leqq CC'm \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_{x_\varepsilon}^{+\infty} (1-u_\varepsilon(y))\,dy.$$

The conclusion (11.2) of the theorem follows then from (11.5) with (11.6)–(11.8) and (11.10) using (11.3). □

Appendix. Here we prove Lemma 11.2, keeping the notation of §11.

LEMMA. *Let* θ_ε, $\theta \leqq \theta_\varepsilon < 1$ *such that*

$$\lim_{\varepsilon\to 0}\theta_\varepsilon = 1, \qquad \frac{\theta_\varepsilon - 1}{\varepsilon} = O(\varepsilon^{-\gamma}),$$

$0<\gamma<\frac{1}{2}$. *Then there exists* $0<\delta<1-2\gamma$ *such that, as* $\varepsilon\to 0$:

$$(\text{A.1}) \qquad \int_{x_\varepsilon}^{+\infty} (1-u_\varepsilon(x))\,dx = O(\varepsilon^{1+\delta}),$$

$$(\text{A.2}) \qquad \int_{x_\varepsilon}^{+\infty} v_\varepsilon(x)\,dx = O(\varepsilon^{1+\delta}),$$

where $(u_\varepsilon, v_\varepsilon)$ *is a solution of* (8.1) *and* x_ε *is defined by* $u_\varepsilon(x_\varepsilon) = \theta_\varepsilon$.

Proof. The proof is lengthy; it consists of breaking the integral in (A.1) into two parts that are estimated separately. Precisely we write

$$(\text{A.3}) \qquad \int_{x_\varepsilon}^{+\infty} (1-u_\varepsilon(y))\,dy = \int_{x_\varepsilon}^{y_\varepsilon} (1-u_\varepsilon(y))\,dy + \int_{y_\varepsilon}^{+\infty} (1-u_\varepsilon(y))\,dy,$$

where we specifically choose y_ε such that i) $u_\varepsilon(y_\varepsilon) = \eta_\varepsilon$, ii) $\lim_{\varepsilon\to 0}\eta_\varepsilon = 1$, iii) $\lim_{\varepsilon\to 0}((\eta_\varepsilon - 1)/\varepsilon) = -L$ where $L>0$ is chosen below. We can find $\varepsilon_0>0$, such that, for each $0<\varepsilon\leqq\varepsilon_0$, $\theta_\varepsilon<\eta_\varepsilon$; therefore, as $u_\varepsilon'(x)>0$, we have $x_\varepsilon<y_\varepsilon$ for $\varepsilon\leqq\varepsilon_0$. Now we prove that y_ε is bounded independently of ε. We proceed as for the proof of (9.15), starting from the identity (9.14), where now $x\leqq y_\varepsilon$. with (9.10), we get:

$$-\frac{1}{2}|u_\varepsilon'(x)|^2 + \frac{1}{2}c_\varepsilon^2\theta^2 \leqq \max(1,\Lambda^{-1})G_\varepsilon(\eta_\varepsilon), \qquad 0<x\leqq y_\varepsilon$$

or, using (9.11):

$$\text{(A.4)}\qquad \min(1,\Lambda^{-1})G_\varepsilon(1)\theta^2-\max(1,\Lambda^{-1})G_\varepsilon(\eta_\varepsilon)\leqq\frac{1}{2}\left|u_\varepsilon'(x)\right|^2,\qquad 0<x\leqq y_\varepsilon.$$

Now there exists α_0, independent of ε, with $0<\alpha_0<\frac{1}{2}$ such that

$$\text{(A.5)}\qquad G_\varepsilon(\eta_\varepsilon)\leqq\frac{\min(1,\Lambda^{-1})}{\max(1,\Lambda^{-1})}\alpha_0 G_\varepsilon(1).$$

Note that $\min(1,\Lambda^{-1})/\max(1,\Lambda^{-1})<1$, if $\Lambda\neq 1$, and that

$$G_\varepsilon(u)=\int_{(\theta-1)/\varepsilon}^{(u-1)/\varepsilon}-\sigma\phi(\sigma)\,d\sigma$$

is strictly increasing; it suffices to choose the constant $L=-\lim_{\varepsilon\to 0}((\eta_\varepsilon-1)/\varepsilon)$ large enough. Therefore, from (A.4), we deduce

$$\min(1,\Lambda^{-1})G_\varepsilon(1)\left(\theta^2-\alpha_0\right)\leqq\frac{1}{2}\left|u_\varepsilon'(x)\right|^2,\qquad 0<x\leqq y_\varepsilon.$$

Taking L eventually larger, we can keep α_0 such that $\theta^2-\alpha_0>0$, and therefore for ε small enough there exists $\alpha'>0$, independent of ε, such that

$$\text{(A.6)}\qquad u_\varepsilon'(x)\geqq\alpha',\qquad 0<x\leqq y_\varepsilon.$$

Integration between 0 and y_ε gives ($\eta_\varepsilon<1$)

$$\text{(A.7)}\qquad y_\varepsilon\leqq\frac{1-\theta}{\alpha'}\equiv y_0.$$

Now we come back to (A.3); for the first term in the right-hand side we have

$$\begin{aligned}\int_{x_\varepsilon}^{y_\varepsilon}(1-u_\varepsilon(y))\,dy&\leqq(y_\varepsilon-x_\varepsilon)\sup_{x_\varepsilon\leqq y\leqq y_\varepsilon}(1-u_\varepsilon(y))\\&\leqq(y_\varepsilon-x_\varepsilon)(1-\theta_\varepsilon)\\&\leqq\frac{1}{\alpha'}(\eta_\varepsilon-\theta_\varepsilon)(1-\theta_\varepsilon)\\&\leqq\frac{1}{\alpha'}(1-\theta_\varepsilon)^2,\end{aligned}$$

where we use (A.6) and the definitions of x_ε, y_ε. By the choice of θ_ε, we have $1-\theta_\varepsilon=O(\varepsilon^{1-\gamma})$, with $0<\gamma<\frac{1}{2}$, so there exists δ, $0<\delta<1-2\gamma$ such that:

$$\text{(A.8)}\qquad \int_{x_\varepsilon}^{y_\varepsilon}(1-u_\varepsilon(y))\,dy=O(\varepsilon^{1+\delta}).$$

For the second term in the right-hand side of (A.3) we will show, using (9.10), that $\int_{y_\varepsilon}^{+\infty}v_\varepsilon(x)\,dx=O(\varepsilon^{1+\delta})$ (indeed (A.1) and (A.2) are equivalent through (9.10)). This will be achieved by using an appropriate energy identity. We multiply $-u_\varepsilon''+c_\varepsilon u_\varepsilon'=f_\varepsilon(u_\varepsilon)v_\varepsilon$ by $1/f_\varepsilon(u_\varepsilon)$ and integrate from y_ε to $+\infty$ (note that $u_\varepsilon(x)>\theta$ for $y_\varepsilon\leqq x\leqq+\infty$; so

$f_\varepsilon(u_\varepsilon(x))>0$ by assumption). The computation is straightforward; we get

$$\text{(A.9)}\quad \int_{y_\varepsilon}^{+\infty} v_\varepsilon(x)\,dx=\frac{c_\varepsilon}{f_\varepsilon(1)}+\frac{1}{f_\varepsilon(u_\varepsilon(y_\varepsilon))}(u'_\varepsilon(y_\varepsilon)-c_\varepsilon u_\varepsilon(y_\varepsilon))$$
$$+c_\varepsilon(F_\varepsilon(1)-F_\varepsilon(u_\varepsilon(y_\varepsilon)))-\int_{y_\varepsilon}^{+\infty}\frac{f'_\varepsilon(u_\varepsilon(x))}{(f_\varepsilon(u_\varepsilon(x)))^2}|u'_\varepsilon(x)|^2\,dx,$$

where, by definition, $F'_\varepsilon(u)=uf'_\varepsilon(u)/f_\varepsilon(u)^2$. We shall bound the four terms in the right-hand side of (A.9). Recall first that $f_\varepsilon(u)=(1/\varepsilon^2)\phi((u-1)/\varepsilon)$. So

$$\text{(A.10)}\qquad \left|\frac{c_\varepsilon}{f_\varepsilon(1)}\right|=\left|\frac{\varepsilon^2 c_\varepsilon}{\phi(0)}\right|=O(\varepsilon^2),$$

since c_ε is bounded using (9.11). Similarly

$$\frac{1}{f_\varepsilon(u_\varepsilon(y_\varepsilon))}=\frac{\varepsilon^2}{\phi((\eta_\varepsilon-1)/\varepsilon)}=\frac{\varepsilon^2}{\phi(-L)}+o(\varepsilon^2)$$

and using (8.22), (8.23), and (9.11):

$$\text{(A.11)}\qquad \left|\frac{1}{f_\varepsilon(u_\varepsilon(y_\varepsilon))}(u'_\varepsilon(y_\varepsilon)-c_\varepsilon u_\varepsilon(y_\varepsilon))\right|=O(\varepsilon^2).$$

Now, by a single computation, we get the expression of $F_\varepsilon(u)$:

$$\text{(A.12)}\quad F_\varepsilon(1)-F_\varepsilon(\eta_\varepsilon)=\varepsilon^3\int_{(\eta_\varepsilon-1/\varepsilon)}^{0}\frac{d\sigma}{\phi(\sigma)}-\varepsilon^2\left[\frac{u-1}{\phi((u-1)/\varepsilon)}+\frac{1}{\phi((u-1)/\varepsilon)}\right]_{u=\eta_\varepsilon}^{u=1}.$$

So, using the choice of η_ε and (9.11), we get again

$$|c_\varepsilon(F_\varepsilon(1)-F_\varepsilon(u_\varepsilon(y_\varepsilon)))|=O(\varepsilon^2).$$

Now, we bound from above the last term, using

$$\int\frac{f'_\varepsilon(u_\varepsilon(x))}{(f_\varepsilon(u_\varepsilon(x)))^2}u'_\varepsilon(x)\,dx=-\varepsilon^2\int d\left(\frac{1}{\phi(\sigma)}\right);$$

precisely

$$I_\varepsilon\equiv\left|\int_{y_\varepsilon}^{+\infty}\frac{f'_\varepsilon(u_\varepsilon(x))}{(f_\varepsilon(u_\varepsilon(x)))^2}\,dx\right|\leqq\varepsilon^2\sup_{x>0}u'_\varepsilon(x)\left[\sum_{j=0}^{N-1}\left|\frac{1}{\phi(\zeta_{j+1})}-\frac{1}{\phi(\zeta_j)}\right|\right],$$

where $\zeta_0\equiv(\eta_\varepsilon-1)/\varepsilon\underset{\varepsilon\to 0}{\to}-L$, $\zeta_N=0$, and ζ_j, $1\leqq j\leqq N-1$, are the extrema of ϕ. (We restrict ourselves to $\phi(\sigma)$ with a finite number of extrema, for $-L\leqq\sigma<0$.) Therefore, using (8.23) and (9.11), we get finally

$$\text{(A.13)}\qquad I_\varepsilon=O(\varepsilon^2).$$

From (A.9)–(A.13), we deduce

$$\text{(A.14)}\qquad \int_{y_\varepsilon}^{+\infty}v_\varepsilon(x)\,dx=O(\varepsilon^2).$$

We complete the proof using (9.10) with (A.8) and (A.14). □

REFERENCES

[1] D. G. Aronson and H. F. Weinberger, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. Math., 30 (1978), pp. 33–76.

[2] H. Berestycki, P. L. Lions, and L. A. Peletier, *An* ODE *approach to the existence of positive solutions for semilinear problems in* $\mathbb{R}^n$, Indiana Univ. Math. J., 30 (1981), pp. 141–157.

[3] W. B. Bush and F. E. Fendell, *Asymptotic analysis of laminar flame propagation for general Lewis numbers*, Combustion Sci. and Tech., 1 (1970), pp. 421–8.

[4] J. Buckmaster and G. S. S. Ludford, *Theory of Laminar Flames*, Cambridge Univ. Press, Cambridge, 1982.

[5] P. C. Fife, *Mathematical Aspects of Reacting and Diffusing Systems*, Lecture Notes in Biomathematics 28, Springer-Verlag, New York, 1979.

[6] P. C. Fife, *Propagating fronts in reactive media*, Nonlinear Problems: Present and Future, Mathematics Studies 61, 267–285, A. Bishop, D. Campbell and B. Nicolaenko, eds., North-Holland, Amsterdam, 1982.

[7] P. C. Fife and B. Nicolaenko, *The singular perturbation approach to flame theory with chain and competing reactions*, Ordinary and Partial Differential Equations, W. N. Everitt and B. D. Sleeman, eds., Lecture Notes in Mathematics 962, Springer-Verlag, New York, 1980, pp. 232–250.

[8] ———, *Asymptotic flame theory with complex chemistry*, Contemporary Math. 17, AMS 235.

[9] ———, *Flame fronts with complex chemical networks*, Proc. CNLS Conference on Fronts, Interfaces, and Patterns, Physica 12, D (1984), pp. 3–18.

[10] W. E. Johnson and W. Nachbar, *Laminar flame theory and the steady linear burning of a monopropellant*, Arch. Rat. Mech. Anal., 12 (1963), pp. 58–91. See also W. E. Johnson, *On a first order boundary value problem for laminar flame theory*, Arch. Rat. Mech. Anal., 13 (1963), pp. 46–54.

[11] G. Joulin, *Existence, stabilité et structuration des flames prémélangées*, Thesis, Univ. de Poitiers, U.E.R.-E.N.S.M.A., France, 1979.

[12] Ja. I. Kanel', *Stabilization of solutions of the Cauchy problem for equations encountered in combustion theory*, Mat. Sbornik, 59(1962), pp. 245–288.

[13] ———, *On steady state solutions to systems of equations arising in combustion theory*, Dokl. Akad. Nauk USSR 149, 2 (1963), pp. 367–369.

[14] M. Marion, *Sur les équations de flamme laminaire sans température d'ignition*, Thesis, Univ. de Paris, France, 1983.

[15] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[16] M. D. Smooke, J. A. Miller, and R. J. Kee, *Determination of adiabatic flame speeds by boundary value methods*, Combustion Sci. and Technology, 34(1983), pp. 79–90.

[17] F. Williams, *Combustion Theory*, Addison-Wesley, Reading MA, 1983.

[18] Ya. B. Zeldovich, *On the theory of flame propagation*, J. Phys. Chem., 22(1948), pp. 27–48.

[19] Ya. B. Zeldovich, G. I. Barenblatt, V. B. Librovich, and G. M. Mahviladze, *Mathematical Theory of Combustion and Detonation*, Nauka, Moscow, 1980. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

007

AN INTRODUCTION TO THE TECHNIQUE OF RECONSTITUTION*

A. J. ROBERTS[†]

Abstract. In many physical problems it is recognised that the solution is dominated by a particular structure. It is usually possible to derive a differential equation which approximately describes the spatial and/or temporal evolution of this dominant structure. Such an evolution equation is valid only for a limited range of the parameters. The technique of reconstitution provides a rationale and a mechanistic method for correcting such evolution equations by including terms representing higher-order physical interactions and thus to significantly extend the parameter range in which the equation is valid. To obtain some feel for how this technique works it is applied to a simple pair of coupled nonlinear differential equations. The results show clearly how the solutions to the approximate evolution equations, of varying accuracy, relate to the exact solution of the full problem.

1. Introduction. In slowly varying wave theory the oscillations on the scale of a wavelength are explicitly known and then asymptotic theory gives an equation which describes the evolution of the wave's amplitude and phase on a longer scale. This derivation of an equation governing the evolution over a long scale of the bulk properties (the wave envelope) of a known structure (periodic progressive waves) is typical of many problems. Shallow water theory expresses the velocities in the fluid as a polynomial in the vertical coordinate and the asymptotic theory then gives a wave equation for the approximate evolution of the (relatively long scale) horizontal structure. Convection with boundary conditions of fixed heat flux, discussed by Chapman and Proctor (1980), also evolves on a long horizontal scale with a known vertical structure. In all these cases assumptions are made which enable part of the solution's structure to be explicitly calculated while the evolution of this structure over space and/or time is governed by some relatively simple differential equations, called an evolution equation. The assumptions typically invoked are those of long-space and/or slow-time scales upon which the known structure (usually of small amplitude) varies. The resulting leading-order evolution equation can only approximate the dynamics that are present in the exact solutions of the original problem. The aim of this section is to illustrate and extend a method, proposed by Spiegel (1981), for correcting such evolution equations by adding extra terms which bring new physics into the equation.

Such corrections are by no means unknown in fluid mechanics. The Navier–Stokes equation itself may either be derived by making plausible assumptions about the stress-strain relationship, or from the kinetic theory of gases (see Vincenti and Kruger (1965)). If the latter course is taken then the first approximation, which assumes that the properties of the gas are in equilibrium, is just the Euler equation. In deriving the second approximation one assumes that the properties are nearly in equilibrium and we find a viscous dissipation term which provides the nontrivial correction to obtain the Navier–Stokes equation. The equations governing the propagation of shallow water waves is another example (Peregrine (1972, §3, 4)). The first approximation results in a nonlinear nondispersive wave equation which essentially rests on the assumption that the horizontal velocity is uniform in the vertical. The second approximation allows for

* Received by the editors November 29, 1983, and in revised form August 29, 1984.

[†] D.A.M.T.P., Silver St., Cambridge CB3 9EW, England. Present address: Department of Applied Mathematics, University of Adelaide, GPO Box 498, South Australia, 5001.

some vertical variation of horizontal velocity and results in the more realistic nonlinear dispersive Boussinesq shallow water equations. In a similar vein, Dysthe (1979) has modified the nonlinear Schrödinger equation which governs the evolution of modulations to a uniform train of water waves. He added terms to produce an equation which is of fourth order accuracy in the wave slope. Stuart (1960) and Watson (1960) investigated the stability of plane Poiseuille and Couette flow by deriving a differential equation for the time dependence of a specific spatial perturbation to the flow. The differential equations they derived contain nonlinear terms of higher asymptotic order than the linear terms which form the equation usually used in a stability analysis. The derivations of the appropriate corrections in these examples have relied upon heuristic arguments. Based on the assumptions made to derive the leading order evolution equation, the technique of reconstitution (which is examined here) provides a rationale for systematically making corrections to this first approximation.

The procedure we adopt is as follows (details can be clarified by reading §§3, 4, 5). We expand the unknowns in a perturbation expansion in some small parameter (for example, the wave steepness), introducing slow space-time scales if appropriate, and then substitute the expansion into the full equation and group terms with like powers of the small parameter (see §3). Considering in turn increasing powers of the small parameter, we can find the dependence upon some of the independent "fast" variables, leaving the dependence upon the other "slow" variables arbitrary. In the calculation of the fast dependence at higher orders, we typically find a solvability condition that gives an equation which governs the evolution of the previously found structure over the slow scales. Conventionally the solutions to each solvability condition are then multiplied by appropriate powers of the small parameter, and added up to give an approximate answer (§3). Instead, reconstitution takes the solvability conditions, multiplies them by appropriate powers of the small parameter, adds them together and writes the resultant equation solely in terms of the original unscaled, unexpanded variables (§§4, 5). Hence we form just one equation which contains all the information that was previously contained separately in the different solvability conditions. The advantage of the procedure of reconstitution is that the physical processes which previously could only slightly modify the leading-order solution can now interact with the dominant dynamics of the leading-order evolution equation.

The guiding principle of reconstitution can now be stated. A reconstituted equation is one which, upon substitution of a scaling and an expansion, gives a set of equations which are exactly equivalent to the solvability conditions derived from the original equations through the same scaling and expansion. Due to algebraic complexity we are restricted in practice to requiring that this equivalence only hold up to some finite order (usually a low order).

The above principle is akin to that embodied in the use of Padé approximates to sum a Taylor's series where a rational function is calculated which has exactly the same Taylor's series (to some order) as the derived series (see Bender and Orszag (1978, Chap. 8)). However, there is a far greater degree of freedom in the use of reconstitution than there is in Padé approximates. In the derivation of the reconstituted equation a choice has to be made between the plethora of permissible forms. Unless other forms can be justified we only consider reconstituted equations which are in the form of the derived solvability conditions. Moreover, only those terms which are forced to be present due to their appearance in the solvability conditions are included in the equation. The only justification of this choice is the idea that this produces a reconstituted equation with the advantage of having the most direct connection to the original full equations (Occam's razor also favours this criterion). But, as we shall see, this still permits a choice from a variety of equations.

Our task here is to elucidate the properties of reconstituted equations by applying the technique to a simple nonlinear problem. The full equations are introduced in §2 together with their exact solutions which may be used later for comparison with the approximate solutions. In §3 the unknowns are expanded in a power series in a small parameter and the first few solvability conditions are derived. These equations are then solved directly to give the conventional solution which is briefly discussed. We then move on to discuss the two principal types of reconstitution. The first type (which was the original proposal of Spiegel (1981)) additionally requires that the reconstituted equation is of the same differential order as the leading-order solvability condition and is examined in §4 via two examples. The higher order derivatives in the higher order solvability conditions are eliminated by using derivatives of the lower order solvability conditions. However, if the solvability conditions are partial differential equations then in general we cannot eliminate higher derivatives in all independent variables. So in §5 we investigate the nature of the second type of reconstitution which allows equations of higher differential order than the leading-order solvability condition.

2. Properties of the model equation. We want to look at an equation where the dependence of the solution can be separated between two disparate scales. One way of achieving this is to look at an equation which possesses a simple bifurcation. Near the bifurcation one component in the solution will be of marginal stability and hence its evolution will take place over a time scale much longer than that of the other components. We consider the following pair of coupled, nonlinear ordinary differential equations

$$\frac{da}{dt}=ra-ab, \tag{2.1a}$$

$$\frac{db}{dt}=-b+a^2, \tag{2.1b}$$

where r is a parameter of the problem. Near the bifurcation and for small amplitudes the b component evolves to zero on a fast time scale of order 1, while the a component evolves on the much slower time scale of order $1/r$.

We first investigate the behaviour of the exact solutions to equations (2.1). Since it is a second order autonomous system the nature of the solutions can be easily understood by looking at the fixed points, their stability and at the trajectories in the (a,b) plane. For $r\leqq 0$ there exists exactly one fixed point at $(a,b)=(0,0)$, which is stable. For $r>0$ there exists three fixed points: an unstable one at $(a,b)=(0,0)$ and two stable fixed points at $(a,b)=(\pm\sqrt{r},r)$.

The detailed nature of the solutions near the fixed points is of interest and so we look at the behaviour of small perturbations to (a,b) away from the fixed points. The time dependence of perturbations have the form $e^{\lambda t}$ where for the fixed point at the origin the two values of λ are

$$\lambda=r \quad\text{and}\quad \lambda=-1; \tag{2.2}$$

while for the two finite amplitude fixed points

$$\lambda=-(1\pm\sqrt{1-8r})/2. \tag{2.3}$$

From (2.3) we see that the finite amplitude fixed points are always stable. However, at finite r the asymptotic nature of the fixed point changes from being a node for $r<1/8$ to being a focus for $r>1/8$. This is an important qualitative change in behaviour which will be referred to in the following analysis.

The above information about the exact solutions is shown in Fig. 1 which displays the trajectories in the (a,b) plane for some sample values of r.

3. The solvability conditions and their direct solution. We now proceed to find approximate solutions to equation (2.1) which may then be compared with the exact solutions. To do this we treat r as a small parameter and define the parameter ε (also small) such that

$$\varepsilon^2=r, \qquad r,\varepsilon\geqq 0. \tag{3.1}$$

We then expand the unknowns in the following Taylor's series in ε

$$a(t)=\varepsilon A_0(s)+\varepsilon^3A_1(s)+\varepsilon^5A_2(s)+\cdots, \tag{3.2a}$$

$$b(t)=\varepsilon^2B_0(s)+\varepsilon^4B_1(s)+\varepsilon^6B_2(s)+\cdots, \tag{3.2b}$$

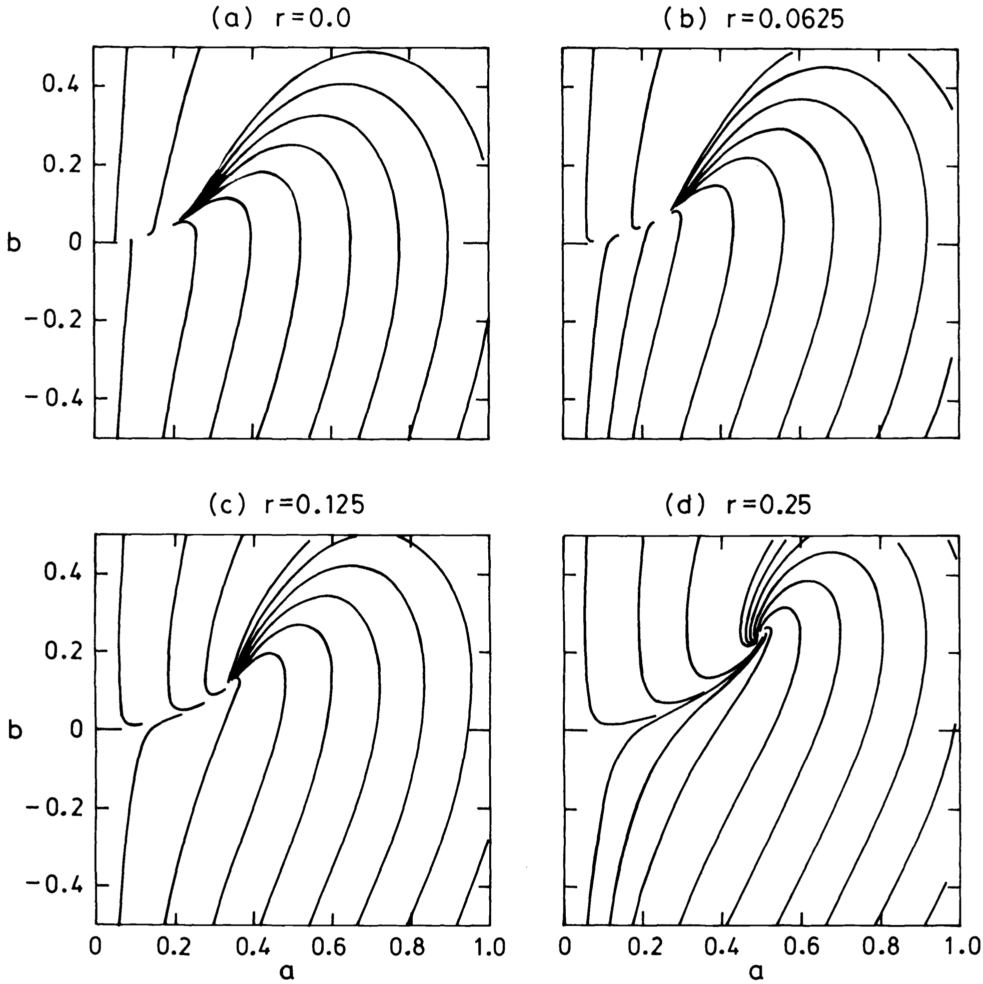


FIG. 1. *Trajectories in the* (a,b) *plane of the exact solutions* (2.1) *for four different values of* r.

where s is a slow time scale defined by

$$s=\varepsilon^2 t. \tag{3.3}$$

There are two ways of motivating the above scalings. The first is to observe that the finite amplitude fixed points are located at $(\pm\varepsilon,\varepsilon^2)$ and so we expect interesting effects to occur on this scale over an appropriate time scale. The second is to expand the solution vector in ascending powers of ε, say

$$(a,b)=\varepsilon(a,b)_0+\varepsilon^2(a,b)_1+\cdots,$$

then at the first order we derive an equation whose general solution is $(a,b)_1=(A_0(s),0)$. Thus, in the terminology used in the introduction, $(1,0)$ is a first approximation of the fast structure in the solution and $A_0(s)$ allows for variations on the slow time. At higher orders in this more general scheme we find that the expansion (3.2) is sufficient.

After substituting the expansions into (2.1) the solvability conditions appear very simply. Alternate powers of ε give alternately an explicit equation for B_n in terms of $A_0,\cdots,A_n$ and a first order differential equation for A_n. The first three members of the set of equations for A_n are

$$-A_0'+A_0-A_0^3=0, \tag{3.4a}$$

$$-A_1'+A_1-3A_0^2A_1+2A_0^2A_0'=0, \tag{3.4b}$$

$$-A_2'+A_2-3A_0^2A_2+4A_0A_1A_0'+2A_0^2A_1'-2A_0A_0'^2-2A_0^2A_0''=0, \tag{3.4c}$$

where $(\ )'$ denotes differentiation with respect to the slow time, i.e. d/ds. Equation (3.4a) is graced with the title of an evolution equation, while equation (3.4b) and (3.4c) merely describe corrections to the evolution equation's solution. The first three equations for B_n are correspondingly

$$B_0=A_0^2, \tag{3.5a}$$

$$B_1=2A_0A_1-2A_0A_0', \tag{3.5b}$$

$$B_2=2A_0A_2+A_1^2-2A_0A_1'-2A_1A_0'+2A_0'^2+2A_0A_0''. \tag{3.5c}$$

Equations (3.4) govern the dynamics of the approximate solutions while equations (3.5) fill out the structure of the solution by giving the forced mode b in terms of the A_n.

Equations (3.4) and (3.5) will form the basis for the reconstitution technique used in the next few sections. But for later comparison we should look first at the direct solutions of (3.4) and (3.5). It is easy to find explicit solutions for A_0 and A_1. Because the equations are autonomous a constant of integration can be absorbed by a suitable choice of the time origin. The solution for A_0 may then be written in one of the two forms

$$A_0^2=1/(1+e^{-2s}), \qquad 0<A_0<1, \tag{3.6a}$$

$$A_0^2=1/(1-e^{-2s}), \qquad 1<A_0. \tag{3.6b}$$

The corresponding general solutions for A_1 are

$$A_1=A_0'\log(e^{2s}+1)+\alpha A_0', \qquad 0<A_0<1, \tag{3.7a}$$

$$A_1=A_0'\log(e^{2s}-1)+\alpha A_0', \qquad 1<A_0, \tag{3.7b}$$

where α is an arbitrary constant of integration. These solutions are illustrated in Fig. 2.

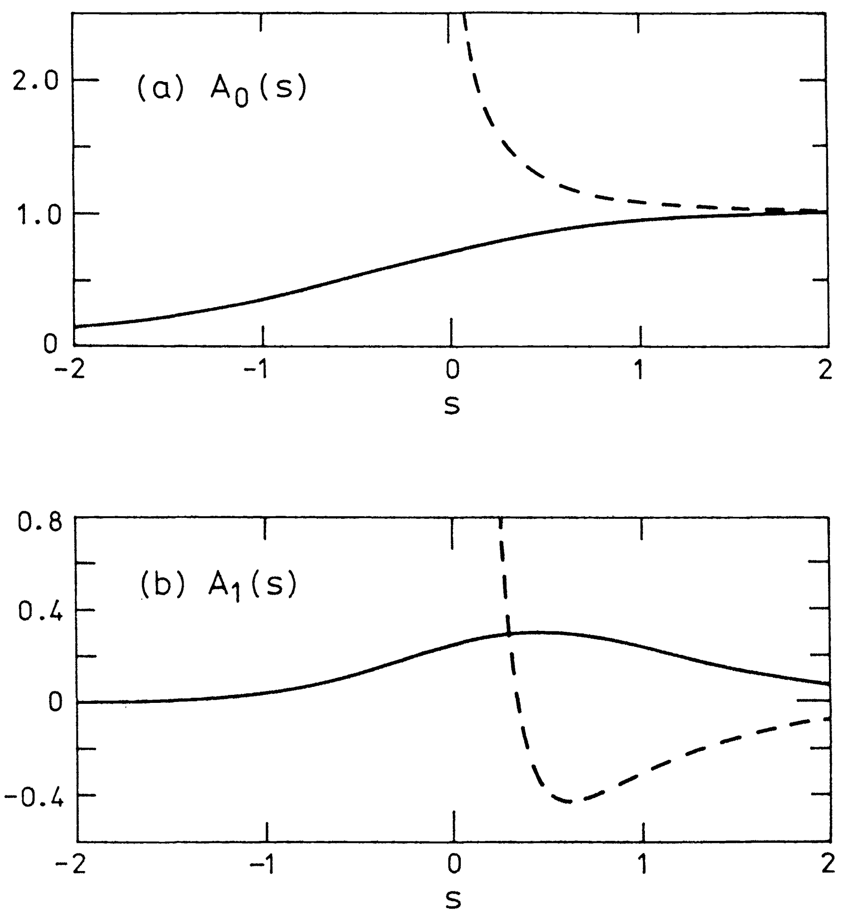


FIG. 2. *Graphs of the direct solutions of the solvability conditions* (3.4a) *and* (3.4b) *as given by* (3.6) *and* (3.7). —, $0<A_0<1$; ---, $1<A_0$. *The graph of* A_1 *is a particular solution, the general solution is obtained by adding an arbitrary multiple of* A_0'.

We now observe some of the problems associated with directly solving a recursive set of equations like equation (3.4). First, if a singularity exists in the leading-order solution then the singularity is compounded in the corrections. From equation (3.6b) $A_0 \sim s^{-1/2}$ for small s, while from equation (3.7b) $A_1 \sim s^{-3/2}\log(s)$ which is far more singular. Second it is usual for the solution's partial sums to be nonuniformly convergent in time. From equations (3.6) and (3.7) we have at large times $A_0+\varepsilon^2A_1 \approx 1 \pm (1/2)(1-4rs)e^{-2s}$ which illustrates the nonuniform convergence (no matter how small r is chosen) by possessing a nonmonotonic approach to 1 occurring on a s-time scale of $1/r$. To avoid the first problem we may use the technique of strained coordinates (see Cole (1968)); while to avoid the second problem we may introduce super-slow time scales (see Jeffrey and Kawahara (1982, §6.2.1)). However, if we wish to calculate higher orders then such techniques have to be compounded ad nauseam.

Another feature of these solutions is that not all initial conditions in the (a,b) plane can be accommodated in the general solutions. Because of the scaling introduced by (3.2) and (3.3) we are restricted to considering a subspace of the (a,b) plane, called the centre manifold. To first order, (3.5a) gives the shape of the manifold to be

$$b=a^2. \tag{3.8}$$

From equations (3.5), (3.6) and (3.7) the next approximation to the centre manifold is

$$b=(1-2r)a^2+2a^4+O(\varepsilon^6); \tag{3.9}$$

here the detailed evolution along the centre manifold (3.9) is given by $a=\varepsilon A_0+\varepsilon^3 A_1$ from equations (3.6) and (3.7), and $b=\varepsilon^2 B_0+\varepsilon^4 B_1$ given from equations (3.5). The $O(\varepsilon^6)$ term in equation (3.9) does not represent any new dynamics in the approach of solutions to the centre manifold. Being the same order of magnitude as the corrections due to A_2, unknown to this order, it represents uncertainty in the precise location of the centre manifold.

Thus we conclude that directly solving the recursive set of equations (3.4) can result in solutions with very large singularities, a nonuniform (qualitatively wrong) convergence and restricts the solution to a somewhat "fuzzy" centre manifold.

4. Type I reconstitution. The technique of reconstitution allows the user many degrees of freedom. In this section we consider two significantly different examples of one type of reconstitution. The reconstitutions considered are restricted by being required to give a differential equation of the same order (here first order) as the leading-order evolution (3.4a).

4.1. Substitute for all known derivatives. Here we use the principle of reconstitution to derive a sequence of more and more accurate evolution equations of the simplest possible form. The first step is to transform equations (3.4) to some exactly equivalent form. We use (3.4a) to eliminate all occurrences of A_0' and A_0'' in (3.4b) and (3.4c); then use the transformed (3.4b) to eliminate all occurrences of A_1' in (3.4c). The transformed equations are

$$A_0'=A_0\text{-}A_0^3, \tag{4.1a}$$

$$A_1'=A_1-3A_0^2A_1+2A_0^3-2A_0^5, \tag{4.1b}$$

$$A_2'=A_2-3A_0^2A_2-3A_0A_1^2+6A_0^2A_1-10A_0^4A_1-4A_0^3+16A_0^5-12A_0^7. \tag{4.1c}$$

Any set of direct solutions to these equations are also solutions to (3.4).

The second step is to use the transformed equations (4.1) to derive successively more accurate evolution equations. As a leading-order approximation we just rewrite (4.1a) in terms of the original variables. Substituting $a=\varepsilon A_0$ and $s=\varepsilon^2 t$ equation (4.1a) becomes

$$\dot{a}=ra-a^3 \tag{4.2}$$

where operator $(\dot{\ })$ denotes d/dt. The first reconstituted equation introduces higher order corrections to this equation. Consider ε^3(4.1a)$+\varepsilon^5$(4.1b)$+O(\varepsilon^7)$ which is just

$$\varepsilon^2(\varepsilon A_0+\varepsilon^3A_1)'=\varepsilon^2(\varepsilon A_0+\varepsilon^3A_1)-(\varepsilon^3A_0^3+3\varepsilon^5A_0^2A_1)+2\varepsilon^5A_0^3-2\varepsilon^5A_0^5+O(\varepsilon^7),$$

where the $O(\varepsilon^7)$ term, as yet unspecified, is introduced for the next step. We now write the equation solely in terms of $a=\varepsilon A_0+\varepsilon^3A_1$, $t=s/\varepsilon^2$ and $r=\varepsilon^2$, incorporating all generated terms of order ε^7 and higher in the $O(\varepsilon^7)$ term, to give the more accurate evolution equation

$$\dot{a}=ra-(1-2r)a^3-2a^5. \tag{4.3}$$

The second reconstituted equation is derived in a similar manner. Consider

$$\varepsilon^3(4.1\text{a})+\varepsilon^5(4.1\text{b})+\varepsilon^7(4.1\text{c})+O(\varepsilon^9)$$

and write it solely in terms of $a=\varepsilon A_0+\varepsilon^3A_1+\varepsilon^5A_2$, $t=s/\varepsilon^2$ and r to give the equation

$$\dot{a}=ra-(1-2r+4r^2)a^3-2(1-8r)a^5-12a^7. \tag{4.4}$$

The justification for considering equations (4.3) and (4.4) is that if the scalings and expansions used for the original problem ((3.2a) and (3.3)) are now applied to these equations then the resultant set of equations are (to some order) identical to equations (4.1) which in turn is exactly equivalent to (3.4). The choice of form for the right-hand sides of equations (4.3) and (4.4) is fairly arbitrary. If enough orders of equations are known then it may be appropriate to use heuristic arguments to pick a more exotic form for the right-hand sides.

An interesting question to ask at this point is whether all problems will produce a set of equations which can be reconstituted? The answer appears to be yes. However, a little care may be needed in the transformation from the original solvability conditions to the equivalent set of equations that are to be combined. Not all transformations are acceptable, for example the set of equations (3.4a), (3.4b) and (4.1c) are not directly reconstitutable. An alternative method to systematically derive improvements to evolution equations, proposed by Coullet and Spiegel (1983), has a bearing on the answer to this question. During the process of reconstitution we carry along some algebraic detail which is lost in the final reconstituted equation. For example, the terms involving A_1 and A_2 in equations (4.1b) and (4.1c) are redundant because they are forced to occur (for reconstitution to work) in combinations dictated by purely A_0 terms in the lower order equations. The method proposed by Coullet and Spiegel avoids this detail by going directly from the original system of equations to the approximate evolution equations which are, of course, the same as the reconstituted equations. However, because their method makes more initial assumptions about the forms of the evolution equation the generalisation of the method to partial differential equations is not trivial and has not yet been worked out, whereas with reconstitution it is simple.

Finding a more accurate evolution equation for a is only part of the reconstitution. To complete the reconstitution an equation giving b in terms of a is also needed; to derive such an equation we proceed much as before. Use (3.4) to eliminate all derivatives from (3.5) to give

$$B_0=A_0^2, \tag{4.5a}$$

$$B_1=2A_0A_1-2A_0^2+2A_0^4, \tag{4.5b}$$

$$B_2=2A_0A_2+A_1^2-4A_0A_1+8A_0^3A_1+4A_0^2-16A_0^4+12A_0^6. \tag{4.5c}$$

Then we write $\varepsilon^2(4.5\text{a})+O(\varepsilon^4)$, $\varepsilon^2(4.5\text{a})+\varepsilon^4(4.5\text{b})+O(\varepsilon^6)$ and

$$\varepsilon^2(4.5\text{a})+\varepsilon^4(4.5\text{b})+\varepsilon^6(4.5\text{c})+O(\varepsilon^8)$$

solely in terms of r, $t=s/\varepsilon^2$ and respectively $b=\varepsilon^2B_0$ and $a=\varepsilon A_0$, $b=\varepsilon^2B_0+\varepsilon^4B_1$ and $a=\varepsilon A_0+\varepsilon^3A_1$, and $b=\varepsilon^2B_0+\varepsilon^4B_1+\varepsilon^6B_2$ and $a=\varepsilon A_0+\varepsilon^3A_1+\varepsilon^5A_2$ to give the successively better approximations

$$b=a^2, \tag{4.6a}$$

$$b=(1-2r)a^2+2a^4, \tag{4.6b}$$

$$b=(1-2r+4r^2)a^2+2(1-8r)a^4+12a^6. \tag{4.6c}$$

These equations describe nothing more than the centre manifold upon which the evolution takes place and are the same as those found in the previous section, compare (4.6b) with (3.9). Here however, once the form of the expressions in (4.6) are chosen, the centre manifold is some definite curve upon which the initial conditions must be chosen; in the previous section it unsatisfactorily depended upon the initial conditions. In Fig. 3 these approximate manifolds are plotted and they can be directly compared to the trajectories of the exact solutions which are plotted in Fig. 1.

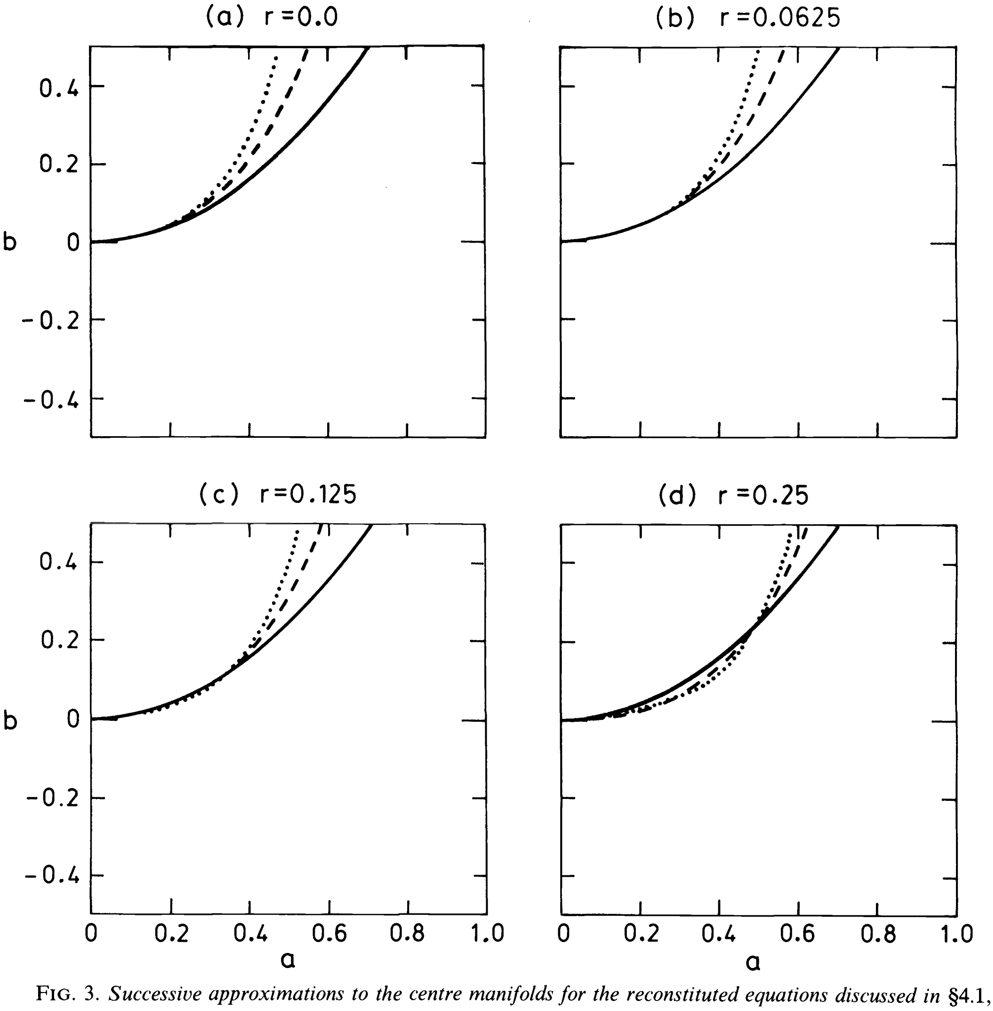


FIG. 3. *Successive approximations to the centre manifolds for the reconstituted equations discussed in* §4.1, *for four different values of r. The curves are*: —, *indicates the leading order approximation* (4.6a); ---, *is the approximation* (4.6b); · · · , *is the approximation* (4.6c).

The dynamics of the solutions to the reconstituted equations (4.3) and (4.4) can be easily understood by looking at the fixed points and their stability. Equation (4.3) has the fixed points $a=0$ for all r and $a=\pm\varepsilon$ for $r>0$. (Throughout this work the simplicity of the original equations (2.1) means that the location of the finite amplitude fixed point is always given exactly, independent of the approximation used). As we would expect, the fixed point $a=0$ is stable if $r<0$ and unstable if $r>0$. The fixed points $a=\pm\varepsilon$ are stable with the growth rate of small disturbances being

$$\lambda=-2r(1+2r).$$

Equation (4.4) also has only the fixed points $a=0$ for all r and $a=\pm\varepsilon$ for $r>0$. The same qualitative results hold with the growth rate of small disturbances from the fixed points $a=\pm\varepsilon$ now being the more accurate expression

$$\lambda=-2r(1+2r+8r^2).$$

The above expressions are just the first few terms of the expansion in small r of the exact growth rate (2.3). They show that the solutions of the reconstituted equation do not have the problem of a nonuniform approach to the stable fixed point which is (wrongly) predicted by the direct solutions of equation (3.4).

Also, these reconstituted equations have solutions that are not as singular as the previous direct solutions (3.6) and (3.7). For large a we find that (4.3) gives $a\sim t^{-1/4}$, and equation (4.4) gives $a\sim t^{-1/6}$; compare these with the $t^{-3/2}\log(t)$ singularity which occurs for A_1. In a few lines of trivial algebra we have produced equations that completely bypass the problems observed in the direction solution of the solvability conditions. Thus we see that the above reconstituted equations clarify the centre manifold on which the solutions are valid and refines the accuracy of the evolution uniformly in time (providing there are no singularities).

