SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1501-0001 $01.25/0

MULTIPLE BIFURCATION PROBLEMS OF CODIMENSION TWO*

JOHN GUCKENHEIMER[†]

Abstract. The term *bifurcation* in this paper refers to changes in the qualitative structure of any solutions to a system of ordinary differential equations with a varying parameter. This paper is about multiple bifurcations for which there is a multiple degeneracy in some feature of the system and a multi-dimensional parameter in its definition. The most immediate motivation for studying these problems is that they occur in the mathematical descriptions of many natural phenomena, but their importance extends beyond the fact that they can be identified in applications. Multiple bifurcations provide both a means of organizing knowledge about simple bifurcations and a powerful analytical tool for locating complicated dynamical behavior in some models. An intuitive reason for these features is that near some multiple bifurcations, the effect of nonlinear interactions is analytically accessible.

1. Introduction. In this section we present a description of the problems studied in this review and an overview of the methods we use. The basic object of interest is a system of ordinary differential equations depending upon parameters λ

$$\dot{x}=f_\lambda(x)=f(x,\lambda). \tag{1.1}$$

Here $x\in\mathbb{R}^n$, $\lambda\in\mathbb{R}^k$ and f_λ: $\mathbb{R}^n\to\mathbb{R}^n$ or f: $\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}^n$. We shall often represent the system (1.1) by the vector field x_λ. The solutions of (1.1) are described by the flow Φ_λ: $\mathbb{R}^n\times\mathbb{R}\to\mathbb{R}^n$ with $\Phi_\lambda(x,t)=x_\lambda(t)$ being the value at time t of the solution to (1.1) with initial condition $x=x_\lambda(0)$. The individual curves x_λ: $\mathbb{R}\to\mathbb{R}^n$ are the *orbits* or *trajectories of the flow.* Our primary attention is devoted to the way in which qualitative properties of the trajectories depend upon the parameters λ. A *bifurcation value* λ_0 of the parameter λ is one for which there are λ_1 in any neighborhood of λ_0 such that the flows Φ_{λ_1} and Φ_{λ_0} are qualitatively different.

An example, here, illustrates these ideas. In the system of equations on $\mathbb{R}^2$

$$\dot{x}_1=\lambda x_1-x_2+x_1(x_1^2+x_2^2),\qquad \dot{x}_2=x_1+\lambda x_2+x_2(x_1^2+x_2^2) \tag{1.2}$$

the origin is a globally stable equilibrium for $\lambda<0$. However, when $\lambda>0$, there is a periodic trajectory which forms the circle $x_1^2+x_2^2=\lambda$. Thus a bifurcation (called the *Hopf* bifurcation, cf. [2]) occurs at $\lambda=0$.

We are interested in two general aspects of a bifurcation problem like (1.2):

(1) To what extent does the geometric structure of the solution set of the system with parameters change if the system is perturbed? In particular, we want to examine perturbations which add higher order terms to an expression like (1.2) which is expanded as a Taylor series of an equilibrium.

(2) To what extent can one use the results of power series expansions to deduce the presence of solutions with complicated asymptotic behavior near an equilibrium? In the Hopf example (1.2), the Hopf bifurcation theorem (2.2) implies the Taylor expansion of degree 3 at an equilibrium is usually sufficient information to determine the presence of limit cycles near the equilibrium. We would like to study analogous problems for which more complicated dynamical phenomena can occur near an equilibrium, in a way that is determined qualitatively by a few terms in the Taylor expansion at the equilibrium.

* Received by the editors May 10, 1982, and in revised form November 18, 1982. This research was partially supported by the National Science Foundation under grant MCS 79-02002. This article is based on an invited address at the SIAM National Meeting, June 1980.

[†] Department of Mathematics, University of California, Santa Cruz, California 95064.

These two issues are closely related to the concept of *structural stability* [113]. We do not attempt here to formulate our results in terms of structural stability, however. Instead, we focus on particular dynamical features that can be described in each of the problems we discuss. There have been efforts to develop a more systematic bifurcation theory within the context of dynamical systems theory, but the fruits of these efforts do not seem ripe for the kinds of applications discussed here. Nonetheless, our attitude is motivated by the analogy with problems of singularity theory for smooth maps [126], where a systematic theory has been developed [84].

The particular bifurcation problems we study are associated with a system (1.1) for which there is an *equilibrium* point (x_0, λ_0) (this means $f(x_0, \lambda_0)=0$) which is not *hyperbolic*. The Jacobian derivative of f with respect to the x variables has zero or pure imaginary eigenvalues. The cases in which there is a single zero eigenvalue or a single pair of pure imaginary eigenvalues are well known and reviewed briefly in §2. The cases with double degeneracy: (1) a two-dimensional nilpotent space, (2) one zero eigenvalue and a pair of pure imaginary eigenvalues, or (3) two pairs of pure imaginary eigenvalues are our principal object of study.