4.2. Substitute for second derivatives. We now investigate another possible form for the reconstituted equations while still requiring them to be first order differential equation in time. The difference between the approach taken here and that taken in §4.1 is that here we transform the solvability conditions (3.4) by only substituting for second and higher derivatives. The only change to equations (3.4) is the substitution from the derivative of equation (3.4) for A_0'' in equation (3.4c), thus the transformed equations are

$$\begin{aligned}
&(4.7\text{a}) && A_0'-A_0+A_0^3=0,\\
&(4.7\text{b}) && A_1'-A_1+3A_0^2A_1-2A_0^2A_0'=0,\\
&(4.7\text{c}) && A_2'-A_2+3A_0A_1^2+3A_0^2A_2-4A_0A_0'A_1-2A_0^2A_1'+2A_0A_0'^2\\
& && \qquad +2A_0^2A_0'-6A_0^4A_0'=0.
\end{aligned}$$

Proceeding as in §4.1, we can form the following sequence of progressively more accurate reconstituted equations

$$\begin{aligned}
&(4.8\text{a}) && \dot a-ra+a^3=0,\\
&(4.8\text{b}) && (1-2a^2)\dot a-ra+a^3=0,\\
&(4.8\text{c}) && 2a\dot a^2+\left[1-2(1-r)a^2-6a^4\right]\dot a-ra+a^3=0.
\end{aligned}$$

Before investigating the nature of the solutions to these equations we should again find expressions for b in terms of a. As before, we transform (3.5) by substituting appropriate expressions for all occurrences of second or higher derivatives of A_n. The resultant set of equations are

$$\begin{aligned}
&(4.9\text{a}) && B_0=A_0^2,\\
&(4.9\text{b}) && B_1=2A_0A_1-2A_0A_0',\\
&(4.9\text{c}) && B_2=2A_0A_2+A_1^2-2A_1A_0'-2A_0A_1'+2A_0A_0'-6A_0^3A_0'+2A_0'^2.
\end{aligned}$$

Combining the above equations we form the following, successively more accurate, reconstituted equations for the centre manifold

$$b=a^2, \tag{4.10a}$$

$$b=a^2-2a\dot{a}, \tag{4.10b}$$

$$b=a^2-\left[2(1-r)a+6a^3\right]\dot{a}+2\dot{a}^2, \tag{4.10c}$$

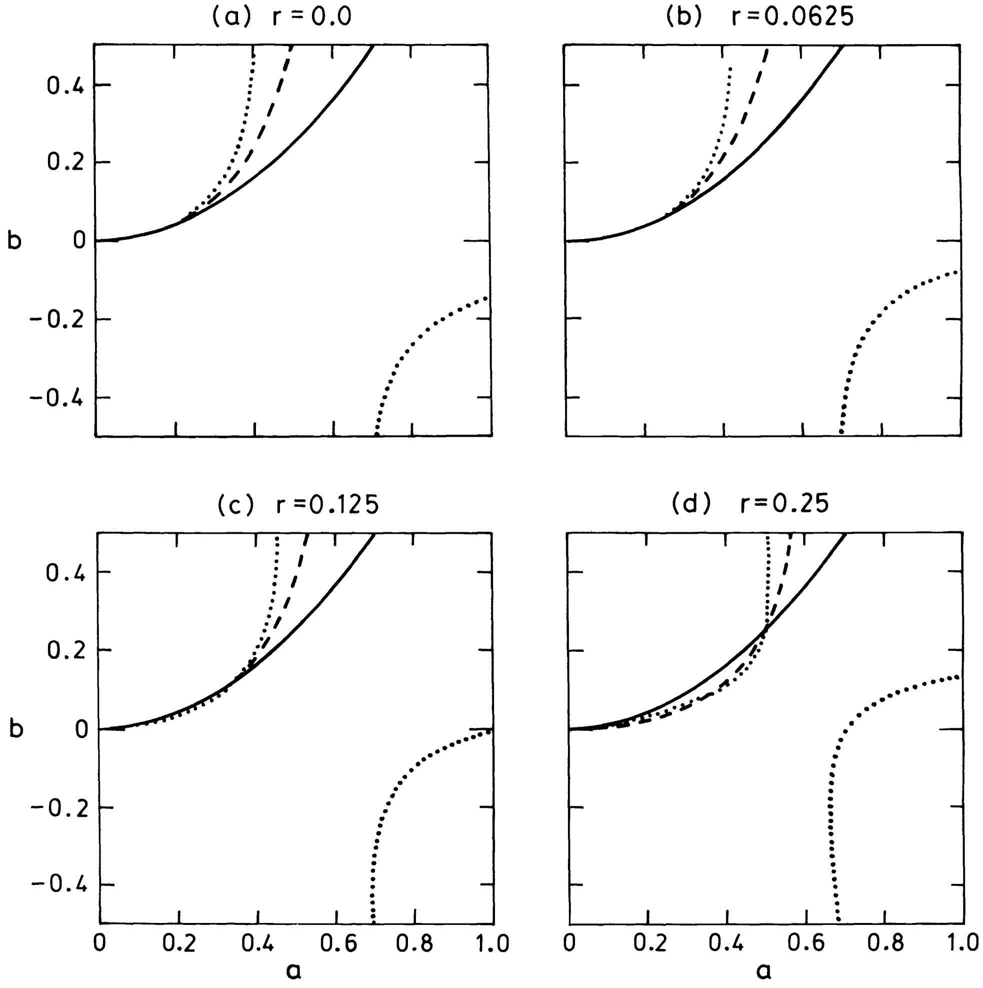


FIG. 4. *Successive approximations to the centre manifolds for the reconstituted equations discussed in* §4.2, *for four different values of r. The curves are*: —, *indicates the curve given by* (4.10a); ---, *is given by* (4.8b) *and* (4.10b); *and* ···, *from* (4.8c) *and* (4.10c).

(Fig. 4). The main feature of the solutions of the first reconstituted equation is the presence of a singularity at $a=\pm 1/\sqrt{2}$ which changes the nature of the fixed points $a=\pm\sqrt{r}$ for $r>\frac{1}{2}$. The breakdown of this approximate solution may be interpreted as an indication that at these sorts of amplitudes the exact solutions of equation (2.1) change their qualitative behaviour. The behaviour of the solutions to the second reconstituted equation (4.8c) are more interesting. The centre manifold is now a double valued function of a (Fig. 4) and for $r>(\sqrt{5}-1)/4$ the finite amplitude fixed point moves up onto the upper branch. This curving back of the centre manifold towards the fixed point is perhaps an indication of the fully developed spiral structure which should be present at these large values of r. However, for $r>(\sqrt{5}-1)/4$ the fixed point is unstable and for r larger than another critical value not much bigger than $(\sqrt{5}-1)/4$ the branch of the centre manifold passing through the origin no longer passes through the fixed point, again indicating the breakdown of the approximate solutions. Such a breakdown appears to be a usual feature of reconstituted equations and may be attributed to the mutual interaction between terms in the equations. These terms would originally have occurred in different orders in the perturbation expansion. It is this very interaction that makes reconstitution so effective, the breakdown of the solutions usefully indicating the extent of validity of the reconstituted equations.

5. Type II reconstitution. In §4 we required that the reconstitution equations be of the same order as the leading-order evolution equation (3.4a). This is usually desirable because appropriate boundary/initial conditions should be known for an equation of that order. However, it does strongly tie the reconstituted equations to the form of the leading-order evolution equation. In this section we widen the scope of the reconstitution equations by allowing second and higher order derivatives in the equivalent transformed version of equations (3.4). In use, this type of reconstitution produces an equation for which boundary layer approximations are often appropriate, rather like the Navier–Stokes equation at high Reynold's number where the Euler equation is a good approximation throughout most of the flow.

The simplest (and presumably soundest) way to include higher derivatives in the evolution equation is not to transform (3.4) at all, but to use them as they are. The first reconstituted equation is identical to (4.8b) discussed in §4.2. The second reconstituted equation is drived by writing

$$\varepsilon^2(3.4\text{a})+\varepsilon^4(3.4\text{b})+\varepsilon^6(3.4\text{c})+O(\varepsilon^8)$$

solely in terms of $a=\varepsilon A_0+\varepsilon^3 A_1+\varepsilon^5 A_2$, $t=s/\varepsilon^2$ and r. It is

$$2a^2\ddot{a}+2a\dot{a}^2+(1-2a^2)\dot{a}-ra+a^3=0. \tag{5.1}$$

The corresponding reconstituted equation for b drived directly from equation (3.5) is

$$b=a^2-2a\dot{a}+2a\ddot{a}+2\dot{a}^2. \tag{5.2}$$

At this point we notice that the reconstituted equation (5.1) can be obtained in a very simple manner. Write (2.1b) as

$$b=\left(1+\frac{d}{dt}\right)^{-1}a=a^2-2a\dot{a}-(2a\ddot{a}+2\dot{a}^2)-(2\dot{a}\ddot{a}+6\dot{a}\ddot{a})+\cdots,$$

then we can substitute some finite truncation of this sum into (2.1a). Truncating after one term we obtain the leading-order evolution equation (4.2), truncating after two terms we get (4.8b) and after three terms we arrive at equation (5.1).

It appears that we have gone to a great deal of trouble to transform our second order differential equation (2.1) to an approximate second order differential (5.1) which will be just as complicated to solve. This apparent fiasco only applies to simple model systems such as (2.1). In more complex, physically interesting, systems this type of reconstitution will produce an equation that is considerably simpler to solve than the original system.

The reconstituted equation (5.1) has the same fixed points with, for $r<\frac{1}{2}$, the same stability as the original problem. The growth rates for small disturbances away from the finite amplitude fixed point is given by

$$\lambda=-\left(1-2r\pm\sqrt{1-4r-12r^2}\right)/4r, \tag{5.3}$$

which has the property that for $r<\frac{1}{6}$ the asymptotic approach monotonic, while $\frac{1}{6}<r<\frac{1}{2}$ the asymptotic approach is oscillatory. This is qualitatively the same as the exact system, the transition occurring at a critical value of r which differs by only 25%. At $r=\frac{1}{2}$ there is a Hopf bifurcation and so for $r>\frac{1}{2}$ the finite amplitude fixed point of the reconstituted equation is unstable and all the solutions tend to a stable limit cycle. The trajectories in the (a, b) plane of the solutions to equation (5.1) and equation (5.2) are plotted in Fig. 5. Comparison between the trajectories of the exact solution (Fig. 1) and the approximate ones (Fig. 5) show a very good qualitative agreement, even far away from where the approximation is valid. A quantitative comparison is most easily made by comparing the two versions of the decay rates of small disturbances to the finite amplitude fixed point (Fig. 6).

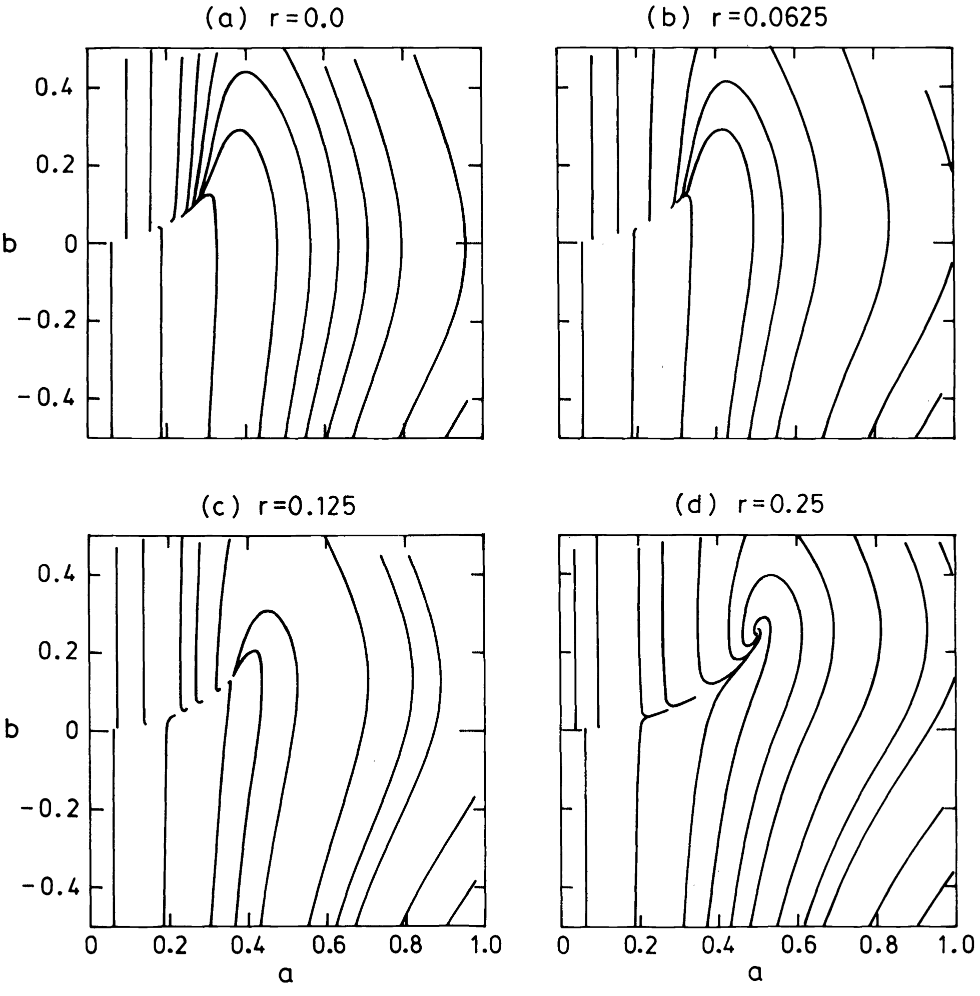


FIG. 5. *Trajectories in the (a, b) plane of the type* II *reconstituted equation* (5.1) *for four values of r.*

The important result of this section is that type II reconstitution produces an equation that is no longer restricted to the original centre manifold. It increases the dimension of the manifold and can produce an equation that is qualitatively correct on the new extended solution space for a wider range of parameter values.

6. Conclusion. We have seen that the technique of reconstitution can significantly improve approximate solutions to differential equations using exactly the same degree of information as other techniques. This is achieved by correcting the leading-order approximate equation itself rather than the more conventional approach of correcting solutions to the equations. The resultant equations describe the evolution of the solution uniformly in time and on a refined centre manifold.

In such a basic system as the simple equation (2.1) reconstitution mainly modifies the quantitative results but in a more complex system it can have a more profound effect (for example, convection with fixed heat flux boundary conditions, see Roberts (1982)). This occurs because reconstitution brings into one equation processes which previously occurred at different orders in an expansion of a set of equations. Hence, instead of higher order processes merely modifying the solution obtained at the lowest order, reconstitution allows all the effects to interact and thus produces an equation that is valid over a wider range of parameter values.

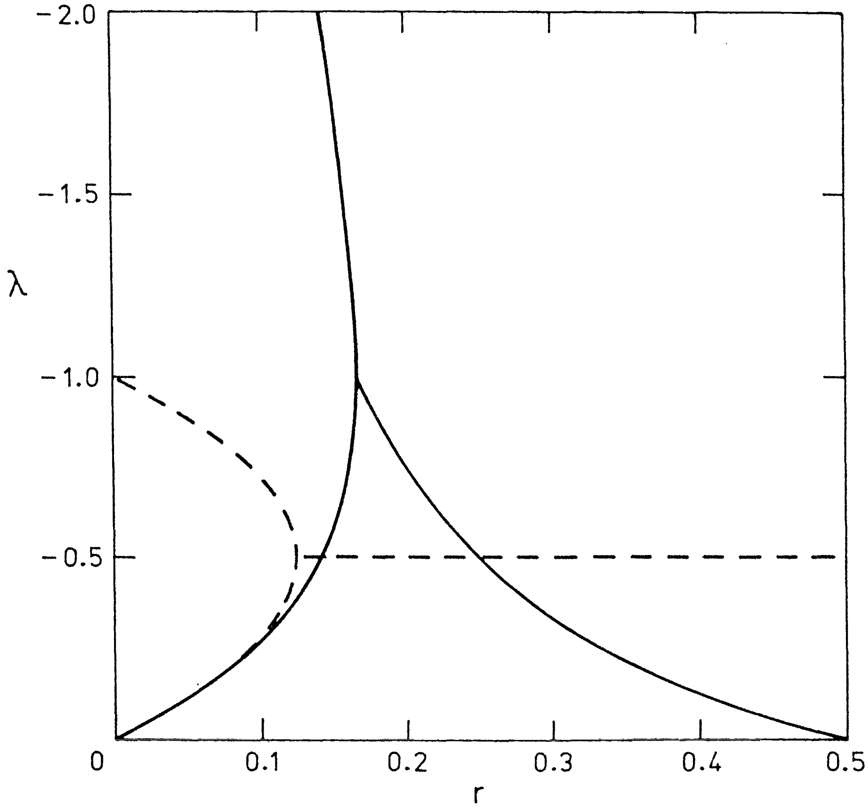


FIG. 6. *The real part of the growth rate of small disturbances to the finite amplitude fixed points of the reconstituted equation* (5.1) (—) *compared with that of the original system* (---).

Acknowledgments. I wish to acknowledge the aid of my supervisor Dr. H. E. Huppert, and the financial assistance of the Commonwealth Scholarship and Fellowship Plan during the course of this research.

REFERENCES

[1] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1978.

[2] C. J. Chapman and M. R. E. Proctor, *Nonlinear Rayleigh-Bernard convection between poorly conducting boundaries*, J. Fluid Mech., 101 (1980), pp. 759–782.

[3] J. D. Cole, *Perturbation Methods in Applied Mathematics*, Blaisdell, Waltham, MA, 1968.

[4] P. Coullet and E. A. Spiegel, *Amplitude equations for systems with competing instabilities*, SIAM J. Appl. Math., 43 (1983), pp. 776–821.

[5] K. B. Dysthe, *Note on a modification to the nonlinear Shcrödinger equation for application to deep water waves*, Proc. Roy. Soc. London A, 369 (1979), pp. 105–114.

[6] A. Jeffrey and T. Kawahara, *Asymptotic Methods in Nonlinear Wave Theory*, Pitman, London, 1982.

[7] D. H. Peregrine, *Equations for water waves and the approximations behind them*, in Waves on Beaches and Resulting Sediment Transport, R. E. Meyer, ed., Academic Press, New York, 1972, pp. 95–121.

[8] A. J. Roberts, *Nonlinear bouyancy effects in fluids*, Ph. D. Thesis, Univ. of Cambridge, Cambridge, 1982.

[9] E. A. Spiegel, *Physics of convection*, Tech. Rep. WHOI-81-102, Woods Hole Oceanographic Inst., Woods Hole, Md., 1981, pp. 43–67.

[10] J. T. Stuart, *On the nonlinear mechanics of wave disturbances in stable and unstable parallel flows, Part*·1. *The basic behaviour in plane Poiseuille flow*, J. Fluid Mech., 9 (1960), pp. 353–370.

[11] W. G. Vincenti and C. H. Kruger, *Introduction to Physical Gas Dynamics*, John Wiley, New York, 1965.

[12] J. Watson, *On the nonlinear mechanics of wave disturbances in stable and unstable parallel flows, Part* 2. *The development of a solution for plane Poiseuille flow and for plane Couette flow*, J. Fluid Mech., 9 (1960), pp. 371–389.

[13] G. B. Whitham, *Linear and Nonlinear Waves*, John Wiley, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

008

A QUASI-LINEAR, SINGULAR PERTURBATION PROBLEM OF HYPERBOLIC TYPE*

A. VAN HARTEN[†] AND R. R. VAN HASSEL[†]

Abstract. Using matched asymptotic expansions, a formal approximation can be constructed for an initial value problem of singularly perturbed, hyperbolic type in two independent variables. Under a time-like condition for the subcharacteristics of the unperturbed operator the correctness of the formal approximation is shown. Because of the nonlinearity of the perturbing hyperbolic operator, this work generalizes Geel [4]. The correctness proof is based on Schauder's fixed point theorem; it uses existence, uniqueness and regularity theory for hyperbolic systems and a priori estimates for a solution analogous to Geel [4] as ingredients.

1. Introduction. In this paper we consider an initial-value problem for a quasi-linear, second order, partial differential equation of hyperbolic type for an unknown function $u(x,t)$, where x denotes the space variable $\in \mathbb{R}$ and t the time variable, $t \geqq 0$;

$$\begin{aligned} &L^\varepsilon(u) \equiv \varepsilon L_2(u) + M_1(u) = h(x,t), \\ &u(x,0;\varepsilon) = f(x) \quad \text{and} \quad u_t(x,0;\varepsilon) = g(x), \end{aligned} \tag{1.1}$$

with

$$L_2(u) \equiv \frac{\partial^2 u}{\partial t^2} - c^2(x,t,u)\frac{\partial^2 u}{\partial x^2}, \qquad c>0$$

and

$$M_1(u) = a(x,t,u)\frac{\partial u}{\partial t} + b(x,t,u)\frac{\partial u}{\partial x} + d(x,t,u)u.$$

In a situation where $\varepsilon > 0$ is a small parameter the problem is of a singularly perturbed type, because ε multiplies the highest order derivatives in the equation. We shall show, how under certain conditions a formal approximation of the solution of (1.1) can be constructed and its correctness can be proven. Our work is a generalization of Geel [4] and de Jager [5].

In [4] Geel considers a subclass of (1.1), where the main difference is that in his case the coefficient c is independent of u. Hence in his work the second order perturbation operator L_2 is linear and only the first order operator $M_1(u)$ contains nonlinearities, while for our problem nonlinearities are also present in the second order terms of the equation. This does not make a dramatic difference for the construction of a formal approximation of the solution. In both cases the method of matched asymptotic expansions (Eckhaus [2]) will do the job if the coefficients are sufficiently regular and $a > 0$. See §2. The formal approximation will then consist of a regular expansion corrected with an initial layer at $t=0$ in the variable $\tau = t/\varepsilon$. However, for the proof of the correctness of the formal approximation a difficulty arises for the following reason. If u is the formal approximation and z the error $u - \tilde{u}$ then z has to satisfy a problem of the following type:

$$\begin{aligned} &L^\varepsilon(\tilde{u}+z) - L^\varepsilon(\tilde{u}) = -r, \\ &z = z_t = 0, \end{aligned} \tag{1.2}$$

* Received by the editors March 1, 1983, and in final revised form March 1, 1984.

[†] Department of Mathematics, State University of Utrecht, Budapestlaan 6, P.O. Box 80.010, 3508 TA Utrecht, The Netherlands.

where $r=-h+L^\varepsilon(\tilde{u})$ is small (say $O(\varepsilon^{N+1})$) in some sense. Now Geel uses a contraction argument in a suitable Banach space to solve the problem for z to estimate its magnitude. In order to do this one writes $L^\varepsilon(\tilde{u}+z)-L^\varepsilon(\tilde{u})=DL^\varepsilon z+\psi^\varepsilon(z)$, where DL^ε is the linearization of the operator $L^\varepsilon(u)$ at $u=\tilde{u}$ and $\psi^\varepsilon(z)$ contains the nonlinear terms. In Geel's situation ψ^ε is an operator, which maps C^1 functions onto C^0 functions in a continuous way, for the nonlinearities are only present in at most first order derivative terms. On the other hand DL^ε with homogeneous initial conditions is invertible and $(DL^\varepsilon)^{-1}$ has the property of mapping C^0 functions onto C^1 functions, continuously. Therefore, if the problem is reformulated for $z=(DL^\varepsilon)^{-1}(r-\psi^\varepsilon(z))$ a contraction principle can indeed be used.

In our case with a nonlinearity also in the second order derivative terms this idea fails for a simple but fundamental reason. Now, ψ^ε maps C^2 onto C^0 functions, but $(DL^\varepsilon)^{-1}$ does not map C^0 onto C^2 functions. The latter fact is clearly illustrated with the following example: $u_{tt}-u_{xx}=(x-t)H(x-t)$, $u=u_t=0$ at $t=0$, H the Heaviside function, has as its solution: $u(x,t)=0$ for $x\leqq -t$, $u(x,t)=(x+t)^3/24$ for $|x|\leqq t$ and $u(x,t)=t^2(x-t/3)/2$ for $x\geqq t$. The conclusion is that a new scheme for the proof of correctness is necessary.

In [1], Douglis proves existence and uniqueness theorems for hyperbolic systems of linear and quasi-linear equations using a different technique than [4], which also yields an estimate for the solution. However, an application of Douglis' method to prove correctness of our formal approximation is not possible. The reason is, that Douglis' method is, from a point of view of singular perturbations, rather rough. No distinction is made between the problems $\varepsilon(v_{tt}-v_{xx})\pm v_t=\exp(-2t)$, $v=v_t=0$ at $t=0$ with the solution $v(x,t)=\varepsilon(1\mp 2\varepsilon)^{-1}\exp(\mp t/\varepsilon)+(1\mp 2\varepsilon)^{-1}$. $\exp(-2t)\pm\frac{1}{2}$. Note, that in the case of a $-$ sign the solution grows exponentially as $\exp(t/\varepsilon)$. Even, in the case where the subcharacteristics of $M_1(u)$ are correctly located with respect to the characteristics of $L_2(u)$, [4], Douglis' method provides us with an exponential estimate $\sim$ $\exp(t/\varepsilon)$, based on the worst case.

Nevertheless, in §3 we prove the correctness of the formal approximation. This is done by a method based on Schauder's fixed point theorem, where a rather delicate combination of techniques and results from [4] and [1] are used as ingredients.

2. Construction of a formal approximation. For simplicity we consider the case where all coefficients and data are smooth, i.e. $a,b,c,d\in C^\infty$ $(\mathbb{R}\times[0,\infty)\times\mathbb{R})$, $h\in C^\infty([0,\infty)\times\mathbb{R})$ and $f,g\in C^\infty(\mathbb{R})$. Further we suppose

$$a>0 \quad \text{on } \mathbb{R}\times[0,\infty)\times\mathbb{R}. \tag{2.1}$$

Our formal approximation $\tilde{u}$ will have the following form:

$$\tilde{u}(x,t;\,\varepsilon)=\sum_{n=0}^{N}\varepsilon^n w_n(x,t)+\sum_{n=1}^{N+1}\varepsilon^n v_n(x,\tau)-\varepsilon^{N+1}v_{N+1}(x,0) \tag{2.2}$$

with $\tau=t/\varepsilon$. The first part of this expansion is of regular type and the second part describes the correction by an initial layer at $t=0$. The last term is introduced in order to satisfy the initial conditions exactly.

We shall now require that $\tilde{u}$ satisfy the equation in (1.1) up to $O(\varepsilon^N)$, i.e. $L^\varepsilon(\tilde{u})-h=O(\varepsilon^{N+1})$. Hence w_0 has to satisfy the reduced equation

$$a(x,t,w_0)\frac{\partial w_0}{\partial t}+b(x,t,w_0)\frac{\partial w_0}{\partial x}+d(x,t,w_0)w_0=h \tag{2.3}$$

and we provide it with the initial condition

$$w_0 = f \quad \text{at } t = 0.$$

Because of (2.1) this problem for w_0 has a unique, smooth solution in some domain $D = \{(x,t) | 0 \leq t < s(x)\}$ with $s \in C(\mathbb{R})$, $s > 0$. For several reasons (such as intersecting characteristics corresponding to (2.3)), it can occur that D cannot be taken equal to $\mathbb{R} \times [0, \infty)$. The first order correction term v_1 in the layer is necessary to satisfy the initial condition for the time derivative. Substitution of $w_0(x, \varepsilon\tau) + \varepsilon v_1(x, \tau)$ in (1.1) leads us to

$$\begin{aligned} &\left[\frac{\partial^2}{\partial\tau^2} + \bar{a}_0(x)\frac{\partial}{\partial\tau}\right] v_1(x,\tau) = 0, \\ (2.4) \qquad &\frac{\partial v_1}{\partial\tau} = g - \frac{\partial w_0}{\partial t} \quad \text{at } \tau = 0, \\ &\lim_{\tau\to\infty} v_1(x,\tau) = 0, \end{aligned}$$

with $\bar{a}_0(x) = a(x, 0, w_0(x,0))$.

The decay condition for $\tau \to \infty$ can be considered as a matching condition between the regular expansion outside the layer and the sum of the regular expansion and the layer correction inside the layer. The solution of (2.4) is given by

$$(2.5) \qquad v_1(x,\tau) = \left.\frac{\partial v_1}{\partial\tau}\right|_{\tau=0}(x) \cdot [\bar{a}_0(x)]^{-1} \exp(-\bar{a}_0(x)\tau).$$

For the higher order terms w_n, $n = 1, 2, \cdots$ we proceed as follows. The equation for w_n is found by putting

$$\left[\left(\frac{d}{d\varepsilon}\right)^n L^{\varepsilon}_{(x,t)}\left(\sum_{n=0}^{N} \varepsilon^n w_n\right)\right]\Bigg|_{\varepsilon=0} = 0;$$

it is of the following form

$$(2.6) \qquad a_0\frac{\partial w_n}{\partial t} + b_0\frac{\partial w_n}{\partial x} + \bar{d}_0 w_n = k_n.$$

Here $a_0(x,t)$ and $b_0(x,t)$ are the functions found from a and b by substituting $u = w_0(x,t)$. The coefficient $\bar{d}_0$ is given by $a_u \partial w_0/\partial t + b_u \partial w_0/\partial x + d + d_u w_0$ with u replaced by $w_0(x,t)$. The right-hand side k_n depends only on $w_0, \cdots, w_{n-1}$.

The equations for the higher order correction terms in the layer are derived from

$$\left[\left(\frac{d}{d\varepsilon}\right)^{n-1} L^{\varepsilon}_{(x,\tau)}\left(\sum_{n=0}^{N} \varepsilon^n w_n(x, \varepsilon\tau) + \sum_{n=1}^{N+1} \varepsilon^n v_n\right)\right]\Bigg|_{\varepsilon=0} = 0.$$

The structure of these equations is

$$(2.7) \qquad \left[\frac{\partial^2}{\partial\tau^2} + \bar{a}_0\frac{\partial}{\partial\tau}\right] v_n = \hat{k}_n,$$

where $\hat{k}_n$ depends on $w_0, \cdots, w_{n-2}$ and $v_1, \cdots, v_{n-1}$. Now we provide (2.6) with the initial condition

$$(2.8) \qquad w_n(x,0) = -v_n(x,0).$$

In addition to (2.7) we require

$$v_n(x,0) = -\frac{\partial w_{n-1}}{\partial t}(x,0), \qquad \lim_{\tau\to\infty} v_n(x,\tau) = 0. \tag{2.9}$$

We observe that the higher order terms can now be calculated in a unique way by the following scheme:

$$\begin{array}{lll}
n=0: & (2.3) \rightarrow & w_0 \xrightarrow[(2.4)]{} \\
n=1: & v_1 \xrightarrow[(2.6),\,(2.8)]{} & w_1 \xrightarrow[(2.7),\,(2.9)]{} \\
n=2: & v_2 \xrightarrow{\qquad} & \cdots \\
 & \vdots & \\
n=k: & \cdots & w_k \xrightarrow[(2.7),\,(2.9)]{} \\
n=k+1: & v_{k+1} \xrightarrow[(2.8),\,(2.8)]{} & \cdots
\end{array}$$

Using induction with respect to n it is not difficult to show that

$$w_n \in C^\infty(D). \tag{2.10}$$

It is also a nice exercise to verify that

$$v_n(x,\tau) = \sum_{j=1}^{n} P_{j,n}(x,\tau)\exp(-j\bar{a}_0(x)\tau) \tag{2.11}$$

with $P_{j,n}$ a polynomial in τ with coefficients $\in C^\infty(\mathbb{R})$, i.e., each of the layer terms v_n vanishes exponentially for $\tau \to \infty$.

The conclusion is that as a result of this construction we obtain a formal approximation $\tilde{u}$ as in (2.2), which satisfies

$$\begin{gathered} L^\varepsilon(\tilde{u}) = h - r, \\ \tilde{u} = f, \qquad \tilde{u}_t = g \quad \text{at } t = 0. \end{gathered} \tag{2.12}$$

where $r \in C^\infty(D)$ and for each ε-independent, compact subset $K \subset D$:

$$\sup_K |r| = O(\varepsilon^{N+1}) \quad \text{for } \varepsilon \downarrow 0. \tag{2.13}$$

In this sense the construction yields a result which approximately satisfies the equation and which satisfies the correct initial conditions.

The question is whether this will imply that (1.1) has a solution u such that $u - \tilde{u}$ is small in some sense. Without additional assumptions we cannot hope that this will be the case. The construction is such that $\tilde{u}(x_0, t_0)$ depends only on the data along the characteristic k of the unperturbed, first order operator M_1 through (x_0, t_0), whereas the solution $u(x_0, t_0)$ is determined by the data is the domain of influence I at (x_0, t_0) of the second order, hyperbolic operator L_2, which is a contradiction, if k lies outside I (cf. Geel [4, pp. 45–50]). Therefore in our proof of correctness in the next section we have to require time-likeness of the characteristics of M_1 with respect to those of the operator L_2.

If the characteristics of M_1 are space-like w.r.t. those of L_2, we encounter here a natural class of singular perturbations problems, for which a formal approximation can be constructed with an arbitrary fractional order, $O(\varepsilon^\nu)$, of accuracy, but without any relation to the genuine solution, compare Eckhaus [2, p. 198, §2] and Geel [4, p. 45].

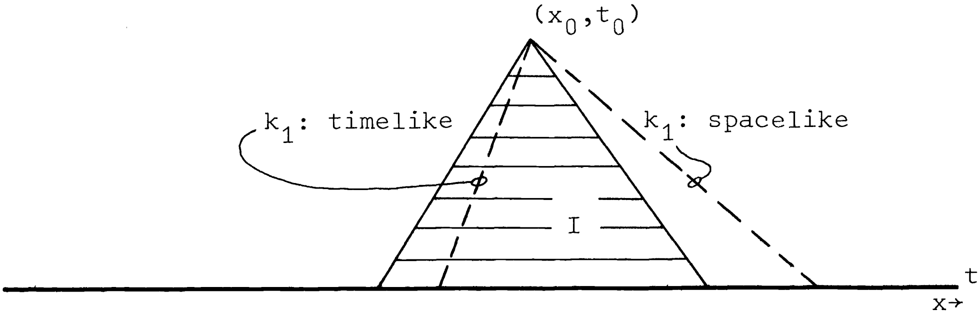


FIG. 1. *a, b, c constant* $a, c>0: k_1=\{(x,t)|0\leqq t\leqq t_0,\ x-x_0=\frac{b}{a}(t-t_0)\}$, $I=\{(x,t)|0\leqq t\leqq t_0,\ |x-x_0|\leqq c(t_0-t)\}$

3. Correctness of the formal approximation. In addition to the smoothness of the coefficients and data and the positivity of a (2.1), we assume that a time-likeness condition is satisfied for the characteristics of M_1 w.r.t. those of L_2 at the 0th order term of the approximation w_0:

$$|b_0a_0^{-1}|<c_0 \quad \text{on } D \tag{3.1}$$

with $c_0(x,t)=c(x,t,w_0(x,t))$; a_0, b_0 are defined in an analogous way, see (2.6).

Let us also introduce some terminology. A point $(x_0,t_0)\in D$ is called $L_{2,0}$-regular, if the trajectories of $dx/dt=\pm c_0(x,t)$; $x(t_0)=x_0$ exist and stay in D for $0\leqq t\leqq t_0$. These trajectories are the characteristics through (x_0,t_0) of the linear, hyperbolic operator $L_{2,0}=\partial^2/\partial t^2-c_0^2\partial^2/\partial x^2$ and we denote them by $x=1_0^{\mp}(t;\ x_0,t_0)$. For a $L_{2,0}$-regular point (x_0,t_0) the domain of dependence w.r.t. $L_{2,0}$, $I_0(x_0,t_0)=\{(x,t)|0\leqq t\leqq t_0,\ l_0^-(t;\ x_0t_0)\leqq x\leqq l_0^+(t;\ x_0,t_0)\}$, is well-defined and $\subset D$. A subset $K\subset D$ will be called $L_{2,0}$-regular, if each point K is $L_{2,0}$-regular.

We shall now formulate our correctness result.

THEOREM 3.1. *Consider a compact, $L_{2,0}$-regular subset $K\subset D$. If the coefficients and data are smooth and* (2.1), (3.1) *hold, then there is an $\varepsilon_0>0$ such that for $0<\varepsilon<\varepsilon_0$ the problem* (1.1) *has a unique, smooth solution u on some neighbourhood Ω of*

$$K^*=\bigcup_{(x_0,t_0)\in K} I_0(x_0,t_0)$$

in D and the formal approximation $\tilde{u}$ in (2.2) *is correct in the following sense*

$$\sup_{\Omega}|u-\tilde{u}|\leqq M\varepsilon^{N+1} \tag{3.2}$$

with $M>0$ and ε independent constants (*but dependent on Ω, $N\geqq 0$, $\varepsilon_0, f, g, h, a, b, c, d$*).

In order to prove this result we consider the problem for the remainder term $z=u-\tilde{u}$. Using (1.1) and (2.12) we find

$$\begin{aligned} &L^\varepsilon(\tilde{u}+z)-L^\varepsilon(\tilde{u})=-r,\\ &z=0, \qquad z_t=0 \quad \text{at } t=0. \end{aligned} \tag{3.3}$$

Instead of attacking (3.3) directly it is more convenient to consider a problem with coefficients and data which are also defined for $(x,t)\in\mathbb{R}\times[0,\infty)$ and which coincide with a, b, c, d and r for (x,t) in a neighbourhood of K^*. Since K is compact and $L_{2,0}$-regular, it is easy to check that K^* is a compact subset of D with the property $(x_0,t_0)\in K^*\Rightarrow I_0(x_0,t_0)\subset K^*$. Now let Ω_1 be a bounded open neighbourhood of K^* in D, such that $\bar{\Omega}_1\subset D$ and take $\chi\in C^\infty(\mathbb{R}\times[0,\infty))$, such that $\chi\equiv 1$ on a neighbourhood Ω_0 of K^* in D with $\bar{\Omega}_0\subset\Omega_1$ and $\chi\equiv 0$ outside Ω_1. We define new coefficients by

$$
\begin{aligned}
&\hat{c}(x,t,z;\,\varepsilon)=\hat{\gamma}(1-\chi)+\chi c(x,t,\tilde{u}(x,t;\,\varepsilon)+z)>0,\\
&\hat{a}(x,t,z;\,\varepsilon)=1-\chi+\chi a(x,t,\tilde{u}+z)>0,\\
&\hat{b}(x,t,z;\,\varepsilon)=\chi b(x,t,\tilde{u}+z),\\
&\hat{d}(x,t,z;\,\varepsilon)=\chi\tilde{d}(x,t,\tilde{u}+z),
\end{aligned}
\tag{3.4}
$$

with

$$
\begin{aligned}
\tilde{d}(x,t,z;\,\varepsilon)=\Big\{&\varepsilon\big(c^2(x,t,\tilde{u})-c^2(x,t,\tilde{u}+z)\big)\cdot\frac{\partial^2\tilde{u}}{\partial x^2}\\
&+d(x,t,\tilde{u}+z).z+\big(a(x,t,\tilde{u}+z)-a(x,t,\tilde{u})\big)\cdot\frac{\partial\tilde{u}}{\partial t}\\
&+\big(b(x,t,\tilde{u}+z)-b(x,t,\tilde{u})\big)\cdot\frac{\partial\tilde{u}}{\partial x}+\big(d(x,t,\tilde{u}+z)-d(x,t,\tilde{u})\big)\cdot\tilde{u}\Big\}\Big/z,
\end{aligned}
$$

where we interpret χa, χb, χc, $\chi\tilde{d}$ as $\equiv 0$ outside Ω_1. As a consequence of (3.1) the constant $\hat{\gamma}>0$ can be chosen in such a way that

$$
|\hat{b}\hat{a}^{-1}|<\gamma\hat{c}\quad\text{on }\mathbb{R}\times[0,\infty)\times[-\hat{\delta},\hat{\delta}]\times[0,\hat{\varepsilon}_0]
\tag{3.5}
$$

for suitable constants, $\hat{\delta}>0$, $\hat{\varepsilon}_0>0$ and with a constant $\gamma\in(0,1)$. Now we consider the following problem, which for $(x,t)\in\Omega_0\supset K^*$ coincides with (3.3):

$$
\begin{aligned}
&\hat{L}^\varepsilon(\hat{z})=-r\chi\underset{\text{def}}{=}\hat{r},\\
&\hat{z}=0,\qquad \hat{z}_t=0\quad\text{at }t=0,
\end{aligned}
\tag{3.6}
$$

with

$$
\hat{L}^\varepsilon(z)=\varepsilon\{z_{tt}-\hat{c}^2(x,t,z;\,\varepsilon)z_{xx}\}+\hat{a}(x,t,z;\,\varepsilon)z_t+\hat{b}(x,t,z;\,\varepsilon)z_x+\hat{d}(x,t,z;\,\varepsilon)z
$$

and $\hat{r}\in C^\infty(\mathbb{R}\times[0,\infty))$, support $(\hat{r})\subset\bar{\Omega}_1$.

For (3.6) we can prove an existence result together with an estimate of the solution $\hat{z}$ on a strip $S_T=\mathbb{R}\times[0,T]$ with $T>0$ an arbitrary constant independent of ε.

LEMMA 3.2. *For given $T>0$, there are constants $\bar{\rho}>0$ and $\varepsilon_0>0$ such that for ε sufficiently small, $0<\varepsilon\leqq\varepsilon_0$ and for $\hat{r}$ sufficiently small:*

$$
\sup_{S_T}|\hat{r}|\leqq\bar{\rho}\varepsilon,
\tag{3.7}
$$

the problem (3.6) has a solution $\hat{z}\in C^\infty(S_T)$ which satisfies the estimate

$$
\sup_{S_T}\big[|\hat{z}|+\varepsilon^{3/4}\{|\hat{z}_x|+|\hat{z}_t|\}\big]\leqq R\varepsilon^{-1/4}\sup_{S_T}|\hat{r}|
\tag{3.8}
$$

with a constant $R>0$ independent of ε and $\hat{r}$.

This lemma shows that if the order N of the formal approximation in (2.2), (2.9), (2.10) is sufficiently high, $N \geqq 1$, then (3.3) has a solution z on Ω_0, $z \in C^\infty(\Omega_0)$ and $\sup_{\Omega_0}|z| \leqq R\varepsilon^{N+3/4}$. Since augmenting the order of the formal approximation form N to $K>N$ adds only terms to the formal approximation of order N of order ε^{N+1} uniformly on $\bar{\Omega}_0$, it follows that given $\tilde{u}$ with $N \geqq 0$ there is a solution $u \in C^\infty(\Omega_0)$ for which (3.2) holds. The proof of uniqueness of such a solution on a suitable subset $\Omega \subset \Omega_0$, Ω a neighbourhood of K^* is a rather subtle matter, which is postponed till after the proof of Lemma 3.2.

Proof of Lemma 3.2. a: First we study the linear equation obtained from (3.6) by freezing the coefficients $\hat{a}$, $\hat{b}$, $\hat{c}$, $\hat{d}$ of $\hat{L}^\varepsilon(z)$ at $z=\omega(x,t;\ \varepsilon)$, where $\omega(\cdot,\cdot;\ \varepsilon)$ for each $\varepsilon \in (0,\varepsilon_0]$ is in $C^1(S_T)$. Hence we define

$$\hat{L}^\varepsilon_\omega(z)=\varepsilon\left\{z_{tt}-\hat{c}^2_\omega z_{xx}\right\}+\hat{a}_\omega z_t+\hat{b}_\omega z_x+\hat{d}_\omega z$$

with $\hat{c}_\omega(x,t;\ \varepsilon) \underset{\text{def}}{=} \hat{c}(x,t,\omega(x,t;\ \varepsilon);\ \varepsilon)$, etc. Now we consider the following problem:

$$(3.9)\qquad \begin{aligned}&\hat{L}^\varepsilon_\omega(\check{z})=\hat{r},\\ &\check{z}=0, \qquad \check{z}_t=0 \quad \text{at}\ t=0,\end{aligned}$$

with $\hat{r}$ as in (3.6). We introduce on $C^1(S_T)$ the norm $|\cdot|_1$ with for $s \in C^1(S_T)$

$$|s|_1 \underset{\text{def}}{=} \sup_{S_T}\left[|s|+\varepsilon^{3/4}\left\{|s_x|+|s_t|\right\}\right],$$

and we suppose

$$(3.10)\qquad |\omega(\cdot,\cdot;\ \varepsilon)|_1 \leqq m\varepsilon^{3/4}$$

with a constant $m>0$. This implies that $|\omega_x|$ and $|\omega_t|$ are bounded by m on S_T. Therefore the coefficients $\hat{c}_\omega$, $\hat{a}_\omega$, $\hat{b}_\omega$ and $\hat{d}_\omega$ and their derivatives with respect to x and t are absolutely bounded by a constant $m_1>0$, independent of ε and ω. Further there are constants $\alpha_0>0$, $\gamma_0>0$ and $p>0$, independent of ε and ω, such that

$$(3.11)\qquad \begin{aligned}&\hat{a}_\omega \geqq \alpha_0>0 && \text{on } S_T,\\ &\hat{c}_\omega \geqq \gamma_0>0 && \text{on } S_T,\\ &\hat{c}_\omega-|\hat{b}\hat{a}^{-1}| \geqq p>0 && \text{on } S_T.\end{aligned}$$

For the last inequality we use (3.5). This is allowed, if ε_0 is sufficiently small, so that $\varepsilon_0 \leqq \hat{\varepsilon}_0$, $m\varepsilon_0^{3/4} \leqq \hat{\delta}$.

ai. Let us investigate the existence, uniqueness and regularity of a solution of (3.9). This can be done by writing (3.9) as an equivalent system in its normal form

$$(3.12)\qquad \begin{aligned}&P\check{Z}_t+\Lambda P\check{Z}_x+Q\check{Z}=\hat{R},\\ &\hat{Z}=0 \quad \text{at}\ t=0,\end{aligned}$$

with

$$Z=\begin{pmatrix}\check{z}\\ \check{z}_x\\ \check{z}_t\end{pmatrix},\quad P=\begin{pmatrix}1&0&0\\0&1&+\hat{c}_\omega\\0&1&-\hat{c}_\omega\end{pmatrix},\quad \Lambda=\begin{pmatrix}1&&\varnothing\\&-\hat{c}_\omega&\\ \varnothing&&+\hat{c}_\omega\end{pmatrix},$$

$$Q=\begin{pmatrix}0&-1&-1\\ \hat{d}_\omega/\varepsilon&\hat{a}_\omega/\varepsilon&\hat{b}_\omega/\varepsilon\\ \hat{d}_\omega/\varepsilon&\hat{a}_\omega/\varepsilon&\hat{b}_\omega/\varepsilon\end{pmatrix},\qquad \hat{R}=\begin{pmatrix}0\\ \hat{r}/\varepsilon\\ \hat{r}/\varepsilon\end{pmatrix}.$$

Using Douglis [1, Thm. 5] we find, that (3.12) has a unique solution $\check{Z}\in\{C^1(S_T)\}^3$. Furthermore, there are constants $A<B$, independent of ε and ω, such that $\check{Z}=0$ on $\{(x,t)\in S_T|x\notin[A,B]\}$. In addition this theorem also provides us with an estimate for $\check{Z}$, $\check{Z}_x$ and $\check{Z}_t$. The transcription of these results in terms of (3.9) yields: there exists a unique solution $\check{z}$ of (3.9) in $C^2(S_T)$ for all $\varepsilon\in(0,\varepsilon_0)$. This solution satisfies

$$(3.13)\qquad \check{z}=0 \quad \text{on } \{(x,t)\in S_T|x\notin[A,B]\},$$

$$(3.14)\qquad \begin{aligned}&\sup_{S_T}\max\left(|\check{z}|,|\check{z}_x|,|\check{z}_t|,|\check{z}_{xx}|,|\check{z}_{xt}|,|\check{z}_{tt}|\right)\leqq\frac{m_2}{\varepsilon}\exp\left(\frac{m_3T}{\varepsilon}\right),\\ &\sup_{S_T}\max\left(|\hat{r}|,|\hat{r}_x|,|\hat{r}_t|\right),\end{aligned}$$

with constants A, B, m_2 and $m_3>0$ independent of ε, ω and $\hat{r}$. However, the estimate in (3.14) can be considerably improved.

aii. An improved estimate for the solution of (3.9) is obtained by using Geel's method [4]. Though in our case the coefficient $\hat{c}_\omega$ depends on ε a calculation analogous to [4, Chap. III, §2] yields

$$(3.15)\qquad |\check{z}(\cdot,\cdot;\varepsilon)|\leqq m_4^{-1/4}\sup_{S_T}|\hat{r}|,$$

with a constant $m_4>0$ independent of ε, ω and $\hat{r}$. In this calculation the properties of the coefficients $\hat{a}_\omega$, $\hat{c}_\omega$, $\hat{b}_\omega$, $\hat{d}_\omega$ as specified in and just above (3.11) are used in an essential way. Geel's method is based on the method of energy integrals, the essence of which can be found in [5].

b. Our next step is to rewrite the problem (3.6) for $\hat{z}$ as a search for a fixed point of a nonlinear operator F_ε. Here F_ε is defined as follows: if $\omega\in C^1(S_T)$ and $|\omega|\leqq m\varepsilon^{3/4}$, then

$$(3.16)\qquad F_\varepsilon(\omega)=\check{z}$$

with $\check{z}$ the solution of (3.9).

Now a solution of (3.6) is found by solving the equation

$$(3.17)\qquad F_\varepsilon(\hat{z})=\hat{z}.$$

In order to show the existence of a solution of (3.17) we shall use Schauder's fixed point theorem (Fucik [3]). The nonlinear operator F_ε is defined on $B_{\rho_0}=\{\omega\in C^1(S_T)|\ |\omega|_1\leqq\rho_0\}$, if $\rho_0\leqq m\varepsilon^{3/4}$. Because of (3.15) F_ε maps B_{ρ_0} into itself, if $\sup_{S_T}|\hat{r}|\leqq m_4^{-1}\varepsilon^{1/4}\rho_0$. The conclusion is if

$$(3.18)\qquad \sup_{S_T}|r|\leqq\tilde{\rho}\varepsilon \text{ with } \tilde{\rho}=mm_4^{-1},$$

then

$$(3.19)\qquad F_\varepsilon(B_\rho)\subset B_\rho \text{ with } \rho=m_4\varepsilon^{-1/4}\sup_{S_T}|\hat{r}|.$$

Next, an application of Douglis [1, Thm. 4] to the system (3.12) shows that F_ε is a continuous operator from B_ρ into B_ρ. If $\{\omega_n;\ n\in\mathbb{N}\}$ is a sequence in B_ρ, then $\{F_\varepsilon(\omega_n);\ n\in\mathbb{N}\}$ has the following properties: (i) $\exists$ compact subset $V\subset S_T$ such that $\forall n\in\mathbb{N}$ $F_\varepsilon(\omega_n)\equiv0$ outside V, (ii) $\exists E>0$ such that $\forall n\in\mathbb{N}$ $F_\varepsilon(\omega_n)$ and its derivatives up to the second order are absolutely bounded by E. The first property follows from (3.13) and the second one from (3.14). As a consequence of the Arzela–Ascoli theorem [6] the sequence $\{F_\varepsilon(w_n);\ n\in\mathbb{N}\}$ has a converging subsequence in the sense of $|\ |_1$.

The conclusion is that F_ε is a continuous, compact operator from B_ρ into B_ρ. Schauder's fixed point theorem implies that (3.17) possesses a solution $\hat{z}\in F_\varepsilon(B_\rho)$.

Hence, (3.6) possesses a solution $\hat{z}\in C^2(S_T)$ with $|\hat{z}|_1\leqq\rho=m_4\varepsilon^{-1/4}\sup_{S_T}|\hat{r}|$.

c. The derivatives $D_\alpha\hat{z}$ with $|\alpha|=k$, $k=1,2,\cdots$ satisfy linear systems of equations analogous to (3.12) with $\omega=\hat{z}$. Using Douglis' results [1] and induction with respect to k it is not difficult to prove the smoothness of $\hat{z}$.

This completes the proof of Lemma 3.2. □

As for the proof of the correctness result given in Theorem 3.1 we still have to demonstrate the uniqueness of the solution u on some suitable neighbourhood Ω of K^* in D, where $\Omega\subset\Omega_0$. This neighbourhood Ω is constructed in the following way. Let Ω_0' be a neighbourhood of K^* in D, such that

(i) $\bar{\Omega}_0'\subset\Omega_0$

and

(ii) $\bar{\Omega}_0'=\cup_{i=1}^N\{\cup_{(x_0,t_0)\in J_i}I_0(x_0,t_0)\}$

with $J_i=\{(x_0,t_0)|t_0=T_i,\ x_-^{(i)}\leqq x_0\leqq x_+^{(i)}\}$, $T_1\leqq T_2\leqq\cdots\leqq T_N$ and $N\in\mathbb{N}$. Next,we define

$$\Omega=\bigcup_{(x_0,t_0)\in\Omega_0'} I_\mu(x_0,t_0). \tag{3.20}$$

Here, $I_\mu(x_0,t_0)$ is defined analogous to $I_0(x_0,t_0)$, but with $\pm c_0(x_0,t_0)$ replaced by $\pm\{c_0(x_0,t_0)-\mu\}$ with a constant $\mu>0$. Note, that for μ sufficiently small $\Omega\subset\Omega_0$. A sketch of the situation is given in Fig. 2.

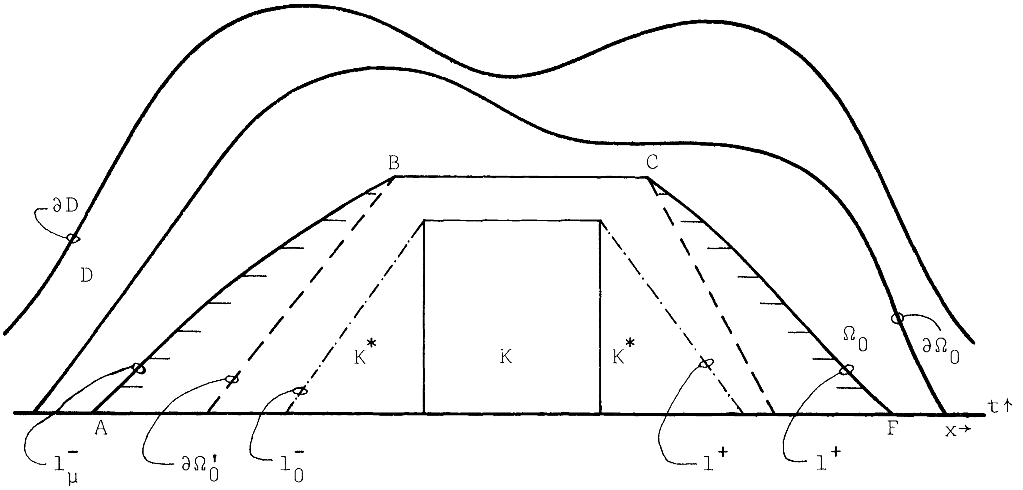
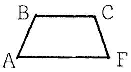


FIG 2. *Ω is the interior of the set*: 

The uniqueness of the solution can now be guaranteed on $\bar{\Omega}$.

LEMMA 3.3. *Suppose* $\mu > 0$ *is, such that* $\Omega \subset \Omega_0$. *Then the solution of problem* (1.1) *is unique on* $\bar{\Omega}$.

Proof of Lemma 3.3. Suppose z and $\hat{z}$, with $\hat{z}$ as in (3.8), are different solutions of problem (3.3) in $C^2(\Omega)$. Define: $T = \inf\{t > 0 \mid \exists (x,t) \in \bar{\Omega},\ z(x,t) \neq \hat{z}(x,t)\}$. If z is not identically equal to $\hat{z}$ on $\bar{\Omega}$, then $T < T_N$. Choose $\nu > 0$ in such a way, that $\forall z_0 \in \mathbb{R}$, $|z_0| \leqq \nu$: $c_0(x,t) - \frac{1}{2}\mu \leqq \hat{c}(x,t,z_0;\ \varepsilon) \leqq c_0(x,t) + \frac{1}{2}\mu$ for $(x,t) \in \bar{\Omega}$ and $\varepsilon \in (0, \varepsilon_n]$.

Because of (3.8) and the continuity of z, $\exists T' > T$ such that $\forall (x,t) \in \bar{\Omega}$, $T \leqq t \leqq T'$: $|z(x,t)| \leqq \nu$ and $|\hat{z}(x,t)| \leqq \nu$. Now, take (x_0, T) such that $\exists \{(x_n, t_n);\ n \in \mathbb{N}\}$ such that: $\lim_{n \to \infty}(x_n, t_n) = (x_0, T)$ and $z(x_n, t_n) \neq \hat{z}(x_n, t_n)$. We consider

$$R = \bigcup_{(y,T'),\ y_- \leqq y \leqq y_+} I_\mu(y, T'),$$

where y_- and y_+ are chosen in such a way, that $R \subset \bar{\Omega}$ and $(x_0, T) \in R$. An application of Douglis [1, Thm. 1] to (3.3) written in system form shows, that $z = \hat{z}$ on R, which is a contradiction with our assumption. Therefore we have to have uniqueness on all of $\bar{\Omega}$. □

REFERENCES

[1] A. DOUGLIS, (1952), *Some existence theorems for hyperbolic systems of partial differential equations in two independent variables*, Comm. Pure and Appl. Math., 5, pp. 119–154.

[2] W. ECKHAUS, (1979), *Asymptotic Analysis of Singular Perturbations*, North-Holland, Amsterdam.

[3] S. FUCIK, et al. (1973), *Spectral Analysis of Non-Linear Operators*, Lecture Notes in Mathematics 346, Springer, Berlin.

[4] R. GEEL, (1978), *Singular perturbations of hyperbolic type*, thesis, Univ. Amsterdam.

[5] E. M. DE JAGER, (1975), *Singular perturbations of hyperbolic type*, Nieuw Archief van de Wiskunde, (3) 23, pp. 145–172.

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

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

A NONLINEAR EIGENVALUE PROBLEM MODELLING THE AVALANCHE EFFECT IN SEMICONDUCTOR DIODES*

PETER A. MARKOWICH[†]

Abstract. This paper is concerned with the analysis of the solution set of the two-point boundary value problem modelling the avalanche effect in semiconductor diodes for negative applied voltage. This effect is represented by a large increase of the absolute value of the current starting at a certain reverse basis. We interpret the avalanche model as a nonlinear eigenvalue problem (with the current as eigenparameter) and show (using a priori estimates and a well-known theorem on the structure of solution sets of nonlinear eigenvalue problems for compact operators) that there exists an unbounded continuum of solutions which contains a solution for every negative voltage. Therefore, the solution branch does not "break down" at a certain threshold voltage (as expected on physical grounds). We discuss the current-voltage characteristic and prove that the absolute value of the current increases at most (and at least) exponentially in the avalanche case as the voltage decreases to minus infinity.

AMS-MOS subject classifications (1980). Primary 34B15, 58C40, 78A25

Key words. nonlinear eigenvalue problems, unbounded solution continua, two-point boundary-value problems, impact ionization, semiconductors

1. Introduction.

We investigate the (one-dimensional) boundary-value problem which describes the performance of a semiconductor diode in the case of avalanche generation. The physical situation is as follows. A semiconductor is doped with donor atoms on the right side (n-side) and with acceptor atoms on the left side (p-side) and a bias is applied to the Ohmic contacts (see Fig. 1).

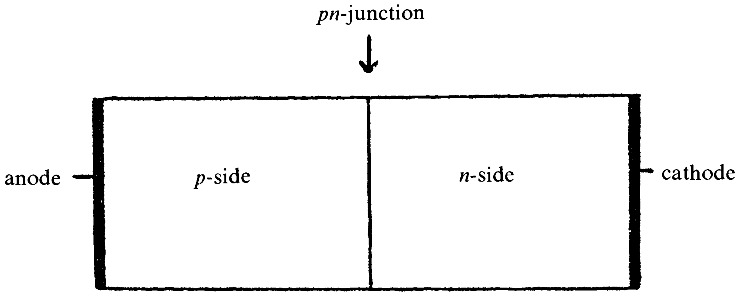


FIG. 1. *Diode.*

For simplicity we assume that the pn-junction is in the middle of the device, that the doping profile (that is the difference of the the concentrations of donors and acceptors) is constant in the n-side as well as in the p-side and odd about the pn-junction.

A well-known phenomenon is the "breakdown" of the diode due to impact ionization (avalanche generation, see Sze (1981)) under sufficiently large negative bias. This "breakdown" is based on a "sudden" increase of the current (as a function of the applied bias).

*Received by the editors November 7, 1983, and in revised form April 2, 1984. The work of this author was sponsored by the United States Army under contract DAAG29-80-C00041.

[†] Technische Universität Wien, Institut fur Angewandte und Numerische Mathematik, Gusshausstrasse 27–29, A-1040 Wien, Austria.

To study the current-voltage ($J-V$) characteristic of the device we investigate the basic semiconductor device equations describing potential and carrier distributions in the diode (see Van Roosbroeck (1950), Sze (1981));

$$\begin{aligned}
&(1.1) \qquad \lambda^2\psi''=n-p-D \quad \text{Poisson's equation},\\
&(1.2) \qquad n'=n\psi'+J_n \quad \text{electron continuity equation}, \qquad -1\leqq x\leqq 1.\\
&(1.3) \qquad p'=p\psi'-J_p \quad \text{hole continuity equation},
\end{aligned}$$

ψ denotes the electrostatic potential, ψ' is the electric field, $n(p)$ the electron (hole) density, $J_n(J_p)$ the electron (hole) current density and D the doping profile. The equations (1.1)–(1.3) are already in dimensionless form, the doping profile is scaled to maximally one and the independent variable x to $[-1,1]$. In our symmetric and piecewise constant case

$$(1.4) \qquad D=\begin{cases} 1, & 0\leqq x\leqq 1 \quad (n\text{-side}),\\ -1, & -1\leqq x<0 \quad (p\text{-side})\end{cases}$$

holds. The pn-junction is at $x=0$. $\lambda^2(\ll 1)$ is a scaling parameter.

Generally the current relations are given by

$$(1.5) \qquad \text{(a)}\ J_n'=R, \qquad \text{(b)}\ J_p'=-R,$$

where the recombination-generation term R is a nonlinear function of n,p, J_n, J_p and ψ'. We assume that R is given by the avalanche-generation term (see Sze (1981), Schütz (1982));

$$(1.6) \qquad R=R(J_n,J_p,\psi')=-\alpha(\psi')\big(|J_n|+|J_p|\big),$$

where $\alpha\geqq 0$ is the electron-hole ionization rate. α is strongly field-dependent. Commonly used α's are $\alpha(\tau)=\gamma e^{-\sigma/|\tau|}$, $\alpha(\tau)=\gamma|\tau|e^{-\sigma/|\tau|}$, γ, $\sigma\geqq 0$. For simplicity we assume that α: $C([0,1])\to[0,\gamma]$, $\gamma\geqq 0$ is the nonnegative functional

$$(1.7) \quad \alpha(f)=\beta\big(\|f\|_{[-1,1]}\big), \qquad \beta\colon [0,\infty]\to[0,\gamma], \quad \beta\in C([0,\infty]) \text{ and nondecreasing}$$

($\|f\|_{[a,b]}:=\sup_{a\leqq x\leqq b}|f(x)|$). We will later on remark on the extension of results to more realistic ionization rates.

The total current J is given by

$$(1.8) \qquad J=J_n+J_p.$$

Note that J is a constant in $[-1,1]$ because of (1.5).

The boundary conditions (at the Ohmic contacts) for (1.1)–(1.5) are

$$(1.8\text{a}) \qquad np=\delta^4, \qquad n-p-D=0 \quad \text{at } x=\pm 1$$

where $\delta^2(\leqq 1)$ also originates from the scaling and

$$(1.8\text{b}) \qquad \psi(1)=\ln\frac{n(1)}{\delta^2}-V, \qquad \psi(-1)=\ln\frac{p(-1)}{\delta^2}+V.$$

$V\in R$ is the (scaled) voltage applied to the diode (details on the scaling can be found in Markowich and Ringhofer (1984) and Markowich (1983)).

Because of our symmetry assumptions we restrain the investigation to "symmetric" solutions, i.e. solutions which fulfill

$$(1.9)\qquad \psi(x)=-\psi(-x),\quad n(x)=p(-x),\quad J_n(x)=J_p(-x),\quad x\in[-1,1].$$

Another simplification is accomplished by employing the substitution

$$(1.10)\qquad n=\delta^2e^{\psi}u,\quad p=\delta^2e^{-\psi}v,\quad J_n=\delta^2e^{\psi}u',\quad J_p=-\delta^2e^{-\psi}v'.$$

The system of equations obtained from (1.1)–(1.8) by using (1.9), (1.10) is

$$\begin{aligned}
&(1.11)\qquad && \lambda^2\psi''=\delta^2e^{\psi}u-\delta^2e^{-\psi}v-1,\\
&(1.12)\qquad && (e^{\psi}u')'=-\alpha(\psi')(|e^{\psi}u'|+|e^{-\psi}v'|),\qquad 0\leqq x\leqq 1,\\
&(1.13)\qquad && (e^{-\psi}v')'=-\alpha(\psi')(|e^{\psi}u'|+|e^{-\psi}v'|),
\end{aligned}$$

subject to the boundary conditions:

$$\begin{aligned}
&(1.14a)\qquad && \psi(0)=0,\qquad \psi(1)=\ln\left(\frac{1+\sqrt{1+4\delta^4}}{2\delta^2}\right)-V,\\
&(1.14b)\qquad && u(0)=v(0),\qquad u(1)=e^V,\\
&(1.14c)\qquad && v'(0)=-u'(0),\qquad v(1)=e^{-V}.
\end{aligned}$$

The boundary conditions for u and v at $x=-1$ are $u(-1)=e^{-V}$, $v(-1)=e^V$. The maximum principle (see Protter and Weinberger (1967)) applied to (1.12), (1.13) gives

$$(1.15)\qquad u\geqq e^{-|V|},\qquad v\geqq e^{-|V|}\quad \text{on }[-1,1].$$

Therefore n and p are positive (as physically required for densities). A solution for $V=0$ is given by $u\equiv 1$, $v\equiv 1$ and by solving

$$(1.16a)\qquad \lambda^2\psi''=\delta^2e^{\psi}-\delta^2e^{-\psi}-1,\qquad 0\leqq x\leqq 1,$$

$$(1.16b)\qquad \psi(0)=0,\qquad \psi(1)=\ln\left(\frac{1+\sqrt{1+4\delta^4}}{2\delta^2}\right).$$

The solution $(V,\psi,u,v)=(0,\psi_e,1,1)$ where ψ_e is the unique solution of (1.6) (existence and uniquenss of ψ_e will be proved later on) is called equilibrium solution. It implies $J=0$ (the whole diode is in thermal equilibrium).

The two-point boundary-value problem (1.11)–(1.14) models the bias-controlled diode. In some cases it is more convenient to investigate the current-controlled device represented by the equations (1.11)–(1.13) subject to the boundary conditions:

$$\begin{aligned}
&(1.17a)\qquad && \psi(0)=0,\qquad \psi(1)=\ln\left(\frac{1+\sqrt{1+4\delta^4}}{2\delta^2}\right)-\ln u(1),\\
&(1.17b)\qquad && u(0)=v(0),\qquad u(1)v(1)=1,\quad u(1)>0,\\
&(1.17c)\qquad && u'(0)=\frac{J}{2\delta^2},\qquad v'(0)=-\frac{J}{2\delta^2}.
\end{aligned}$$

Note that (1.17c) follows from

$$J=J_n(0)+J_p(0)=\delta^2e^{\psi(0)}u'(0)-\delta^2e^{-\psi(0)}v'(0)=2\delta^2u'(0)=-2\delta^2v'(0).$$

The problems (1.11)–(1.14) and (1.11)–(1.13), (1.17) are equivalent in the following sense. A solution (V_1, ψ_1, u_1, v_1) of (1.11)–(1.14) yields the solution (J_1, ψ_1, u_1, v_1) of (1.11)–(1.13), (1.17) where $J_1 = \delta^2 e^{\psi_1} u_1' - \delta^2 e^{-\psi_1} v_1'$ and a solution (J_2, ψ_2, u_2, v_2) of (1.11)–(1.13), (1.17) yields the solution (V_2, ψ_2, u_2, v_2) of (1.11)–(1.14) where $V_2 = \ln u_2(1)$.

There are numerous analytical and numerical investigations of the (even multi-dimensional) semiconductor device equations in the nonavalanche case (i.e. the recombination-generation rate R only depends on n and p) (see Mock (1983) for a rather complete presentation of the results as well as for a collection of references). For the avalanche problem, however, there are (to the author's knowledge) only a few numerical studies (see Schütz (1982); Schütz, Selberherr and Pötzl (1982)).

In this paper we regard (1.11)–(1.14) and (1.11)–(1.13), (1.17) as nonlinear eigenvalue problems (in the sense of Rabinowitz (1971), Krasnoselskii (1964) and investigate the solution set for nonpositive current:

$$(1.18)\qquad C^- = \{(I,\psi,u,v) \in (-\infty,0] \times (C^2([0,1]))^3 | (\psi,u,v) \text{ solves } (1.11)\text{–}(1.13), (1.17) \text{ with } J = I\}$$

and the properties of the current-voltage $(J - V)$ characteristic:

$$(1.19)\qquad J^- = \{(V,J) \in \mathbb{R} \times (-\infty,0])| \text{there is } (\psi,u,v) \text{ such that } (J,\psi,u,v) \in C^- \text{ and } V = \ln u(1)\}.$$

The main theorem of this paper states that C^- contains an unbounded continuum (i.e. a closed and connected set (in the $(-\infty,0] \times (C^2([0,1]))^3$-topology) emanating from the equilibrium solution $(0, \psi_e, 1, 1)$ whose projection into $(-\infty,0]$ equals $(-\infty,0]$ (that means C^- contains solutions for all $J \leqq 0$) and that the voltage $V \to -\infty$ as $J \to -\infty$. Therefore (1.11)–(1.14) has a solution (ψ,u,v) for every $V \leqq 0$.

This result holds independently of the upper bound γ of the ionization rate α and carries over to more realistic α's than given by (1.7). Therefore the conjecture that the branch of solutions of (1.11)–(1.14) breaks down if $\gamma > \frac{1}{2}$ holds (see Sze (1981)) is mathematically rejected at least for this model problem. We show, however, that the magnitude of γ has a decisive impact on the $J - V$-characteristic. For $\alpha \equiv 0$ (non-avalanche case) the current fulfills $c_2 V \leqq J \leqq c_1 V$ for $V \leqq 0$ while $|J|$ increases exponentially as $V \to -\infty$ for $\gamma > \frac{1}{2}$. ($c_1, c_2 > 0$ only depend on λ and δ). The exponential growth of the current represents the "avalanche effect" and the diode "breaks down" in real life when the current gets too large. We also show nonuniqueness for $V = 0$ for all α for which $\alpha(\psi_e')$ is sufficiently large.

The paper is organized as follows. Section 2 deals with the a priori estimates needed to prove existence for all $J \leqq 0$ and §3 contains the existence proof and conclusions.

2. A priori estimates. For the following we take $J \leqq 0$.

At first we solve the continuity equations (1.12)–(1.13) for fixed $\psi \in C^1([0,1])$. We rewrite them as

$$\begin{aligned} &(2.1\text{a})\qquad J_n' = -\alpha(\psi')(|J_n| + |J_p|), \\ &(2.1\text{b})\qquad J_p' = \alpha(\psi')(|J_n| + |J_p|), \end{aligned} \qquad 0 \leqq x \leqq 1,$$

with the initial values

$$J_n(0)=\frac{J}{2},\qquad J_p(0)=\frac{J}{2}. \tag{2.2c}$$

(2.1a) implies that J_n is nonincreasing; since $J_n(0)\leqq 0$, we get $J_n\leqq 0$ on $[0,1]$. $J_p(0)\leqq 0$ holds and therefore we (initially) solve

$$J_n'=-\alpha|J|,\qquad J_p'=\alpha|J|$$

(we often drop the argument ψ' of α) and get

(2.3a)

$$J_n=-\frac{|J|}{2}(2\alpha x+1),\qquad J_p=-\frac{|J|}{2}(-2\alpha x+1)\quad\text{for}\begin{cases}x\in[0,1] & \text{if } 0\leqq\alpha<\dfrac{1}{2},\\ x\in\left[0,\dfrac{1}{2\alpha}\right] & \text{if } \alpha>\dfrac{1}{2}.\end{cases}$$

For $\alpha>\frac{1}{2}$ we have to solve

$$J_n'=\alpha(J_n-J_p),\quad J_p'=-\alpha(J_n-J_p),\quad \frac{1}{2\alpha}\leqq x\leqq 1,$$
$$J_n\left(\frac{1}{2\alpha}\right)=-|J|,\qquad J_p\left(\frac{1}{2\alpha}\right)=0$$

and obtain

$$J_n=-\frac{|J|}{2}(e^{2\alpha x-1}+1),\qquad J_p=-\frac{|J|}{2}(-e^{2\alpha x-1}+1)\quad\text{for } x\in\left[\frac{1}{2\alpha},1\right]. \tag{2.3b}$$

u and v are computed from (1.10), (1.14b, c):

(2.4a)

$$u=e^V+\frac{|J|}{2\delta^2}\cdot\begin{cases}\displaystyle\int_x^1 e^{-\psi(s)}(2\alpha s+1)\,ds \quad \text{for } x\in[0,1]\quad \text{if } 0\leqq\alpha\leqq\frac{1}{2},\\ \displaystyle\int_x^1 e^{-\psi(s)}(e^{2\alpha s-1}+1)\,ds \quad \text{for } x\in\left[\frac{1}{2\alpha},1\right]\quad \text{if } \alpha>\frac{1}{2},\\ \displaystyle\int_x^{1/2\alpha} e^{-\psi(s)}(2\alpha s+1)\,ds+\int_{1/2\alpha}^1 e^{-\psi(s)}(e^{2\alpha s-1}+1)\,ds\\ \qquad\qquad\qquad \text{for } x\in\left[0,\frac{1}{2\alpha}\right]\quad \text{if } \alpha>\frac{1}{2},\end{cases}$$

$$v=e^{-V}-\frac{|J|}{2\delta^2}\cdot\begin{cases}\displaystyle\int_x^1 e^{\psi(s)}(-2\alpha s+1)\,ds \quad \text{for } x\in[0,1]\quad \text{if } 0\leqq\alpha\leqq\frac{1}{2},\\ \displaystyle\int_x^1 e^{\psi(s)}(-e^{2\alpha s-1}+1)\,ds \quad \text{for } x\in\left[\frac{1}{2\alpha},1\right]\quad \text{if } \alpha\geqq\frac{1}{2},\\ \displaystyle\int_x^{1/2\alpha} e^{\psi(s)}(-2\alpha s+1)\,ds+\int_{1/2\alpha}^1 e^{\psi(s)}(-e^{2\alpha s-1}+1)\,ds\\ \qquad\qquad\qquad \text{for } x\in\left[0,\frac{1}{2\alpha}\right]\quad \text{if } \alpha>\frac{1}{2}.\end{cases}$$

We use the condition $u(0)=v(0)$ to relate V and J and get

$$V=\text{area}\sinh\left(\frac{J\cdot I(\psi)}{4\delta^2}\right),\qquad J\leqq 0 \tag{2.5}$$

where the functional I: $C^1([0,1])\to\mathbb{R}$ is given by

$$I(\psi)=\int_0^1\left[e^{\psi(s)}g_\alpha(s)+e^{-\psi(s)}f_\alpha(s)\right]ds \tag{2.6a}$$

with

(2.6b)

$$g_\alpha(x)=\begin{cases}-2\alpha x+1 & \text{for } x\in[0,1] \text{ if } 0\leqq\alpha\leqq\dfrac{1}{2} \text{ and for } x\in\left[0,\dfrac{1}{2\alpha}\right] \text{ if } \alpha>\dfrac{1}{2},\\ -e^{2\alpha x-1}+1 & \text{for } x\in\left[\dfrac{1}{2\alpha},1\right] \text{ if } \alpha>\dfrac{1}{2},\end{cases}$$

$$f_\alpha(x)=\begin{cases}2\alpha x+1 & \text{for } x\in[0,1] \text{ if } 0\leqq\alpha\leqq\dfrac{1}{2} \text{ and for } x\in\left[0,\dfrac{1}{2\alpha}\right] \text{ if } \alpha>\dfrac{1}{2},\\ e^{2\alpha x-1}+1 & \text{for } x\in\left[\dfrac{1}{2\alpha},1\right] \text{ if } \alpha>\dfrac{1}{2}.\end{cases}$$

$I\in C(C^1([0,1])\to\mathbb{R})$ holds. For the estimates of the current J in terms of the voltage V we use

$$J=\frac{4\delta^2\sinh V}{I(\psi)}\quad\text{if } I(\psi)\neq 0. \tag{2.7}$$

We collect the properties of u, v, J_n, J_p and J in:

LEMMA 2.1. *Assume that* $0\leqq\alpha(\psi')\leqq\frac{1}{2}$ *holds. Then*

(i) $I(\psi)\geqq(1-2\alpha)\int_0^1 e^{\psi(s)}ds+\int_0^1 e^{-\psi(s)}ds>0$;

(ii) $J<0\Leftrightarrow V<0$, $J=0\Leftrightarrow V=0\Leftrightarrow u\equiv v\equiv 1$.

Let $V<0$ *hold. Then*

(iii) $J_n<0$, $J_p<0$ *on* $[0,1]$;

(iv) u *is decreasing on* $[0,1]$, $e^V\leqq u\leqq e^{-V}$;

(v) v *is increasing on* $[0,1]$, $e^V\leqq v\leqq e^{-V}$.

LEMMA 2.2. *Assume that* $\alpha(\psi')>\frac{1}{2}$ *is fixed. Then*

(i) *there is* $\psi\in C^1([0,1])$ *such that* $I(\psi)<0$;

(ii) $J=0\Rightarrow(V=0\Leftrightarrow u\equiv v\equiv 1)$, $J<0\Leftrightarrow(V<0, I(\psi)>0$ *or* $V>0, I(\psi)<0)$;

(iii) $V=0\Leftrightarrow J=0$ *or* $I(\psi)=0$.

Let $J<0$ *hold. Then*

(iv) $J_n\leqq 0$ *on* $[0,1]$; $J_p\leqq 0$ *on* $[0,1/2\alpha)$, $J_p\geqq 0$ *on* $(1/2\alpha,1]$,

(v) u *is decreasing on* $[0,1]$, $u\geqq e^V$,

(vi) v *is increasing on* $[0,1/2\alpha]$ *and decreasing on* $[1/2\alpha,1]$, $v\geqq e^V$.

Therefore, for any solution $(V=0,\psi^*,u^*,v^*)\neq(V=0,\psi_e,1,1)$ *of* (1.11)–(1.14) (*with* $\gamma>\frac{1}{2}$) $I(\psi^*)=0$ *has to hold.*

We now turn to Poisson's equation (1.11) subject to the boundary conditions (1.14a). Differentiating (1.11) gives

$$\lambda^2(\psi')''=(n+p)\psi'+J,\qquad 0\leqq x\leqq 1 \tag{2.8a}$$

and (1.14) implies

$$(\psi'(0))' = -\frac{1}{\lambda^2} \qquad (\psi'(1))' = 0. \tag{2.8b}$$

The maximum principle (see Protter and Weinberger (1967)) yields

LEMMA 2.3. *$J \leqq 0$ implies $\psi' \geqq 0$ on $[0,1]$ and $V \leqq \psi_e(1)$. Thus the a priori bound*

$$0 \leqq \psi(x) \leqq \psi(1) = \psi_e(1) - V, \qquad x \in [0,1] \tag{2.9}$$

follows from $J \leqq 0$.

Differentiating (2.8a) and using (1.11a) implies

$$\lambda^2(\psi'')'' = \left[(n+p) + \lambda^2\psi'^2\right]\psi'' + (\psi')^2 + (J_n - J_p)\psi', \qquad 0 \leqq x \leqq 1, \tag{2.10a}$$

$$\psi''(0) = -\frac{1}{\lambda^2}, \qquad \psi''(1) = 0. \tag{2.10b}$$

From (2.3) we conclude that $J_n - J_p \leqq 0$ in $[0,1]$ for $J \leqq 0$. Since $\psi' \geqq 0$, we obtain

LEMMA 2.4. *$J \leqq 0$ implies $\psi'' \geqq 1/\lambda^2$ on $[0,1]$, and therefore*

$$n(x) \geqq p(x), \qquad 0 \leqq x \leqq 1 \tag{2.11}$$

holds.

Proof. $z = -(1/\lambda^2)$ is a lower solution of (2.10) and (2.11) follows from (1.1) (with $D \equiv 1$ on $[0,1]$).

We now derive upper bounds for n.

LEMMA 2.5. *Let $J \leqq 0$ hold. Then*

$$n(x) \leqq \frac{1+\sqrt{1+4\delta^4}}{2} + \frac{|J|}{2} \cdot \begin{cases} (2\alpha+1)(1-x) & \text{for } x \in [0,1] \quad \text{if } 0 \leqq \alpha \leqq \frac{1}{2}, \\ (e^{2\alpha-1}+1)(1-x) & \text{for } x \in \left[\frac{1}{2\alpha}, 1\right] \quad \text{if } \alpha > \frac{1}{2}, \\ \frac{1}{2} - 2x + (e^{2\alpha-1}+1)\left(1 - \frac{1}{2\alpha}\right) & \\ \qquad \text{for } x \in \left[0, \frac{1}{2\alpha}\right] \quad \text{if } \alpha > \frac{1}{2}. & \end{cases} \tag{2.12}$$

Proof. We multiply (2.4a) by $\delta^2 e^{\psi}$ (getting n) and estimate $e^{\psi(x)-\psi(s)} \leqq 1$ for $s \geqq x$ (ψ is nondecreasing). (2.12) is then obtained by integration and estimating

$$\frac{1}{2\alpha}\left(1 - e^{-2\alpha(1-x)}\right) \leqq 1 - x.$$

In the case of zero generation ($\alpha \equiv 0$) we obtain upper bounds for n and p which are independent of J and V.

LEMMA 2.6. *Let $\alpha \equiv 0$ and $J \leqq 0$ hold. Then $\psi'' \leqq 0$ on $[0,1]$ and*

$$p(x) \leqq n(x) \leqq p(x) + 1, \qquad 0 \leqq x \leqq 1, \tag{2.13}$$

$$0 < n(x) + p(x) \leqq \sqrt{1+4\delta^4}, \qquad 0 \leqq x \leqq 1 \tag{2.14}$$

hold.

Proof. $\alpha = 0$ implies $J_n \equiv J_p \equiv J/2$ and therefore $\bar{z} = 0$ is an upper solution of (2.10). (2.13) follows from (1.1a). Also $(n+p)' = \psi'(n-p)$ holds. Thus $n+p$ is nondecreasing and $n + p \leqq n(1) + p(1) = \sqrt{1+4\delta^4}$ follows.

We now derive lower and upper bounds for ψ using the estimates of n and p.

LEMMA 2.7. *Let $J \leqq 0$ hold. Then*

$$\psi(x) \leqq \left(\psi(1)+\frac{1}{2\lambda^2}\right)x-\frac{x^2}{2\lambda^2}, \qquad 0 \leqq x \leqq 1 \tag{2.15}$$

follows. If $\alpha \equiv 0$

$$\psi(x) \geqq -\frac{x^2}{2\lambda^2}+\psi(1)x \tag{2.16}$$

holds.

Proof. $\bar{\psi}(x)=(\psi(1)+1/2\lambda^2)x-x^2/2\lambda^2$ solves

$$\lambda^2\bar{\psi}''=-1, \quad \bar{\psi}(0)=0, \quad \bar{\psi}(1)=\psi(1).$$

Lemma 2.4 implies that $\bar{\psi}$ is an upper solution of (1.11a). For $\alpha \equiv 0$ we obtain by integrating $\psi'' \leqq 0$: $\psi(1) \leqq \psi'(0)$ and from $\psi'' \geqq -1/\lambda^2$

$$\psi(x) \geqq -\frac{x^2}{2\lambda^2}+\psi'(0)x \geqq -\frac{x^2}{2\lambda^2}+\psi(1)x.$$

To get a lower bound for ψ in the avalanche case we prove

LEMMA 2.8. *Let $J \leqq 0$, $|J| \geqq K_1$ hold. Also let $0<\varepsilon<1$ be such that $|J|\varepsilon^5 \geqq K_2$ where K_1, K_2 only depend on λ, δ and γ. Then*

$$\psi'(x) \geqq \frac{1}{\varepsilon+g_\alpha(x)}+\sigma(x)\frac{1}{\sqrt{|J|\varepsilon^5}}, \qquad 0 \leqq x \leqq 1 \tag{2.17}$$

holds where $\|\sigma\|_{[0,1]} \leqq K_2$ (K_2 only depends on λ, δ and γ) and

$$g_\alpha(x)=\begin{cases} (2\alpha+1)(1-x) & \textit{for } x \in [0,1] \quad \textit{if } 0 \leqq \alpha \leqq \frac{1}{2}, \\ (e^{2\alpha-1}+1)(1-x) & \textit{for } x \in \left[\frac{1}{2\alpha},1\right] \quad \textit{if } \alpha > \frac{1}{2}, \\ \frac{1}{\alpha}-2x+(e^{2\alpha-1}+1)\left(1-\frac{1}{2\alpha}\right) & \textit{for } x \in \left[0,\frac{1}{2\alpha}\right] \quad \textit{if } \alpha>\frac{1}{2}. \end{cases} \tag{2.18}$$

Proof. Lemmas 2.4 and 2.5 imply

$$\frac{n+p}{|J|} \leqq \frac{2n}{|J|} \leqq \frac{1+\sqrt{1+4\delta^4}}{|J|}+g_\alpha(x).$$

We now choose ε such that $1+\sqrt{1+4\delta^4} \leqq |J|\varepsilon$. Thus $(n+p)/|J| \leqq \varepsilon+g_\alpha(x)$ holds. Obviously the solution $y(\geqq 0)$ of

$$\frac{\lambda^2}{|J|}y''=(\varepsilon+g_\alpha(x))y-1, \tag{2.19a}$$

$$y'(0)=-\frac{1}{\lambda^2}, \qquad y'(1)=0 \tag{2.19b}$$

is a lower solution of (2.8), which means that $0 \leqq y \leqq \psi'$ holds on $[0,1]$. For large $|J|$ the problem constitutes a linear singularly perturbed (Neumann-type) boundary value problem with the reduced solution (obtained by setting $\lambda^2/|J|$ to zero):

$$y_r=\frac{1}{\varepsilon+g_\alpha(x)}.$$

A standard singular perturbation analysis (see Howes (1978)) which takes the possible smallness of ε into account ($y_r(1)=1/\varepsilon$!!) gives

$$\|y-y_r\|_{[0,1]}\leqq K_3\frac{1}{\sqrt{|J|\varepsilon^5}}$$

whenever $|J|\geqq K_1$, $|J|\varepsilon^5\geqq K_2$. This implies (2.17).

A lower bound for ψ follows by integrating (2.17):

LEMMA 2.9. *Let the assumption of Lemma 2.7 hold. Then*

$$\psi(x)\geqq\mu(x)\frac{1}{\sqrt{|J|\varepsilon^5}}+\begin{cases} L_1+\dfrac{1}{2\alpha+1}\ln\left(\dfrac{1}{\varepsilon+(2\alpha+1)(1-x)}\right) \\ \qquad\qquad \textit{for } x\in[0,1] \quad \textit{if } 0\leqq\alpha\leqq\dfrac{1}{2} \\ L_2 \quad \textit{for } x\in\left[0,\dfrac{1}{2\alpha}\right] \quad \textit{if } \alpha>\dfrac{1}{2} \\ L_3+\dfrac{1}{e^{2\alpha-1}+1}\ln\left(\dfrac{1}{(e^{2\alpha-1}+)(1-x)+\varepsilon}\right) \\ \qquad\qquad \textit{for } x\in\left[\dfrac{1}{2\alpha},1\right] \quad \textit{if } \alpha>\dfrac{1}{2}\end{cases} \tag{2.20}$$

holds where $\|\mu\|_{[0,1]}\leqq K_2$ *and* L_1, L_2, L_3 *only depend on* γ.

Since ε can be made arbitrarily small when $J\to-\infty$ (still keeping $|J|\varepsilon^5$ large), Lemmas 2.8 and 2.9 imply that $\|\psi\|_{[0,1]}$ ($\geqq\psi'(1)\geqq K_4(1/\varepsilon)$) and $\psi(1)$ ($\geqq K_5\ln(1/\varepsilon)$) become unbounded as $J\to-\infty$.

We also need an upper bound of ψ':

LEMMA 2.10. $J\leqq 0$ *implies*

$$(0\leqq)\,\psi'\leqq K_6e^{|V|}(|J|+1) \tag{2.21}$$

where K_6 *only depends on* λ *and* δ.

Proof. (2.4a) implies $n+p\geqq n\geqq\delta^2e^V$. Thus the solution w of

$$\lambda^2w''=\delta^2e^Vw-|J|,\quad w'(0)=-\frac{1}{\lambda^2},\quad w'(1)=0$$

is an upper solution of (2.8). Therefore

$$\psi'(x)\leqq w(x)=e^{-V}\frac{|J|}{\delta^2}+\frac{\cosh\left((\delta/\lambda)^{V/2}(1-x)\right)}{\lambda\delta e^{V/2}\sinh\left((\delta/\lambda)e^{V/2}\right)}$$

and (2.21) follows since $V\leqq\psi_e(1)$.

We now employ the derived bounds to get a priori estimates on the current-voltage characteristic.

THEOREM 2.1. *Assume that* $(J,\psi,u,v)\in C^-$ *and* $(V,J)\in J^-$. *Then*

$$C_1|V|\leqq|J|\leqq C_2|V|\quad \textit{if } \alpha(\psi')=0, \tag{2.22}$$

$$C_1|V|\leqq|J|\leqq C_3e^{|V|}(1-e^{-2|V|})\quad \textit{if } 0<\alpha(\psi')\leqq\frac{1}{2} \tag{2.23}$$

holds. If $\alpha(\psi')>\frac{1}{2}$ *then*

$$C_4(1-e^{-2|V|})e^{|V|(1-1/2\alpha)}\leqq|J|\leqq D\exp\left((5+\mu)(e^{2\alpha-1}+1)|V|\right) \tag{2.24}$$

holds for $\frac{1}{2} < \alpha(\psi') < \alpha_0$ *with some* $\alpha_0 < 1$ *and every* $\mu > 0$. *If* $\alpha(\psi') \geqq \alpha_0$ *then*

$$C_5(1-e^{-2|V|})e^{|V|} \leqq |J| \leqq D\exp\big((5+\mu)(e^{2\alpha-1}+1)|V|\big). \tag{2.25}$$

The constants $C_1, \cdots, C_5$ *only depend on* λ, δ *and* γ. D *depends on* λ, δ, γ *and* μ.

Proof. $0 \leqq \alpha \leqq \frac{1}{2}$ and $\psi \geqq 0$ imply $0 < I(\psi) \leqq 2\int_0^1 \cosh \psi(s)\, ds$ and (2.15), (2.7) yield:

$$0 < I(\psi) \leqq \frac{2\sinh(\psi(1)+1/2\lambda^2)}{\psi(1)+1/2\lambda^2}.$$

Thus

$$|J| \geqq \frac{\delta^2(1-e^{-2|V|})(\psi(1)+1/2\lambda^2)}{e^{-|V|}\sinh(\psi(1)+1/2\lambda^2)},$$

and the lower bounds for $|J|$ in (2.22), (2.23) follows. $\alpha \equiv 0$ implies $I(\psi) = 2\int_0^1 \cosh \psi(s)\, ds$ and the estimate (2.16) gives

$$I(\psi) \geqq 2\frac{\sinh(\psi(1)-1/2\lambda^2)}{\psi(1)-1/2\lambda^2}$$

when $|V|$ is so large that $\psi(1)-1/2\lambda^2 > 0$. We derive (using (2.7)):

$$|J| \leqq \frac{\delta^2(1-e^{-2|V|})(\psi(1)-1/2\lambda^2)}{e^{-|V|}\sin(\psi(1)-1/2\lambda^2)},$$

and (2.22) is proven. For $0 < \alpha \leqq \frac{1}{2}$ we estimate

$$I(\psi) \geqq \int_0^1 (-2\alpha s+1)\, ds = 1-\alpha.$$

Thus

$$|J| \leqq \frac{2\delta^2 e^{|V|}(1-e^{-2|V|})}{1-\alpha}$$

and the upper bound in (2.23) follows. Now let $\alpha(\psi') > \frac{1}{2}$. Then, since ψ is nondecreasing and since $g_\alpha(x)$ is positive in $[0, 1/2\alpha)$ and negative in $(1/2\alpha, 1]$:

$$\begin{aligned} I(\psi) &\leqq \int_0^1 f_\alpha(s)\, ds + e^{\psi(1/2\alpha)}\int_0^1 g_\alpha(s)\, ds \\ &= 1-\frac{1}{4\alpha}+\frac{1}{2\alpha}e^{2\alpha-1}+e^{\psi(1/2\alpha)}\left(1+\frac{1}{4\alpha}-\frac{1}{2\alpha}e^{2\alpha-1}\right) \end{aligned}$$

($f_\alpha \geqq 0$ in $[0,1]$!) hold. The function $h(\alpha) = 1+1/4\alpha-(1/2\alpha)e^{2\alpha-1}$ has a unique zero $\alpha_0 \in (\frac{1}{2}, 1)$. Thus

$$I(\psi) \leqq \begin{cases} 2, & \alpha \geqq \alpha_0, \\ 1-\dfrac{1}{4\alpha}+\dfrac{1}{2\alpha}e^{2\alpha-1}+\dfrac{1}{2}e^{\psi(1/2\alpha)}, & \dfrac{1}{2} < \alpha < \alpha_0 \end{cases}$$

holds. If $I(\psi) > 0$, the lower bound in (2.25) follows and the lower bound in (2.24) (also for $I(\psi) > 0$) is implied by (2.15) which gives

$$\psi\left(\frac{1}{2\alpha}\right) \leqq c + |V|\frac{1}{2\alpha}.$$

Lemma 2.9 implies '$V\to-\infty$ as $J\to-\infty$' and therefore $I(\psi)\leqq 0$ can (for $J\leqq 0$) only hold for $|J|\leqq F$, where F depends on λ,δ and γ. This proves the lower bounds in (2.24), (2.25).