In problems with a double degeneracy at an equilibrium, at least two parameters are needed to capture all of the qualitative features which are present in perturbations of the original system. Expressed in terms of Taylor series expansions, we view a problem with multiple degeneracy in the following way. The vector fields which yield a certain type of bifurcation problem are those which satisfy special conditions. For example, the vector fields with a two-dimensional nilpotent subspace at an equilibrium are those for which $P(0)=(dP/d\xi)(0)=0$, where $P(\xi)$ is the characteristic polynomial of the Jacobian derivative at the equilibrium. The number of independent conditions (two in the example) on the Taylor series is the *codimension* of the set of vector fields satisfying the special conditions. We also call this the codimension of the bifurcation. Thus, a vector field defined near an equilibrium with no eigenvalues on the imaginary axis has codimension zero. The problems addressed in this paper have codimension two. In addition to the special conditions which determine the type of bifurcation problem, there will be additional inequalities that the Taylor series is required to satisfy. In the example of a two-dimensional nilpotent subspace, we require $(d^2P/d\xi^2)(0)\neq 0$, to prevent the occurrence of a three-dimensional nilpotent subspace, as well as other inequalities that will be specified later.

Given a vector field x with a bifurcation of codimension k, we embed x in a k-dimensional family x_λ which is transverse to the set of vector fields satisfying the k special conditions determining the bifurcation. For example, a two-parameter family x_λ transverse to the vector fields with an equilibrium having a two-dimensional nilpotent space is defined by

$$(1.3) \qquad \dot{x}_1 = x_2, \qquad \dot{x}_2 = \lambda_1 + \lambda_2 x_1.$$

The process of determining these transversal families is discussed in §3. If we are lucky then a transversal family x_λ with $x_0 = x$ will contain all of the qualitative dynamical features that exist in perturbations of x. (The family (1.3) does not have this property.) In all cases, inequalities must be imposed upon nonlinear terms in the Taylor expansion at the equilibrium of x in order to have the desired properties of stability with respect to perturbations. When one locates a transversal family x_λ which is stable to perturbations, one says (loosely) by way of analogy with singularity theory that x_λ is a *universal unfolding* of x.

There are two remarks which we make about this approach to bifurcation problems. First, the applications of the theory often involve systems which possess a symmetry. Symmetries of the physical problem should be carried over to its mathematical description as a bifurcation problem. In discussing stability with respect to perturbations, it is necessary to specify whether the perturbation should be restricted to those which also satisfy the given symmetry. When only symmetric perturbations are considered, then one obtains a new list of codimension k bifurcations for each new symmetry group that is studied. We shall consider results obtained for a few very simple symmetry groups in this paper.

The second remark concerns the relationship between the results described in this paper and other approaches to bifurcation theory. There are three issues to be considered: 1) static vs. dynamic bifurcations, 2) the imposition of trivial solutions, and 3) the presence of a distinguished bifurcation parameter.

(1) One part of the solution of a bifurcation problem is the determination of how equilibrium solutions depend upon the parameter values. This static problem is a substantial subject in its own right and is much more amenable to systematic study (using singularity theory) than the *dynamic problem* of determining the nonequilibrium solutions as well. In the static problem one is interested in the zero set of families of maps $f_\lambda\colon \mathbb{R}^n \to \mathbb{R}^n$. The qualitative structure of the zero sets is preserved by coordinate transformations of the form $g_\lambda = \phi_\lambda \circ f_\lambda \circ \psi_\lambda$ with $\phi_\lambda, \psi_\lambda\colon \mathbb{R}^n \to \mathbb{R}^n$ invertible mappings depending smoothly on λ and subject to the condition that $\phi_\lambda(0) = 0$. These coordinate transformations can be used to make the linear parts of the Taylor expansion at two equilibria the same if the two Jacobian derivatives have the same rank. On the other hand a smooth change of coordinates $y = \phi(x)$ transforms the differential equation $\dot{x} = f(x)$ into $\dot{y} = D\phi_{\phi^{-1}(y)} f(\phi^{-1}(y))$. This yields a similarity transformation on the linear part of the Taylor series at an equilibrium. Thus the eigenvalues of the Jacobian derivative are unchanged by the type of coordinate change and problems that are indistinguishable in the static theory appear very different when considered dynamically.

(2) The second issue involved in comparing our results with other approaches to bifurcation theory involves the hypothesis that there be a "trivial" solution. Perturbation methods frequently assume that there is an equilibrium fixed at 0, and nonzero solutions with a specified dependence on a small parameter ε are sought. Customarily, the equilibrium at 0 remains an equilibrium for all values of all the parameters in the problem. In our setting, we can easily treat both the general case and this case in which there is the *constraint* that $f_\lambda(0) = 0$ for all λ.

(3) The issue of a distinguished parameter arises when one replaces individual vector fields as the basic object of study by one parameter families of vector fields. If one adopts the latter point of view, then one wants to study all of the perturbations of a given one-parameter family which has a degenerate bifurcation. With luck, one hopes that there is a finite dimensional family of one-parameter families of vector fields that contains all of the qualitative features found in perturbations of the degenerate one-parameter family. Here there is a two-tier structure in which the single parameter of the original system is *distinguished* from the additional parameters in the problem. Asymptotic methods usually distinguish a single parameter in terms of which nonzero solutions are expanded.