(2.20) yields $\psi(1)=\psi_e(1)-V\geqq L_3+(1/(e^{2\alpha-1}+1))\ln(1/\varepsilon)$. We set $\varepsilon=1/|J|^{1/5-\sigma}$ for $0<\sigma<\frac{1}{5}$ and obtain the upper bounds in (2.24), (2.25) for $|J|$ sufficiently large since $V\leqq\psi_e(1)$ holds.

Since Lemma 2.8 implies that $\alpha(\psi')\to\gamma$ as $J\to-\infty$ (or as $V\to-\infty$), the theorem proves the current J increases (in absolute value) at least (and at most) exponentially as $V\to-\infty$ in the avalanche case $\gamma>\frac{1}{2}$. For zero generation ($\gamma=0$) the increase is at most (and at least) linear. For the intermediate case $0<\gamma<\frac{1}{2}$ the increase is at least linear and at most exponential. The author conjectures that the distinction of the cases $\frac{1}{2}\leqq\gamma<\alpha_0$ and $\gamma\geqq\alpha_0$ only comes in for technical reasons and that (2.25) holds for all $\gamma>\frac{1}{2}$.

3. Existence theorems. We need the following:

LEMMA 3.1. *Assume that* $a,b,f\in C([0,1])$ *and* $a,\ b>0$ *on* $[0,1]$. *Then the two point boundary value problem*

$$w''=a(x)e^{w+\eta}-b(x)e^{-w-\eta}-f(x),\qquad 0\leqq x\leqq 1,\tag{3.1a}$$

$$w(0)=\mu_0,\qquad w(1)=\mu_1\tag{3.1b}$$

has a unique solution $w=w(\eta,\mu_0,\mu_1,f,x)$ *for all* $\eta,\ \mu_0,\ \mu_1\in\mathbb{R}$, $f\in C([0,1])$. *The map*

$$w:\begin{cases}(\eta,\mu_0,\mu_1,f)\to w(\eta,\mu_0,\mu_1,f,\cdot),\\ \mathbb{R}^3\times C^1([0,1])\to C^1([0,1])\end{cases}$$

is completely continuous.

Proof. To prove existence of a solution of (3.1) we consider the operator M: $C([0,1])\times[0,1]\to C[0,1]$ defined by $M(z,\sigma)=y$ where y is the solution of

$$L(y,z,\sigma):=\ y''-\sigma\big(a_1(x)e^z-b_1(x)e^{-z}+f(x)\big)=0,\qquad 0\leqq x\leqq 1,$$
$$y(0)=\sigma\mu_0,\qquad y(1)=\sigma\mu_1$$

with $a_1(x)=a(x)e^{\eta}$, $b_1(x)=b(x)e^{-\eta}$. Every fixed point of $M(\cdot,1)$ is a solution of (3.1) and vice versa.

Clearly, M is completely continuous. Also $M(z,0)=0$ for all $z\in C([0,1])$. We denote

$$\bar{a}_1:=\max_{0\leqq x\leqq 1}a_1(x),\qquad \alpha_1:\ \min_{0\leqq x\leqq 1}a_1(x)(>0),$$
$$\bar{b}_1:=\max_{0\leqq x\leqq 1}b_1(x),\qquad \underline{b}_1:=\min_{0\leqq x\leqq 1}b_1(x)(>0)$$
$$y_1=\min\big(-|\mu_0|,-|\mu_1|,\underline{y}\big),\qquad y_2=\max\big(|\mu_0|,|\mu_1|,\bar{y}\big)\ \cdot$$

where $\underline{y},\bar{y}$ are the (unique) solution of the equations

$$0=\bar{a}_1e^{\underline{y}}-\underline{b}_1e^{-\underline{y}}+\|f\|_{[0,1]},$$

$$0=\underline{a}_1e^{\bar{y}}-\underline{b}_1e^{-\bar{y}}-\|f\|_{[0,1]}.$$

Then $L(y_1,y_1,\sigma)\geqq 0$, $L(y_2,y_2,\sigma)\leqq 0$ holds for all $\sigma\in[0,1]$. Thus y_1,y_2 are lower and upper solutions respectively of the equation $L(y,y,\sigma)=0$ for arbitrary $\sigma\geqq 0$ $((d/dy)(a_1e^y-b_1e^{-y}+f(x))=a_1e^y+b_1e^{-y}>0$ holds!).

Therefore all possible fixed points y of $M(\cdot,\sigma)$ fulfill the a priori estimate $\|y\|_{[0,1]} \leqq \max(|y_1|,|y_2|)$ for all $\sigma \in [0,1]$ and the Leray–Schauder theorem implies the existence of a fixed point w of $M(\cdot,1)$.

Assume now that (3.1) has the solutions w_1 and w_2. Then $g = w_1 - w_2$ fulfills

$$\begin{aligned} &g'' = \left(a(x)e^{\xi_1(x)+\eta} + b(x)e^{-\xi_2(x)-\eta}\right)g, \qquad 0 \leqq x \leqq 1, \\ &g(0) = g(1) = 0 \end{aligned}$$

where $\xi_1(x)$, $\xi_2(x)$ are between $w_1(x)$ and $w_2(x)$. $g \equiv 0$ follows immediately. The problem (3.1) is therefore uniquely soluble.

To prove the complete continuity of the map w, we take $\eta \in [\underline{\eta},\bar{\eta}]$, $\mu_0 \in [\underline{\mu}_0,\bar{\mu}_0]$, $\mu_1 \in [\underline{\mu}_1,\bar{\mu}_1]$ and denote

$$\underline{a} := \min_{0 \leqq x \leqq 1} a(x), \qquad \bar{a} := \max_{0 \leqq x \leqq 1} a(x)$$

(analogously for b). Then the unique solution w_1 of

$$\begin{aligned} &w_1'' = \bar{a}e^{w_1+\bar{\eta}} - \underline{b}e^{-w_1-\underline{\eta}} + \|f\|_{[0,1]}, \qquad 0 \leqq x \leqq 1, \\ &w_1(0) = \underline{\mu}_0, \qquad w_1(1) = \underline{\mu}_1 \end{aligned}$$

is a lower solution of (3.1) and the unique solution w_2 of

$$\begin{aligned} &w_2'' = \underline{a}e^{w_2+\underline{\eta}} - \bar{b}e^{-w_2-\bar{\eta}} - \|f\|_{[0,1]}, \qquad 0 \leqq x \leqq 1 \\ &w_2(0) = \bar{\mu}_0, \qquad w_2(1) = \bar{\mu}_1 \end{aligned}$$

is an upper solution of (3.1), i.e. $w_1 \leqq w \leqq w_2$ on $[0,1]$ holds. Since

$$\|w''\|_{[0,1]} \leqq \bar{a}e^{w_2+\bar{\eta}} + \bar{b}e^{-w_1-\underline{\eta}} + \|f\|_{[0,1]}$$

holds, Ascoli's theorem implies that w: $\mathbb{R}^3 \times C^1([0,1]) \to C^1([0,1])$ maps bounded sets into precompact sets. The continuity of w is immediate.

Now we prove the basic

THEOREM 3.1. *For any* $\gamma \geqq 0$ *the solution set* C^- *of* (1.11)–(1.13), (1.17) *contains an unbounded continuum* $\tilde{C}^-$ (*in the* $(-\infty,0] \times (C^2([0,1]))^3$-*topology*) *emanating from the equilibrium solution* $(0,\psi_e,1,1)$ *whose projection into* $(-\infty,0]$ *equals* $(-\infty,0]$ (i.e. $\tilde{C}^-$ *contains a solution* (J,ψ,u,v) *or every* $J \leqq 0$).

Proof. We regard $V = V(J,\psi)$ (given by (2.5)) as functional V: $(-\infty,0] \times C^1([0,1]) \to R$. The continuity of I implies the continuity of V. Using (2.3)–(2.5) we rewrite Poisson's equation (1.11) as

$$\lambda^2\psi'' = \delta^2 e^{\psi + V(J,\psi)} - \delta^2 e^{-\psi - V(J,\psi)} - 1 + |J|G(\psi)(x), \qquad 0 \leqq x \leqq 1, \tag{3.2a}$$

$$\psi(0) = 0, \psi(1) = \psi_e(1) - V(J,\psi) \tag{3.2b}$$

with

$$G(\psi)(x) = e^{\psi(x)}\int_x^1 e^{-\psi(s)}\left|\frac{J_n(s)}{J}\right|ds + e^{-\psi(x)}\int_x^1 e^{\psi(s)}\left|\frac{J_p(s)}{J}\right|ds \tag{3.3}$$

(note that $J_n(x)/J|$, $|J_p(x)/J|$ are independent of J, they only depend on $\alpha(\psi')$ and on x). G: $C^1([0,1]) \to C^1([0,1])$ is continuous since α: $C^1([0,1]) \to [0,\gamma]$ is continuous.

We set $\psi=\psi_e+\phi$ and rewrite (3.2) as the fixed point problem $\phi=T(J,\phi)$ where $y=T(J,z)$ is defined as the unique solution of the problem

$$\begin{aligned}
&(3.4\text{a}) \quad \lambda^2 y''=\delta^2\exp(y+\psi_e+V(J,\psi_e+z))\\
&(3.4\text{b}) \qquad -\delta^2\exp(-y-\psi_e-V(J,\psi_e+z))-1-\lambda^2\psi_e''+|J|G(\psi_e+z), \qquad 0\leqq x\leqq 1\\
&\qquad\quad y(0)=0,\; y(1)=-V(J,\psi_e+z)
\end{aligned}$$

(T is well defined since Lemma 3.1 implies the unique solvability of (3.4)). Lemma 3.1 and the continuity of V and G imply that T: $(-\infty,0]\times C^1([0,1])$ is completely continuous. $V(0,w)\equiv 0$ and therefore $y=T(0,z)$ is given by the solution of

$$\begin{aligned}
&(3.5\text{a}) \qquad \lambda^2 y''=\delta^2 e^{y+\psi_e}-\delta^2 e^{-y-\psi_e}-1-\lambda^2\psi_e'', \qquad 0\leqq x\leqq 1,\\
&(3.5\text{b}) \qquad y(0)=y(1)=0.
\end{aligned}$$

$y\equiv 0$ follows. From Rabinowitz (1971, Thm. 3.2) we conclude that the solution set of $T(J,\rho)=\phi$ contains an unbounded continuum E^-(in the $(-\infty,0]\times C^1([0,1])$-topology). Theorem 2.1 and Lemma 2.3 imply

$$|V|\leqq\max\left(\psi_e(1),\frac{|J|}{C_1}\right)$$

and

$$-\psi_e\leqq\phi\leqq\psi_\varepsilon(1)-\psi_e+|V|.$$

Lemma 2.10 yields

$$-\psi_e'\leqq\phi'\leqq K_6 e^{|V|}(|J|+1)-\psi_e'.$$

We conclude from these estimates

$$\|\phi\|_{[0,1]}+\|\phi'\|_{[0,1]}\leqq M(1+|J|+e^{|J|/C_1}+e^{|J|/C_1}|J|)$$

where $M>0$ is independent of J (and V). Since E^- is unbounded, it has to contain solutions (J,ϕ) for all $J\leqq 0$. The statement of the theorem follows by observing that u and v as given by (2.4) are continuous as functions of (J,ψ) in the $(-\infty,0]\times C^1([0,1])$-topology.

The most important implication of Theorem 3.1 is:

THEOREM 3.2. *For any $\gamma\geqq 0$ the current-voltage characteristic J^- contains a continuous curve Γ^- emanating from $(0,0)$ whose projection Γ_J^- into the $J-$axis equals $(-\infty,0]$ and whose projection Γ_V^- into the V-axis fulfills*

$$\begin{aligned}
&(3.6\text{a}) \qquad \Gamma_V^-=(-\infty,0] \quad \textit{for}\; 0\leqq\gamma\leqq\frac{1}{2},\\
&(3.6\text{b}) \qquad (-\infty,0]\subseteq\Gamma_V^-\subseteq(-\infty,\psi_e(1)] \quad \textit{for}\;\gamma>\frac{1}{2}.
\end{aligned}$$

Proof. Theorem 3.1 implies that J^- contains a continuum 0^- emanating from $(0,0)$ whose projection into the J-axis equals $(-\infty,0]$. From Lemma 2.9 we conclude that $\psi(1)=\psi_e(1)-V$ is positive and unbounded as $J\to-\infty$. Lemma 2.1, (ii) implies that $V\leqq 0$ and $J\leqq 0$ and $0\leqq\gamma\leqq\frac{1}{2}$. Therefore (3.6a) follows. (3.6b) is concluded by noting that $\psi(1)=\psi_e(1)-V\geqq 0$ for $J\leqq 0$ holds (see Lemma 2.3).

The obvious consequence for the solution set

(3.7)

$$D^{-}=\left\{(U,\psi,u,v)\in(-\infty,0]\times\left(C^{2}([0,1])\right)^{3}\,|\,(\psi,u,v)\text{ solves (1.11)–(1.14) for }V=U\right\}$$

of the voltage-controlled diode is

COROLLARY 3.1. *D^{-} contains an unbounded continuum $\tilde{D}^{-}$ emanating from $(0,\psi^{*},v^{*},u^{*})$ whose projection into $(-\infty,0]$ equals $(-\infty,0]\cdot(\psi^{*},u^{*},v^{*})=(\psi_{e},1,1)$ if the equilibrium solution is unique (e.g. for $0\leqq\gamma\leqq\frac{1}{2}$). $J\leqq 0$ holds for every $(V,\psi,u,v)\in\tilde{D}^{-}$. ($\tilde{D}^{-}$ contains a solution for every $V\leqq 0$).*

We now show that multiple solutions of (1.11)–(1.14) for $V=0$ can occur.

THEOREM 3.3. *Assume that $\alpha(\psi_{e}')>H(>\frac{1}{2})$ where $H>0$ only depends on λ and δ. Then there is a solution (ψ^{*},u^{*},v^{*}) of (1.11)–(1.14) for $V=0$ which is different from the equilibrium solution $(\psi_{e},1,1)$ and $J^{*}=\delta^{2}e^{\psi^{*}}(u^{*})'-\delta^{2}e^{-\psi^{*}}(v^{*})'<0$ holds.*

Proof. We obtain from (2.6)

$$I(\psi)=2\left(\int_{0}^{1}\cosh\psi(s)\,ds-\int_{0}^{1}h_{\alpha}(s)\sinh\psi(s)\,ds\right)$$

where

$$h_{\alpha}(x)=\begin{cases}2\alpha x, & 0\leqq x\leqq\dfrac{1}{2\alpha},\\ e^{2\alpha x-1}, & \dfrac{1}{2\alpha}\leqq x\leqq 1,\end{cases}\qquad\left(\text{for }\alpha(\psi')>\frac{1}{2}\right).$$

Obviously $h_{\alpha}(x)\geqq 2\alpha x$ on $[0,1]$ and therefore

$$I(\psi)\leqq 2\left(\int_{0}^{1}\cosh\psi(s)\,ds-2\alpha\int_{0}^{1}s\sinh\psi(s)\,ds\right)$$

holds. Choosing α such that

$$\alpha(\psi_{e}')>\frac{1}{2}\,\frac{\int_{0}^{1}\cosh\psi_{e}(s)\,ds}{\int_{0}^{1}s\sinh\psi_{e}(s)\,ds}$$

implies $I(\psi_{e})<0$. Thus there is a neighbourhood N of 0 in $C^{1}([0,1])$ with $I(\psi_{e}+\phi)<0$ for $\phi\in N$. (2.5) implies $V(J,\psi_{e}+\phi)>0$ for $(J,\phi)\in[-\infty,0)\times N$. Since the continuum E^{-} (used in the proof of Theorem 3.1) emanates in $(0,0)$, the intersection of E^{-} and $(-\infty,0)\times N$ is not empty. This implies that there are solutions of (1.11)–(1.13), (1.17) for $J<0$ for which $V>0$ holds. Since V is negative for $J<0$, $|J|$ sufficiently large, there has to be $(J,\phi^{*})\in E^{-}$, $J\neq 0$, with $I(\psi_{e}+\phi^{*})=0$. This gives $V(J,\psi_{e}+\phi^{*})=0$ and Theorem 3.3 follows.

The condition $I(\psi_{e})<0$ implies that J^{-} is not contained in $(-\infty,0]^{2}$ (see Fig. 2). However, $I(\psi_{e})\leqq 0$ is physically unreasonable and it is not clear whether the nonuniqueness of the equilibrium solution prevails if $I(\psi_{e})>0$.

We conclude from the Theorems 2.1 and 3.2 that the avalanche case $\gamma>\frac{1}{2}$ is distinguished from the nonavalanche case $0\leqq\gamma\leqq\frac{1}{2}$ by a more rapid decrease of the current J as $V\to-\infty$ (see Fig. 3).

We remark that an investigation of (1.11)–(1.13), (1.17) for nonnegative current can be done in a similar fashion. A relation of the form (2.5) (with a different functional I) holds and the existence of an unbounded continuum of solutions E^{+} follows as in the proof of Theorem 3.1. To conclude that E^{+} contains solutions for all

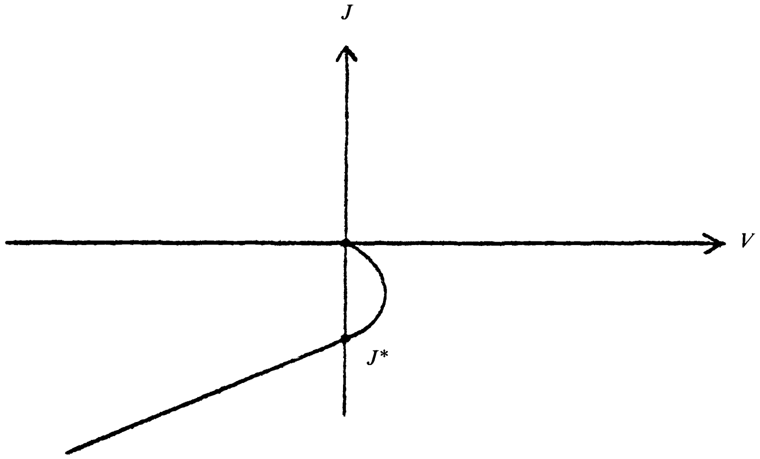


FIG. 2. *Qualitative structure of the $J-V$-characteristic for $I(\psi_e)<0$.*

$J \geqq 0$, additional a priori estimates have to be obtained (since the estimates in §2 only hold for $J \leqq 0$). For the case $\alpha \equiv 0$ these estimates are given in Mock (1983), Markowich (1983).

Finally, assume that α is not a functional on $C^1([0,1])$ but simply a continuous and nondecreasing function α: $\mathbb{R} \to [0,\gamma]$, such that the ionization rate $\alpha(\psi'(x))$ is space-dependent. Then (2.4)–(2.6) have to be modified by substituting 'αs' in the integrands by $\int_0^s \alpha(\psi'(s))\,ds$ and "$1/2\alpha$" in the integration intervals by that value $\bar{x} \in [0,1]$ for which $\int_0^{\bar{x}} \alpha(\psi'(s))\,ds = \frac{1}{2}$ holds. Theorem 3.1 still holds. By estimating (2.12) in terms of γ, an analogue of (2.18), (2.20) (also in terms of γ) implying '$\psi(1) \to \infty$ as $J \to \infty$' is obtained and Theorem 3.2 and Corollary 3.1 follow. The estimates of the current-voltage characteristic given in Theorem 2.1 for $0 \leqq \gamma < \frac{1}{2}$ still hold. The avalanche-estimate (2.25) (with α in the upper bound substituted by γ) holds if for example $\alpha(\tau) \geqq \sigma_1 e^{-\sigma_2/|\tau|}$ with σ_1 sufficiently large and σ_2 sufficiently small.

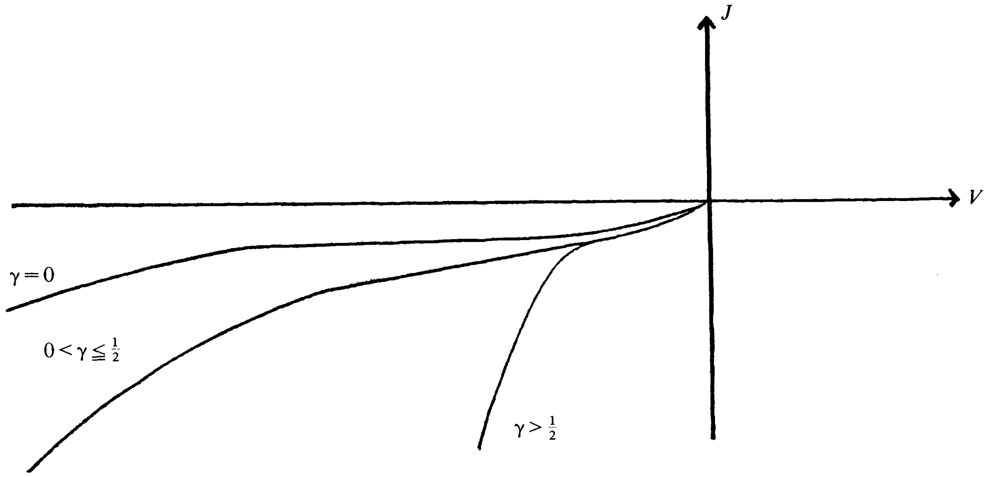


FIG. 3. *Qualitative structure of the $J-V$-characteristic for various γ's ($I(\psi_e)>0$ is assumed) for reverse bias.*

REFERENCES

A. HOWES (1978), *An asymptotic theory for a class of nonlinear Robin problems*, J. Differential Equations, 30, pp. 192–234.

M. A. KRASNOSELSKII (1964), *Positive Solutions of Operator Equations*, Noordhoff, Groningen.

P. A. MARKOWICH AND C. A. RINGHOFER (1984), *A singularly pertrubed value problem modelling a semiconductor device*, SIAM J. Appl. Math., 44 (1984), pp. 231–256.

P. A. MARKOWICH (1983), *A singular perturbation analysis of the fundamental semiconductor device equations*, submitted.

M. S. MOCK (1983), *Analysis of Mathematical Models of Semiconductor Devices*, Boole Press, Dublin.

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

P. H. RABINOWITZ (1971), *Some global results for nonlinear eigenvalue problems*, J. Funct. Analysis, 7, pp. 487–513.

A. SCHÜTZ (1982), *Simultation des Lawinendurchburchs in MOS-Transistoren*, Ph. D. thesis, Technical University of Vienna.

A. SCHÜTZ, S. SELBERHERR AND H. W. PÖTZL (1982), *A two-dimensional model of the avalanche effect in MOS-transistors*, Solid-State Electronics, 25, pp. 177–183.

S. M. SZE (1981), *Physics of Semiconductor Devices*, second edition, Wiley Interscience, New York.

W. V. VAN ROOSBRECK (1950), *Theory of flow of electrons and holes in gemanium and other semiconductors*, Bell System Tech. J., 29, pp. 560–607.

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

010

EXTREMAL PROBLEMS FOR EIGENVALUE FUNCTIONALS*

DAVID C. BARNES[†]

Abstract. We consider the eigenvalues, $\lambda_n(\rho)$, of self-adjoint Sturm–Liouville systems to be real valued functionals of certain coefficient functions in the differential equation. We introduce a classical (in general nonlinear) functional $K(\rho)$ which is tangent to $\lambda_n(\rho)$ at a fixed function ρ^*. That is, $\lambda_n(\rho^*)=K(\rho^*)$ and $\delta\lambda_n=\delta K$ at ρ^*. Then by using classical calculus of variations on $K(\rho)$ we show how to find extremals of $\lambda_n(\rho)$ over certain classes of functions ρ.

1. Introduction. Consider the eigenvalue problem

$$(fy')'+(\lambda g+q)y=0, \quad 0\le x\le l, \tag{1.1}$$

with the self-adjoint boundary conditions of the form

$$\begin{aligned}\alpha_1 y(0)+\alpha_2 y'(0)+\alpha_3 y(l)+\alpha_4 y'(l)=0,\\ \beta_1 y(0)+\beta_2 y'(0)+\beta_3 y(l)+\beta_4 y'(l)=0.\end{aligned} \tag{1.2}$$

The coefficient functions f,g,q may depend on x as well as some other function $\rho(x)$ and its derivative $\rho'(x)$. Following Troutman [21], we will sometimes abbreviate (abuse?) notation so that, for example,

$$f=f(x)=f[\rho(x)]=f(x,\rho(x),\rho'(x)),$$

with similar meanings for $g[\rho(x)]$, $q(x)$, etc. We will also allow the coefficients α_i, β_i in (1.2) to depend on the values of ρ and ρ' at $x=0$ and l. Thus boundary conditions like $\rho'^2(x)y'(x)\to 0$ as $x\to l$ may be used.

Assume that f,g,q are continuously differentiable functions of the three independent variables (x,ρ,ρ'). We define a class $\mathfrak{C}$ of functions $\rho(x)$ by the following conditions:

1. $\rho(x)$ is piecewise continuous on $0\le x\le l$;
2. $f[\rho(x)]>0$ and $g[\rho(x)]>0$ for $0<x<l$;
3. self-adjointness, for any functions y,z which satisfy (1.2),

$$f[\rho(x)](yz'-zy')\big|_0^l=0; \tag{1.3}$$

4. two boundary conditions on the values of $\rho(x)$ and $\rho'(x)$ may (or may not) be prescribed at $x=0,1$.

Thus for any $\rho\in\mathfrak{C}$ the eigenvalues of (1.1) are real valued functionals on $\mathfrak{C}$ and we denote them by $\lambda_n(\rho)$.

In this paper we will show how to use many of the tools of classical calculus of variations to study extremals of the functionals $\lambda_n(\rho)$ for $\rho\in\mathfrak{C}$. For example, we will develop an Euler equation for $\lambda_n(\rho)$ and show how to use natural boundary conditions, Weierstrass–Erdmann corner conditions, Lagrange multipliers and so on to obtain information about the extremals of $\lambda_n(\rho)$.

* Received by the editors March 1, 1984, and in final revised form January 14, 1985.

[†] Mathematics Department, Washington State University, Pullman, Washington 99164-2930.

In an earlier work D. C. Barnes [8] considered a special case of this problem where the coefficients were allowed to depend only on $x, \rho(x)$ but not on $\rho'(x)$. See also E. R. Barnes [9], Banks [4], Keller and Niordson [14] and Nehari [17].

2. Some examples of eigenvalue problems. Consider a horizontal vibrating string, having density $a(x)$, total mass M, and constant tension T. The characteristic frequencies of vibration are determined by the eigenvalues of [10] the equation;

$$Ty''+\lambda a(x)y=0, \qquad 0\le x\le l.$$

If the string is held in a vertical position, then the tension is no longer constant and the frequencies are determined by

$$((T+\rho(x))y')'+\lambda\rho'(x)y=0, \tag{2.1}$$

where

$$\rho(x)=\int_0^x a(t)\,dt.$$

The function $\rho(x)$ will satisfy boundary conditions $\rho(0)=0$, $\rho(l)=M$. Many different boundary conditions can be imposed on y at $x=0$ and l. Equation (2.1) is of the form (1.1) with $f=T+\rho$, $g=\rho'$, $q=0$.

Consider a slender untwisted vertical column, clamped at $x=0$, and free at $x=l$, subject to an axial compressive load. The critical buckling load of the column is determined by the first eigenvalue of the system [13], [14], [18], [20],

$$[a(x)^m y']'+\lambda\left[K+\int_x^l a(t)\,dt\right]y=0, \quad y(0)=a^m(l)y'(l)=0. \tag{2.2}$$

The interesting values of m are 1, 2, and 3, where (see [18, p. 136]) $m=3$ corresponds to a column made of thinwall tubing. The case $m=2$, which was considered by Keller and Niordson [14], assumes the column is solid but that all cross-sections have a similar shape. The case $m=1$ corresponds to a "profile" column which is formed from a flat slab of material having constant thickness, but variable width $a(x)$. Columns of the last type have been considered by D. C. Barnes [8]. Introducing $\rho(x)$ defined by

$$\rho(x)=\int_x^l a(t)\,dt \tag{2.3}$$

into (2.2) yields

$$[(-\rho'(x))^m y']'+\lambda[K+\rho(x)]y=0, \qquad y(0)=[\rho'(l)]^m y'(l)=0. \tag{2.4}$$

Assuming a given mass M yields $\rho(0)=M$, $\rho(1)=0$.

Now (2.4) is of the form (1.1) with $g[\rho]=[-\rho'(x)]^m$, $g[\rho]=K+\rho$, $q[\rho]=0$. The extremals of $\lambda_1(\rho)$ determine the shape of the tallest column having a fixed total mass M.

Finally, we introduce a constrained extremal problem. Consider the simple equation

$$y''+\lambda\rho(x)y=0, \qquad y(0)=y(l)=0, \qquad 0\le x\le l, \tag{2.5}$$

where the coefficient ρ is subjected to a constraint of the form

$$\int_0^l \Phi(x,\rho(x),\rho'(x))\,dx=M.$$

A large number of works have maximized or minimized eigenvalues of this problem using various constraints on $\rho(x)$. See for example [2]–[9], [13], [14], [16], [17], [19], [23]. In §4 below we will show how to find extremals of these eigenvalues using a Lagrange multiplier.

3. Tangent functionals. Let ρ be any fixed element of $\mathfrak{C}$ and let y^* be an eigenfunction corresponding to $\lambda_n(\rho^*)$. Define functionals $J(\rho)$ and $K(\rho)$ on $\mathfrak{C}$ by

$$\begin{aligned} J(P)&=\int_0^l (\lambda_n(\rho^*)g[\rho]+q[\rho])Y^{*2}-f[\rho](Y^{*\prime})^2dx,\\ K(\rho)&=\lambda_n(\rho^*)-f[\rho^*]Y^{*\prime}Y^*\big|_0^l-J(\rho). \end{aligned}\tag{3.1}$$

We will show that under certain conditions, the functional $K(\rho)$ is "tangent" to $\lambda_n(\rho)$ at $\rho=\rho^*$. To be more precise about this, let ε be a small parameter and let $\rho_\varepsilon(x)\in\mathfrak{C}$ be any one parameter family of functions in $\mathfrak{C}$ of the general form

$$\rho_\varepsilon(x)=\rho^*(x)+\varepsilon h(x)+O(\varepsilon^2)=\rho^*(x)+\delta\rho^*(x)+O(\varepsilon^2).\tag{3.2}$$

We will show in §5 below that

$$-\Delta\lambda_n=f[\rho_\varepsilon]y'y^*-f[\rho^*]y^{*\prime}y\big|_0^l+\Delta J+O(\varepsilon^2)\tag{3.3}$$

where

$$\Delta\lambda_n=\lambda_n(\rho_\varepsilon)-\lambda_n(\rho^*),\qquad \Delta J=J(\rho_\varepsilon)-J(\rho^*).$$

It follows from (3.3) that

$$-\Delta\lambda_n=f[\rho^*](y'y^*-y^{*\prime}y)\big|_0^l+\ \Delta fy^{*\prime}y^*\big|_0^l+\Delta J+O(\varepsilon^2)\tag{3.4}$$

where $\Delta f=f[\rho_\varepsilon]-f[\rho^*]$. The two relationships (3.3) and (3.4) are the basis of our entire analysis.

If the boundary terms were not contained in (3.3), then dividing by ε and letting $\varepsilon\to 0$ shows that $-\delta\lambda_n=\delta J$, so a function $\rho^*\in\mathfrak{C}$ is an extremal of $\lambda_n(\rho)$ if and only if it is also an extremal of $J(\rho)$. Now, however, $J(\rho)$ is a standard example of a functional and the entire theory of calculus of variations applies directly to $J(\rho)$, and thus indirectly to $\lambda_n(\rho)$. There are various ways to eliminate the boundary terms in (3.3) and to implement this idea.

THEOREM 1. *Let ρ^* be an extremal of $\lambda_n(\rho)$ for $\rho\in\mathfrak{C}$. Then, except at its corner points, ρ^* satisfies the first and second Euler equations*

$$\begin{aligned} &F_\rho[\rho^*(x)]-d/dxF_{\rho'}[\rho^*(x)]=0,\\ &F[\rho^*(x)]-\rho^*(x)F_{\rho'}[\rho^*(x)]=\int_0^x F_x[\rho^*(t)]\,dt+C, \end{aligned}\tag{3.5}$$

where $F[\rho]$ is defined by

$$F[\rho]=(\lambda_n(\rho^*)g[\rho]+q[\rho])y^{*2}-f[\rho](y^{*\prime})^2.\tag{3.6}$$

At each corner point, both functions

$$F_{\rho'}\left[\rho^*(x)\right] \text{ and } F_\rho\left[\rho^*(x)\right]-\rho^*(x)F_{\rho'}\left[\rho^*(x)\right] \tag{3.7}$$

are continuous.

To prove Theorem 1 we take a variation in ρ^* which does not move the end points. Thus we assume that if either $x\to 0$ or $x\to l$, then

$$\rho_\varepsilon\to\rho^* \text{ and } \rho_\varepsilon^{*\prime}\to\rho^{*\prime}. \tag{3.8}$$

Thus the coefficients α_i and β_i in (1.2) are the same for either function ρ_ε or ρ^*. Therefore, the self-adjoint condition (1.3) may be used and it implies that the first boundary term in (3.4) drops out. Furthermore, (3.8) implies that $\Delta f = f[\rho_\varepsilon(x)] - f[\rho^*(x)]\to 0$ if $x\to 0$ or l, so (3.4) yields $-\Delta\lambda_n=\Delta J+O(\varepsilon^2)$. Therefore, $\delta J=0$ at $\rho=\rho^*$ so ρ^* is an extremal of $J(\rho)$ and standard calculus of variations [21, Chap. 6] finishes the proof.

As an example of Theorem 1, consider Keller and Niordson's [14] tallest column problem (2.4) with $m=2$. We see $g[\rho]=\rho$, $f[\rho]=\rho'^2$, $q[\rho]=0$, so

$$F[\rho]=\lambda_n(\rho^*)\rho(x)y^{*2}-(\rho'(x))^2(y^{*\prime})^2.$$

Then (3.5) implies that, between corner points,

$$\lambda_1(\rho^*)y^{*2}+2d/dx\left[\rho^{*\prime}(x)(y^{*\prime})^2\right]=0. \tag{3.9}$$

The optimality condition used by Keller and Niordson [14, p. 437, Eq. 8] is equivalent to (3.9) (using a change of notation $y^*\leftrightarrow\varphi$, $a\leftrightarrow-\rho^{*\prime}$ and taking a derivative). The corner condition (3.7) implies that $2\rho^{*\prime}(x)(y^{*\prime})^2$ is continuous. Since $y^{*\prime}\neq 0$ for $0\le x<l$, we see that $\rho^*(x)$ has a continuous derivative and furthermore, [21, p. 217], [10, p. 200] it even follows that ρ^* has a continuous second derivative and satisfies (3.9) for all x.

Although Theorem 1 shows that an extremal of $\lambda_n(\rho)$ is also an extremal of $J(\rho)$ that relationship assumes no variation in $\rho^*(x)$ at the end points, that is (3.8). We would now like to investigate the behavior of ρ^* at the end points so we need to relax the assumption (3.8). This can be done if we specialize the boundary conditions a little bit.

THEOREM 2. *Suppose the coefficients α_i, β_i in the boundary conditions* (1.2) *are constants independent of $\rho(x)$. Let $\rho^*\in\mathfrak{C}$ and let y^* be the eigenfunction corresponding to $\lambda_n(\rho^*)$. Then ρ^* is an extremal of $\lambda_n(\rho)$ for $\rho\in\mathfrak{C}$ if and only if ρ^* is an extremal of $J(\rho)$ and in addition ρ^* satisfies the four "compatibility" conditions*:

$$\begin{aligned} f_\rho\left[\rho^*(x)\right]y^*y^{*\prime}&=0 \quad \textit{at } x=0,l,\\ f_{\rho'}\left[\rho^*(x)\right]y^*y^{*\prime}&=0 \quad \textit{at } x=0,l. \end{aligned} \tag{3.10}$$

To prove Theorem 2, we note that since α_i, β_i are constants, the self-adjoint condition (1.3) can be used with (3.4) and any family of functions $\rho_\varepsilon(x)$ to obtain

$$-\Delta\lambda_n=\Delta f y^*y^{*\prime}\big|_0^l+\Delta J+O(\varepsilon^2). \tag{3.11}$$

Suppose that ρ^* is an extremal of $\lambda_n(\rho)$. Dividing (3.11) by ε and letting $\varepsilon \to 0$ and using (3.5) shows that at $\varepsilon = 0$, $d/d\varepsilon f[\rho_\varepsilon] y^* y^{*\prime}|_0^l = 0$. This implies that at $\varepsilon = 0$,

$$f_\rho[\rho^*]\, y^* y^{*\prime} d/d\varepsilon\, \rho_\varepsilon(x) + f_{\rho'}[\rho^*]\, y^* y^{*\prime} d/d\varepsilon\, \rho_\varepsilon'(x)\Big|_0^l = 0,$$

which yields the four conditions (3.10) since $\rho_\varepsilon(x)$ was arbitrary. This proves the "only if" part of Theorem 2. The "if" part follows from (3.11).

The four compatibility conditions (3.10) are trivially satisfied if $f[\rho]$ is independent of ρ and ρ'. If, however, $f[\rho]$ involves ρ or ρ' then they restrict the kinds of boundary conditions that one may expect to use in extremal problems for $\lambda_n(\rho)$.

Suppose, for example, that $f[\rho] = \rho^2 + \rho'^2$. Then (3.10) gives 4 distinct end point conditions. In general, the differential equation (1.1) has two boundary conditions associated with it and two more boundary conditions (natural or otherwise) are associated with ρ for a total of 8 conditions. The Euler equation (3.5), together with (1.1), forms a coupled system of two second order equations which will have 4 constants in the general solution and one more constant is available in the eigenvalue parameter λ. This total of 5 constants cannot in general be used to fit 8 distinct end point conditions.

These difficulties can be overcome if boundary conditions like y or $y' = 0$ at $x = 0, l$ are used. It is, however, not unusual for extremal problems for $\lambda_n(\rho)$ to have no extremal function ρ^*.

As another example of nonexistence of ρ^*, consider the tallest profile column problem, (2.4) with $m = 1$,

$$(-\rho' y')' + \lambda \rho y = 0, \quad y(0) = 0, \quad y'(1)\rho'(1) = 0. \tag{3.12}$$

In this case,

$$J(\rho) = \int_0^1 \lambda_1(\rho^*)\rho(x) y^{*2} + \rho'(x)(y^{*\prime})^2 dx.$$

The Euler equation is $\lambda_1(\rho^*) y^{*2} - 2 y^{*\prime} y^{*\prime\prime} = 0$. Normalizing y^* so that $y^{*\prime}(0) = 1$, we integrate to find $y^{*\prime} = [\lambda_1(\rho^*) y^{*^3} + 1]^{1/3}$. Thus $y^{*\prime} \geq 1 > 0$. Now (3.12) implies that $\rho^{*\prime}(1) = 0$ and (2.3) implies $\rho^*(1) = 0$. However, $\rho^*(x)$ is the solution of a second order linear homogeneous equation (3.12), so $\rho^*(x) \equiv 0$, a contradiction. Thus such a function $\rho^*(x)$ cannot exist. This curious state of affairs is still under investigation, but it might be the case that the tallest column is obtained by concentrating most of the mass at the bottom of the column and erecting a tall thin spike on top of the mass, then letting the width of the spike go to zero while increasing its length.

Theorem 2 deals with the case in which α_i, β_i are constants. In many applications $\rho(x)$ may in fact occur in the boundary condition. Separated boundary conditions of the form

$$\begin{aligned} &\alpha_1 y(0) + \alpha_2 f[\rho(0)]\, y'(0) = 0, \\ &\beta_3 y(l) + \beta_4 f[\rho(l)]\, y'(l) = 0, \end{aligned} \tag{3.13}$$

where α_i, β_i are constants, are sometimes used. See (2.2). In other applications, periodic boundary conditions of the form

$$\begin{aligned} &y(0) - y(l) = 0, \\ &f[\rho(0)]\, y'(0) - f[\rho(l)]\, y'(l) = 0 \end{aligned} \tag{3.14}$$

might be used. For these kinds of boundary conditions we are able to give an analogue of Theorem 2.

THEOREM 3. *Suppose the boundary conditions are either of the form* (3.13) *or* (3.14). *Then a function* $\rho^* \in \mathfrak{C}$ *is an extremal of* $\lambda_n(\rho)$ *for* $\rho \in \mathfrak{C}$ *if and only if it is also an extremal of the functional* $J(\rho)$ *defined by* (3.1).

The proof of Theorem 3 is based on (3.3). Let $\rho_\varepsilon(x) \in \mathfrak{C}$. Using either (3.13) or (3.14), it follows that the boundary term in (3.3) is zero. Therefore, $-\Delta\lambda_n = \Delta J + O(\varepsilon^2)$. Dividing by ε and letting $\varepsilon \to 0$ proves the theorem

A relationship of the form (2.3) automatically provides a boundary condition on $\rho(x)$. However, not all extremal problems must incorporate such a relationship and for some problems, no boundary condition may be given on $\rho(x)$. In such a case we look for a natural boundary condition. Theorems 2 and 3 show that ρ^* must be an extremal of $J(\rho)$, so classical theory implies:

THEOREM 4. *Suppose the boundary conditions on y are either of the form* (3.13) *or* (3.14) *or else of the form* (1.2) *where* α_i *and* β_i *are constants independent of* $\rho(x)$. *If boundary conditions are not prescribed on* $\rho^*(x)$ *at one of the end points, then* $\rho^*(x)$ *will satisfy the natural boundary condition*

$$F_{\rho'}[\rho^*(x)] = 0 \tag{3.15}$$

at that end point. Here $F[\rho]$ *is given by*

$$F[\rho] = (\lambda_n(\rho^*)g[\rho] + q[\rho])y^{*2} - f[\rho](y^{*\prime})^2.$$

4. Constrained extremal problems. We now consider the problem of finding extremals of $\lambda_n(\rho)$ when $\rho \in \mathfrak{C}$ and in addition ρ is constrained by a condition of the form

$$H(\rho) = \int_0^l \phi(x, \rho(x), \rho'(x))\,dx = M. \tag{4.1}$$

Using ideas similar to those used for Theorem 1 and introducing a Lagrange multiplier, it follows that

THEOREM 5. *Let* ρ^* *be an extremal of* $\lambda_n(\rho)$ *for* $\rho \in \mathfrak{C}$ *and suppose that* ρ^* *is not an extremal of* $H(\rho)$. *Then there exists a constant* μ *such that, between corner points,* ρ^* *satisfies*

$$\frac{\partial}{\partial\rho}G[\rho] - \frac{d}{dx}\frac{\partial}{\partial\rho'}G[\rho] = 0 \tag{4.2}$$

where

$$G[\rho] = (\lambda_n[\rho^*]g[\rho] + q[\rho])y^{*2} - f[\rho](y^{*\prime})^2 + \mu\phi[\rho].$$

At a corner point $G_{\rho'}[\rho^*]$ *and* $G[\rho^*] - \rho^* G_{\rho'}[\rho^*]$ *are continuous functions of x.*

The other theorems can also be modified to take account of (4.1). Thus it follows that the compatibility conditions (3.10) hold if α_i, β_i are constants and the natural boundary condition is $G_{\rho'}[\rho^*(x)] = 0$ at the end points even if ρ^* is required to satisfy (4.1).

As an example of Theorem 5 we will calculate the extremals of $\lambda_n(\rho)$ for the system

$$y'' + \lambda\rho(x)y = 0, \qquad y(0) = y(1) = 0 \tag{4.3}$$

subject to the constraints

$$H(\rho)=\int_0^1 (\rho'(x))^2 dx=1 \quad \text{and} \quad \rho(0)=0. \tag{4.4}$$

Although there can be little doubt that the extremal ρ^* of $\lambda_n(\rho)$ is, in this case, a global minimum, we will not prove this.

No boundary condition on ρ at $x=1$ is prescribed so we will use a natural one. In this case $G(\rho)=\lambda_n(\rho^*)y^{*2}\rho(x)+\mu(\rho'(x))^2$. The natural boundary condition is $\rho^{*\prime}(1)=0$ and the Euler equation (4.2) implies $\lambda^* y^{*2}-2\mu\rho^{*\prime\prime}(x)=0$. Using $C=-1/2\lambda^*/\mu$ and (4.3), we obtain the coupled system:

$$\begin{aligned} &\rho^{*\prime\prime}+Cy^{*2}=0, && \rho^*(0)=\rho^{*\prime}(1)=0, \\ &y^{*\prime\prime}+\lambda^*\rho^*y^*=0, && y^*(0)=y^*(1)=0, \quad y^{*\prime}(0)=.75. \end{aligned} \tag{4.5}$$

Here we have chosen to normalize y^* so that $y^{*\prime}(0)=.75$. This gives a "pretty" graph of y^* and ρ^* and also 3 boundary conditions on ρ^* and y^* at $x=0$. Now in order to start the Runge–Kutta numerical solution, we need a boundary condition for $\rho^{*\prime}(0)$. We will arbitrarily take $\rho^{*\prime}(0)=1$ and, after the solution is complete, we will renormalize $\rho^*(x)$ so that (4.4) is satisfied. This leaves the two constants C, λ to be determined in such a way as to satisfy the two equations $\rho^{*\prime}(1)=y^*(1)=0$. A 4-point Runge–Kutta method was used to solve (4.5), and a two-dimensional version of Newton's method

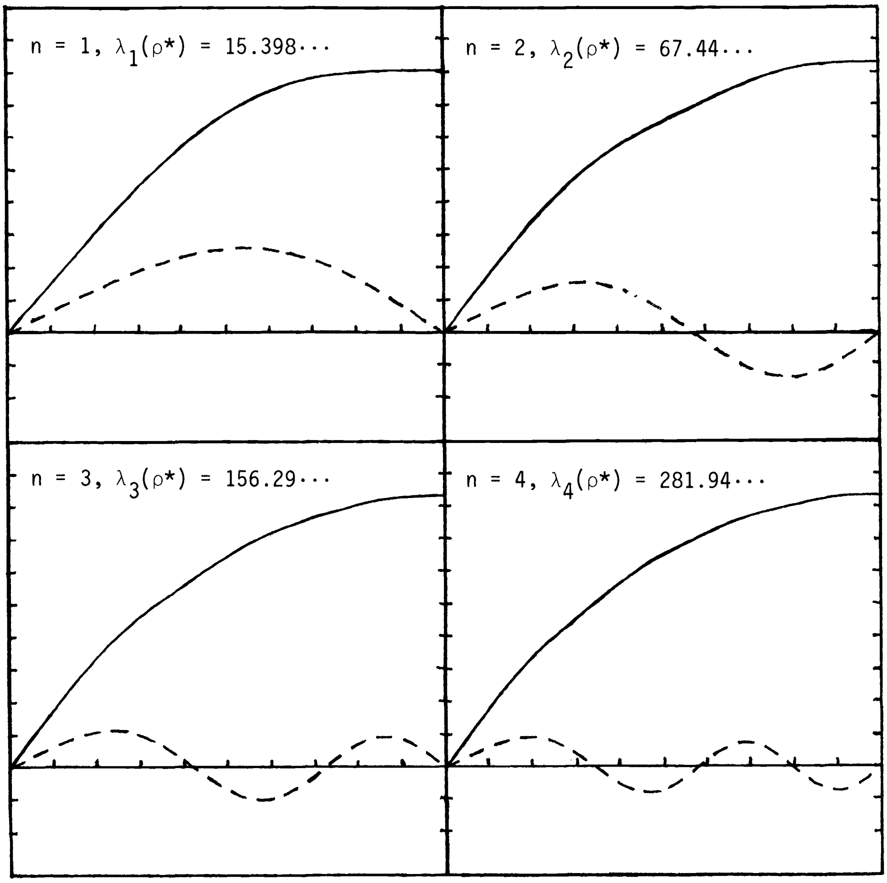


FIG. 1. *Plots of the extremals y^* and ρ^* for $n=1,2,3,4$.*

was used to solve for C, λ (see [1, p. 92]). The Jacobians involved were approximated with difference quotients. The first guess for C, λ determined which of the eigenvalues $\lambda_n(\rho^*)$ the method would converge to. The equations were solved for step sizes ranging from 2^{-4} to 2^{-9} and strong convergence was observed. Computer-generated plots of the extremals ρ^*, y^* for $n=1, 2, 3, 4$ are reproduced in Fig. 1.

These results may also be expressed as isoperimetric inequalities which hold for any $\rho\in\mathfrak{C}$ in the following form:

$$\rho(0)=0 \text{ and } \rho\in\mathfrak{C}\Rightarrow\lambda_n(\rho)\left[\int_0^l(\rho'(x))^2dx\right]^{1/2}\geq\Lambda_n,$$

$$\Lambda_1=15.398\cdots,\quad \Lambda_2=67.44\cdots,\quad \Lambda_3=156.29\cdots,\quad \Lambda_4=281.94\cdots.$$

5. Equation (3.3) and tangent functionals. Let $\rho_\varepsilon(x)$, $\rho^*(x)\in\mathfrak{C}$ and let y and y^* be eigenfunctions corresponding to $\lambda_n(\rho)$, $\lambda_n(\rho^*)$. Thus

$$(5.1)\qquad (f[\rho^*]y^{*\prime})'+(\lambda_n(\rho^*)g[\rho^*]+q[\rho^*])y^*=0,$$

$$(5.2)\qquad (f[\rho_\varepsilon]y')'+(\lambda_n(\rho_\varepsilon)g[\rho_\varepsilon]+q[\rho_\varepsilon])y=0.$$

Suppose y and y^* are normalized so

$$(5.3)\qquad \int_0^l g[\rho^*]y^{*2}dx=\int_0^l g[\rho_\varepsilon]y^2dx=1.$$

Multiply (5.1) by y and (5.2) by y^*. Subtract the two equations and integrate. A little manipulation yields

$$f[\rho_\varepsilon]y'y^*-f[\rho^*]y^*y'\big|_0^l+\int_0^l(\Delta\lambda_ng[\rho^*]+\lambda_n(\rho^*)\Delta g+\Delta q)yy^*-\Delta fy'y^{*\prime}dx=0,$$

where we have used the notation

$$\Delta\lambda_n=\lambda_n(\rho_\varepsilon)-\lambda_n(\rho^*),\quad \Delta g=g(\rho_\varepsilon)-g(\rho^*),\quad \Delta q=q(\rho_\varepsilon)-q(\rho^*),$$
$$\Delta f=f(\rho_\varepsilon)-f(\rho^*),\qquad \Delta y=y-y^*.$$

Solving for $\Delta\lambda_n$ yields

$$(5.4)\qquad -\Delta\lambda_n\int_0^l g[\rho^*]yy^*dx$$
$$=f[\rho_\varepsilon]y'y^*-f[\rho^*]y^*y\big|_0^l+\int_0^l(\lambda_n(\rho^*)\Delta g+\Delta q)yy^*-\Delta y'y^{*\prime}dx.$$

Replacing y with $y^*+\Delta y$ and collecting some terms which are $O(\varepsilon^2)$ and using (5.3) yields

$$-\Delta\lambda_n=f[\rho_\varepsilon]y'y^*-f[\rho^*]y^*y\big|_0^l+\int_0^l(\lambda_n(\rho^*)\Delta g+\Delta q)y^{*2}-\Delta f(y^{*\prime})^2dx+O(\varepsilon^2).$$

Interchanging the order of the operations Δ and $\int$ and using (3.1) gives (3.3).

Similar manipulations can be used to show that $J(\rho^*)+f[\rho^*]Y^*Y^*|_0^l=0$, which implies that $K(\rho^*)=\lambda_n(\rho^*)$. Furthermore, when subject to appropriate boundary conditions, it follows from (3.1) and (3.3) that $\Delta\lambda_n=\Delta K+O(\varepsilon^2)$. Thus the two functionals $K(\rho)$ and $\lambda_n(\rho)$ are indeed tangent to each other at $\rho=\rho^*$.

6. Some extensions. Many generalizations of these results are possible. For example, one might allow the coefficient functions f, g, q or the constraint function Φ to depend on $x, \rho(x)$, $\rho'(x)$ and $\rho''(x)$. Theorems 1–5 can be generalized in obvious ways.

Theorem 4 can be generalized in many different ways by appealing to more general isoperimetric theorems for classical functionals. See, for example, Hestenes [12, Chap. 7]. Thus one might consider multiple constraints on $\rho(x)$ of the forms

$$\begin{aligned}
&\int_0^l \Phi_i[\rho(x)]\,dx = M_i, && i=1,2,\cdots,p,\\
&\int_0^l \Phi_i[\rho(x)]\,dx \le M_i, && i=p+1,p+2,\cdots,q,\\
&\psi_j[\rho(x)]=0, && j=1,2,\cdots,r,\\
&\psi_j[\rho(x)]\le 0, && j=r+1,r+2,\cdots,s.
\end{aligned}$$

It appears that such an approach might be used to solve problems similar to those considered by Bandle [2], Banks [3], [4], [5], D. C. Barnes [6], [7] and others.

One could also consider 4th (and higher) order eigenvalue problems

$$-[r[\rho]y'']''+[f[\rho]y']'+[\lambda g[\rho]+q[\rho]]y=0, \tag{6.1}$$

where, for example, $r[\rho]=r(x,\rho(x),\rho'(x),\rho''(x))$. The function $F[\rho]$ becomes

$$F[\rho]=(\lambda_n(\rho^*)g[\rho]+q[\rho])y^{*2}-f[\rho](y^{*\prime})^2-r[\rho](y^{*\prime\prime})^2,$$

and, with $H(\rho)$ defined by (4.1),

$$\begin{aligned}
J[\rho]&=\int_0^l F[\rho(x)]\,dx,\\
G[\rho]&=F[\rho]+\mu\Phi[\rho].
\end{aligned}$$

Generalizing Theorem 1, we see that an extremal ρ^* of $J[\rho]$ will, between corner points, satisfy:

$$\frac{\partial F}{\partial\rho}-\frac{d}{dx}\left(\frac{\partial F}{\partial\rho'}\right)+\frac{d^2}{dx^2}\left(\frac{\partial F}{\partial\rho''}\right)=0.$$

The other theorems can also be modified to deal with (6.1).

Eigenvalue problems for partial differential equations can also be dealt with using these methods. Consider the equation

$$(fu_x)_x+(fu_y)_y+(\lambda g+q)u=0, \qquad (x,y)\in\mathfrak{D}$$

where $\mathfrak{D}$ is a two-dimensional domain and self-adjoint boundary conditions are given on $\partial\mathfrak{D}$

$$u+\sigma\frac{\partial u}{\partial \bar{n}}=0 \quad \text{on } \partial\mathfrak{D}.$$

In this case the coefficients f, g, q, σ depend on $x, y, \rho(x,y)$, $\rho_x(x,y)$ and $\rho_y(x,y)$. The appropriate analogues of the preceding formulas are:

$$\begin{aligned}
&F[\rho]=(\lambda_n(\rho^*)g[\rho]+q[\rho])u^{*2}-f[\rho](u_x^{*2}+u_y^{*2}),\\
&J(\rho)=\iint_{\mathfrak{D}} F[\rho]\,dx\,dy,\\
&G[\rho]=F[\rho]+\mu\Phi[\rho],\\
&H(\rho)=\iint_{\mathfrak{D}} \Phi[\rho]\,dx\,dy.
\end{aligned}$$

The compatibility conditions corresponding to (3.10) are in this case

$$f_\rho[\rho^*]=0, \quad f_{\rho_x}[\rho^*]=0, \quad f_{\rho_y}[\rho^*]=0 \quad \text{on } \partial\mathfrak{D}.$$

Another extension would be to consider functions of eigenvalues along the lines of the works by Keller [15], Gentry [11], Willner and Mahor [22]. This problem is under investigation and will, perhaps, be published later.

REFERENCES

[1] K. E. ATKINSON, *An Introduction to Numerical Analysis*, John Wiley, New York, 1978.

[2] C. BANDLE, *Bounds for the frequencies of an inhomogeneous string*, SIAM J. Appl. Math., 25 (1973), pp. 634–639.

[3] D. O. BANKS, *Bounds for the eigenvalues of some vibrating systems*, Pacific J. Math., 10 (1960), pp. 439–474.

[4] ______, *Upper bounds for the eigenvalues of some vibrating systems*, Pacific J. Math., 11 (1961), pp. 1183–1203.

[5] ______, *Lower bounds for the eigenvalues of a vibrating string whose density satisfies a Lipschitz condition*, Pacific J. Math., (1) 20 (1967), pp. 393–410.

[6] D. C. BARNES, *Some isoperimetric inequalities for the eigenvalues of vibrating strings*, Pacific J. Math., 29 (1969), pp. 43–61.

[7] ______, *Lower bounds for eigenvalues of Sturm–Liouville systems*, Indiana J. Math., 30 (1981), pp. 193–198.

[8] ______, *Extremal problems for eigenvalues with applications to buckling, vibration, and sloshing*, this Journal, 16 (1985), pp. 341–357.

[9] E. R. BARNES, *The shape of the strongest column and some related extremal eigenvalue problems*, Quart. Appl. Math., 34 (1977), pp. 393–409.

[10] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. 1, Interscience, New York, 1953.

[11] R. D. GENTRY AND D. O. BANKS, *Bounds for functions of eigenvalues*, J. Math. Anal. Appl., 51 (1975), pp. 100–128.

[12] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[13] J. B. KELLER, *The shape of the strongest column*, Arch. Rat. Mech. Anal., 5 (1960), pp. 275–285.

[14] J. B. KELLER AND F. I. NIORDSON, *The tallest column*, J. Math. Mech., 16 (1966), pp. 433–446.

[15] J. B. KELLER, *The minimum ratio of two eigenvalues*, SIAM J. Appl. Math., 31 (1976), pp. 485–491.

[16] M. G. KREIN, *On certain problems on the maximum and minimum of characteristic values and on Lyapunov zones of stability*, Trans. Amer. Math. Soc., 2 (1955), pp. 163–187.

[17] Z. NEHARI, *Extremal problems for a class of functionals defined on convex sets*, Bull. Amer. Math. Soc., 73 (1967), pp. 584–591.

[18] G. J. SIMITSES, *An Introduction to the Elastic Stability of Structures*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[19] I. TADJBAKHSH AND J. B. KELLER, *Strongest columns and isoperimetric inequalities for eigenvalues*, J. Appl. Mech., 29 (1962), pp. 159–164.

[20] S. P. TIMOSHENKO AND J. N. GERE, *Theory of Elastic Stability*, McGraw-Hill, New York, 1961.

[21] J. L. TROUTMAN, *Variational Calculus with Elementary Convexity*, Springer-Verlag, New York, 1983.

[22] B. E. WILLNER AND T. J. MAHOR, *The two-dimensional eigenvalue range and extremal eigenvalue problems*, this Journal, 13 (1982), pp. 621–631.

[23] I. M. RAPAPORT, *On a variational problem in the theory of ordinary differential equations with boundary conditions*, Doklady Akad. Nauk. SSSR (N.S.), 73 (1950), pp. 889–890.

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

011

A LYAPUNOV FUNCTIONAL FOR A RETARDED DIFFERENTIAL EQUATION*

J. C. F. DE OLIVEIRA[†] AND L. A. V. CARVALHO[‡]

Dedicated to the memory of Professor Dr. Walter de Bona Castelan

Abstract. Consider the autonomous linear retarded differential equation

$$\dot{x}(t)=A_0x(t)+A_1x(t-r)+\int_{-r}^{0}A_{01}(\theta)x(t+\theta)\,d\theta.$$

A positive definite quadratic functional is given that yields an equivalent norm in the phase space, which measures the exact asymptotic behavior of its solutions, providing the best estimate for the rate of growth or decay of the said solutions.

1. Introduction.

For ordinary differential equations

$$\dot{x}(t)=A_0x(t), \tag{1.1}$$

where A_0 is an $n\times n$ constant real matrix, the following result, due to A. M. Lyapunov, is true.

THEOREM. *Let $\gamma=\max\{\operatorname{Re}\lambda:\det(\lambda I-A_0)=0\}$. For any $\varepsilon>0$, let $\delta=-(\gamma+2\varepsilon)$ and let W be any positive definite matrix. Then, there exists a unique positive definite matrix M such that*

$$(A_0+\delta I)^TM+M(A_0+\delta I)=-W, \tag{1.2}$$

where $(A_0+\delta I)^T$ denotes the transpose of $A_0+\delta I$.

In fact, M is given by

$$M=\int_0^\infty e^{2\delta t}e^{A_0^Tt}We^{A_0t}\,dt.$$

Moreover, if $V:R^n\to R$ is defined by

$$V(x)=x^TMx,$$

then:

(i) $(V(x))^{1/2}$ is an Euclidean norm on R^n, equivalent to the canonical norm;

(ii) the derivative of V with respect to (1.1) satisfies the inequality

$$\dot{V}(x)\leqq-2\delta V(x). \tag{1.3}$$

Therefore, from Gronwall's inequality, it follows that

$$\{V(x(t))\}^{1/2}\leqq e^{-\delta t}\{V(x(0))\}^{1/2},\qquad t>0, \tag{1,4}$$

for any solution $x(t)$ of (1.1); this last inequality gives the best estimate for the rate of growth or decay of these solutions as t goes to infinity.

*Received by the editors December 9, 1981, and in revised form May 10, 1984.

[†] I.M.E.—Universidade de São Paulo, São Paulo (SP), Brasil 05508. The work of this author was partially supported by CNPq (Brazil).

[‡] Present address, Brown University, Providence, Rhode Island 02912. I.M.F.—Universidade Federal de Goiâs, Goiânia (Go), Brasil 74000. The work of this author was partially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brasil.

The above theorem was extended to retarded differential equations of the form

$$\dot{x}(t)=A_0x(t)+A_1x(t-r) \tag{1.5}$$

by Infante and Castelan [2].

Our purpose here is to complete this extension to equations of the form

$$\dot{x}(t)=A_0x(t)+A_1x(t-r)+\int_{-r}^{0}A_{01}(\theta)x(t+\theta)\,d\theta, \tag{1.6}$$

where r is a positive real constant,A_0, A_1 and $A_{01}(\theta)$ are $n\times n$ real matrices with A_{01} square integrable in $[-r,0]$.

We construct for (1.6) a positive definite quadratic functional V on an appropriate phase space, satisfying bounds analogous to (1.3) and (1.4); therefore, the functional V gives the best estimate for the asymptotic behavior of the solutions of (1.6).

As in [2], the construction of V relies upon the computation of the solutions of the matrix functional differential equation

$$\begin{aligned}(1.7)\quad \dot{Q}(\alpha)=&(A_0+\delta I)^TQ(\alpha)+e^{\delta r}A_1^TQ^T(r-\alpha)\\&+\int_{-r}^{-\alpha}e^{-\delta\theta}A_{01}^T(\theta)Q^T(-\alpha-\theta)\,d\theta+\int_{-\alpha}^{0}e^{-\delta\theta}A_{01}(\theta)Q(\alpha+\theta)\,d\theta,\\&\qquad\qquad 0\leqq\alpha\leqq r,\end{aligned}$$

satisfying the properties that $Q(0)$ is positive definite and

$$\dot{Q}(0)+\dot{Q}^T(0)=-W, \tag{1.8}$$

where W is a given positive definite matrix.

Equation (1.8) is the analogue of (1.2). Equation (1.7), despite its functional appearence, is finite dimensional. In fact, we will prove in §3 that (1.7) has a unique solution for any prescribed value of $Q(0)$. Moreover, in several cases, (1.7) can be reduced to a linear system of ordinary differential equations with constant coefficients and, therefore, it can be, in theory, completely integrated. This happens, for example, when the kernel $A_{01}(\theta)$ is a finite sum of terms of the form $P(\theta)e^{\lambda\theta}$, where λ is a complex number and $P(\theta)$ is a matrix with polynomial entries.

We apply the results to the scalar equation

$$\dot{x}(t)=-a\int_{-r}^{0}x(t+\theta)\,d\theta,$$

where a is a positive constant, after we make a detailed analysis of its characteristic equation. Another application is presented to the equation

$$\dot{x}(t)=-a\int_{-r}^{0}(1+\theta)x(t+\theta)\,d\theta,$$

where a is a positive constant.

2. Preliminaries. Let $r>0$ be a given real constant, denote by L_2 the space of the Lebesgue square integrable functions defined on $[-r,0]$ with values in R^n, and consider the Hilbert space $\chi=R^n\times L_2$ with inner product

$$\langle(\xi_1,\psi_1),(\xi_2,\psi_2)\rangle=\xi_1^T\xi_2+\int_{-r}^{0}\psi_1^T(\theta)\psi_2(\theta)\,d\theta.$$

If $x:[-r,0]\to R^n$ and $t\geqq 0$, we denote by x_t the function $x_t:[-r,0]\to R^n$ given by $x_t(\theta)=x(t+\theta)$.

For a given $(\xi,\psi)\in\chi$, a solution of (1.6) with initial value

$$x(0)=\xi, \qquad x_0=\psi \quad \text{a.e.,} \tag{2.1}$$

is, by definition a function $x:[-r,\infty)\to R^n$ such that x satisfies (2.1), is absolutely continuous and satisfies (1.6) almost everywhere on any interval $[0,t]$, $t>0$.

It is known [1], [2], that (1.6)–(2.1) has a unique solution, which we denote by $x(\cdot,\xi,\psi)$, and defines a C_0-semigroup of linear operators on χ given by

$$T(t)(\xi,\psi)=(x(t,\xi,\psi),x_t(\cdot,\xi,\psi)). \tag{2.2}$$

The infinitesimal generator of this semigroup is the operator A defined by

$$A(\xi,\psi)=\left(A_0\xi+A_1\psi(-r)+\int_{-r}^{0}A_{01}(\theta)\psi(\theta)\,d\theta,\dot{\psi}\right)$$

whose domain is the dense subspace of χ given by

$$D(A)=\{(\xi,\psi)\in\chi:\psi \text{ is absolutely continuous in } [-r,0], \dot{\psi}\in L_2 \text{ and } \psi(0)=\xi\}.$$

The spectrum $\sigma(A)$ of A consists of those complex numbers λ which satisfy the characteristic equation $\det\Delta(\lambda)=0$, where

$$\Delta(\lambda)=\lambda I-A_0-A_1e^{-\lambda r}-\int_{-r}^{0}A_{01}(\theta)e^{\lambda\theta}\,d\theta.$$

Moreover, there exists a maximum $\gamma\in R$ of the set $\operatorname{Re}\sigma(A)$, and for every $\varepsilon>0$ there exists a constant $K=K(\varepsilon)\geqq 1$ such that

$$\|T(t)\|\leqq Ke^{(\gamma+\varepsilon)t} \tag{2.3}$$

for all $t\geqq 0$, where $\|T(t)\|$ is the uniform norm of the operator $T(t)$.

Relation (2.3) gives the best estimate for the rate of growth or decay of the solutions of (1.6), or, in other words, it gives the order of the semigroup defined by (2.2).

Finally, we observe that the solutions of (1.6)–(2.1) can be written in the following form

$$\begin{aligned} x(t,\xi,\psi)=S(t)\xi&+\int_{-r}^{0}S(t-\alpha-r)A_1\psi(\alpha)\,d\alpha\\ &+\int_{-r}^{0}\left\{\int_{0}^{\alpha+r}S(t-u)A_{01}(\alpha-u)\,du\right\}\psi(\alpha)\,d\alpha \end{aligned} \tag{2.4}$$

for any $t\geqq 0$, where the matrix S is the unique solution of the initial value problem

$$\begin{aligned} &\dot{S}(u)=S(u)A_0+S(u-r)A_1+\int_{-r}^{0}S(u+\theta)A_{01}(\theta)\,d\theta, \qquad u>0,\\ &S(0)=I\\ &S(u)=0 \quad \text{for } u<0. \end{aligned} \tag{2.5}$$

3. A quadratic functional. Our objective in this section is to find a quadratic positive definite functional for (1.6) on χ in the spirit of the direct method of Lyapunov.

By analogy with the case of the ordinary differential equations, we define, for each $\varepsilon > 0$, $(\xi,\psi)\in\chi$ and a given positive definite $n\times n$ matrix W,

$$U(\xi,\psi)=\int_0^\infty e^{2\delta u}x^T(u,\xi,\psi)Wx(u,\xi,\psi)\,du, \tag{3.1}$$

where $\delta=-(\gamma+2\varepsilon)$ and $\gamma=\max\{\mathrm{Re}\lambda : \lambda\in\sigma(A)\}$.

In view of (2.3), it is easy to see that the above integral converges. Also, it defines a continuous nonnegative quadratic form on χ.

The time derivative of (3.1) with respect to equation (1.6) is given by

$$\dot U(\xi,\psi)=-2\delta U(\xi,\psi)-\xi^TW\xi, \qquad (\xi,\psi)\in D(A), \tag{3.2}$$

and, therefore, $\dot U(\xi,\psi)\leqq -2\delta U(\xi,\psi)$ for all $(\xi,\psi)\in\chi$ [4].

This inequality implies that for any $(\xi,\psi)\in\chi$ we have

$$U(x(t,\xi,\psi),x_t(\cdot,\xi,\psi))\leqq U(\xi,\psi)e^{-2\delta t}. \tag{3.3}$$

If we write

$$|(\xi,\psi)|=\{U(\xi,\psi)\}^{1/2}, \tag{3.4}$$

we have

$$|T(t)(\xi,\psi)|\leqq e^{-2\delta t}|(\xi,\psi)|, \qquad (\xi,\psi)\in\chi. \tag{3.5}$$

We now define

$$Q(\alpha)=\int_0^\infty e^{\delta u}S^T(u)We^{\delta(u-\alpha)}S(u-\alpha)\,du, \qquad \alpha\in R, \tag{3.6}$$

where S^T denotes the transpose of S, S defined as in (2.4)–(2.5). Again, the integral is well defined and it is easy to check that $Q(\alpha)=Q^T(-\alpha)$ for all $\alpha\in R$ and that Q is continuously differentiable in $\alpha\in(0,\infty)$ with $Q(0)$ positive definite.

We wish now to express the functional in (3.1) in terms of Q in (3.6). In order to do that, we will find a differential equation that is satisfied by $Q(\alpha)$, $\alpha>0$.

By taking the derivative of Q at $\alpha\in(0,r)$, we get, after some rearrangements of integrals and repeated use of (3.6) together with the property $Q^T(\alpha)=Q(-\alpha)$, for $0<\alpha<r$,

$$\begin{aligned}\dot Q(\alpha)&=(A_0+\delta I)^TQ(\alpha)+e^{\delta r}A_1^TQ^T(r-\alpha)\\&\quad+\int_{-r}^{-\alpha}e^{-\delta\theta}A_{01}^T(\theta)Q^T(-\alpha-\theta)\,d\theta+\int_{-\alpha}^0 e^{-\delta\theta}A_{01}(\theta)Q(\alpha+\theta)\,d\theta.\end{aligned} \tag{3.7}$$

THEOREM 3.1. *Equation* (3.7) *has a unique continuously differentiable matrix solution* $Q(\alpha)$ *on* $[0,r]$ *for each prescribed initial value* $Q(0)$ *and for each* $r>0$ *sufficiently small.*

Proof. Finding a solution of (3.7) with $Q(0)$ prescribed is equivalent to finding a fixed point of the transformation $Q\to TQ$ given by

$$\begin{aligned}TQ(\alpha)&=Q(0)+\int_0^\alpha\left[(A_0+\delta I)^TQ(u)+e^{\delta r}A_1^TQ^T(r-u)\right]du\\&\quad+\int_0^\alpha\int_{-r}^{-u}e^{-\delta\theta}A_{01}^T(\theta)Q^T(-\theta-u)\,d\theta\,du+\int_0^\alpha\int_{-u}^0 e^{-\delta\theta}A_{01}(\theta)Q(u+\theta)\,d\theta\,du,\end{aligned}$$

from $C([0,r], L(R^n))$ into itself, where $L(R^n)$ is the space of all $n\times n$ real matrices and $C([0,r], L(R^n))$ is the Banach space of the continuous functions from $[0,r]$ into $L(R^n)$ with the supremum norm. It is easy to see that T^m is a contraction for m sufficiently large.

Therefore, we are led to consider the solutions of (3.7) with initial condition

$$Q(0)=\int_0^\infty e^{2\delta u}S^T(u)WS(u)\,du,$$

which is nothing but the function given by relation (3.6).

Let us now write the functional U in terms of Q. Using the representation formula (2.4), we have for $t\geqq 0$ and any $(\xi,\psi)\in\chi$,

$$U(\xi,\psi)=\int_0^\infty e^{-2\delta t}\left\{\left[S(t)\xi+\cdots+\int_{-r}^0\left(\int_0^{\alpha+r}S(t-u)A_{01}(\alpha-u)\,du\right)\psi(\alpha)\,d\alpha\right]^T W\right.$$
$$\left.\cdot\left[S(t)\xi+\cdots+\int_{-r}^0\left(\int_0^{\beta+r}S(t-u)A_{01}(\beta-u)\,du\right)\psi(\beta)\,d\beta\right]\right\}dt,$$

or

(3.8)
$$\begin{aligned}U(\xi,\psi)=\ &\xi^TQ(0)\xi+2\xi^T\int_{-r}^0 e^{\delta(\alpha+r)}Q(\alpha+r)A_1\psi(\alpha)\,d\alpha\\
&+2\xi^T\int_{-r}^0\int_0^{\alpha+r}e^{\delta u}Q(u)A_{01}(\alpha-u)\psi(\alpha)\,du\,d\alpha\\
&+\int_{-r}^0\int_{-r}^0[\psi(\alpha)]^TA_1^Te^{\delta(\alpha+\beta+2r)}Q(\beta-\alpha)A_1\psi(\beta)\,d\alpha\,d\beta\\
&+2\int_{-r}^0\int_{-r}^0[\psi(\alpha)]^TA_1^T\left\{\int_0^{\beta+r}e^{\delta(u+\alpha+r)}Q(u-\alpha-r)A_{01}(\beta-u)\,du\right\}\psi(\beta)\,d\alpha\,d\beta\\
&+\int_{-r}^0\int_{-r}^0[\psi(\alpha)]^T\left\{\int_0^{\alpha+r}\int_0^{\beta+r}A_{01}^T(\alpha-u)e^{\delta(u+v)}\right.\\
&\qquad\left.\cdot Q(v-u)A_{01}(\beta-v)\,du\,dv\right\}\psi(\beta)\,d\alpha\,d\beta.\end{aligned}$$

If we compute $\dot U(\xi,\psi)$ from (3.8), we get

$$\dot U(\xi,\psi)=-2\delta U(\xi,\psi)+\xi^T[\dot Q(0)+\dot Q^T(0)]\xi,$$

so that (3.2) implies that

$$\dot Q(0)+\dot Q^T(0)=-W. \tag{3.9}$$

Remarks. We observe that whenever $A_{01}(\theta)$ is a finite sum of terms of the form $P(\theta)e^{\lambda\theta}$, where λ is a complex number and $P(\theta)$ is a matrix with polynomial entries, (3.7) can always be reduced to a homogeneous system of ordinary differential equations with constant coefficients and, therefore, can be, in theory, explicitly integrated. We also observe that, in general, U is not positive definite. In fact, if $A_1=0$ and $A_{01}\equiv 0$, $U(0,\psi)=0$ for arbitrary ψ.

Let us now proceed to complete the functional U given by (3.8) in order to obtain V. Let M, R and N be $n \times n$ positive definite matrices and define the functional V as follows:

(3.10)

$$V(\xi,\psi)=U(\xi,\psi)+\xi^T M\xi+\int_{-r}^{0} e^{2\delta\theta}\psi^T(\theta)R\psi(\theta)\,d\theta+\int_{-r}^{0}\int_{\theta}^{0}e^{2\delta s}\psi^T(s)N\psi(s)\,ds\,d\theta.$$

It is easy to see that there exist constants $C_1>0$ and $C_2>0$ such that

$$C_1\|(\xi,\psi)\|_X^2 \leqq V(\xi,\psi)\leqq C_2\|(\xi,\psi)\|_X^2 \tag{3.11}$$

for all $(\xi,\psi)\in\chi$.

Relation (3.11) shows that $(V(\xi,\psi))^{1/2}$ is a Hilbertian norm in χ which is equivalent to $\|\cdot\|_\chi$.

We now compute the derivative of V with respect to (1.6). We have that, for any $(\xi,\psi)\in D(A)$,

$$\dot V(\xi,\psi)=-2\delta V(\xi,\psi)+Z(\xi,\psi),$$

where,

$$Z(\xi,\psi)=\xi^T(-W+2\delta M+R+A_0^TM+MA_0+rN)\xi-e^{-2\delta r}\psi^T(-r)R\psi(-r)$$
$$+2\xi^TMA_1\psi(-r)+2\xi^TM\int_{-r}^{0}A_{01}(\theta)\psi(\theta)\,d\theta-\int_{-r}^{0}e^{-2\delta\theta}\psi^T(\theta)N\psi(\theta)\,d\theta.$$

We claim that the positive definite matrices W, M, R and N can be chosen such that

$$Z(\xi,\psi)\leqq 0$$

for all $(\xi,\psi)\in D(A)$.

Suppose W, M, R and N are scalar multiples of the identity matrix. Using the same letters W, M, R and N to denote both the matrices and the corresponding scalars and using the inequality

$$a^T\cdot b\leqq\frac{1}{2}\left(|a|^2+|b|^2\right)$$

for all $a,b\in R^n$, we get

$$\xi^TA_1\psi(-r)\leqq\frac{1}{2}\left(|A_1|^2|\xi|^2+|\psi(-r)|^2\right),$$
$$\int_{-r}^{0}\xi^TA_{01}(\theta)\psi(\theta)\,d\theta\leqq\frac{1}{2}\left(|\xi|^2\int_{-r}^{0}|A_{01}(\theta)|^2d\theta+\int_{-r}^{0}|\psi(\theta)|^2d\theta\right),$$

which imply that

$$Z(\xi,\psi)\leqq-|\xi|^2\left\{W-M\left(2\delta+2|A_0|+|A_1|^2+\int_{-r}^{0}|A_{01}(\theta)|^2d\theta\right)-rN-R\right\}$$
$$-|\psi(-r)|^2(e^{-2\delta r}R-M)-\int_{-r}^{0}(Ne^{-2\delta\theta}-M)|\psi(\theta)|^2d\theta.$$

We can choose $N>0$ arbitrary. Then, we take $M>0$ such that $M<\min\{Ne^{-2\delta\theta}\colon -r\leqq\theta\leqq 0\}$. Once this is done, we take $R>0$ such that $R>Me^{2\delta r}$ and, finally, we take $W>0$ such that

$$W>M\left(2\delta+2|A_0|+|A_1|^2+\int_{-r}^{0}|A_{01}(\theta)|^2 d\theta\right)-rN-R.$$

With these choices, we have $Z(\xi,\psi)\leqq 0$ for all $(\xi,\psi)\in D(A)$ and, as a consequence, it follows that $\dot{V}(\xi,\psi)\leqq -2\delta V(\xi,\psi)$ for all $(\xi,\psi)\in D(A)$. By [4, Thm. 3.9] this last bound for $\dot{V}$ extends to all χ. Gronwall's inequality then implies that

$$\dot{V}(x(t),x_t)\leqq e^{-2\delta t}V(x(0),x_0)$$

for any solution of (1.6) and so, we can state the following extension of the theorem given in [5]:

THEOREM 3.2. *Consider the retarded equation*

$$\dot{x}(t)=A_0x(t)+A_1x(t-r)+\int_{-r}^{0}A_{01}(\theta)x(t+\theta)\,d\theta$$

and the functional given in (3.10). *If*

$$\gamma=\max\left\{\operatorname{Re}\lambda:\det\left[\lambda I-A_0-A_1e^{-\lambda r}-\int_{-r}^{0}A_{01}(\theta)e^{\lambda\theta}d\theta\right]=0\right\}$$

and $\varepsilon>0$, *then there exist constant positive definite matrices* M, R *and* N, *and a differentiable matrix* $Q(\alpha)$, $0\leqq\alpha\leqq r$ *with* $Q(0)=Q^T(0)$ *such that the functional* V *is positive definite, bounded above and* $\dot{V}\leqq 2(\gamma+\varepsilon)V$. *Of course, if* $\gamma<0$, *the above result implies exponential asymptotic stability*; *moreover, the rate of decay is precisely as expected.*

4. Applications. Let us now study the scalar equation

$$\dot{x}(t)=-a\int_{-r}^{0}x(t+\theta)d\theta,\qquad t\geqq 0,\tag{4.1}$$

where $a\in R$ is a constant. We will investigate its asymptotic behavior through an analysis of its characteristic equation

$$\lambda+a\int_{-r}^{0}e^{\lambda\theta}d\theta=0,\tag{4.2}$$

and the solution of the corresponding equation (3.7).

Equation (4.1) is a special case of the scalar equation

$$\dot{x}(t)=-\alpha\int_{-\infty}^{0}x_t(\theta)\,d\eta(\theta),\qquad t\geqq 0,\tag{4.3}$$

where $\alpha\in R$ is a constant and η is a nondecreasing function. In [3], it is shown that if

$$\int_{-\infty}^{0}\theta\,d\eta(\theta)>\frac{-1}{\alpha},$$

then all roots of the characteristic equation of (4.3), namely,

$$\lambda+\alpha\int_{-\infty}^{0}e^{\lambda\theta}d\eta(\theta)=0,\tag{4.4}$$

satisfy $\operatorname{Re}\lambda\leqq 0$. This result, when applied to (4.1) states that when $0<ar^2<2$, the roots of (4.2) satisfy $\operatorname{Re}\lambda\leqq 0$. We improve this result by showing that a necessary and sufficient condition for all roots of (4.2) to have negative real part is that $0<ar^2<\pi^2/2$.

In order to do that, we first note that (4.2) can be written as

$$\lambda^2 = -a + ae^{-\lambda r}, \qquad \lambda \neq 0, \tag{4.5}$$

or, letting $\lambda = \alpha + \beta i$, as the system

$$\begin{aligned} &\alpha^2 - \beta^2 = -a + ae^{-\alpha r}\cos\beta r, \\ &2\alpha\beta = -ae^{-\alpha r}\sin\beta r, \\ &(\alpha,\beta) \neq (0,0). \end{aligned} \tag{4.6}$$

Then, we have:

THEOREM 4.1. *If* $0 < ar^2 < \pi^2/2$ *then every root* λ *of* (4.2) *has negative real part, and conversely.*

Proof. Suppose that $0 < ar^2 < \pi^2/2$ and that $\lambda = \alpha + \beta i$, with $\alpha \geqq 0$, is a root of (4.2). Then, in view of (4.6), β cannot be zero, for otherwise we would have $\alpha^2 + a = ae^{-\alpha r}$, which implies $\alpha = 0$, a contradiction. Therefore, $\beta \neq 0$ and we can suppose, without loss of generality, that $\beta > 0$. Suppose then that $\alpha = 0$. In this case, the second of equations (4.6) tells us that $\beta r = k\pi$, where $k \geqq 1$ is an integer. Hence, the first of equations (4.6) becomes

$$-\frac{k^2\pi^2}{r^2} = -a + a(-1)^k,$$

which implies that k is an odd integer and $(2ar^2)/\pi^2 = k^2 \geqq 1$, a contradiction.

Suppose now that $\alpha > 0$. Then, since $\sin\beta r = -(2\alpha\beta e^{\alpha r})/a < 0$, we must have $\beta r > \pi$. Therefore,

$$\pi^4 < \beta^4 r^4 \leqq |\lambda|^4 r^4 = a^2 r^4 (1 - 2e^{-\alpha r}\cos\beta r + e^{-2\alpha r}) \leqq 4a^2 r^4,$$

which implies that $\pi^2 < 2ar^2$, a contradiction.

To prove the converse, we note first that for $(2ar^2)/\pi^2 = k^2$, k odd, there is a pair of conjugate purely imaginary roots of (4.5) given by $\lambda = \pm(k\pi/r)i$. Also, these are the only purely imaginary roots of (4.5). Secondly, we note that $\lambda = \alpha + \beta i$ is a root of (4.6) if and only if λ/r is a root of

$$\alpha^2 - \beta^2 + ar^2 = ar^2 e^{-\alpha}\cos\beta, \qquad 2\alpha\beta = -ar^2 e^{-\alpha}\sin\beta.$$

Next, we introduce the parameter $\rho = ar^2$ and consider the equations

$$F(\alpha,\beta,\rho) \equiv \alpha^2 - \beta^2 + \rho - \rho e^{-\alpha}\cos\beta = 0, \qquad G(\alpha,\beta,\rho) \equiv 2\alpha\beta + \rho e^{-\alpha}\sin\beta = 0.$$

We have, for any odd k,

$$\begin{aligned} &F\left(0, k\pi, \frac{k^2\pi^2}{2}\right) = 0, \\ &G\left(0, k\pi, \frac{k^2\pi^2}{2}\right) = 0, \\ &\frac{\partial(F,G)}{\partial(\alpha,\beta)}\left(0, k\pi, \frac{k^2\pi^2}{2}\right) = \frac{k^4\pi^4}{4} + 4k^2\pi^2 \neq 0. \end{aligned}$$

Therefore, the system $F = 0$, $G = 0$ defines, for each odd k, a unique curve $\Gamma_k : \alpha = \alpha(\rho)$, $\beta = \beta(\rho)$ about the point $\rho = (k^2\pi^2)/2$ with $\alpha((k^2\pi^2)/2) = 0$ and $\beta((k^2\pi^2)/2) = k\pi$.

It is not difficult to verify that

$$\frac{d\alpha}{d\beta}\left(\frac{k^2\pi^2}{2}\right)=\frac{4}{16+k^2\pi^2}>0. \tag{4.7}$$

Hence, (4.5) has roots with positive real part for ρ in a neighborhood of $(k^2\pi^2)/2$, for k odd.

Let us see now that for any $\rho>\pi^2/2$ there is at least one root of (4.5) with positive real part. Since, by (4.6), we have

$$\alpha^2(\rho)-\beta^2(\rho)\leqq 2\rho$$

for any of the curves Γ_k, we see that Γ_k can be continued to $\rho=+\infty$. Each of these curves must stay in the first quadrant for ρ in the maximal interval of existence, because they cannot touch or cross the imaginary axis in view of the fact that the only purely imaginary roots of (4.5) are $\lambda=k\pi i$, for odd k, at $\rho=(k^2\pi^2)/2$, and (4.7) shows that at these roots the curves cross the imaginary axis from the second to the first quadrant.

Finally, it is easy to verify that for $\rho<0$ there is always a positive real root to (4.5). This finishes the proof of Theorem 4.1.

Remarks. It follows from the above proof that every root of (4.2) has nonpositive real part if and only if $0\leqq ar^2<\pi^2/2$.

We now proceed to find the solution $Q(\alpha)$ to the equation corresponding to (3.7). In this case, (3.7) reduces to the scalar equation

(4.8)

$$\dot{Q}(\alpha)=\delta Q(\alpha)-a\int_{-r}^{-\alpha}e^{-\delta\theta}Q(-\alpha-\theta)\,d\theta-a\int_{-\alpha}^{0}e^{-\delta\theta}Q(\alpha+\theta)\,d\theta,\qquad 0\leqq\alpha\leqq r.$$

Any solution of (4.8) also solves

$$Q^{(2)}(\alpha)=2\delta Q(\alpha)-(a+\delta^2)Q(\alpha)+ae^{\delta r}Q(r-\alpha),\qquad 0\leqq\alpha\leqq r \tag{4.9}$$

and

$$Q^{(4)}(\alpha)+2(a-\delta^2)Q^{(2)}(\alpha)+\left[(a+\delta^2)^2-a^2e^{2\delta r}\right]Q(\alpha)=0,$$

whose characteristic equation is

$$m^4+2(a-\delta^2)m^2+(a+\delta^2)^2-a^2e^{2\delta r}=0. \tag{4.10}$$

Letting $\pm\lambda$ and $\pm\mu$ denote the roots of (4.10), we can write the general (real) solution of (4.9) as

$$Q(\alpha)=C_1e^{\lambda(\alpha-r/2)}+C_2e^{-\lambda(\alpha-r/2)}+C_3e^{\mu(\alpha-r/2)}+C_4e^{-\mu(\alpha-r/2)}, \tag{4.11}$$

if λ and μ are nonzero and distinct, where $C_1,\cdots,C_4$ are arbitrary constants with $C_2=\bar{C}_1$ and $C_4=\bar{C}_3$ if λ and μ are nonreal complex numbers. On the other hand, if $\lambda=0$ and $\mu\neq 0$, which happens when $\delta=0$, then the general real solution of (4.9) is given by

$$Q(\alpha)=C_1+C_2(\alpha-r/2)+C_3e^{\mu(\alpha-r/2)}+\bar{C}_3e^{-\mu(\alpha-r/2)}.$$

The remaining cases can be handled similarly.

Now, in order to use (4.11) to solve (4.8), it is necessary and sufficient that

$$C_2=N(\delta,\lambda)C_1,\quad C_4=N(\delta,\mu)C_3\quad\text{and}\quad C_3=-M(\delta)C_1,$$

where

$$N(\delta,\mu)=\frac{(u-\delta)^2+a}{ae^{\delta r}} \quad \text{and} \quad M(\delta)=\frac{(\lambda/(\lambda^2-\delta^2))\left[e^{-\lambda r/2}-N(\delta,\lambda)e^{\lambda r/2}\right]}{(\mu/(\mu^2-\delta^2))\left[e^{-\mu r/2}-N(\delta,\mu)e^{\mu r/2}\right]}$$

so that the general solution of (4.8) is given by

$$Q(\alpha)=C\left\{e^{\lambda(\alpha-r/2)}+N(\delta,\lambda)e^{-\lambda(\alpha-r/2)}-M(\delta)\left[e^{\mu(\alpha-r/2)}+N(\delta,\mu)e^{-\mu(\alpha-r/2)}\right]\right\},$$
$$0\leqq\alpha\leqq r,$$

where C is an arbitrary real constant.

The particular solution we are interested in is the one for which the value of C is determined by condition (3.9), i.e., $2\dot{Q}(0)=-w$, w a positive real number. If this solution is inserted into (3.8), one gets a functional U, which, along the solutions of (4.1) is given by (3.1) and $\dot{U}$ is given by (3.2). Hence, if $\gamma<0$ then we can choose $\delta>0$ and positive scalars M, N, R and W so that V as in (3.10) is a Lyapunov functional, which proves the asymptotic stability of (4.1). But, if $\gamma<0$, we can also take $\delta=0$ in (3.8) and this choice yields a simpler functional which, together with the invariance principle [1] proves the asymptotic stability of (4.1). In this case, U assumes the form

$$U(\xi,\psi)=Q(0)\xi^2-2a\xi\int_{-r}^{0}\left[\int_{0}^{\beta+r}Q(u)\,du\right]\psi(\alpha)^1d\alpha$$
$$+a^2\int_{-r}^{0}\int_{-r}^{0}\left[\int_{0}^{\alpha+r}\int_{0}^{\beta+r}Q(u-v)\,du\,dv\right]\psi(\alpha)\psi(\beta)\,d\alpha\,d\beta,$$

where Q is now the solution of the integro-differential equation

$$\dot{Q}(\alpha)=-a\int_{0}^{r-\alpha}Q(u)\,du-a\int_{0}^{\alpha}Q(u)\,du,$$

that satisfies the condition $2\dot{Q}(0)=-w$.

Proceeding as before, one gets

$$Q(\alpha)=C\left\{-\frac{\sqrt{2a}}{ar}\cos\sqrt{2a}\,\frac{r}{2}+\sin\sqrt{2a}\left(\frac{\alpha-r}{2}\right)\right\}, \qquad 0\leqq\alpha\leqq r,$$

where $C=-w(2\sqrt{2a}\,\cos\sqrt{2a}\,(r/2))$.

As another application, consider the scalar equation

$$\text{(4.12)} \qquad \dot{x}(t)=-a\int_{-1}^{0}(1+\theta)x(t+\theta)\,d\theta, \qquad t\geqq0,$$

where $a>0$ is a constant.

A Lyapunov functional was presented by Levin and Nohel in [5] for this equation, namely,

$$W(\xi,\psi)=\frac{1}{2}\xi^2+\frac{a}{2}\int_{-1}^{0}\left[\int_{\theta}^{0}\psi(s)\,ds\right]^2d\theta,$$

whose derivative along the solutions of (4.12) is given by

$$W(\xi,\psi)=-\frac{a}{2}\left[\int_{-1}^{0}\psi(\theta)\,d\theta\right]^2.$$

By using this functional and the invariance principle, it was shown in [4] that all solutions of (4.12) are bounded and that, if $a \neq 4m^2\pi^2$ for all integers m, then (4.12) is asymptotically stable.

Hence, if $a \neq 4m^2\pi^2$ for all integers m, Theorem 3.2 can be applied to provide a positive definite quadratic Lyapunov functional V for (4.12) with $\dot{V}$ strongly negative definite. Equation (3.7), with $\delta = 0$, becomes, in this case,

$$\dot{Q}(\alpha) = -a\int_0^{1-\alpha}(1-\alpha-u)Q(u)\,du - a\int_0^{\alpha}(1-\alpha+u)Q(u)\,du, \qquad 0 \leqq \alpha \leqq 1. \tag{4.13}$$

Any solution of (4.13) also satisfies

$$\begin{aligned} Q^{(2)}(\alpha) &= -aQ(\alpha) + a\int_0^{1-\alpha} Q(u)\,du + a\int_0^{\alpha} Q(u)\,du, \\ Q^{(3)}(\alpha) &= -a\dot{Q}(\alpha) + aQ(\alpha) - aQ(1-\alpha). \end{aligned} \tag{4.14}$$

Putting $g(\alpha) = Q(\alpha) - Q(1-\alpha)$ and $h(\alpha) = Q(\alpha) + Q(1-\alpha)$, we have

$$g^{(3)}(\alpha) = -ag^{(1)}(\alpha), \qquad h^{(3)}(\alpha) = -ah^{(1)}(\alpha) + 2ag(\alpha). \tag{4.15}$$

Using the relations $g(\frac{1}{2}) = g^{(2)}(\frac{1}{2}) = h^{(1)}(\frac{1}{2}) = 0$ in the general solution of the system (4.15) and the fact that $2Q(\alpha) = g(\alpha) + h(\alpha)$, it follows that the general solution of (4.14) is given by

$$Q(\alpha) = E_1 + E_2 \cos\sqrt{a}\left(\alpha - \frac{1}{2}\right) + E_3\left(\alpha - \frac{3}{2}\right)\sin\sqrt{a}\left(\alpha - \frac{1}{2}\right), \tag{4.16}$$

where E_1, E_2 and E_3 are arbitrary real constants. Upon forcing (4.16) to satisfy (4.13) one gets the final expression for the solution of (4.13), depending on just one multiplicative real constant. This multiplicative constant can be chosen such that $2\dot{Q}(0) = -w$, $w > 0$, since $\sqrt{a}/2\pi$ is not an integer.

REFERENCES

[1] J. K. HALE, *Theory of Functional Differential Equations*, Applied Mathematical Sciences, Vol. 3, Springer-Verlag, New York, 197.

[2] E. F. INFANTE AND W. B. CASTELAN, *A Lyapunov functional for a matrix difference-differential equation*, J. Differential Equations, 29 (1978), pp. 439–451.

[3] H. S. STECH, *The effect of time lags on the stability of the equilibrium state of a population growth equation*, J. Math. Biol., (1977/78), 5, no. 2, pp. 115–120.

[4] J. A. WALKER, *On the applications of Lyapunov's direct method to linear dynamical systems*, Math. Anal. Appl., 53 (1976), pp. 187–220.

[5] J. J. LEVIN AND J. A. NOHEL, *On a nonlinear delay equation*, J. Math. Anal. Appl., 8 (1964), pp. 31–34.

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

012

A THEOREM FOR THE UNIFORM BOUNDEDNESS OF A FAMILY OF COMPACT OPERATORS*

LI RONGHUA†

Abstract. A kind of sufficient condition for uniform boundedness of a family of matrices $G^n(\theta,\Delta t)$, which was given by the present author and Zhou Changlin, has been generalized to a family of compact operators $T^n(\theta,\Delta t)$.

1. Introduction. It is well known that the stability problem of difference schemes for an initial value problem with constant coefficients can be reduced by means of separation of variables to the problem of uniform boundedness of a family of matrices

$$\{G^n(\theta,\Delta t), a\leqq\theta\leqq b, 0\leqq n\Delta t\leqq T\}, \tag{1}$$

where $\Delta t>0$ is the time stepsize. A necessary condition for the family (1) to be uniformly bounded is the von Neumann condition, namely, that the spectral radius $\rho(\theta,\Delta t)$ of $G(\theta,\Delta t)$ satisfies

$$|\rho(\theta,\Delta t)|\leqq 1+M\Delta t, \tag{2}$$

where M is a constant independent of n and Δt. Another necessary condition is that the family of matrices

$$\{G^n(\theta,0), a\leqq\theta\leqq b\} \tag{3}$$

is uniformly bounded. Yet another is the following condition, for the boundedness of powers of each individual matrix, by Oldenburger [4]:

The eigenvalues $\lambda_i(\theta,0)$ of $G(\theta,0)$ satisfy

$$\begin{aligned} &|\lambda_i(\theta,0)|\leqq 1 && \text{if } \lambda_i(\theta,0) \text{ only has linear elementary divisors,}\\ &|\lambda_i(\theta,0)|<1 && \text{otherwise.}\end{aligned}$$

The present author and Zhou Changlin, considered in [3] (cf. [5]) a kind of sufficient condition for the stability in which only the eigenvalues of $G(\theta,\Delta t)$ are involved. This condition is formulated in the following theorem.

THEOREM 1. *The family of matrices* (1) *is uniformly bounded if the following conditions* (i), (ii) *and* (iii) *are satisfied*:

(i) *The matrices $G(\theta,\Delta t)$ of order p satisfy the following Lipschitz condition with index (α,β)*:

$$\|G(\theta_2,\Delta t)-G(\theta_1,\Delta t)\|\leqq M_1|\theta_2-\theta_1|^{\alpha}+M_2(\Delta t)^{\beta}, \qquad \alpha,\beta>0.$$

(ii) *The eigenvalues $\lambda_1(\theta,t),\cdots,\lambda_p(\theta,\Delta t)$ of $G(\theta,\Delta t)$ and $\lambda_1(\theta,0),\cdots,\lambda_p(\theta,0)$ of $G(\theta,0)$ satisfy the von Neumann condition and the R. Oldenburger condition respectively.*

(iii) *For any given $\theta_0\in[a,b]$, if $G(\theta_0,0)$ has a k-ple eigenvalue with modulus* 1 *and linear elementary divisors, for example $\lambda_1(\theta_0,0)=\cdots=\lambda_k(\theta_0,0)$, then when $\theta\to\theta_0$ and*

*Received by the editors September 25, 1981, and in final revised form March 1, 1984.

† Department of Mathematics, Jilin University, Changchun, People's Republic of China.

$\Delta t \to 0$, *the following condition holds*:

(4)

$$|\theta-\theta_0|^{\alpha}+(\Delta t)^{\beta}=O\left(\Delta t+\left|1-\max_{1\leqq i\leqq k}|\lambda_i(\theta,\Delta t)|\right|^{+}\min_{\substack{i\neq j\\ 1\leqq i,j\leqq k}}|\lambda_i(\theta,\Delta t)-\lambda_j(\theta,\Delta t)|\right).$$

In this paper we shall generalize the result of Theorem 1 to the case of a family consisting of compact operators.

2. Uniform boundedness of the compact operator family. In this section we shall make use of some known results in the spectral and perturbation theory of compact operators which can be found in [2, Chap. 7].

Let H be a Hilbert space; T is a compact operator on H. It is well known that the eigenvalue set of T is finite or countable, and in the latter case its only accumulation point is zero. Thus the elements of this set can be written as a sequence $\lambda_1,\lambda_2,\cdots,\lambda_n,\cdots$, decreasing in absolute value (with multiple eigenvalues repeated). We denote the spectrum of T by $\sigma(T)=\{\lambda_i\}\cup\{0\}$ and τ_N is the complement of $\sigma_N=\{\lambda_1,\lambda_2,\cdots,\lambda_N\}$. Let $N(\sigma_N)$ and $N(\tau_N)$ denote some neighborhoods of σ_N and τ_N with empty intersection respectively. Set

$$e_{\sigma_N}(\mu)=\begin{cases}1 & \text{if } \mu\in N(\sigma_N),\\ 0 & \text{if } \mu\in N(\tau_N),\end{cases}\qquad e_{\tau_N}(\mu)=\begin{cases}1 & \text{if } \mu\in N(\tau_N),\\ 0 & \text{if } \mu\in N(\sigma_N).\end{cases}$$

Then $E(\sigma_N)=e_{\sigma_N}(T)$ and $E(\tau_N)=e_{\tau_N}(T)$ are projection operators from H into $E(\sigma_N)H$ and $E(\tau_N)H$ respectively, and H can be written as a direct sum of $E(\sigma_N)H$ and $E(\tau_N)H$, where the subspaces $E(\sigma_N)H$ has finite dimension and is an invariant subspace of T.

Let $f(z)$ be analytic in a neighborhood of $\sigma(T)$. Then

$$f(T)=f(TE(\sigma_N))+f(TE(\tau_N)).$$

Choosing a circle $C:|Z|=R<1$ such that $\lambda_i \overline{\in} C(i=1,2,\cdots)$ and an integer N sufficiently big such that $|\lambda_n|\leqq\rho<R$ for $n>N$, we thus have

$$f(T)=f(TE(\sigma_N))+\frac{1}{2\pi i}\int_C f(z)(zI-TE(\tau_N))^{-1}dz. \tag{5}$$

In particular,

$$T^n=(TE(\sigma_N))^n+\frac{1}{2\pi i}\int_C z^n(zI-TE(\tau_N))^{-1}dz. \tag{6}$$

It can be seen that the second term on the right-hand side of (6) tends to zero when $n\to\infty$, and the first term is a power of the operator $TE(\sigma_N)$ which maps the finite-dimensional space $E(\sigma_N)H$ into itself. Therefore by the Oldenburger theorem we obtain the following result.

THEOREM 2. *A necessary condition for the family $\{T^n\}$ of compact operators to be uniformly bounded is that $|\lambda_i|\leqq 1$ and when the index of λ_i is greater than* 1, $|\lambda_i|<1$.

We now consider the uniform boundedness of the family

$$\{T^n(\theta,\Delta t),\, a\leqq\theta\leqq b,\, 0\leqq n\Delta t\leqq T\}. \tag{7}$$

For this purpose it suffices to prove that the family (7) is uniformly bounded in a neighborhood of any $\theta_0 \in [a,b]$. Assume that $T(\theta,\Delta t)$ is continuous in θ and Δt. Then $\lambda_i = \lambda_i(\theta,\Delta t)$ also will be continuous in θ and Δt. Let N be sufficiently big such that $|\lambda_n(\theta_0,0)| < \rho < 1$ for $n > N$. Then there exists a $\delta > 0$ such that when $|\theta - \theta_0| < \delta$, $0 < \Delta t < \delta$ and $n > N$, $|\lambda_n(\theta,\Delta t)| \leqq \rho$. Since $|\lambda_i(\theta,\Delta t)|$ decreases with respect to i, we can choose R_1, R_2 such that $\rho < R_1 < R_2 < 1$ and that the region: $\{z : R_1 < |z| < R_2\}$ contains no $\lambda_i(\theta,\Delta t)$ when $|\theta_0 - \theta_0| < \delta$ and $0 < \Delta t < \delta$. Hence we have by formula (6)

(8)

$$T^n(\theta,\Delta t) = (T(\theta,\Delta t))E(\sigma_N(\theta,\Delta t))^n + \frac{1}{2\pi i}\int_C z^n(zI - T(\theta,\Delta t)E(\tau_N(\theta,\Delta t)))^{-1}dz.$$

Obviously, the second term on the right-hand side is uniformly bounded (in fact, when $n \to \infty$ it tends to zero uniformly). Further, from perturbation theory (see [2, pp. 584–587]) we know that when δ is sufficiently small, the dimension of the subspaces $E(\sigma_N(\theta,\Delta t))H$ equals a finite integer $p \geqq 0$. Hence the operator $T(\theta;\ \Delta t)E(\sigma_N(\theta,\Delta t))$ can be expressed as a $p \times p$ matrix whose eigenvalues are $\lambda_1(\theta,\Delta t),\cdots,\lambda_N(\theta,\Delta t)$, $N \leqq p$. If $T(\theta,\Delta t)E(\sigma_N(\theta,\Delta t))$ satisfies the conditions of Theorem 1, the first term on the right hand side of (8) will be uniformly bounded. We have then proven the following theorem.

THEOREM 3. *Assume that the operator* $T(\theta,\Delta t)E(\sigma_N(\theta,\Delta t)) : E(\sigma_N(\theta,\Delta t))H \to E(\sigma_N(\theta,\Delta t))H$ *and its eigenvalues* $\lambda_1(\theta,\Delta t),\cdots,\lambda_p(\theta,\Delta t)$ *satisfy the conditions of Theorem* 1, *where N is sufficiently big that* $|\lambda_n(\theta,0)| < \rho < 1$ *for* $n > N$. *Then the family* (7) *of compact operators is uniformly bounded.*

Remark 1. From the proof of [2, Lemma 6, Chap. 7, §6] (consider the restriction of $T(\theta,\Delta t)$ to $E(\sigma_N(\theta,\Delta t))H$) we see that $E(\sigma_N(\theta,\Delta t))$ has the same smoothness as $T(\theta,\Delta t)$. Therefore, we can check the condition (i) of Theorem 1 for $T(\theta,\Delta t)$ instead for $E(\sigma_N(\theta,\Delta t))$.

Remark 2. In particular, if the compact operators $T = T(\theta,\Delta t)$ do not depend on θ and Δt, then we can deduce from (6) that $\lim_{n\to\infty} T^n = T^\infty$ exists if and only if all eigenvalues λ_i $(i = 1,2,\cdots)$ of T satisfy $|\lambda_i| \leqq 1$; if there is an eigenvalue, for example λ_1, with modulus 1, then all its elementary divisors must be linear. By using this fact we can obtain the convergence theorem of stationary iterative process for the operator equation $x - Tx = b$. In the case of matrix equations, corresponding results can be found in [6].

REFERENCES

[1] R. D. RICHTMYER AND K. W. MORTON, *Difference Methods for Initial Value Problems*, Interscience, New York, 1967.

[2] D. SCHWARTZ, *Linear Operators, Part* I: *General Theory*, Interscience, New York, 1958.

[3] LI RONGHUA AND ZHOU CHANGLIN, *The uniform boundedness of a family of matrices* $\{G^n(\theta,\Delta t)\}$ *and the stability condition of difference schemes*, (I), (II), (III), Acta Scientiarum Naturalium Universitatis Jilinensis, 2 (1963), pp. 319–349; 2 (1964), pp. 85–97; 3 (1964), pp. 15–29.

[4] R. OLDENBURGER, *Infinite powers of matrices and characteristic roots*, Duke Math. J., 16 (1940), pp. 357–361.

[5] MA SILIANG AND LI RONGHUA, *On the locally condition*(J) *for the uniform boundedness of a family of matrices* $G^n(\theta,\Delta t)$, Acta Math. Sinica, 26 (1983), pp. 723–730.

[6] G. I. MARCUK AND JU. A. KUZNECOV, *Iteration methods and quadratic functionals*, Methods of Numerical Mathemastics "Nauka" Sibirsk. Otdel., Novsibirsk, 1975, pp. 4–143, 279. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

013

A NEW SET OF POLYNOMIAL MEANS RELATED TO THE MEANS OF MACLAURIN*

J. L. BRENNER†

A comparison theorem is proved for a one-parameter set of symmetric means in k variables, homogeneous of degree 1. If the parameter has the value 1, the inequalities specialize to certain inequalities proved by Maclaurin. Generalizations are given.

1. Introduction. The means M_1, $M_2, \cdots$ studied in this article are all symmetric and homogeneous of degree 1. The program proposed is to find all comparison theorems of the form $M_1 > M_2$ among them. Special cases of comparison theorems are Theorem 1.1: The theorem of the (power) means; Theorem 1.2: Maclaurin's theorem concerning the elementary symmetric functions; Theorem 1.3: Muirhead's theorem.

Let $a_1, \cdots, a_k$ be a set of k positive numbers, not all equal. Let $\alpha(1), \cdots, \alpha(k)$ be a set of real numbers with nonzero sum $|\alpha| = \Sigma\alpha(i)$. The mean

$$\left[\left(\sum_{\text{symmetric}} a_1^{\alpha(1)} \cdots a_k^{\alpha(k)}\right)/k!\right]^{1/|\alpha|} = M(a_1 \cdots a_k | \alpha(1) \cdots \alpha(k))$$

is symmetric and homogeneous of degree 1: $M(\lambda a | \cdots) = \lambda M(a | \cdots)$. An example is the power mean, $\alpha \equiv (r, 0, \cdots, 0)$, with the well-known comparison theorem:

THEOREM 1.1. THEOREM OF THE (POWER) MEANS. $M(a|r, 0, \cdots, 0) < M(a|s, 0, \cdots, 0)$ *if* $r < s$ [4]. *If* $r = 1$, *the mean is the arithmetic mean. If* $r = 0$, *the mean is the geometric mean.*

Another set of examples are the means based on the elementary symmetric functions: $\alpha \equiv (1, 1, \cdots, 1_t, 0, \cdots, 0)$. If $t = 1$, this mean is the arithmetic mean; if $t = k$, this mean is the geometric mean. The comparison theorem of Maclaurin is:

THEOREM 1.2. (THEOREM OF MACLAURIN). $M(a|1, 1, \cdots, 1_t, 0, \cdots, 0) > M(a|1, 1, \cdots, 1_s, 0, \cdots,)$ *if* $1 \leqq t < s \leqq k$ [4].

A further example is the theorem of Muirhead; this theorem relates two means $M(a|\alpha)$, $M(a, \beta)$ provided $|\alpha| = |\beta|$.

THEOREM 1.3. (THEOREM OF MUIRHEAD). *If the components of* α, β *are all nonnegative and nonincreasing, and if there is a doubly stochastic matrix* $D \neq I$ *such that* $\beta = \alpha D$, *then* $M(a|\alpha) > M(a|\beta)$ [4].

A further comparison theorem, new at the time it was published, is based on the vectors $\alpha = (1, 1, \cdots, 1_t, 0, \cdots, 0)$, $\beta = (1, 1, \cdots, 1_t, \varepsilon, 0, \cdots, 0)$ *with* $0 < \varepsilon \leqq 1$.

THEOREM 1.4 [2]. *With* α, β *defined as in the preceding paragraph, the relation* $M(a|\alpha) > M(a|\beta)$ *holds.*

We conclude this section with some general remarks. We call the means $M(a|\alpha)$ *polynomial means*, since if $\alpha(i)$ are nonnegative integers, $M^{|\alpha|}$ is an ordinary polynomial. A comparison theorem ($M_1 > M_2$ for all positive vectors a with $a_1 \neq a_2$) is rare. It can be shown that a pair of polynomial means is usually incomparable. For example

*Received by the editors January 27, 1984.

†10 Phillips Road, Palo Alto, California 94303.

with $\beta \equiv \beta(\varepsilon)$ defined as in Theorem 1.4, two means $M(a|\beta)$ will be incomparable in general, when ε takes two different values. See [2] for a detailed analysis.

Ordinary mathematical experience also suggests that (true) theorems in analysis are hard to come by. In view of the remark at the end of the preceding paragraph, the results of this article are unexpected. A special case of what will be proved is:

THEOREM 1.5. *Define* $\beta_{t,\varepsilon}=(1,1,\cdots,1_t,\varepsilon,0,\cdots,0)$, $\beta_{t+1,\varepsilon}=(1,1,\cdots,1_{t+1},\varepsilon,0,\cdots,0)$, $0<\varepsilon\leqq 1$. *Then* $L_{t,\varepsilon}\equiv M(a|\beta_{t,\varepsilon})\geqq M(a|\beta_{t+1,\varepsilon})=L_{t+1,\varepsilon}$.

The inequality is undoubtedly strict; I can prove strict inequality only if ε is rational. The method of proof is somewhat novel. According to a suggestion of Hardy, Littlewood, and Pólya [4], it is entirely "elementary." (So also are the methods of [2].) In fact the method used here appears rarely, if at all, in the literature. (Correspondence [1] indicates that the method has been applied before, for instance to prove the Cauchy inequality.)

The conjecture that $L_{t-1,\varepsilon}$ decreases monotonically on the range $0\leqq\varepsilon\leqq 1$ is false. Take $t=2$, $0.5\leqq\varepsilon<1$, $a_1=a_2=1$, $a_3=1/2$, $a_4=a_5=\cdots=a_k=0^+$.

Note that if $\mathbb{1}\equiv(1,1,\cdots,1)$, then $M(\mathbb{1}|\alpha)=1$.

2. Discursum. Special results when $k=2$. The results discussed here are special. They cannot be generalized to $k>2$.

LEMMA 2.1. *If* $k=2$, $\frac{1}{3}\leqq\varepsilon$ (*but not if* $0<\varepsilon<\frac{1}{3}$) *the relation* $L_{0,\varepsilon}>L_{1,\varepsilon}$ *holds* [3].

COUNTEREXAMPLE 2.2. If $k=3$, $\varepsilon=\frac{1}{2}$, the relation $L_{0,\varepsilon}>L_{1,\varepsilon}$ no longer holds. Take $a_1=a_2=1$, $a_3=\frac{1}{4}$. Then $L_{0,\varepsilon}=\frac{25}{36}$, $L_{1,\varepsilon}=(\frac{7}{12})^{2/3}$. But $L_{0,\varepsilon}^{3/2}=\frac{125}{216}<\frac{7}{12}=L_{1,\varepsilon}^{3/2}$, because $(\frac{7}{12})216=126$. Counterexamples also exist if $k>3$.

LEMMA 2.3. *If* $k=2$, $0<\varepsilon<1$, *or if* $\varepsilon<-1$, $y=(1-\varepsilon+\varepsilon^2)/(\varepsilon-\varepsilon^3)$, *then* $L_{1,\varepsilon}<2^yL_{0,\varepsilon}$.

Proof. The inequality of Hölder gives

$$\frac{1}{2}\left(a_1a_2^{\varepsilon}+a_1^{\varepsilon}a_2\right)\leqq\frac{1}{2}\left(a_1^{\varepsilon}+a_2^{\varepsilon}\right)^{1/\varepsilon}\left(a_1^{\varepsilon(1-\varepsilon)}+a_2^{\varepsilon(1-\varepsilon)}\right)^{1/(1-\varepsilon)}.$$

Therefore $L_{1,\varepsilon}^{1+\varepsilon}\leqq\frac{1}{2}(2^{1/\varepsilon}L_{0,\varepsilon})(2^{1/(1-\varepsilon)}L_{0,\varepsilon}^{\varepsilon})$. □

LEMMA 2.4. *If* $k=2$, $-1<\varepsilon<0$, *the relation* $L_{1,\varepsilon}>L_{0,\varepsilon}$ *holds.*

Proof. $L_{1,\varepsilon}>G>L_{0,\varepsilon}$.

LEMMA 2.5. *If* $k=2$, $\varepsilon>1$, *the relation* $L_{0,1}>L_{1,\varepsilon}$ *holds.*

REMARK. This assertion is stronger than 2.1, since $L_{0,\varepsilon}>L_{0,1}$.

Proof of Lemma 2.5. The two inequalities

$$(a_1a_2)^{\varepsilon}<\left[\frac{1}{2}(a_1+a_2)\right]^{2\varepsilon},$$

$$\left[\frac{1}{2}\left(a_1^{1-\varepsilon}+a_2^{1-\varepsilon}\right)\right]^{(1-\varepsilon)/(1-\varepsilon)}<\left[\frac{1}{2}(a_1+a_2)\right]^{1-\varepsilon}$$

are special cases of the power-mean inequality. If they are multiplied, the result is

$$L_{1,\varepsilon}^{1+\varepsilon} \quad < \quad L_{0,1}^{1+\varepsilon}.$$

□

REMARK 2.7. Lemma 2.5 is not valid if $k=3$. Take $a_1=a_2=1$, $a_3=4$, $\varepsilon=\frac{5}{2}$. Then

$$L_{0,1}=2,\quad L_{1,\varepsilon}=\left[\tfrac{37}{3}\right]^{2/7},\quad 9L_{0,1}^7 \;=\; 1152<1369 \;=\; 9L_{1,\varepsilon}^7.$$

3. Proof that $L_{1,\varepsilon}>L_{2,\varepsilon}$. An outline of the proof that $L_{1,\varepsilon}>L_{2,\varepsilon}$ is as follows. First the putative inequality is transformed to an inequality between powers of ratios: $[K(\varepsilon)/P(\varepsilon)]^{r+s} \overset{?}{>} [6/K(\varepsilon)]^s$, where $K(\varepsilon)$, $P(\varepsilon)$ are certain polynomials and $\varepsilon=r/s$, r,s being positive integers, $0<r<s$. Then each ratio is written as a (collapsing) product of a number of fractions. Finally, each fraction in the left member is compared with a corresponding fraction in the right member. The details are given in full in case $k=3$; these details are followed by an explanation of the modifications needed when $k>3$.

The symbols used in the sequel are defined as follows:

$$\text{(3.01)}\qquad A_1:=a_1^{1/s},\quad A_2:=a_2^{1/s},\quad A_3:=a_3^{1/s};$$

$$\text{(3.02)}\qquad \begin{aligned} H(t):={}&A_1^sA_2^rA_3^t+A_1^rA_2^sA_3^t+A_1^sA_2^tA_3^r+A_1^rA_2^tA_3^s\\ &+A_1^tA_2^sA_3^r+A_1^tA_2^rA_3^s,\quad t=1,2,\cdots,s;\end{aligned}$$

$$\text{(3.03)}\qquad Q(i):=2\left(A_1^i+A_2^i+A_3^i\right),\quad i=1,2,\cdots,s;$$

$$\text{(3.04)}\qquad \begin{aligned} T(j):={}&A_1^sA_2^j+A_1^jA_2^s+A_1^sA_3^j+A_1^jA_3^s+A_2^jA_3^s,\\ &\qquad\qquad j=1,2,\cdots,r.\end{aligned}$$

Recall that $0<r<s$. The first set of inequalities is given by 3.05.

LEMMA 3.05. *The inequalities*

$$\text{(3.06)}\qquad H(t-1)Q(i)>H(t)Q(i-1)$$

hold, provided $r+i\geqq t$; in particular if $i=t$.

Proof.

$$\begin{aligned} &\frac{1}{2}H(t-1)Q(i)-\frac{1}{2}H(t)Q(i-1)\\ &\quad=A_1^{t-1}A_2^rA_3^{t-1}(A_1-A_3)\left(A_1^{s+i-t}-A_3^{s+i-t}\right)\\ &\qquad+A_1^{t-1}A_2^sA_3^{t-1}(A_1-A_3)\left(A_1^{r+i-t}-A_3^{r+i-t}\right)\\ &\qquad+A_1^rA_2^{t-1}A_3^{t-1}(A_2-A_3)\left(A_2^{s+i-t}-A_3^{s+i-t}\right)\\ &\qquad+A_1^sA_2^{t-1}A_3^{t-1}(A_2-A_3)\left(A_2^{r+i-t}-A_3^{r+i-t}\right)\\ &\qquad+A_1^{t-1}A_2^{t-1}A_3^r(A_1-A_2)\left(A_1^{s+i-t}-A_2^{s+i-t}\right)\\ &\qquad+A_1^{t-1}A_2^{t-1}A_3^s(A_1-A_2)\left(A_1^{r+i-t}-A_2^{r+i-t}\right).\end{aligned}$$

Each line is nonnegative, and at least one line is positive. □

The next inequalities needed are given by 3.07.

LEMMA 3.07. *The inequalities*

$$H(t-1)T(j) > H(t)T(j-1) \text{ hold}, \tag{3.08}$$

provided $r+t \geqq j$; *thus for all* j, t.

Proof. $H(t-1)T(j)-H(t)T(j-1)$ is a sum of 72 terms, since each of $H(t-1)$, $T(j)$, $H(t)$, $T(j-1)$ is a sum of 6 terms. These 72 terms, when properly combined, give

$$\begin{aligned}
&A_1^{2s}A_2^{j-1}A_3^{j-1}(A_2-A_3)\left(A_2^{r+t-j}A_3^{r+t-j}\right)\\
&+A_1^{r+s}A_2^{j-1}A_3^{j-1}(A_2-A_3)\left(A_3^{s+t-j}-A_3^{s+t-j}\right)\\
&+A_1^{j-1}A_2^{j-1}A_3^{2s}(A_1-A_2)\left(A_1^{r+t-j}-A_2^{r+t-j}\right)\\
&+A_1^{j-1}A_2^{j-1}A_3^{r+s}(A_1-A_2)\left(A_1^{s+t-j}-A_2^{s+t-j}\right)\\
&+A_1^{j-1}A_2^{2s}A_3^{j-1}(A_1-A_3)\left(A_1^{r+t-j}-A_2^{r+t-j}\right)\\
&+A_1^{j-1}A_2^{r+s}A_3^{j-1}(A_1-A_3)\left(A_1^{s+t-j}-A_3^{s+t-j}\right).
\end{aligned}$$

Again each line is nonnegative, and at least one line is positive. □

The inequality $L_{1,\varepsilon} > L_{2,\varepsilon}$ can now be proved when $k=3$. This inequality comes down to the relation

$$D \quad = \quad [E_{1,\varepsilon}/E_{2,\varepsilon}]^{r+s} \quad > \quad [6/E_{1,\varepsilon}]^{s} \quad = \quad F, \tag{3.09}$$

where $E_{2,\varepsilon} = a_1a_2a_3^{\varepsilon} + a_1a_2a_3^{\varepsilon} + a_1a_2^{\varepsilon}a_3 + a_1a_2^{\varepsilon}a_3 + a_1^{\varepsilon}a_2a_3 + a_1^{\varepsilon}a_2a_3$.

The ratio $E_{1,\varepsilon}/E_{2,\varepsilon}$ can be written as a collapsing product:

$$E_{1,\varepsilon}/E_{2,\varepsilon} = \prod_{t=1}^{s} H(t-1)/H(t); \quad \text{thus} \tag{3.10}$$

$$D = \left[\prod_{t=1}^{s} H(t-1)/H(t)\right]^{r}\left[\prod_{t=1}^{s} H(t-1)/H(t)\right]^{s} = G^{r}U^{s}, \quad \text{say.} \tag{3.11}$$

On the other hand, the fraction $6/E_{1,\varepsilon}$ can be written as a collapsing product of $r+s$ partial fractions:

$$6/E_{1,\varepsilon} = \left[\prod_{1}^{s} Q(i-1)/Q(i)\right]\cdot\left[\prod_{j=1}^{r} T(j-1)/T(j)\right] \quad = \quad V\cdot W, \quad \text{say.} \tag{3.12}$$

Thus $F = V^{s}W^{s}$. It must be proved that $D > F$. Lemma 3.05 compares corresponding factors of U and V, so that $U > V$ and $U^{s} > V^{s}$. Lemma 3.07 states that each (partial) factor of G exceeds every (partial) factor of W, so that the product of the rs factors in G^{r} exceeds the product of the rs factors in W^{s}: $G^{r} > W^{s}$. This completes the proof that $L_{1,\varepsilon} > L_{2,\varepsilon}$ in case $k=3$.

The extension to the case $k>3$ requires no new ideas; the modifications needed are as follows. Definition (3.01) is expanded to read

$$A_{\nu} := a_{\nu}^{1/s}, \qquad \nu = 1, \cdots, k; \tag{3.13}$$

(3.02) becomes

$$H(t) := \sum A_{\mu}^{s}A_{\nu}^{r}A_{\rho}^{t}; \tag{3.14}$$

there are $6C_3^k = k(k-1)(k-2)$ terms in the sum. (3.03) is changed to read

$$Q(i) := 2\sum A_\nu^i. \tag{3.15}$$

Finally, (3.04) metamorphoses to

$$T(j) := \sum_{\mu \neq \nu} A_\mu^s A_\nu^j; \tag{3.16}$$

there are $k(k-1)$ terms in this sum. The proof of Lemma 3.05 is

$$\begin{aligned} &\frac{1}{2}H(t-1)Q(i) - \frac{1}{2}H(t)Q(i-1) \\ &= \sum A_\mu^{t-1} A_\nu^r A_\rho^{t-1} (A_\mu - A_\rho)\left(A_\mu^{s+i-t} - A_\rho^{s+i-t}\right) \\ &\quad + \sum A_\mu^{t-1} A_\nu^s A_\rho^{t-1} (A_\mu - A_\rho)\left(A_\mu^{r+i-t} - A_\rho^{r+i-t}\right). \end{aligned}$$

The proof of Lemma 3.07 is

$$\begin{aligned} &H(t-1)T(j) - H(t)T(j-1) \\ &= \sum A_\mu^{r+s} A_\nu^{j-1} A_\rho^{j-1} (A_\nu - A_\rho)\left(A_\nu^{s+t-j} - A_\rho^{s+t-j}\right) \\ &\quad + \sum A_\mu^{2s} A_\nu^{j-1} A_\rho^{j-1} (A_\nu - A_\rho)\left(A_\nu^{r+t-j} - A_\rho^{r+t-j}\right). \end{aligned}$$

(The sum is extended over all 3-sets μ, ν, ρ.) After these modifications, the formal definitions of D, F, G, U, V, W remain unchanged. The inequality $L_{1,\varepsilon} > L_{2,\varepsilon}$ is established for $0 \leqq \varepsilon \leqq 1$.

Lemma 3.17. *If $\varepsilon > 1$, the relation $L_{1,\varepsilon} > L_{2,\varepsilon}$ is valid.*

Proof. Only a sketch of the argument is given. The definitions of D, G, U, are unchanged. The definitions of F, V, W must be changed to read: $F = V^s W^s$, $V = \prod_{i=1}^r Q(i-1)/Q(i)$, $W = \prod_1^s T(j-1)/T(j)$, where

$$T(j) = A_1^r A_2^j + A_1^j A_2^r + A_1^r A_3^j + A_1^j A_3^r + A_2^r A_3^j + A_2^j A_3^r.$$

Remaining details are omitted. □

Lemma 3.18. *If $\varepsilon > 0$, $k \geqq 2$, then $L_{1,\varepsilon} > L_{0,1-\varepsilon}$.*

Proof. The inequality $(a_1^\varepsilon a_2^\varepsilon)^{1/2} > [\frac{1}{2}(a_1^{1-\varepsilon} + a_2^{1-\varepsilon})]^{\varepsilon/(1-\varepsilon)}$ leads to the result $a_1^\varepsilon a_2 + a_2^\varepsilon a_1 > [\frac{1}{2}(a_1^{1-\varepsilon} + a_2^{1-\varepsilon})]^{(1+\varepsilon)/(1-\varepsilon)}$. The inequality $\Sigma b_i/l > [\Sigma b_i^{(1-\varepsilon)/(1+\varepsilon)}/l]^{(1+\varepsilon)/(1-\varepsilon)}$ establishes the result $(l = C_2^k)$

$$L_{1,\varepsilon}^{1+\varepsilon} > \left[\frac{1}{l}\sum_{i>j}\frac{1}{2}\left(a_i^{1-\varepsilon} + a_j^{1-\varepsilon}\right)\right]^{(1+\varepsilon)/(1-\varepsilon)} = L_{0,1-\varepsilon}^{1+\varepsilon}.$$

Corollary 3.19. *If $0 < \varepsilon < 1$, then*

$$A = L_{0,1} > L_{1,\varepsilon} > L_{0,1-\varepsilon}.$$

This assertion is an improvement (near $\eta = 1$) over the known fact that the power mean $[\Sigma a_i^\eta / k]^{1/\eta}$ increases with η.

Lemma 3.20. *If* $-1<\varepsilon<0$, $k\geqq 2$, *the relation* $L_{1,\varepsilon}<2^{1/(1+\varepsilon)}L_{1,0}$ *holds.*
Proof.

$$C_2^k L_{1,\varepsilon}^{1+\varepsilon}=\sum_{i<j}\left(a_i^\varepsilon a_j+a_i a_j^\varepsilon\right)/2$$

$$=\sum_{i<j} a_i a_j\left(\frac{a_i^{\varepsilon-1}+a_j^{\varepsilon-1}}{2}\right)<\sum_{i<j}\left(\frac{a_i+a_j}{2}\right)^2\left(\frac{a_i+a_j}{2}\right)^{\varepsilon-1}$$

$$=\sum_{i<j}\left(\frac{a_i+a_j}{2}\right)^{1+\varepsilon}<k(k-1)\left\{\frac{2}{k(k-1)}\sum\frac{a_i+a_j}{2}\right\}^{1+\varepsilon}$$

$$=k(k-1)\left\{\sum a_i/k\right\}^{1+\varepsilon}. \qquad \square$$

Lemma 3.21. *If* $\varepsilon<-1$, $k\geqq 2$, *the relation* $L_{1,\varepsilon}>L_{1,0}$ *holds.*

Theorem 3.22. *If* $\varepsilon>1$ *is rational, the relation* $L_{0,\varepsilon}>L_{1,\varepsilon}$ *holds.*

Proof. It is enough to illustrate the proof for the case $k=3$, since the proof for $k>3$ follows the same lines. Set $\varepsilon=r/s$, $A_i=a_i^{1/s}$. Then

$$L_{0,\varepsilon}^{\varepsilon(1+\varepsilon)s}=\left[\sum A_i^r/k\right]^{s+r},$$
$$L_{1,\varepsilon}^{\varepsilon(1+\varepsilon)s}=\left[\sum\left(A_i^s A_j^r+A_i^r A_j^s\right)/(k(k-1))\right]^r.$$

To prove $L_{0,\varepsilon}^{\varepsilon(1+\varepsilon)s}>L_{1,\varepsilon}^{\varepsilon(1+\varepsilon)s}$, transform the latter into the inequality

$$\left[\frac{\Sigma A_i^r}{\frac{1}{k-1}\Sigma\left(A_i^s A_j^r+A_i^r A_j^s\right)}\right]^r>\left[\frac{k}{\Sigma A_i^r}\right]^s. \tag{3.23}$$

The first step in establishing this is to write the bracket on the left-hand side as the collapsing product of s factors, namely

$$\frac{\Sigma A_i^r}{\frac{1}{2}\Sigma\left(A_i^s A_j^r+A_i^r A_j^s\right)}=\prod_1^s H(t-1)/H(t),$$

with

$$H(t)=\frac{1}{k-1}\sum_{i\neq j}\left(A_i^t A_j^r+A_i^r A_j^t\right),\qquad t=1,2,\cdots,s.$$

(Here, $k-1=2$.) Similarly, the bracket on the right-hand side can be written as the collapsing product of r factors: $k/\Sigma A_i^r=\Pi_1^r K(u-1)/K(u)$, with $K(u)=\Sigma A_i^u$, $u=1,2,\cdots,r$. (Here, $k=3$.) The next step is to prove that, if $u\geqq t$, the inequality

$$H(t-1)/H(t)>K(u-1)/K(u)\quad \text{holds if } u\geqq t. \tag{3.24}$$

To see this, compute $D=(k-1)[H(t-1)K(u)-H(t)K(u-1)]$. This difference can be written in the form

$$\begin{aligned}(3.25)\qquad & A_1^{t-1}A_2^{t-1}(A_1-A_2)\left(A_1^{r+u-t}-A_2^{r+u-t}\right)\\ & +A_1^{t-1}A_2^{t-1}A_3^{r}(A_1-A_2)\left(A_1^{u-t}-A_2^{u-t}\right)\\ & +A_1^{t-1}A_2^{r}A_3^{t-1}(A_1-A_3)\left(A_1^{u-t}-A_3^{u-t}\right)\\ & +A_1^{t-1}A_3^{t-1}(A_1-A_3)\left(A_1^{r+u-t}-A_3^{r+u-t}\right)\\ & +A_2^{t-1}A_3^{t-1}(A_2-A_3)\left(A_2^{r+u-t}-A_3^{r+u-t}\right)\\ & +A_1^{r}A_2^{t-1}A_3^{t-1}(A_2-A_3)\left(A_2^{u-t}-A_3^{u-t}\right).\end{aligned}$$

This proves (3.24).

Now note that the rs factors of $[\prod H(t-1)/H(t)]^r$ can be matched with the rs factors of $[K(u-1)/K(u)]^s$ in such a way that each factor of the left-hand side of (3.23) exceeds the corresponding factor of the right-hand side. The way to do this is to use the result, just established, that the relation $H(t-1)/H(t)>K(u-1)/K(u)$ holds as long as $u \geqq t$. First, for each u, $1 \leqq u \leqq s$, match $[K(u-1)/K(u)]^s$ with $[H(u-1)/H(u)]^s$. The remaining factors on the left-hand side are $[\prod_1^s H(t-1)/H(t)]^{r-s}$, and on the right-hand side the remaining factors are $[\prod_{s+1}^r K(u-1)/K(u)]^s$. The matching is complete, and all details have been given for the case $k=3$. If $k>3$, the modification of (3.25) that is needed is clear to anyone who inspects (3.25). □

The relations $L_{2,\varepsilon}>L_{3,\varepsilon}>L_{4,\varepsilon}>\cdots$ can be proved by an elaboration of the methods already given. The argument still involves collapsing products; however each collapsing product involves not just one type of fraction (such as $K(u-1)/K(u)$) but several types of fraction. See [2].

Here is a further extension.

DEFINITION 3.26.

$$L_{0,\varepsilon,\eta}=\left[\frac{1}{2C_2^k}\sum_{i<j}\left(a_i^{\varepsilon}a_j^{\eta}+a_i^{\eta}a_j^{\varepsilon}\right)\right]^{1/(\varepsilon+\eta)},$$

$$L_{1,\varepsilon,\eta}=\left[\frac{1}{6C_3^k}\sum\left(a_ia_j^{\varepsilon}a_t^{\eta}+a_ia_j^{\eta}a_t^{\varepsilon}+a_i^{\varepsilon}a_ja_t^{\eta}+a_i^{\eta}a_ja_t^{\varepsilon}+a_i^{\varepsilon}a_j^{\eta}a_t+a_i^{\eta}a_j^{\varepsilon}a_t\right)\right]^{1/(1+\varepsilon+\eta)}$$

THEOREM 3.27. *If* $0<\varepsilon,\eta$, *then* $L_{0,\varepsilon,\eta} \geqq L_{1,\varepsilon,\eta}$, *with strict inequality if* ε,η *are rational.*

The proof follows lines already adumbrated.

PROBLEM 3.30. If $0<\varepsilon<1$, do the relations $L_{0,\varepsilon}/L_{1,\varepsilon}<L_{1,\varepsilon}/L_{2,\varepsilon}<\cdots$ hold? (These inequalities are valid for $\varepsilon \to 1$.)

Acknowledgment. I acknowledge with thanks extensive numerical calculations that H. P. Robinson carried out at my request.

REFERENCES

[1] J. ACZEL, Private communication.

[2]. J. L. BRENNER, *A unified treatment and extension of some means of classical analysis*—I: *Comparison theorems*, J. Combin. Inform. System Sci., 3 (1972), pp. 175–199.

[3] J. L. BRENNER AND W. A. NEWCOMB, *Proposal number* E 3040, American Math. Monthly, 91 (1984), p. 203.

[4] G. H. HARDY, J. E. LITTLEWOOD AND G. PÓLYA, *Inequalities*, Cambridge Univ. Press, Cambridge, 1934.

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

WEIGHTED ZERO DISTRIBUTION FOR POLYNOMIALS ORTHOGONAL ON AN INFINITE INTERVAL*

WALTER VAN ASSCHE[†]

Abstract. The distribution of the zeros of orthogonal polynomials on an infinite interval is studied by means of a distribution function Z_n that makes a jump at each zero of the nth polynomial. The jumps are chosen properly in order that the function Z_n converges as $n \to \infty$. The asymptotic behaviour is given for the special case of (generalized) Laguerre and Hermite polynomials.

1. Introduction. The study of the distribution of the zeros of orthogonal polynomials has already led to a great variety of results. For orthogonal polynomials associated with a compact interval (say $[-1,1]$) it was found that, for a great class of polynomials, the zeros behave according to the arcsine-law ([3], [4], [9]). The study of the zeros of orthogonal polynomials on an infinite interval is somewhat more complicated, due to the fact that the zeros spread out over the entire interval and cannot be kept within a compact subset. This means that the sequence of distribution functions $\{J_n\}$ $(n=1,2,\cdots)$, where $J_n(t)$ makes a jump of size $1/n$ at each zero of the nth orthogonal polynomial, is not uniformly tight. A sequence of distribution functions $\{F_n\}$ $(n=1,2,\cdots)$ with $F_n(-\infty)=0$ and $F_n(\infty)=1$ is *uniformly tight* if for every $\varepsilon>0$ there exists a compact interval $[a,b]$ such that $F_n(b)-F_n(a)>1-\varepsilon$ for every $n>0$. It is well-known that a necessary and sufficient condition for a sequence of distribution functions to contain subsequences that converge weakly to a proper distribution function is that this sequence is uniformly tight ([1, Thms. 6.1 and 6.2]). The fact that the above defined sequence $\{J_n\}$ is not uniformly tight therefore means that there are no convergent subsequences. The problem can be solved as follows: there is a unique linear function which maps the first zero to -1 and the last zero to 1. Denote the images of the zeros by $y_{k,n}$, then

$$-1=y_{1,n}<y_{2,n}<\cdots<y_{n,n}=1.$$

Then we define $J_n(t)$ as the distribution function that makes a jump of size $1/n$ at each $y_{k,n}$ $(k=1,\cdots,n)$, and this function is called the *contracted zero distribution*. Clearly this sequence $\{J_n\}$ $(n=1,2,\cdots)$ is uniformly tight. For this sequence Ullman [10] found the limiting distribution for the orthogonal polynomials associated with the weight functions

$$w_{p,m}(x)=|x|^p\exp(-|x|^m), \qquad p>-1, \quad m=2,4,6.$$

Rakhmanov [7] obtained the limiting contracted zero distribution for polynomials orthogonal with a weight function $w(x)$ on $(-\infty,\infty)$ that has a particular behaviour at infinity

$$\lim_{|x|\to\infty}|x|^{-\lambda}\log w(x)=-r \qquad (r>0,\lambda>1).$$

*Received by the editors December 6, 1983, and in revised form August 6, 1984.

[†] Katholieke Universiteit Leuven, Departement Wiskunde, B-3030 Heverlee, Belgium. Aspirant van het Belgisch Nationaal Fonds voor Wetenschappelijk Onderzoek.

Similar results have been obtained by Mhaskar and Saff [6] for the weight function $w(x)=\exp(-|x|^{\lambda})$ with $\lambda>0$ (notice that these last authors therefore also have results for $0<\lambda\leqq 1$, which is a case not included in Rakhmanov's results).

In this paper we give another method to obtain a sequence of distribution functions that is uniformly tight. Again let $Z_n(t)$ be a distribution function that makes a jump at each zero of the nth orthogonal polynomial. Instead of making equal jumps of size $1/n$ we let the jumps be dependent of the zeros in such a way that large zeros have small weights. In §3 appropriate weights are given for polynomials orthogonal on $(-\infty,\infty)$ and $(0,\infty)$ with respect to a weight function with a particular behaviour at infinity and the limit distribution is found. In §4 the rate of convergence is given for the case of Laguerre and Hermite polynomials.

2. Definitions and preliminary results.

Let $w(x)$ be a weight function on an infinite interval I such that all the moments exist. Then there exists a unique sequence of orthogonal polynomials $\{p_n(x)\}$ $(n=0,1,2,\cdots)$ with

$$\int_I p_n(x)p_m(x)w(x)\,dx=\delta_{mn},$$

$$p_n(x)=k_n\prod_{j=1}^{n}(x-x_{j,n}),\qquad k_n>0.$$

The zeros of p_n are real, simple and belong to I:

$$x_{1,n}<x_{2,n}<\cdots<x_{n,n}.$$

We define the *weighted zero distribution* as

$$Z_n(x)=\sum_{j=1}^{n}\beta_{j,n}U(x-x_{j,n}),\tag{2.1}$$

where

$$U(x)=\begin{cases}0, & x<0,\\ 1, & x\geqq 0.\end{cases}$$

The proof of our results is based on the *Stieltjes transform*

$$S(F;z)=\int_{-\infty}^{+\infty}\frac{1}{z-t}\,dF(t),\qquad z\in\mathbb{C}\backslash\mathbb{R},\tag{2.2}$$

where F is a function of bounded variation. This transform is a very useful one, due to the following continuity theorem of *Grommer and Hamburger*:

THEOREM 2.1. *Suppose that $\{F_n(x)\}$ $(n=1,2,\cdots)$ is a sequence of functions of bounded variation such that their total variations are uniformly bounded.*

i) *If $F_n(x)$ converges weakly to $F(x)$, then $S(F_n;z)$ converges to $S(F;z)$ uniformly on every compact subset of $\mathbb{C}\backslash\mathbb{R}$.*

ii) *If $S(F_n;z)$ converges to a function $S(z)$ uniformly on every compact subset of $\mathbb{C}\backslash\mathbb{R}$, then $S(z)$ is the Stieltjes transform of a function $F(x)$ of bounded variation and $F_n(x)$ converges weakly to $F(x)$.*

A proof of this theorem can be found in [11, pp. 104–105]. *Weak convergence* of $F_n(x)$ to $F(x)$ means that for every bounded continuous function $f(x)$ the relation

$$\int_{-\infty}^{+\infty}f(x)\,dF_n(x)\to\int_{-\infty}^{+\infty}f(x)\,dF(x)\tag{2.3}$$

holds as $n\to\infty$. This is denoted by $F_n(x)\Rightarrow F(x)$. Theorem 2.1 means that the Stieltjes transform behaves, in a way, better than the Fourier transform since the limit of Stieltjes transforms is again a Stieltjes transform, which is in general not true for Fourier transforms (take for example $F_n(x)=U(x-n)$). Another consequence of this theorem is that we do not need to investigate the tightness of a sequence of distribution functions explicitly: once we have the limiting Stieltjes transform, we can see whether mass has flown to infinity or not. For the Stieltjes transform there is an explicit inversion formula, namely

$$\tag{2.4} \begin{aligned} &\tfrac{1}{2}\{F(x+)+F(x-)\}-\tfrac{1}{2}\{F(y+)+F(y-)\}\\ &\qquad=-\frac{1}{\pi}\lim_{v\to 0}\int_y^x \operatorname{Im} S(F;u+iv)\,du \end{aligned}$$

[11, pp. 93-96].

Recently Rakhmanov [7] proved some asymptotic formulas for polynomials orthogonal with respect to a weight function on an infinite interval that has some sort of regular behaviour at infinity. We recall the result in

THEOREM 2.2. *Let $w(x)$ be a weight function on $(-\infty,\infty)$ such that for some $\lambda>1$*

$$\lim_{|x|\to\infty}|x|^{-\lambda}\log w(x)=-1$$

and let $\{p_n(x)\}$ be the orthogonal polynomials belonging to this weight; then the following limit relation is valid uniformly with respect to z in any compact subset of $\mathbb{C}\setminus\mathbb{R}$

$$\tag{2.5} \lim_{n\to\infty}\frac{\log|p_n(z)|}{n^{1-1/\lambda}}=D(\lambda)|\operatorname{Im} z|,$$

where

$$D(\lambda)=\frac{\lambda}{\lambda-1}\left\{\frac{1}{\sqrt{\pi}}\frac{\Gamma((\lambda+1)/2)}{\Gamma(\lambda/2)}\right\}^{1/\lambda}$$

From this theorem one can easily obtain an asymptotic formula for orthogonal polynomials with a weight function on $(0,\infty)$ with the same kind of behaviour at infinity. For the classical Laguerre polynomials there èxists a stronger limit relation, known as *Perron's formula*:

THEOREM 2.3. *Let $\alpha>-1$; then*

$$\tag{2.6} \begin{aligned} L_n^{(\alpha)}(z)&=\frac{1}{2\sqrt{\pi}}e^{z/2}(-z)^{-(2\alpha+1)/4}n^{(2\alpha-1)/4}\exp\{2\sqrt{-nz}\}\\ &\qquad\cdot\left\{\sum_{j=0}^{p-1}C_j(\alpha;z)n^{-j/2}+O(n^{-p/2})\right\}, \end{aligned}$$

where the bound for the remainder holds uniformly on every compact subset of $\mathbb{C}\setminus[0,\infty)$; $(-z)^{-(2\alpha+1)/4}$ and $\sqrt{-z}$ must be taken real and positive if $z<0$.

This theorem can be found in Szegö's book [8, Thm. 8.22.3], and in this formula we have $C_0(\alpha;z)=1$. However, to prove the results in §5 we also need to know what $C_1(\alpha;z)$ is. In the appendix we use the method of steepest descend to obtain

$$\tag{2.7} C_1(\alpha;z)=\frac{1}{4\sqrt{-z}}\left\{-3z+\frac{1}{3}z^2+\frac{1}{4}-\alpha^2\right\}.$$

In this paper we will use two "roots": let $z=re^{i\theta}$; then

$$z^{1/2}=r^{1/2}e^{i\theta/2} \quad \text{if } \theta\in[0,2\pi), \tag{2.8}$$
$$\sqrt{z}=r^{1/2}e^{i\theta/2} \quad \text{if } \theta\in[-\pi,\pi).$$

Notice that one always has $\sqrt{-z}=-iz^{1/2}$.

3. Weighted zero distribution for weights on an infinite interval. In this section we will use the asymptotic formula of Rakhmanov (Theorem 2.2) to obtain the weighted zero distribution of polynomials orthogonal on either $(-\infty,\infty)$ or $(0,\infty)$. The contracted zero distribution for such polynomials has already been obtained by Rakhmanov himself [7] and special cases were found by Mhaskar and Saff [6] and Ullman [10].

THEOREM 3.1. *Let $w(x)$ be a weight function on $(-\infty,\infty)$ such that*

$$\lim_{|x|\to\infty}|x|^{-\lambda}\log w(x)=-1 \qquad (\lambda>1) \tag{3.1}$$

and let $\{p_n(x)\}$ be the orthonormal polynomials corresponding to this weight and $x_{1,n}<x_{2,n}<\cdots<x_{n,n}$ its zeros. Put

$$Z_n(x)=\frac{1}{D(\lambda)n^{1-1/\lambda}}\sum_{j=1}^{n}\frac{U(x-x_{j,n})}{1+x_{j,n}^2} \tag{3.2}$$

where

$$D(\lambda)=\frac{\lambda}{\lambda-1}\left\{\frac{1}{\sqrt{\pi}}\frac{\Gamma((\lambda+1)/2)}{\Gamma(\lambda/2)}\right\}^{1/\lambda},$$

then as $n\to\infty$

$$Z_n(x)\Rightarrow Z(x)=\frac{1}{\pi}\int_{-\infty}^{x}\frac{dt}{1+t^2}=\frac{1}{\pi}\arctan x+\frac{1}{2}, \tag{3.3}$$

so that for every bounded continuous function $f(x)$ on $(-\infty,\infty)$

$$\lim_{n\to\infty}\frac{1}{D(\lambda)n^{1-1/\lambda}}\sum_{j=1}^{n}\frac{f(x_{j,n})}{1+x_{j,n}^2}=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{f(t)}{1+t^2}\,dt.$$

Proof. Denote by ν_n a discrete measure defined by

$$\nu_n(\{x_{j,n}\})=\frac{1}{n^{1-1/\lambda}}, \qquad j=1,2,\cdots,n,$$
$$\nu_n(A)=0 \qquad \text{if } A \text{ contains no zeros of } p_n(x);$$

then the relation (2.5) is equivalent to

$$\int_{-\infty}^{\infty}\log|z-x|\,d\nu_n(x)\to D(\lambda)|\operatorname{Im}z| \tag{3.4}$$

uniformly on compact subsets of $\mathbb{C}\backslash\mathbb{R}$. Both sides of this asymptotic relation are harmonic functions both in $\{\operatorname{Im}z<0\}$ and $\{\operatorname{Im}z>0\}$. If we decompose z into $z=u+iv$ $(u,v\in\mathbb{R})$, then the partial derivatives $\partial/\partial u$ and $\partial/\partial v$ of the left-hand side of (3.4) will tend to the partial derivatives of the limit, uniformly on compact subsets of $\mathbb{C}\backslash\mathbb{R}$ [5, p.

249], so that

$$\int_{-\infty}^{\infty} \frac{u-x}{(u-x)^2+v^2}\, d\nu_n(x) \to 0,$$
$$\int_{-\infty}^{\infty} \frac{v}{(u-x)^2+v^2}\, d\nu_n(x) \to D(\lambda)\operatorname{sign} v,$$

uniformly on compact subsets of $\mathbb{C}\backslash\mathbb{R}$. Therefore we obtain that

$$\int_{-\infty}^{\infty} \frac{1}{z-x}\, d\nu_n(x) = \int_{-\infty}^{\infty} \frac{u-x-iv}{(u-x)^2+v^2}\, d\nu_n(x) \to -iD(\lambda)\operatorname{sign} v$$

and in particular we find

$$\int_{-\infty}^{\infty} \frac{1}{i\pm x}\, d\nu_n(x) \to -iD(\lambda), \tag{3.5}$$

from which

$$Z_n(\infty) = \frac{1}{D(\lambda)} \int_{-\infty}^{\infty} \frac{1}{1+x^2}\, d\nu_n(x) \to 1.$$

Let us now calculate the Stieltjes transform of Z_n:

$$S(Z_n;z) = \frac{1}{D(\lambda)} \int_{-\infty}^{\infty} \frac{1}{1+x^2} \frac{d\nu_n(x)}{z-x}$$
$$= \frac{1}{D(\lambda)} \frac{1}{1+z^2} \left\{ \int_{-\infty}^{\infty} \frac{d\nu_n(x)}{z-x} + z\int_{-\infty}^{\infty} \frac{d\nu_n(x)}{1+x^2} + \int_{-\infty}^{\infty} \frac{x}{1+x^2}\, d\nu_n(x) \right\}.$$

Because of (3.5) the last term will tend to zero as n increases, so that

$$\lim_{n\to\infty} S(Z_n;z) = \frac{1}{1+z^2}\{z - i\operatorname{sign} v\} = \begin{cases} \dfrac{1}{z+i} & \text{if } \operatorname{Im} z > 0, \\[2mm] \dfrac{1}{z-i} & \text{if } \operatorname{Im} z < 0. \end{cases}$$

This limit is the Stieltjes transform of the distribution function Z given in (3.3) (see the following lemma), and Theorem 2.1 then leads to the conclusion of this theorem. □

It remains to prove that the Stieltjes transform of Z is indeed the function we found above:

LEMMA 3.2. *Let* $Z(t) = 1/\pi \arctan t + \frac{1}{2}$; *then*

$$S(Z;z) = \begin{cases} \dfrac{1}{z+i} & \textit{if } \operatorname{Im} z > 0, \\[2mm] \dfrac{1}{z-i} & \textit{if } \operatorname{Im} z < 0. \end{cases}$$

Proof. We will use the inversion formula (2.4). Take $0<v<1$, then

$$\operatorname{Im}\frac{1}{u+iv+i}=-\frac{v+1}{u^2+(v+1)^2},$$

$$\left|\frac{v+1}{u^2+(v+1)^2}\right|\leqq\frac{2}{u^2+1},$$

so Lebesgue's theorem is applicable and

$$-\frac{1}{\pi}\lim_{v\downarrow 0}\int_{-\infty}^{x}\operatorname{Im}\frac{1}{u+iv+i}du=\frac{1}{\pi}\int_{-\infty}^{x}\frac{1}{1+u^2}du=Z(x).$$

The same reasoning holds for $-1<v<0$. □

Now we prove a similar result for orthogonal polynomials associated to a weight function on $(0,\infty)$.

THEOREM 3.3. *Let $w^+(x)$ be a weight function on $(0,\infty)$ such that*

$$\lim_{x\to\infty}x^{-\gamma}\log w^+(x)=-1\qquad(\gamma>\tfrac{1}{2})\tag{3.6}$$

and let $\{p_n(x)\}$ be the orthogonal polynomials corresponding to this weight and $x_{1,n}^+<x_{2,n}^+<\cdots<x_{n,n}^+$ its zeros. Put

$$Z_n^+(x)=\frac{2}{D(2\gamma)(2n)^{1-1/2\gamma}}\sum_{j=1}^{n}\frac{U(x-x_{j,n}^+)}{1+x_{j,n}^+};\tag{3.7}$$

then as $n\to\infty$

$$Z_n^+(x)\Rightarrow Z^+(x)=\frac{1}{\pi}\int_0^x\frac{dt}{\sqrt{t}(1+t)}\tag{3.8}$$

and as a consequence one finds that for every bounded continuous function $f(x)$ on $[0,\infty)$

$$\lim_{n\to\infty}\frac{2}{D(2\gamma)(2n)^{1-1/2\gamma}}\sum_{j=1}^{n}\frac{f(x_{j,n}^+)}{1+x_{j,n}^+}=\frac{1}{\pi}\int_0^\infty\frac{f(t)}{\sqrt{t}(1+t)}dt.$$

Proof. Consider the weight $w(x)=|x|w^+(x^2)$ on $(-\infty,\infty)$. Obviously this weight satisfies condition (3.1) of Theorem 3.1 with $\lambda=2\gamma$. The zeros $x_{j,n}$ corresponding to the weight $w(x)$ are related to the zeros $x_{j,n}^+$ by

$$x_{j,2n}=\begin{cases}-(x_{n+1-j,n}^+)^{1/2} & \text{if } j\leqq n,\\ (x_{j,n}^+)^{1/2} & \text{if } j>n,\end{cases}$$

so that

$$Z_n^+(x)=\begin{cases}0 & \text{if } x\leqq 0,\\ \dfrac{2}{D(2\gamma)(2n)^{1-1/2\gamma}}\displaystyle\sum_{j=n+1}^{2n}\frac{U(\sqrt{x}-x_{j,2n})}{1+x_{j,2n}^2} & \text{if } x>0.\end{cases}$$

For $x>0$ we thus have that $Z_n^+(x)=2Z_{2n}(\sqrt{x})-1$ and by Theorem 3.1

$$Z_n^+(x)\to 2Z(\sqrt{x})-1.$$

A simple change of variable then gives the result. □

4. Rate of convergence for Laguerre and Hermite polynomials. Let $w(x)=x^\alpha e^{-x}$ ($\alpha>-1; x\in[0,\infty)$); then the orthogonal polynomials associated with w are (up to a factor) the (generalized) *Laguerre polynomials* $\{L_n^{(\alpha)}(x)\}$ ($n=0,1,2,\cdots$). Obviously this weight satisfies condition (3.6) of Theorem 3.3 with $\gamma=1$, so that the weighted zero distribution (3.7) converges weakly to the distribution function $Z^+(x)$. The following theorem gives the rate of convergence for this asymptotic relation in terms of Stieltjes transforms:

THEOREM 4.1. *Let $Z^+(x)$ be as in* (3.8) *and $Z_n^+(x)$ be the weighted zero distribution of the Laguerre polynomial $L_n^{(\alpha)}(x)$, given by* (3.7) *with $\gamma=1$; then as $n\to\infty$*

$$S(\sqrt{n}\{Z_n^+-Z^+\};z)\to-\frac{2\alpha+1}{4}\frac{1}{z} \tag{4.1}$$

uniformly on compact subsets of $\mathbb{C}\backslash[0,\infty)$,

Proof. We will write $p_n(x)=L_n^{(\alpha)}(x)$. Start with the well-known identity

$$\frac{d}{dz}L_n^{(\alpha)}(z)=-L_{n-1}^{(\alpha+1)}(z).$$

With the use of Perron's formula (2.6) we have for $z\in\mathbb{C}\backslash[0,\infty)$

$$\frac{1}{\sqrt{n}}\frac{p_n'(z)}{p_n(z)}=-\frac{1}{\sqrt{-z}}\left(1-\frac{1}{n}\right)^{(2\alpha+1)/4}\exp\left\{2\sqrt{-nz}\left[\left(1-\frac{1}{n}\right)^{1/2}-1\right]\right\}$$
$$\cdot\left\{1+\frac{C_1(\alpha+1;z)}{\sqrt{n-1}}\right\}\left\{1-\frac{C_1(\alpha;z)}{\sqrt{n}}\right\}+o\left(\frac{1}{\sqrt{n}}\right).$$

Now

$$\left(1-\frac{1}{n}\right)^{(2\alpha+1)/4}=1-\frac{2\alpha+1}{4n}+o\left(\frac{1}{n}\right),$$
$$\exp\left\{2\sqrt{-nz}\left[\left(1-\frac{1}{n}\right)^{1/2}-1\right]\right\}=1-\frac{\sqrt{-z}}{\sqrt{n}}+o\left(\frac{1}{\sqrt{n}}\right),$$

so that we easily find that

$$\frac{1}{\sqrt{n}}\frac{p_n'(z)}{p_n(z)}=-\frac{1}{\sqrt{-z}}\left\{1+\frac{1}{\sqrt{n}}\left[C_1(\alpha+1;z)-C_1(\alpha;z)-\sqrt{-z}\right]\right\}+o\left(\frac{1}{\sqrt{n}}\right). \tag{4.2}$$

We will calculate the Stieltjes transform of Z_n:

$$S(Z_n^+;z)=\int_0^\infty\frac{dZ_n^+(t)}{z-t}$$
$$=\frac{1}{1+z}\int_0^\infty\frac{1+t+z-t}{z-t}\,dZ_n^+(t)$$
$$=\frac{1}{1+z}\left\{Z_n(\infty)+\int_0^\infty\frac{1+t}{z-t}\,dZ_n^+(t)\right\}.$$

But we also have

$$\frac{p_n'(z)}{p_n(z)}=\sum_{j=1}^{n}\frac{1}{z-x_{j,n}^+}=\sqrt{n}\int_0^\infty \frac{1+t}{z-t}\,dZ_n^+(t)$$

and if we put $z=-1$ in this expression

$$Z_n^+(\infty)=-\frac{1}{\sqrt{n}}\frac{p_n'(-1)}{p_n(-1)}.$$

The Stieltjes transform of Z^+ is given by

$$S(Z^+;z)=\frac{1}{1+z}-\frac{1}{\sqrt{-z}\,(1+z)},\qquad z\in\mathbb{C}\backslash[0,\infty)$$

(see Lemma 4.2) so that by combining these equations we obtain

$$S\left(\sqrt{n}\,[Z_n^+-Z^+];z\right)=\frac{\sqrt{n}}{1+z}\left\{\frac{1}{\sqrt{n}}\frac{p_n'(z)}{p_n(z)}+\frac{1}{\sqrt{-z}}-\frac{1}{\sqrt{n}}\frac{p_n'(-1)}{p_n(-1)}-1\right\}.$$

Now we use (4.2) to get

$$S\left(\sqrt{n}\,[Z_n^+-Z^+];z\right)=\frac{1}{1+z}\Big\{C_1(\alpha+1;\,-1)-C_1(\alpha;\,-1)$$

$$-\frac{1}{\sqrt{-z}}C_1(\alpha+1;z)+\frac{1}{\sqrt{-z}}C_1(\alpha;z)\Big\}+o(1).$$

Next we can use the expression of C_1 obtained in (2.7) to conclude that as $n\to\infty$

$$S\left(\sqrt{n}\,[Z_n^+-Z^+];z\right)\to\frac{1}{\sqrt{-z}\,(1+z)}\left\{\frac{2\alpha+1}{4\sqrt{-z}}+\sqrt{-z}-\sqrt{-z}\left(\frac{2\alpha+1}{4}+1\right)\right\}$$

$$=-\frac{2\alpha+1}{4}\frac{1}{z}.$$

□

LEMMA 4.2. *Let $Z^+(x)$ be as in* (3.8); *then*

$$S(Z^+;z)=\frac{1}{1+z}-\frac{1}{\sqrt{-z}\,(1+z)},\qquad z\in\mathbb{C}\backslash[0,\infty)$$

where the square root is as in (2.8).

Proof. By Lemma 3.2 we have that for $\operatorname{Im} w>0$

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{1}{1+x^2}\frac{dx}{w-x}=\frac{1}{w+i}$$

and this integral on the left can easily be rewritten as

$$\int_{-\infty}^{\infty}\frac{1}{1+x^2}\frac{dx}{w-x}=w\int_0^\infty\frac{1}{\sqrt{x}\,(1+x)}\frac{dx}{w^2-x}$$

so that

$$S(Z^+;w^2)=\frac{1}{w}\frac{1}{w+i}\qquad(\operatorname{Im} w>0).$$

Now let $z \in \mathbb{C}\backslash[0,\infty)$; then there exists a unique w with $\operatorname{Im} w > 0$ such that $w^2 = z$. This w is equal to $z^{1/2}$ and therefore

$$S(Z^+;z) = \frac{1}{z^{1/2}}\frac{1}{z^{1/2}+i} = \frac{1}{z+1} - \frac{i}{z^{1/2}(1+z)}.$$

The result now follows since $\sqrt{-z} = -iz^{1/2}$. □

Let $w(x) = |x|^{2\alpha}e^{-x^2}$ ($\alpha > -\frac{1}{2}$; $x \in \mathbb{R}$); then the orthogonal polynomials associated to w are (up to a factor) the (generalized) *Hermite polynomials* $\{H_n^{(\alpha)}(x)\}$ ($n = 0, 1, \cdots$). It is easy to prove the following relations to the Laguerre polynomials:

$$\begin{aligned} &H_{2m}^{(\alpha)}(x) = C_m L_m^{(\alpha-1/2)}(x^2),\\ (4.3)\qquad &H_{2m+1}^{(\alpha)}(x) = D_m x L_m^{(\alpha+1/2)}(x^2),\\ &\frac{d}{dx}H_{2m}^{(\alpha)}(x) = -2C_m x L_{m-1}^{(\alpha+1/2)}(x^2) \end{aligned}$$

where C_m and D_m are constants depending on m. It is clear that this weight function satisfies condition (3.1) of Theorem 3.1 (with $\lambda = 2$) so that the weighted zero distribution converges weakly to the Cauchy distribution $Z(x)$ given by (3.3). The rate of convergence is given by

THEOREM 4.3. *Let* $Z(x)$ *be as in* (3.3) *and* $Z_n(x)$ *be the weighted zero distribution of the Hermite polynomial* $H_n^{(\alpha)}(x)$, *given by* (3.2) (*with* $\lambda = 2$); *then as* $n \to \infty$

$$(4.4)\qquad S\left(\sqrt{2n}\{Z_n - Z\};z\right) \to -\frac{\alpha}{z}$$

uniformly on compact subsets of $\mathbb{C}\backslash\mathbb{R}$.

Proof. We will prove the theorem only for $n = 2m$; the proof is only slightly different for $n = 2m+1$. Write $p_{2m}(x) = H_{2m}^{(\alpha)}(x)$; then from (4.3) we get

$$(4.5)\qquad \frac{p'_{2m}(z)}{p_{2m}(z)} = \frac{-2zL_{m-1}^{(\alpha+1/2)}(z^2)}{L_m^{(\alpha-1/2)}(z^2)}.$$

We will first consider those z in $\mathbb{C}\backslash\mathbb{R}$ for which $\operatorname{Im} z > 0$, so that $0 < \arg z < \pi$. Now because of the definition (2.8) we then have $\sqrt{-z^2} = -iz$. Combining (4.5) with Perron's formula (2.6) yields

$$\begin{aligned} \frac{1}{2\sqrt{m}}\frac{p'_{2m}(z)}{p_{2m}(z)} &= -i\left(1-\frac{1}{m}\right)^{\alpha/2}\exp\left\{-2iz\sqrt{m}\left[\left(1-\frac{1}{m}\right)^{1/2}-1\right]\right\}\\ &\quad\cdot\left\{1+\frac{C_1\left(\alpha+\frac{1}{2};z^2\right)}{\sqrt{m-1}}\right\}\left\{1-\frac{C_1\left(\alpha-\frac{1}{2};z^2\right)}{\sqrt{m}}\right\}+o\left(\frac{1}{\sqrt{m}}\right). \end{aligned}$$

Since

$$\left(1-\frac{1}{m}\right)^{\alpha/4} = 1-\frac{\alpha}{2m}+o\left(\frac{1}{m}\right),$$

$$\exp\left\{-2iz\sqrt{m}\left[\left(1-\frac{1}{m}\right)^{1/2}-1\right]\right\} = 1+\frac{iz}{\sqrt{m}}+o\left(\frac{1}{\sqrt{m}}\right),$$

we obtain for $\operatorname{Im} z > 0$

(4.6)

$$\frac{1}{2\sqrt{m}}\frac{p_{2m}'(z)}{p_{2m}(z)} = -i\left\{1+\frac{1}{\sqrt{m}}\left[C_1\left(\alpha+\frac{1}{2};z^2\right)-C_1\left(\alpha-\frac{1}{2};z^2\right)+iz\right]\right\}+o\left(\frac{1}{\sqrt{m}}\right).$$

When $\operatorname{Im} z < 0$ then $\sqrt{-z^2} = iz$. Modifying the calculations for this z gives for $\operatorname{Im} z < 0$

$$\frac{1}{2\sqrt{m}}\frac{p_{2m}'(z)}{p_{2m}(z)} = i\left\{1+\frac{1}{\sqrt{m}}\left[C_1\left(\alpha+\frac{1}{2};z^2\right)-C_1\left(\alpha-\frac{1}{2};z^2\right)-iz\right]\right\}+o\left(\frac{1}{\sqrt{m}}\right). \tag{4.7}$$

Now we calculate the Stieltjes transform of Z_{2m}:

$$\begin{aligned} S(Z_{2m};z) &= \int_{-\infty}^{+\infty}\frac{dZ_{2m}(t)}{z-t} \\ &= \frac{1}{1+z^2}\int_{-\infty}^{+\infty}\frac{1+t^2+z^2-t^2}{z-t}\,dZ_{2m}(t) \\ &= \frac{1}{1+z^2}\left\{zZ_{2m}(\infty)+\int_{-\infty}^{+\infty}t\,dZ_{2m}(t)+\int_{-\infty}^{+\infty}\frac{1+t^2}{z-t}\,dZ_{2m}(t)\right\}. \end{aligned}$$

The second term on the right vanishes because the zeros of p_{2m} are symmetric with respect to the origin; therefore

$$S(Z_{2m};z) = \frac{1}{1+z^2}\left\{zZ_{2m}(\infty)+\int_{-\infty}^{+\infty}\frac{1+t^2}{z-t}\,dZ_{2m}(t)\right\}.$$

For the last integral we have

$$\int_{-\infty}^{\infty}\frac{1+t^2}{z-t}\,dZ_{2m}(t) = \frac{1}{2\sqrt{m}}\frac{p_{2m}'(z)}{p_{2m}(z)} = \frac{1}{2\sqrt{m}}\sum_{j=1}^{2m}\frac{1}{z-x_{j,n}}.$$

By considering this last expression at $\pm i$ we easily find

$$Z_{2m}(\infty) = \frac{1}{4i\sqrt{m}}\left\{\frac{p_{2m}'(-i)}{p_{2m}(-i)}-\frac{p_{2m}'(i)}{p_{2m}(i)}\right\} = \frac{1}{2\sqrt{m}}\sum_{j=1}^{2m}\frac{1}{1+x_{j,n}^2}.$$

If we combine all this, together with Lemma 3.2, then we find for $\operatorname{Im} z > 0$

$$\begin{aligned} S\left(2\sqrt{m}\left[Z_{2m}-Z\right];z\right) = 2\sqrt{m}\Bigg\{ & \frac{z}{1+z^2}\frac{1}{4i\sqrt{m}}\frac{p_{2m}'(-i)}{p_{2m}(-i)} \\ & -\frac{z}{1+z^2}\frac{1}{4i\sqrt{m}}\frac{p_{2m}'(i)}{p_{2m}(i)} \\ & +\frac{1}{1+z^2}\frac{1}{2\sqrt{m}}\frac{p_{2m}'(z)}{p_{2m}(z)}-\frac{1}{z+i}\Bigg\}. \end{aligned}$$

Now we use (4.6) and (4.7) to obtain

$$S(2\sqrt{m}[Z_{2m}-Z];z)=\frac{2}{1+z^2}\left\{z\left[C_1\left(\alpha+\frac{1}{2};-1\right)-C_1\left(\alpha-\frac{1}{2};-1\right)-1\right]\right.$$
$$\left.-i\left[C_1\left(\alpha+\frac{1}{2};z^2\right)-C_1\left(\alpha-\frac{1}{2};z^2\right)+iz\right]\right\}+o(1).$$

Again we take the explicit form (2.7) of C_1, and recalling that for $\operatorname{Im} z>0$ the equality $\sqrt{-z^2}=iz$ holds, we finally find that as $m\to\infty$

$$S(2\sqrt{m}[Z_{2m}-Z];z)\to\frac{2}{1+z^2}\left\{-z\left(\frac{\alpha}{2}+1\right)-i\left(\frac{\alpha}{2iz}+iz\right)\right\}=-\frac{\alpha}{z}$$

where $\operatorname{Im} z>0$. A similar reasoning holds when $\operatorname{Im} z<0$. □

We should be careful and not conclude from the result of Theorem 4.1 and Theorem 4.3 that the functions $R^+(x)=\sqrt{n}\{Z_n^+(x)-Z^+(x)\}$ and $R(x)=\sqrt{2n}\{Z_n(x)-Z(x)\}$ converge weakly to some function. This is not true. Although we can identify the limits in (4.1) and (4.4) as Stieltjes transforms of functions that make only one jump at the origin, the statement of the Grommer–Hamburger theorem is not applicable, since the functions $R_n^+(x)$ and $R_n(x)$ are not uniformly bounded in total variation. However, we can conclude a result that looks like (2.3), but which is weaker:

THEOREM 4.4. i) *Let $Z_n^+(x)$ and $Z^+(x)$ be as in the case of generalized Laguerre polynomials and put $R_n^+(x)=\sqrt{n}\{Z_n^+(x)-Z^+(x)\}$. Let $f(x)$ be such that $f((1+y)/(1-y))$ is analytic in some open set containing $[-1,1]$; then as $n\to\infty$*

$$\sum_{j=1}^{n}\frac{f(x_{j,n}^+)}{1+x_{j,n}^+}-\frac{\sqrt{n}}{\pi}\int_0^\infty\frac{f(x)}{\sqrt{x}(1+x)}dx=\int_0^\infty f(x)\,dR_n^+(x)\to-\frac{2\alpha+1}{4}f(0)-f(\infty).$$

ii) *Let $Z_n(x)$ and $Z(x)$ be as in the case of generalized Hermite polynomials and put $R_n(x)=\sqrt{2n}\{Z_n(x)-Z(x)\}$. Let $f(x)$ be such that $f(\pm((1+y)/(1-y))^{1/2})$ is analytic in some open set containing $[-1,1]$, where the root is as defined in (2.8); then as $n\to\infty$*

$$\sum_{j=1}^{n}\frac{f(x_{j,n})}{1+x_{j,n}^2}-\frac{\sqrt{2n}}{\pi}\int_{-\infty}^{\infty}\frac{f(x)}{1+x^2}dx=\int_{-\infty}^{\infty}f(x)\,dR_n(x)\to-\alpha f(0)-f(\infty)-f(-\infty).$$

Proof. i) Define the function $g(y)=f((1+y)/(1-y))$; then g will be analytic in some open set containing $[-1,1]$. An obvious substitution and Cauchy's theorem leads to

$$\int_0^\infty f(x)\,dR_n^+(x)=\int_{-1}^{1}g(y)\,dR_n^+\left(\frac{1+y}{1-y}\right)$$
$$=\frac{1}{2\pi i}\int_\Gamma g(z)\int_{-1}^{1}\frac{dR_n^+\left(\frac{1+y}{1-y}\right)}{z-y}dz,$$

where Γ is a closed graph encircling $[-1,1]$ in the open set mentioned. A new substitution yields

$$\int_0^\infty f(x) R_n^+(x) = \frac{1}{2\pi i}\int_\Gamma \frac{g(z)}{z-1}\int_0^\infty \frac{(1+x)\,dR_n^+(x)}{x-(1+z)/(1-z)}\,dz$$

$$= \frac{2}{2\pi i}\int_\Gamma \frac{g(z)}{(z-1)^2} S\left(R_n^+;\frac{1+z}{1-z}\right)dz + \frac{R_n^+(\infty)}{2\pi i}\int_\Gamma \frac{g(z)}{z-1}\,dz.$$

Now use (4.1) and let $n\to\infty$ (note that $S(R_n^+;(1+z)/(1-z))$ converges uniformly for $z\in\Gamma$):

$$\int_0^\infty f(x)\,dR_n^+(x) \to -\frac{2\alpha+1}{2}\frac{1}{2\pi i}\int_\Gamma \frac{g(z)}{1-z^2}\,dz + \lim_{n\to\infty} R_n^+(\infty)\frac{1}{2\pi i}\int_\Gamma \frac{g(z)}{z-1}\,dz.$$

An application of the theorem of residues and the fact that

$$\begin{aligned}
R_n^+(\infty) &= \sqrt{n}\,\{Z_n^+(\infty) - Z^+(\infty)\}\\
&= -\sqrt{n}\left\{\frac{1}{\sqrt{n}}\frac{p_n'(-1)}{p_n(-1)} + 1\right\}\\
&= C_1(\alpha+1;\,-1) - C_1(\alpha;\,-1) - 1 + o(1)\\
&\to -\frac{2\alpha+5}{4},
\end{aligned}$$

then leads to

$$\int_0^\infty f(x)\,dR_n^+(x) \to -\frac{2\alpha+1}{4}[g(-1)-g(1)] - \frac{2\alpha+5}{4}g(1).$$

Now $g(1)=f(\infty)$ and $g(-1)=f(0)$, which establishes case i.

ii) We give the proof only for $n=2m$, again the other case is similar. Suppose first that f is an even function: $f(x)=f(-x)$. Since $R_n(-x)=R_n(\infty)-R_n(x-)$ it follows that

$$\int_{-\infty}^{+\infty} f(x)\,dR_n(x) = 2\int_0^\infty f(x)\,dR_n(x).$$

Define $g(y)=f(((1+y)/(1-y))^{1/2})$; then as in the previous case

$$\int_{-\infty}^{+\infty} f(x)\,dR_n(x) = \frac{2}{2\pi i}\int_\Gamma g(z)\int_{-1}^{1}\frac{dR_n\left(\left(\frac{1+y}{1-y}\right)^{1/2}\right)}{z-y}\,dz.$$

This is easily seen to lead to

$$\int_{-\infty}^{\infty} f(x)\,dR_n(x) = 2\int_\Gamma \frac{g(z)}{z-1}\int_0^\infty \frac{(1+x^2)\,dR_n(x)}{x^2-\dfrac{1+z}{1-z}}\,dz$$

$$= \int_\Gamma \frac{g(z)}{z-1}\int_{-\infty}^{+\infty}\frac{(1+x^2)\,dR_n(x)}{x^2-\dfrac{1+z}{1-z}}\,dz;$$

the last equation follows from the symmetry of $R_n(x)$. Easy algebra gives

$$\int_{-\infty}^{+\infty} f(x)\,dR_n(x)=R_n(\infty)\int_\Gamma \frac{g(z)}{z-1}\,dz-2\int_\Gamma \frac{g(z)}{(z-1)^2}\int_{-\infty}^{+\infty}\frac{dR_n(x)}{x^2-(1+z)/(1-z)}\,dz.$$

Use

$$\frac{1}{x^2-(1+z)/(1-z)}=\frac{1}{2}\left(\frac{1-z}{1+z}\right)^{1/2}$$
$$\cdot\left\{\frac{1}{x-((1+z)/(1-z))^{1/2}}-\frac{1}{x+((1+z)/(1-z))^{1/2}}\right\}$$

with the root as in (2.8), to obtain

$$\int_{-\infty}^{\infty} f(x)\,dR_n(x)=R_n(\infty)\int_\Gamma \frac{g(z)}{z-1}\,dz+\int_\Gamma\left(\frac{1-z}{1+z}\right)^{1/2}\frac{g(z)}{(z-1)^2}S\left(R_n;\left(\frac{1+z}{1-z}\right)^{1/2}\right)dz$$
$$-\int_\Gamma\left(\frac{1-z}{1+z}\right)^{1/2}\frac{g(z)}{(z-1)^2}S\left(R_n;-\left(\frac{1+z}{1-z}\right)^{1/2}\right)dz.$$

Now let $n\to\infty$, where we notice that $S(R_n;\pm((1+z)/(1-z))^{1/2})$ converges uniformly for $z\in\Gamma$ to the limit given in (4.4):

$$\int_{-\infty}^{+\infty} f(x)\,dR_n(x)\to\lim_{n\to\infty}R_n(\infty)g(1)+2\alpha\int_\Gamma\frac{g(z)}{z^2-1}\,dz.$$

Now since $n=2m$

$$R_n(\infty)=2\sqrt{m}\,\{Z_{2m}(\infty)-Z(\infty)\}$$
$$=2\sqrt{m}\left\{\frac{1}{4i\sqrt{m}}\left[\frac{p'_{2m}(-i)}{p_{2m}(-i)}-\frac{p'_{2m}(i)}{p_{2m}(i)}\right]-1\right\}$$
$$=2\left\{C_1\left(\alpha+\frac{1}{2};-1\right)-C_1\left(\alpha-\frac{1}{2};-1\right)-1\right\}+o(1)$$
$$\to-(\alpha+2)$$

so that, together with the theorem of residues,

$$\int_{-\infty}^{+\infty} f(x)\,dR_n(x)\to-(\alpha+2)g(1)+\alpha[g(1)-g(-1)]$$
$$=-\alpha f(0)-2f(\infty),$$

and this is true whenever f is an even function. When f is odd, $f(-x)=-f(x)$; then it is easily seen that

$$\int_{-\infty}^{+\infty} f(x)\,dR_n(x)=0.$$

The general case follows by defining

$$f^+(x)=f(x)+f(-x),$$
$$f^-(x)=f(x)-f(-x),$$

so that $f^+(x)$ is even, $f^-(x)$ is odd and $f(x)=\frac{1}{2}\{f^+(x)+f^-(x)\}$. □

5. Concluding remarks. In §3 we proved that the weighted zero distribution, making a jump of size $\beta_{j,n}$ at the zero $x_{j,n}$ of the nth degree orthogonal polynomial, converges weakly for weight functions on $(-\infty,\infty)$ and $(0,\infty)$ having a particular behaviour at infinity. For both cases the limit was found to be independent of the weight function from which we started, so that we can say that the zeros have *invariant* or *weight-free* asymptotic behaviour. Such an invariant zero behaviour was already known for orthogonal polynomials on a finite interval.

In the second order asymptotic behaviour more information on the weight function $w(x)$ was found, more precisely the index α, which appears in the weight function for Laguerre and Hermite polynomials, can be found in investigating the second order zero behaviour. This has been shown in §4.

Now, if we compare in Theorem 3.1 the function $Z_n(x)$ with the Riemann sum of the limit $Z(x)$, then we can heuristically say that the sequences of measures $\{\nu_n\}$ $(n=1,2,\cdots)$ defined by

$$\begin{array}{ll} \nu_n(\{x_{j,n}\})=\dfrac{1}{n^{1-1/\lambda}} & \text{if } x_{j,n} \text{ is a zero of } p_n(x), \\ \nu_n(A)=0 & \text{if A contains no zeros of } p_n(x) \end{array}$$

behave in the limit like Lebesgue measure on $\mathbb{R}$. In the same way we can reason from Theorem 3.3 that the sequence of measures $\{\mu_n\}$ $(n=1,2,\cdots)$ defined by

$$\begin{array}{ll} \mu_n(\{x_{j,n}^+\})=\dfrac{1}{n^{1-1/2\gamma}} & \text{if } x_{j,n}^+ \text{ is a zero of } p_n(x), \\ \mu_n(A)=0 & \text{if A contain no zeros of } p_n(x) \end{array}$$

behaves in the limit like the measure

$$\mu(A)=\int_A \frac{dt}{\sqrt{t}}, \qquad A\subset[0,\infty).$$

These assertions cannot be proved properly because of the fact that these sequences of measures are not uniformly tight. Hence our approach of compactifying these measures using appropriate weights, which enabled us to prove convergence.

Appendix. In Szegö's book [8] only the first term in Perron's asymptotic expansion for the Laguerre polynomials is explicitly given. By using the same method of proof we will give the second term in Perron's formula (Theorem 2.3) which was needed in §4 to find the rate of convergence for the weighted zero distribution of Laguerre and Hermite polynomials.

LEMMA. *In Perron's formula* (*Theorem* 2.3) *we have*

$$C_1(\alpha;z)=\frac{1}{4\sqrt{-z}}\left\{-3z+\frac{1}{3}z^2+\frac{1}{4}-\alpha^2\right\}.$$

Proof. We need the following results: if $z\in\mathbb{C}\setminus[0,\infty)$ then

$$e^{-z}z^{\alpha/2}L_n^{(\alpha)}(z)=\frac{1}{n!}\int_0^\infty e^{-t}t^{n+\alpha/2}J_\alpha\left\{2(tz)^{1/2}\right\}dt$$

where J_α is the Bessel function with index α ([8, formula (5.4.1)]). For the Bessel function J_α the asymptotic relation

$$J_\alpha(z)=\left(\frac{2}{\pi}\right)^{1/2}\frac{1}{\sqrt{z}}\cos\left(z-\frac{\alpha\pi}{2}-\frac{\pi}{4}\right)\left\{\sum_{j=0}^{p-1}a_jz^{-2j}+O(z^{-2p})\right\}$$

$$+\left(\frac{2}{\pi}\right)^{1/2}\frac{1}{\sqrt{z}}\sin\left(z-\frac{\alpha\pi}{2}-\frac{\pi}{4}\right)\left\{\sum_{j=0}^{p-1}b_jz^{-2j-1}+O(z^{-2p-1})\right\}$$

holds for $z\in\mathbb{C}\setminus(-\infty,0]$ ([8, formula (1.71.8)]). Here

$$a_j=(-1)^j(\alpha,2j),$$
$$b_j=(-1)^{j+1}(\alpha,2j+1)$$

where (α,ν) is Hankel's symbol

$$(\alpha,\nu)=\frac{2^{-2\nu}}{\nu!}\left\{(4\alpha^2-1^2)\cdots\left[4\alpha^2-(2\nu-1)^2\right]\right\}$$
$$=\frac{\Gamma(\alpha+\nu+1/2)}{\nu!\Gamma(\alpha-\nu+1/2)}$$

([2, p. 85 and p. 23]).

Substituting the asymptotic expression in the integral representation for the Laguerre polynomials leads to

$$\text{(A.1)}\qquad e^{-z}z^{\alpha/2}L_n^{(\alpha)}(z)=\frac{1}{\sqrt{\pi}}\left\{\sum_{m=0}^{2p-1}A_m(z)+R_p(z)\right\},$$

where

$$A_{2k}(z)=a_kz^{-k-1/4}\frac{2^{-2k}}{n!}\int_0^\infty e^{-t}t^{n-k+(2\alpha-1)/4}\cos\left\{2(tz)^{1/2}-\frac{\alpha\pi}{2}-\frac{\pi}{4}\right\}dt,$$

$$A_{2k+1}(z)=b_kz^{-k-3/4}\frac{2^{-2k}}{2n!}\int_0^\infty e^{-t}t^{n-k+(2\alpha-3)/4}\sin\left\{2(tz)^{1/2}-\frac{\alpha\pi}{2}-\frac{\pi}{4}\right\}dt,$$

$$|R_p(z)|=O\left\{\frac{|z|^{p-1/4}}{n!}\int_0^\infty e^{-t}t^{n-p+(2\alpha-1)/4}\exp Re\left[2i(tz)^{1/2}\right]dt\right\}.$$

The $A_m(z)$ are integrals of the type

$$\frac{1}{r!}\int_0^\infty e^{-t}t^r\exp[t^{1/2}\xi]\,dt\qquad(\xi\neq0).$$

If we put $t=ry$, then this gives

$$\frac{r^{r+1}e^{-r}}{r!}\int_0^\infty(e^{1-y}y)^r\exp[r^{1/2}y^{1/2}\xi]\,dy.$$

Now $(e^{1-y}y)^r = \exp[r(\log y + 1 - y)]$, and the "essential" saddle point is the solution of $(\log y + 1 - y)' = 0$, so that $y = 1$. The critical directions are those $y \in \mathbb{C}$ such that $\gamma(y-1)^2 < 0$, where $\gamma = (\log y + 1 - y)''|_{y=1} = -1$, so that the real axis is the critical direction. Therefore we put $y = 1 + \rho/\sqrt{r}$, where $-r^\delta \leqq \rho \leqq r^\delta$ $(\delta < 1/6)$ and integrate over the positive real axis. Furthermore we have ($q_1(\rho)$ and $q_2(\rho)$ being polynomials in ρ)

$$(e^{1-y}y)^r = e^{-\rho^2/2}\left\{1 + \frac{\rho^3}{3\sqrt{r}} + \frac{1}{r}q_1(\rho) + \cdots\right\},$$

$$\exp[r^{1/2}y^{1/2}\xi] = \exp\left[\sqrt{r}\,\xi + \frac{\rho\xi}{2}\right]\left\{1 - \frac{\rho^2\xi}{8\sqrt{r}} + \frac{1}{r}q_2(p) + \cdots\right\}.$$

Using these considerations, we find that

$$\int_0^\infty (e^{1-y}y)^r \exp[r^{1/2}y^{1/2}\xi]\,dy$$

$$= \frac{1}{\sqrt{r}}\int_{-r^\delta}^{r^\delta} e^{-\rho^2/2}\left\{1 + \frac{\rho^3}{3\sqrt{r}}\right\} e^{\sqrt{r}\xi + \rho\xi/2}\left\{1 - \frac{\rho^2\xi}{8\sqrt{r}}\right\}\left\{\frac{1}{r}q(\rho) + \cdots\right\} d\rho$$

$$= \frac{1}{\sqrt{r}} e^{r^{1/2}\xi}\int_{-\infty}^{+\infty} e^{-\rho^2/2 + \rho\xi/2}\left\{1 + \frac{1}{\sqrt{r}}\left(\frac{\rho^3}{3} + \frac{\rho^2\xi}{8}\right)\right\}\left\{\frac{1}{r}q(\rho) + \cdots\right\} d\rho$$

where $q(\rho)$ is a polynomial in ρ. Putting

$$I_q = \int_{-\infty}^\infty \exp\left[-\frac{\rho^2}{2} + \frac{\rho\xi}{2}\right]\rho^q\,d\rho$$

then

$$I_0 = e^{\xi^2/8}\sqrt{2\pi},$$

$$I_1 = e^{\xi^2/8}\sqrt{2\pi}\,\frac{\xi}{2},$$

$$I_2 = e^{\xi^2/8}\sqrt{2\pi}\left(1 + \frac{\xi^2}{4}\right),$$

$$I_3 = e^{\xi^2/8}\sqrt{2\pi}\left(\frac{3\xi}{2} + \frac{\xi^3}{8}\right)$$

so that

$$\int_0^\infty (e^{1-y}y)^r \exp[r^{1/2}y^{1/2}\xi]\,dy$$

$$= \sqrt{\frac{2\pi}{r}}\exp\left[\sqrt{r}\,\xi + \frac{\xi^2}{8}\right]\left\{1 + \frac{1}{8\sqrt{r}}\left(3\xi + \frac{\xi^3}{12}\right) + O\left(\frac{1}{r}\right)\right\}.$$

Stirling's formula then leads to

$$\text{(A.2)}\qquad E(r;\xi) = \frac{1}{r!}\int_0^\infty e^{-t}t^r \exp[t^{1/2}\xi]\,dt$$

$$= \exp\left[\sqrt{r}\,\xi + \frac{\xi^2}{8}\right]\left\{1 + \frac{\xi}{8\sqrt{r}}\left(3 + \frac{\xi^2}{12}\right) + O\left(\frac{1}{r}\right)\right\}.$$

Now if we take $z \in \mathbb{C}\setminus[0,\infty)$ and if we recall that $\sqrt{-z} = -iz^{1/2}$, then

$$\cos\left\{2(tz)^{1/2}-\frac{\alpha\pi}{2}-\frac{\pi}{4}\right\}=\frac{1}{2}(-1)^{(2\alpha+1)/4}\left\{\exp(-2t^{1/2}\sqrt{-z})+\exp(2t^{1/2}\sqrt{-z})\right\},$$
$$\sin\left\{2(tz)^{1/2}-\frac{\alpha\pi}{2}-\frac{\pi}{4}\right\}=\frac{1}{2i}(-1)^{(2\alpha+1)/4}\left\{\exp(-2t^{1/2}\sqrt{-z})-\exp(2t^{1/2}\sqrt{-z})\right\};$$

furthermore we find that $\operatorname{Re}(\sqrt{-z})=\operatorname{Im} z^{1/2}>0$ so that $|\exp(-2r^{1/2}\sqrt{-z})|\to 0$ as $r\to\infty$. Calculating A_0 in (A.1), we obtain

$$A_0(z)=\frac{a_0}{2}z^{-1/4}(-1)^{(2\alpha+1)/4}\frac{(n+\alpha/2-1/4)!}{n!}$$
$$\cdot\left\{E\left(n+\frac{\alpha}{2}-\frac{1}{4};-2\sqrt{-z}\right)+E\left(n+\frac{\alpha}{2}-\frac{1}{4};2\sqrt{-z}\right)\right\}$$

where $r!$ is to be interpreted as $\Gamma(r+1)$. If we use (A.2), Stirling's formula and the considerations just taken, then

$$A_0(z)=\frac{1}{2}(-z)^{-1/4}(-1)^{\alpha/2}n^{(2\alpha-1)/4}\exp(2\sqrt{-nz})e^{-z/2}$$
$$\cdot\left\{1+\frac{\sqrt{-z}}{4}\left(3-\frac{z}{3}\right)n^{-1/2}+O\left(\frac{1}{n}\right)\right\}.$$

Similarly we find

$$A_1(z)=\frac{b_0}{4i}z^{-3/4}(-1)^{(2\alpha+1)/4}\frac{(n+\alpha/2-3/4)!}{n!}$$
$$\cdot\left\{E\left(n+\frac{\alpha}{2}-\frac{3}{4};-2\sqrt{-z}\right)-E\left(n+\frac{\alpha}{2}-\frac{3}{4};2\sqrt{-z}\right)\right\}$$
$$=-\frac{b_0}{4i}z^{-3/4}(-1)^{(2\alpha+1)/4}n^{(2\alpha-3)/4}\exp(2\sqrt{-nz})e^{-z/2}$$
$$\cdot\left\{1+2\sqrt{-z}\left(3-\frac{z}{3}\right)n^{-1/2}+O\left(\frac{1}{n}\right)\right\}.$$

Now $b_0=-(4\alpha^2-1)/4$ so that

$$A_1(z)=\frac{1-4\alpha^2}{16}(-z)^{-1/4}(\sqrt{-z})^{-1}(-1)^{\alpha/2}n^{(2\alpha-1)/4}\exp(2\sqrt{-nz})e^{-z/2}$$
$$\cdot\left\{n^{-1/2}+O\left(\frac{1}{n}\right)\right\}.$$

Also we have

$$|R_1|=O\left(\frac{n^{(2\alpha-1)/4}}{n}\right)$$

so that finally

$$L_n^{(\alpha)}(z)=\frac{1}{2\sqrt{\pi}}(-z)^{-(2\alpha+1)/4}n^{(2\alpha-1)/4}\exp(2\sqrt{-nz})e^{z/2}$$

$$\cdot\left\{1+\frac{1}{4\sqrt{-z}}\left(-3z+\frac{1}{3}z^2+\frac{1}{4}-\alpha^2\right)n^{-1/2}+O\left(\frac{1}{n}\right)\right\},$$

and if we compare this with (2.6) we get (2.7) and the lemma is proved. □

Acknowledgments. The author would like to thank Professor J. L. Teugels, who suggested the idea of using weights in order to obtain tightness. I am also very grateful to the referee for drawing my attention to Rakhmanov's results, which enabled me to write a revised version with more general results than in the original version.

REFERENCES

[1] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[2] A. Erdélyi and associates, *Higher Transcendental Functions*, Vol. II, McGraw-Hill, New York, 1953.

[3] P. Erdös and G. Freud; *On orthogonal polynomials with regularly distributed zeros*, Proc. London Math. Soc. (3), 29 (1974), pp. 521–537.

[4] P. Erdös and P. Turan, *On interpolation* III, Ann. Math., 41 (1940), pp. 510–555.

[5] O. D. Kellogg, *Foundations of Potential Theory*, Springer-Verlag, Berlin, 1967.

[6] H. N. Mhaskar and E. B. Saff, *Extremal problems for polynomials with exponential weights*, Trans. Amer. Math. Soc., 285 (1984), pp. 203–234.

[7] E. A. Rakhmanov, *On asymptotic properties of polynomials orthogonal on the real axis*, Mat. Sbornik 119 (161) (1982), pp. 163–203 (in Russian); English translation in Math. USSR Sb., 47 (1984), pp. 155–193.

[8] G. Szegö, *Orthogonal Polynomials*, 4th edition, AMS Colloquium Publications 23, American Mathematical Society, Providence, RI, 1975.

[9] J. L. Ullman, *On the regular behavior of orthogonal polynomials*, Proc. London Math. Soc. (3), 24 (1972), pp. 119–148.

[10] ______, *Orthogonal polynomials associated with an infinite interval*, Michigan Math. J., 27 (1980), pp. 353–363.

[11] A. Wintner, *Spektraltheorie der Unendlichen Matrizen*, Hirzel, Leipzig, 1929.

SIAM J. MATH. ANAL.
Vol. 16, No. 6, November 1985

015

A GENERALISED HANKEL CONVOLUTION*

J. DE SOUSA PINTO[†]

Abstract. The classical Hankel convolution defined by Hirschman and Cholewinski is extended to a class of generalised functions. Algebraic properties of the convolution are examined and the existence and significance of an identity element are discussed.

1. Introduction.

1.1. A Hankel-convolution operation has been defined in the classical sense by Hirschman, [4], and by Cholewinski, [5]. We consider here an extension of that definition to a class of generalised functions analogous to that introduced by Zemanian, [2]. This extension has useful applications when dealing with continuous linear systems which can be characterized by a Hankel-convolutional representation; such systems, which we call "Hankel-translation invariant continuous linear systems", may thereafter be considered when developing sampling expansions for inverse J_ν-Hankel transforms of distributions of compact support on the positive half-line.

1.2. Throughout we work in terms of the Hankel transform of order zero; straightforward generalization can be made as required to deal with the case of the transform of order ν, for any $\nu > -\frac{1}{2}$.

We use the following definition for the (classical) Hankel transform of order $\nu > -\frac{1}{2}$

$$(01) \qquad F(\tau) \equiv H_\nu[f](\tau) = \int_0^\infty xf(x)J_\nu(\tau x)\,dx, \qquad 0 < \tau < \infty,$$

$$(02) \qquad f(x) \equiv H_\nu^{-1}[F](x) = \int_0^\infty \tau F(\tau)J_\nu(\tau x)\,d\tau, \qquad 0 < x < \infty.$$

Since this differs from the definition used by Zemanian, we begin with a brief review and translation of some of the essential results obtained by him for the generalised Hankel transform. We remark that it is a classical result (cf. Sneddon, [1]) that if $x^{1/2}f(x)$ is piecewise continuous and belong to $L^1(0, \infty)$, then the direct transform is well defined by (01), and the inversion formula (02) holds almost everywhere. Further, for $f(x)$ and $g(x)$ both satisfying these conditions we have the Parseval relation

$$(03) \qquad \int_0^\infty xf(x)g(x)\,dx = \int_0^\infty \tau F(\tau)G(\tau)\,d\tau.$$

Finally we shall need results involving the linear differential operator, $N_{\nu+1}$, $\nu > -\frac{1}{2}$, defined by

$$(04) \qquad N_{\nu+1}[f(x)] \equiv x^\nu D[x^{-\nu}f(x)]$$

and the Bessel operator of order $\nu = 0$, Δ, defined by

$$(05) \qquad \Delta[f(x)] \equiv D^2 f(x) + x^{-1}Df(x)$$

where D stands here for the usual differentiation operator.

* Received by the editors April 26, 1983, and in revised form August 8, 1984.

[†] Departamento de Matematica, Universidade de Aveiro, 3800 Aveiro, Portugal.

If $x^{\nu+2}f(x)\to 0$ as $x\to\infty$, where $f(x)$ is a sufficiently smooth H_0-transformable function, then integration by parts shows that

$$H_{\nu+1}N_{\nu+1}[f](\tau)=-\tau H_\nu[f](\tau) \tag{06}$$

or, setting $g=H_\nu[f]$ and changing τ into x,

$$H_{\nu+1}[(-x)g(x)](\tau)=N_{\nu+1}H_\nu[g](\tau). \tag{07}$$

In general, for sufficiently well behaved $\phi(x)$ and nonnegative integers i,j we can obtain from (06) and (07)

$$H_{i+j}N_{i+j}\cdots N_{i+1}\big[(-x)^i\phi(x)\big](\tau)=(-\tau)^jN_i\cdots N_1[\Phi(\tau)] \tag{08}$$

or, taking the defining formula (04) into consideration,

$$H_{i+j}\big[x^{i+j}(x^{-1}D_x)^j\phi(x)\big](\tau)=(-1)^{i+j}\tau^{i+j}(\tau^{-1}D_\tau)^i[\Phi(\tau)]. \tag{09}$$

Similarly, for any sufficiently smooth function $f(x)$ on $(0,\infty)$ it can be shown that

$$H_0[\Delta f(x)](\tau)=-\tau^2H_0[f](\tau) \tag{10}$$

provided that f is H_0-transformable and that $x^2f(x)$ and $xf'(x)$ both tend to 0 as $x\to\infty$.

2. Spaces of fundamental and generalized functions.

2.1. A complex-valued function ϕ, defined and infinitely differentiable on $(0,\infty)$, is said to belong to the space $\mathbb{H}_0(0,\infty)$ if and only if the numbers $\gamma_{ij}(\phi)$ defined by

$$\gamma_{ij}(\phi)=\sup_{0<x<\infty}\left|x^i(x^{-1}D)^j\phi(x)\right| \tag{11}$$

are finite for every pair i,j of nonnegative integers.

$\mathbb{H}_0(0,\infty)$ is a testing-function space with the topology generated by the multinorm $(\gamma_{ij})_{i,\,j=0}^{\infty}$ and we have

$$\mathscr{D}(0,\infty)\subset\mathbb{H}_0(0,\infty)\subset\mathscr{E}(0,\infty) \tag{12}$$

where $\mathscr{D}(0,\infty)$ and $\mathscr{E}(0,\infty)$ denote respectively the restrictions of $\mathscr{D}(\mathbb{R})$ and $\mathscr{E}(\mathbb{R})$ to the positive real axis. Using (09) and following the same lines of Zemanian [2], it can readily be shown that the H_0-transformation is a topological isomorphism of $\mathbb{H}_0(0,\infty)$ onto $\mathbb{H}_0(0,\infty)$.

Denote by $\mathbb{M}(0,\infty)$ the linear space of all infinitely smooth functions $\theta(x)$, $0<x<\infty$ such that for each nonnegative integer m there exists a nonnegative integer $k=k(m)$ for which

$$(1+x^k)^{-1}(x^{-1}D)^m\theta(x) \tag{13}$$

is bounded on $(0,\infty)$. By using the generalised Leibnitz formula it can be shown that the map $\phi\to\theta\phi$ is an endomorphism of $\mathbb{H}_0(0,\infty)$ for each $\theta\in\mathbb{M}(0,\infty)$; $\mathbb{M}(0,\infty)$ is the space of multipliers on $\mathbb{H}_0(0,\infty)$.

2.2. We denote by $\mathbb{H}_0^*(0,\infty)$ the space of all complex-valued functions ψ, defined and infinitely smooth on $(0,\infty)$ which are of the form

$$\psi(x)=x\phi(x). \tag{14}$$

$\mathbb{H}_0^*(0,\infty)$ is again a (complete) testing-function space, with topology generated by the sequence of multinorms

$$\gamma_{ij}^*(\psi)=\gamma_{ij}(x^{-1}\psi(x)). \tag{15}$$

As usual we denote the dual of $\mathbb{H}_0^*(0,\infty)$ by $\mathbb{H}_0^{*\prime}(0,\infty)$.

For any $\psi(x)=x\phi(x)\in\mathbb{H}_0^*(0,\infty)$, and any nonnegative integer r, set

$$\alpha_r^*(\psi)=\max_{0\le i,j\le r}\gamma_{ij}^*(\psi)=\max_{0\le i,j\le\infty}\gamma_{ij}(\phi)=\alpha_r(\phi).$$

Then, for each $\mu\in\mathbb{H}_0^{*\prime}(0,\infty)$ there will exist constants C and r such that

$$\phi\in\mathbb{H}_0(0,\infty)\Rightarrow|\langle\mu(x),x\phi(x)\rangle|\le C\cdot\alpha_r(\phi). \tag{16}$$

In particular, let $f(x)$ be any locally integrable function on $(0,\infty)$ which is such that $xf(x)\in L^1(0,\infty)$ and $f(x)$ does not grow more rapidlly than a polynomial when $x\to\infty$. Then $f(x)$ generates a regular generalised function in $\mathbb{H}_0^{*\prime}(0,\infty)$ by the formula

$$\langle f(x),x\phi(x)\rangle=\int_0^\infty xf(x)\phi(x)\,dx. \tag{17}$$

Any generalised function in $\mathbb{H}_0^{*\prime}(0,\infty)$ not generated by a formula of the type (17) will be described as singular.

In general, the derivative of a generalised function in $\mathbb{H}_0^{*\prime}(0,\infty)$ (defined in the usual sense of Schwartz), is not a generalised function in $\mathbb{H}_0^{*\prime}(0,\infty)$. However, in certain cases the result of applying a differential operator to a generalised function in $\mathbb{H}_0^{*\prime}(0,\infty)$ does yield a generalised function in $\mathbb{H}_0^{*\prime}(0,\infty)$. In particular, using for differential operators in a generalised sense the same notation as the one used for the corresponding operators when applied in a classical sense, we have the results:

(i) $\mu\in\mathscr{E}'(0,\infty)\subset\mathbb{H}_0^{*\prime}(0,\infty)\Rightarrow D\mu\in\mathbb{H}_0^{*\prime}(0,\infty)$;
(ii) $\mu\in\mathbb{H}_0^{*\prime}(0,\infty)\Rightarrow(x^{-1}D)^j\mu\in\mathbb{H}_0^{*\prime}(0,\infty)$;
(iii) $\mu\in\mathbb{H}_0^{*\prime}(0,\infty)\Rightarrow\Delta^\nu\mu\in\mathbb{H}_0^{*\prime}(0,\infty)$;
for any nonnegative integers j and ν.

2.3. We can now define the generalised H_0-transform of any $\mu\in\mathbb{H}_0^{*\prime}(0,\infty)$ by using the analogue of the Parseval relation:

$$\langle\mu(x),x\phi(x)\rangle=\langle H_0[\mu](\tau),\tau\Phi(\tau)\rangle \tag{18}$$

and clearly we have that

$$\mu\in\mathbb{H}_0^{*\prime}(0,\infty)\Rightarrow H_0[\mu]\in\mathbb{H}_0^{*\prime}(0,\infty).$$

Moreover, we can establish that

$$H_0[\Delta^\nu\mu](\tau)=(-\tau^2)^\nu H_0[\mu] \tag{19}$$

for any nonnegative integer ν.

The generalised H_0-transform of any distribution $\sigma\in\mathscr{E}'(0,\infty)$, in the sense of (18), is a regular generalised function in $\mathbb{H}_0^{*\prime}(0,\infty)$ generated by a smooth function $f(x)$ defined on $(0,\infty)$ by

$$f(x)=\langle\sigma(\tau),\tau J_0(\tau x)\rangle=\langle\sigma(\tau),\tau\Lambda(\tau)J_0(\tau x)\rangle \tag{20}$$

where $\Lambda \in \mathscr{D}(0,\infty)$ is such that $\Lambda(\tau)=1$ on the support of σ. The function f extends onto the finite complex-plane as an entire function of exponential type which grows no faster than a polynomial on the positive real axis; it is easy to show that $f(x) \in \mathbb{M}(0,\infty)$.

3. H_0-convolution.

3.1. We denote by $\mathbb{L}_0^p(0,\infty)$, $1 \le p < \infty$ the space of measurable functions on $(0,\infty)$ such that

$$\|f\|_{0,p} = \left[\int_0^\infty x|f(x)|^p\,dx\right]^{1/p} < \infty$$

and by $\mathbb{L}_0^\infty(0,\infty)$ the space of measurable functions on $(0,\infty)$ such that

$$\|f\|_{0,\infty} = \|f\|_\infty = \sup_{0<x<\infty} |f(x)| < \infty.$$

Consider the kernel $\mathbb{D}_0(x,y,z)$, $0<x,y,z<\infty$, defined by

$$\mathbb{D}_0(x,y,z) = \int_0^\infty \xi J_0(\xi x)J_0(\xi y)J_0(\xi z)\,d\xi \tag{21}$$

for which (cf. Watson [3], Hirschman [4], and Cholewinski [5]) we can establish the following properties:

(i) for $0<x,y<\infty$ and $0 \le \tau < \infty$, we have

$$\int_0^\infty zJ_0(\tau z)\mathbb{D}_0(x,y,z)\,dz = J_0(\tau x)J_0(\tau y) \tag{22}$$

and, in particular, taking $\tau=0$, gives

(23) (ii) $$\int_0^\infty z\mathbb{D}_0(x,y,z)\,dz = 1$$

that is, for fixed $x,y>0$, $\mathbb{D}_0(x,y,z)$ as a function of z belongs to $\mathbb{L}^1{}_0(0,\infty)$;

(iii) for $0<x,y,z<\infty$, $\mathbb{D}_0(x,y,z) \ge 0$, and

(iv) $\mathbb{D}_0(x,y,z) = \mathbb{D}_0(z,x,y) = \mathbb{D}_0(y,z,x) =$ etc.

The classical H_0-convolution is now defined, for any two functions $f(x)$, $g(x)$, $0<x<\infty$, as

$$f\#_0 g(x) = \int_0^\infty \int_0^\infty yzf(y)g(z)\mathbb{D}_0(x,y,z)\,dy\,dz \tag{24}$$

whenever the integral exists. In fact, it can be shown that (Hirschman [4], Cholewinski [5]) if $1 \le p,q,r \le \infty$, $r^{-1}=p^{-1}+q^{-1}-1$ and $f \in \mathbb{L}_0^p(0,\infty)$ and $g \in \mathbb{L}_0^q(0,\infty)$, then the integral in (24) converges for almost all $x \in (0,\infty)$, and

$$\|f\#_0 g\|_{0,r} \le \|f\|_{0,p} \cdot \|g\|_{0,q}. \tag{25}$$

Furthermore,

$$f\#_0 g = g\#_0 f$$

and, for $f,g,h \in \mathbb{L}^1{}_0(0,\infty)$,

$$(f\#_0 g)\#_0 h = f\#_0(g\#_0 h)$$

while if f and g are such that both $H_0[f]$ and $H_0[g]$ exist, we have the convolution-product property

$$H_0[f\#_0 g]=H_0[f]\cdot H_0[g].$$

3.2. If the H_0-convolution $f\#_0 g$ exists, then using Fubini's theorem we can write it in the form

$$f\#_0 g=\int_0^\infty yf(y)\left[\int_0^\infty zg(z)\mathbb{D}_0(x,y,z)\,dz\right]dy=\int_0^\infty yf(y)g(x\circ y)\,dy \tag{26}$$

where we write

$$g(x\circ y)=\int_0^\infty zg(z)\mathbb{D}_0(x,y,z)\,dz \tag{27}$$

with $x\circ y$ denoting the H_0-translation on the positive real line (the analogue of the translation considered for the definition of the usual convolution, $*$). The function $g(x\circ y)$ will be called the H_0-translate of $g(x)$; provided $g(x)$ is locally bounded on $0<x<\infty$, $g(x\circ y)$ is well defined and continuous on $(0,\infty)\times(0,\infty)$, (Nussbaum [7]). The H_0-translation is a particular case of the translations of Delsarte [6], subsequently studied by Braaksma [8].

If $g\in\mathbb{L}_0^1(0,\infty)\cap\mathbb{L}^\infty(0,\infty)$ and $a\in[0,\infty)$, then a simple calculation using Fubini's theorem shows that

$$H_0[g(x\circ a)](\tau)=J_0(a\tau)H_0[g](\tau). \tag{28}$$

4. Generalised H_0-convolution.

4.1. For fixed $x,y\in(0,\infty)$, the function $\mathbb{D}_0(x,y,z)$, $0<z<\infty$ defines a regular generalised function in $\mathbb{H}_0^{*\prime}(0,\infty)$ which we denote by $\mathbb{D}_0(x\circ y,z)$. In fact for fixed $x,y\in(0,\infty)$ and $\phi\in\mathbb{H}_0(0,\infty)$ we have that

$$\begin{aligned}\langle\mathbb{D}_0(x\circ y,z),z\phi(z)\rangle&=\langle\mathbb{D}_0(x,y,z),z\phi(z)\rangle\\&=\int_0^\infty z\phi(z)\mathbb{D}_0(x,y,z)\,dz=\phi(x\circ y)\end{aligned} \tag{29}$$

and since

$$|\phi(x\circ y)|\le\gamma_{00}(\phi)\int_0^\infty z\mathbb{D}_0(x,y,z)\,dz=\gamma_{00}(\phi), \tag{30}$$

then $\mathbb{D}_0(x\circ y,z)$, $0<z<\infty$ truly generates a continuous linear functional on $\mathbb{H}_0^*(0,\infty)$ through (29).

Moreover, since

$$\phi(x\circ y)=\langle J_0(\tau(x\circ y)),\tau\Phi(\tau)\rangle=\langle J_0(\tau x)J_0(\tau y),\tau\Phi(\tau)\rangle,$$

then we can write

$$H_0[\mathbb{D}_0(x\circ y,z)]=J_0(\tau x)J_0(\tau y),\qquad 0<x,y<\infty, \tag{31}$$

in the sense of $\mathbb{H}_0^{*\prime}(0,\infty)$ and even in the classical sense.

We now show that, for any fixed $y>0$, the following implication

$$\phi(x)\in\mathbb{H}_0(0,\infty)\Rightarrow\phi(x\circ y)\in\mathbb{H}_0(0,\infty) \tag{32}$$

holds. In fact, since $\phi \in \mathbb{H}_0(0,\infty)$ then $\Phi = H_0[\phi] \in \mathbb{H}_0(0,\infty)$. On the other hand,

$$H_0[\phi(x\circ y)] = J_0(\tau y)\Phi(\tau);$$

but $J_0(\tau y) \in \mathbb{M}_\tau(0,\infty)$ and so

$$J_0(\tau y)\Phi(\tau) \in \mathbb{H}_0(0,\infty).$$

Hence, since the H_0-transformation is an automorphism on $\mathbb{H}_0(0,\infty)$, the function of x given by

$$H_0^{-1}[J_0(\tau y)\Phi(\tau)] = \phi(x\circ y)$$

also belongs to $\mathbb{H}_0(0,\infty)$.

Next, for any $\phi \in \mathbb{H}_0(0,\infty)$ and $0 < x,y < \infty$, following from a well-known property of the Delsarte translation, we also have that

$$\Delta_x^\nu \phi(x\circ y) = \Delta_y^\nu \phi(x\circ y) \tag{33}$$

for any nonnegative integer ν.

4.2. If $\phi_1, \phi_2 \in \mathbb{H}_0(0,\infty)$ then

$$\begin{aligned}&\text{(a) } \phi_1 \#_0 \phi_2(x) \text{ exists for all } 0<x<\infty;\\ &\text{(b) } \phi_1 \#_0 \phi_2(x) \in \mathbb{H}_0(0,\infty);\\ &\text{(c) } \Delta^\nu[\phi_1 \#_0 \phi_2] = [\Delta^\nu\phi_1]\#_0\phi_2 = \phi_1 \#_0[\Delta^\nu\phi_2].\end{aligned} \tag{34}$$

In fact (34a) follows since $\phi_1, \phi_2 \in \mathbb{L}_0^p(0,\infty)$ for any p such that $1 \le p \le \infty$; (34b) is justified by the fact that the function

$$H_0[\phi_1 \#_0 \phi_2] = \Phi_1 \cdot \Phi_2$$

belongs to $\mathbb{H}_0(0,\infty)$ and similarly for its H_0^{-1}-transform; (34c) follows from (33) and differentiation under integral sign.

Note finally that for any $\phi_1, \phi_2 \in \mathbb{H}_0(0,\infty)$

$$\begin{aligned}\phi_1 \#_0 \phi_2(x) &= \langle \phi_1(y), y\phi_2(x\circ y)\rangle\\ &= \int_0^\infty y\phi_1(y)\phi_2(x\circ y)\,dy\\ &= \int_0^\infty y\phi_1(y)\int_0^\infty z\phi_2(z)\mathbb{D}_0(x\circ y,z)\,dz\,dy\\ &= \int_0^\infty z\phi_2(z)\int_0^\infty y\phi_1(y)\mathbb{D}_0(x\circ z,y)\,dy\,dz\\ &= \int_0^\infty z\phi_2(z)\phi_1(x\circ z)\,dz\\ &= \langle \phi_2(z), z\phi_1(x\circ z)\rangle = \phi_2 \#_0 \phi_1(x)\end{aligned}$$

where the interchange of the order of integrations is justified by Fubini's theorem in view of the fact that both ϕ_1 and ϕ_2 belong to $L_0^1(0,\infty)$.

4.3. If $\lambda \in \mathbb{H}_0(0,\infty)$, then for each fixed $x \in (0,\infty)$, we have $\lambda(x \circ y) \in \mathbb{H}_0(0,\infty)$; it follows that for any $\sigma \in \mathscr{E}'(0,\infty)$ the convolution $\sigma \#_0 \lambda(x)$ is well defined by

$$\sigma \#_0 \lambda(x) = \langle \sigma(y), y\lambda(x \circ y)\rangle. \tag{35}$$

Further,

$$\begin{aligned} H_0[\sigma \#_0 \lambda(x)](\tau) &= \langle \sigma(y), yH_{0x}[\lambda(x \circ y)]\rangle \\ &= \langle \sigma(y), yJ_0(\tau y)\Lambda(\tau)\rangle \\ &= \langle \sigma(y), yJ_0(\tau y)\rangle \cdot \Lambda(\tau) = H_0[\sigma](\tau) \cdot \Lambda(\tau) \end{aligned} \tag{36}$$

where $\Lambda = H_0[\lambda]$. Now $H_0[\sigma \#_0 \lambda] \in \mathbb{H}_0(0,\infty)$, and therefore $\sigma \#_0 \lambda(x) \in \mathbb{H}_0(0,\infty)$. Hence $\sigma \#_0 \lambda$ generates a regular generalised function in $\mathbb{H}_0^{*\prime}(0,\infty)$, and for any $\phi \in \mathbb{H}_0(0,\infty)$ we get

$$\begin{aligned} \langle \sigma \#_0 \lambda(x), x\phi(x)\rangle &= \langle H_0[\sigma](\tau) \cdot \Lambda(\tau), \tau\Phi(\tau)\rangle \\ &= \langle H_0[\sigma](\tau), \tau\Lambda(\tau)\Phi(\tau)\rangle = \langle \sigma(x), x\lambda \#_0 \phi(x)\rangle. \end{aligned} \tag{37}$$

This could be taken as the definition of the generalised H_0-convolution, and this in turn allows another form analogous to the direct product definition of the generalised ordinary convolution:

$$\begin{aligned} \langle \sigma \#_0 \lambda(x), x\phi(x)\rangle &= \langle \sigma(x), x\lambda \#_0 \phi(x)\rangle \\ &= \langle \sigma(x), x\langle \lambda(y), y\phi(x \circ y)\rangle\rangle \\ &= \langle \sigma(x) \otimes \lambda(y), xy\phi(x \circ y)\rangle. \end{aligned} \tag{38}$$

4.4. For $\mu \in \mathbb{H}_0^{*\prime}(0,\infty)$ and $\lambda \in \mathbb{H}_0(0,\infty)$ the convolution is again well defined as a generalised function in $\mathbb{H}_0^{*\prime}(0,\infty)$ by

$$\langle \mu \#_0 \lambda(x), x\phi(x)\rangle = \langle \mu(x), x\lambda \#_0 \phi(x)\rangle$$

since $\lambda \#_0 \phi \in \mathbb{H}_0(0,\infty)$ by (34b). Using (18), we get

$$\begin{aligned} \langle H_0[\mu \#_0 \lambda](\tau), \tau\Phi(\tau)\rangle &= \langle \mu \#_0 \lambda(x), x\phi(x)\rangle \\ &= \langle \mu(x), x\lambda \#_0 \phi(x)\rangle \\ &= \langle H_0[\mu](\tau), \tau H_0[\lambda \#_0 \phi](\tau)\rangle \\ &= \langle H_0[\mu](\tau), \tau\Lambda(\tau)\Phi(\tau)\rangle \\ &= \langle H_0[\mu](\tau)\Lambda(\tau), \tau\Phi(\tau)\rangle \end{aligned}$$

so that, in the sense of $\mathbb{H}_0^{*\prime}(0,\infty)$

$$H_0[\mu \#_0 \lambda] = H_0[\mu] \cdot H_0[\lambda]. \tag{39}$$

4.5. Finally, let $\mu \in \mathbb{H}_0^{*\prime}(0,\infty)$ and $\sigma \in \mathscr{E}'(0,\infty)$. Since, for any $\phi \in \mathbb{H}_0(0,\infty)$, we have $\sigma \#_0 \phi(x) \in \mathbb{H}_0(0,\infty)$, it follows that $\mu \#_0 \sigma$ is well defined as a generalised function in $\mathbb{H}_0^{*\prime}(0,\infty)$ by

$$\langle \mu \#_0 \sigma(x), x\phi(x)\rangle = \langle \mu(x), x\sigma \#_0 \phi(x)\rangle. \tag{40}$$

As before, this may also be expressed in the form

$$\langle \mu\#_0\sigma(x), x\phi(x)\rangle = \langle \mu(x)\otimes\sigma(y), xy\phi(x\circ y)\rangle \tag{41}$$

and, using (18) again, we can derive the analogue of (39)

$$H_0[\mu\#_0\sigma] = H_0[\mu]\cdot H_0[\sigma] \tag{42}$$

(note that $H_0[\sigma]\in \mathbb{M}(0,\infty)$, so that the product in (42) makes sense in $\mathbb{H}_0^{*\prime}(0,\infty)$).

5. Algebraic properties of the generalised H_0-convolution. As already remarked, the classical H_0-convolution defined in $\mathbb{L}^1_0(0,\infty)$ is commutative and associative; however, it possesses no identity element. We consider in turn these properties with respect to generalised H_0-convolution.

5.1. Commutativity.

(i): $\sigma\in\mathscr{E}'(0,\infty)$, $\lambda\in\mathbb{H}_0(0,\infty)$.

We have

$$\begin{aligned}\langle \sigma\#_0\lambda(x), x\phi(x)\rangle &= \langle \sigma(x), x\lambda\#_0\phi(x)\rangle\\ &= \langle H_0(\sigma)(\tau)\Lambda(\tau), \tau\Phi(\tau)\rangle\\ &= \langle \Lambda(\tau), \tau H_0[\sigma](\tau)\Phi(\tau)\rangle\\ &= \langle \lambda(x), x\sigma\#_0\phi(x)\rangle = \langle \lambda\#_0\sigma(x), x\phi(x)\rangle\end{aligned}$$

where the last manipulations make sense since $H_0[\sigma]\in\mathbb{M}(0,\infty)$ and $\sigma\#_0\phi\in\mathbb{H}_0(0,\infty)$.

(ii): $\mu\in\mathbb{H}_0^{*\prime}(0,\infty)$, $\lambda\in\mathbb{H}_0(0,\infty)$.

Then

$$\begin{aligned}\langle \mu\#_0\lambda(x), x\phi(x)\rangle &= \langle \mu(x), x\lambda\#_0\phi(x)\rangle\\ &= \langle H_0[\mu](\tau), \tau\Lambda(\tau)\Phi(\tau)\rangle\\ &= \langle H_0[\mu](\tau)\cdot\Phi(\tau), \tau\Lambda(\tau)\rangle\\ &= \langle \mu\#_0\phi(x), x\lambda(x)\rangle = \langle \lambda\#_0\mu(x), x\phi(x)\rangle\end{aligned}$$

where $\lambda\#_0\mu$ is understood in the sense of the last equality. This is justified because every function in $\mathbb{H}_0(0,\infty)$ is also a multiplier in $\mathbb{H}_0^{*\prime}(0,\infty)$.

(iii): $\mu\in\mathbb{H}_0^{*\prime}(0,\infty)$, $\sigma\in\mathscr{E}'(0,\infty)$.

The same kind of argument gives:

$$\begin{aligned}\langle \mu\#_0\sigma(x), x\phi(x)\rangle &= \langle \mu(x), x\sigma\#_0\phi(x)\rangle\\ &= \langle H_0[\mu](\tau)\cdot H_0[\sigma](\tau), \tau\Phi(\tau)\rangle.\end{aligned}$$

But since $H_0[\mu](\tau)$ does not necessarily belong to $\mathbb{M}(0,\infty)$, no general commutativity property can be deduced. If, in addition, we have $\mu\in\mathscr{E}'(0,\infty)$, then $H_0[\mu]\in\mathbb{M}(0,\infty)$, and the argument to establish commutativity proceeds as before.

5.2. Associativity.

(i): $\sigma\in\mathscr{E}'(0,\infty)$, $\lambda_1, \lambda_2\in\mathbb{H}_0(0,\infty)$.

We can establish the result

$$(\sigma\#_0\lambda_1)\#_0\lambda_2 = \sigma\#_0(\lambda_1\#_0\lambda_2) \tag{43}$$

in the following sense, for any $\phi\in\mathbb{H}_0(0,\infty)$

$$\begin{aligned}\langle(\sigma\#_0\lambda_1)\#_0\lambda_2(x),x\phi(x)\rangle&=\langle\sigma\#_0\lambda_1(x),x\lambda_2\#_0\phi(x)\rangle\\&=\langle\sigma(x),x\lambda_1\#_0(\lambda_2\#_0\phi(x))\rangle\\&=\langle\sigma(x),x(\lambda_1\#_0\lambda_2)\#_0\phi(x)\rangle\\&=\langle\sigma\#_0(\lambda_1\#_0\lambda_2(x)),x\phi(x)\rangle.\end{aligned}$$

The equality $\lambda_1\#_0(\lambda_2\#_0\phi)=(\lambda_1\#_0\lambda_2)\#_0\phi$ is justified by the fact that λ_1, λ_2 and ϕ belong to $\mathbb{L}^1_0(0,\infty)$.

(ii): $\mu\in\mathbb{H}_0^{*\prime}(0,\infty)$, $\sigma\in\mathscr{E}'(0,\infty)$, $\lambda\in\mathbb{H}_0(0,\infty)$.

We have that

$$(\mu\#_0\sigma)\#_0\lambda=\mu\#_0(\sigma\#_0\lambda)\tag{44}$$

since, for any $\phi\in\mathbb{H}_0(0,\infty)$

$$\begin{aligned}\langle(\mu\#_0\sigma)\#_0\lambda(x),x\phi(x)\rangle&=\langle\mu\#_0\sigma(x),x\lambda\#_0\phi(x)\rangle\\&=\langle\mu(x),x\sigma\#_0(\lambda\#_0\phi(x))\rangle\\&=\langle\mu(x),x(\sigma\#_0\lambda)\#_0\phi(x)\rangle\\&=\langle\mu\#_0(\sigma\#_0\lambda(x)),x\phi(x)\rangle\end{aligned}$$

where the equality $\sigma\#_0(\lambda\#_0\phi)=(\sigma\#_0\lambda)\#_0\phi$ is justified by (43).

(iii): $\mu\in\mathbb{H}_0^{*\prime}(0,\infty)$, σ_1, $\sigma_2\in\mathscr{E}'(0,\infty)$.

We show, finally, that

$$(\mu\#_0\sigma_1)\#_0\sigma_2=\mu\#_0(\sigma_1\#_0\sigma_2)\tag{45}$$

since we have

$$\begin{aligned}\langle(\mu\#_0\sigma_1)\#_0\sigma_2(x),x\phi(x)\rangle&=\langle\mu\#_0\sigma_1(x),x\sigma_2\#_0\phi(x)\rangle\\&=\langle\mu(x),x\sigma_1\#_0(\sigma_2\#_0\phi(x))\rangle\\&=\langle\mu(x),x(\sigma_1\#_0\sigma_2)\#_0\phi(x)\rangle\\&=\langle\mu\#_0(\sigma_1\#_0\sigma_2(x)),x\phi(x)\rangle\end{aligned}$$

where the equality $\sigma_1\#_0(\sigma_2\#_0\phi)=(\sigma_1\#_0\sigma_2)\#_0\phi$ is justified by (44) in the particular case when both generalised functions belong to $\mathscr{E}'(0,\infty)$.

5.3. Identity element. For a, b strictly positive we know that $\mathbb{D}_0(a,b,z)$ defines a regular generalised function $\mathbb{D}_0(a\circ b,z)$ in $\mathbb{H}_0^{*\prime}(0,\infty)$. If either of a, b takes the value zero then $\mathbb{D}_0(a,b,z)$ is no longer defined as an ordinary function since

$$\mathbb{D}_0(a,0,z)=\int_0^\infty \xi J_0(a\xi)J_0(\xi z)\,d\xi,\qquad a>0,$$

is only a formal identity because the integral fails to converge for any z. Instead, for any fixed $a>0$, we consider the integral

$$\int_0^R \xi J_0(a\xi)J_0(\xi z)\,d\xi\tag{46}$$

which, for each $R>0$ is uniformly convergent on $0<z<\infty$.

Define the generalised function $\mathbb{D}_0(a,z)$ in $\mathbb{H}_0^{*\prime}(0,\infty)$ by

$$\mathbb{D}_0(a,z)=\lim_{R\to\infty}\int_0^R \xi J_0(a\xi)J_0(\xi z)\,d\xi$$

in the sense that for any $\phi\in\mathbb{H}_0(0,\infty)$,

$$\langle\mathbb{D}_0(a,z),z\phi(z)\rangle=\lim_{R\to\infty}\left\langle\int_0^R \xi J_0(a\xi)J_0(\xi z)\,d\xi, z\phi(z)\right\rangle. \tag{47}$$

For each finite $R>0$ the integral (44) defines a function which generates a regular generalised function in $\mathbb{H}_0^{*\prime}(0,\infty)$ (cf. Sneddon, [1, p. 303]). Therefore,

$$\left\langle\int_0^R \xi J_0(a\xi)J_0(\xi z)\,d\xi, z\phi(z)\right\rangle=\int_0^\infty z\phi(z)\int_0^R \xi J_0(a\xi)J_0(\xi z)\,d\xi\,dz$$

or, by Fubini's theorem

$$\left\langle\int_0^R \xi J_0(a\xi)J_0(\xi z)\,d\xi, z\phi(z)\right\rangle=\int_0^R \xi J_0(a\xi)\int_0^\infty z\phi(z)J_0(\xi z)\,dz\,d\xi$$
$$=\int_0^R \xi\Phi(\xi)J_0(a\xi)\,d\xi.$$

Thus

$$\langle\mathbb{D}_0(a,z),z\phi(z)\rangle=\lim_{R\to\infty}\int_0^R \xi\Phi(\xi)J_0(a\xi)\,d\xi=\phi(a) \tag{48}$$

and so

$$|\langle\mathbb{D}_0(a,z),z\phi(z)\rangle|\leq\gamma_\infty(\phi)$$

which shows that $D_0(a,z)\in\mathbb{H}_0^{*\prime}(0,\infty)$. Moreover, since

$$\langle\mathbb{D}_0(a,z),z\phi(z)\rangle=\phi(a)=\langle J_0(a\tau),\tau\Phi(\tau)\rangle,$$

we obtain

$$H_0[\mathbb{D}_0(a,z)](\tau)=J_0(\tau). \tag{49}$$

Now let $(a_n)_{n=1}^\infty$ be a monotone decreasing sequence of positive real numbers, tending to zero as $n\to\infty$, and consider the sequence of generalised functions $(\mathbb{D}_0(a_n,z))_{n=1}^\infty$ in $\mathbb{H}_0^{*\prime}(0,\infty)$. Since $\mathbb{H}_0^{*\prime}(0,\infty)$ is complete, this limit is again a generalised function in $\mathbb{H}_0^{*\prime}(0,\infty)$. For each n and any $\phi\in\mathbb{H}_0(0,\infty)$

$$\langle\mathbb{D}_0(a_n,z),z\phi(z)\rangle=\phi(a_n)$$

and therefore we define the generalised function $\mathbb{D}_0(z)$ by:

$$\langle\mathbb{D}_0(z),z\phi(z)\rangle=\lim_{n\to\infty}\langle\mathbb{D}_0(a_n,z),z\phi(z)\rangle=\lim_{n\to\infty}\phi(a_n)=\phi(0^+) \tag{50}$$

(independently of the particular sequence $(a_n)_{n=1}^\infty$ chosen).

Moreover, since

$$\langle\mathbb{D}_0(z),z\phi(z)\rangle=\phi(0^+)=\langle 1,\tau\Phi(\tau)\rangle$$

we have

$$H_0[D_0(z)](\tau)=1 \tag{51}$$

the equality being understood in the sense of $\mathbb{H}_0^{*\prime}(0,\infty)$.

The generalised function $\mathbb{D}_0(x)\in\mathbb{H}_0^{*\prime}(0,\infty)$ is the required identity element with respect to the generalised H_0-convolution. In fact, it is easy to show that $\mathbb{D}_0(x)\in\mathscr{E}'(0,\infty)$ and therefore for any $\mu\in\mathbb{H}_0^{*\prime}(0,\infty)$ and every $\phi\in\mathbb{H}_0(0,\infty)$, by using the results in §4.5 we obtain

$$\begin{aligned}\langle\mu\#_0\mathbb{D}_0(x),x\phi(x)\rangle&=\langle\mu(x)\otimes\mathbb{D}_0(y),xy\phi(x\circ y)\rangle\\&=\langle\mu(x),x\langle\mathbb{D}_0(y),y\phi(x\circ y)\rangle\rangle=\langle\mu(x),x\phi(x)\rangle\end{aligned}$$

which shows that

$$\mu\#_0\mathbb{D}_0(x)=\mu(x) \tag{52}$$

in the sense of $\mathbb{H}_0^{*\prime}(0,\infty)$, as asserted.

6. Differentiability properties of the H_0-convolution. We conclude with a brief remark on the differentiabilty properties of the generalised H_0-convolution. Let ν be any nonnegative integer, $\mu\in\mathbb{H}_0^{*\prime}(0,\infty)$ and $\lambda\in\mathbb{H}_0(0,\infty)$. Then, since for any $\phi\in\mathbb{H}_0(0,\infty)$

$$\begin{aligned}\langle\Delta^\nu[\mu\#_0\lambda(x)],x\phi(x)\rangle&=\langle\mu\#_0\lambda(x),x\Delta^\nu\phi(x)\rangle\\&=\langle\mu(x),x\lambda\#_0\Delta^\nu\phi(x)\rangle\\&=\langle\mu(x),x[\Delta^\nu\lambda]\#_0\phi(x)\rangle\\&=\langle\mu\#_0[\Delta^\nu\lambda(x)],x\phi(x)\rangle\\&=\langle\mu(x),x\Delta^\nu[\lambda\#_0\phi(x)]\rangle\\&=\langle\Delta^\nu[\mu(x)],x\lambda\#_0\phi(x)\rangle\\&=\langle[\Delta^\nu\mu]\#_0\lambda(x),x\phi(x)\rangle,\end{aligned}$$

we have that

$$\Delta^\nu[\mu\#_0\lambda]=\mu\#_0[\Delta^\nu\lambda]=[\Delta^\nu\mu]\#_0\lambda \tag{53}$$

in the sense of $\mathbb{H}_0^{*\prime}(0,\infty)$.

If now $\mu\in\mathbb{H}_0^{*\prime}(0,\infty)$ and $\sigma\in\mathscr{E}'(0,\infty)$, then by the same kind of argument, and using (53), we derive the double equality

$$\Delta^\nu[\mu\#_0\sigma]=\mu\#_0\Delta^\nu[\sigma]=[\Delta^\nu\mu]\#_0\sigma \tag{54}$$

in the sense of $\mathbb{H}_0^{*\prime}(0,\infty)$.

REFERENCES

[1] I. N. SNEDDON, *The Use of Integral Transforms*, TATA McGraw-Hill, New York, 1979.

[2] A. H. ZEMANIAN, *Generalized Integral Transformations*, Pure and Applied Mathematics, Vol. XVIII, Interscience, New York,

[3] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Cambridge Univ. Press, Cambridge, 1944.

[4] I. I. HIRSCHMANN, JR., *Variation diminishing Hankel transforms*, J. Anal. Math., 8 (1960/61), pp. 307–336.

[5] F. M. CHOLEWINSKI, *Hankel complex inversion theory*, Mem. Amer. Math. Soc., 58, 1965.

[6] J. DELSARTE, *Une extension nouvelle de la théorie des fonctions presque-périodiques de Bohr*, Acta Math., 69 (1938), pp. 259–317.

[7] A. E. NUSSBAUM, *On functions positive definite relative to the orthogonal group and the representation of functions on Hankel-Stieltjes transforms*, Trans. Amer. Math. Soc., 175 (1973), pp. 389–408.

[8] B. L. J. BRAAKSMA AND H. S. V. DE SNOO, *Generalised translation operators associated with a singular differential operator, ordinary and partial differential equations*, Dundee Conference, Lecture Notes in Mathematics 415, Springer-Verlag, Berlin, 1974, pp. 62–77.