In this paper we take a definite stand on issues (1) and (3). *Dynamic* bifurcation problems *without* distinguished parameters are studied. This contrasts with the recent work of Golubitsky and Schaeffer [38] in which static problems with distinguished parameters are considered.

Sections 3–5 are organized around the common strategy which can be used to "solve" bifurcation problems of codimension two. We give here an outline of the steps in this analysis.

I. The first step in analyzing a bifurcation problem involves the use of smooth coordinate changes to reduce the arbitrariness in the Taylor expansion of a vector field with a degenerate equilibrium. The *normal form theorem* (3.2) of Takens [125] is a procedure for transforming coordinates near an equilibrium so that the Taylor series in the transformed variables is particularly simple. This gives us a much smaller collection of problems whose dynamics have to be explicitly analyzed. The theme, which recurs later, is that the nonlinear terms of the normal forms control the interaction of the degenerate modes which are undergoing bifurcation. A second, technical bonus of applying the normal form theorem to a problem is that is allows one to reduce many of the dynamical questions for codimension two bifurcations to considerations involving two dimensional vector fields.

II. The second step in analyzing a bifurcation problem is the computation of a transversal family containing the normal form of the vector field with a nonhyperbolic equilibrium. This computation is done in terms of the first degree Taylor expansion at the equilibrium. Arnold [6] gives a more comprehensive treatment of this aspect of the analysis.

III. The third step of the analysis is the determination of the dynamics contained in these transversal families. We work here with systems obtained by truncating the normal forms with the terms of a certain degree. Even for these truncated systems, some aspects of the dynamics are subtle. As remarked above, the truncated normal forms for codimension two bifurcations at equilibria all separate so that much of the dynamics can be deduced from a planar subsystem. Some of these planar systems have periodic solutions and care is required for their study.

IV. We examine structural stability properties of solutions obtained in III for the planar subsystems of the normal form families. If the phase portraits of these families are insensitive to perturbations (including the addition of higher order terms in the Taylor series), then we consider the analysis of the planar flows complete. However, we do encounter cases in which the normal forms truncated with nonlinear terms of only one degree lead to individual flows which have a family of periodic solutions. To remedy this structurally unstable situation, we truncate the normal form at a higher degree and repeat the analysis of step III. Our treatment here is incomplete in that we do not evaluate certain integrals. We *presume* that most choices of the additional nonlinear terms lead to structurally stable planar families, but the proof of this relies on additional study of the integrals which we do not evaluate.

V. For systems with imaginary eigenvalues, another difficult step remains before one has obtained a complete picture of the dynamics of the corresponding bifurcation problems. This step is the description of the flows in three and four dimensions which correspond to the planar flows for the reduced systems studied in §§3 and 4. The reduction was made on the basis that the truncated normal forms have a rotational symmetry, but the full system may be only approximately symmetric. By using a C^∞ change of coordinates, the original system can be made to differ from a symmetric one by a function which is *flat* (has zero Taylor series) at the bifurcation point. Section 5 explores the consequences of asymmetry. For quasiperiodic orbits, the asymmetry introduces *small divisor* problems which require the apparatus of the KAM (Kolmogorov–Arnold–Moser) theory for their reduction. There are also questions about *transversal homoclinic* orbits and *hyperbolic* invariant sets which arise. We refer the reader to [5] for descriptions of these *resonance* phenomena.

The results of the analysis are summarized by the planar *bifurcation diagrams* located at the end of the paper. These diagrams partition a neighborhood of the origin in the two-dimensional parameter plane into sectors. Each of these sectors represents a set of systems which have similar dynamics. Some of the sectors subtend large angles at the origin, but in some cases there are thin regions with boundaries that approach the origin tangentially. Each sector boundary is associated with a simpler *elementary* or *codimension one* bifurcation that is described in §2. Some of the boundaries which involve the resonance phenomena discussed in §5 are in fact "fuzzy." These fuzzy boundaries represent small portions of the parameter plane near the bifurcation in which there are dynamical phenomena that we cannot fully describe.

Section 6 is devoted to four representative examples to which the theory has been applied.

1. The variational Van der Pol equations [21] include a codimension two bifurcation to which the results of §§3 and 4 can be applied directly. These equations arise in the study of a sinusoidally forced, weakly-nonlinear oscillator.

2. Following Holmes [56], the motions of an elastic panel in response to a fluid flow across it can be studied with the techniques in this paper. Explicit calculations require an initial finite dimensional approximation to the defining system of partial differential equations.

3. A nonlinear reaction-diffusion problem is discussed for which the normal form associated with a system with infinite degrees of freedom can be calculated [41]. Apart from the issue of showing that there is no "hidden" symmetry, this application gives the first analytic demonstration of "chaotic" solutions to a system of partial differential equations.

4. Thermohaline convection is a bifurcation problem of fluid mechanics for which codimension two bifurcations occur and normal forms can be calculated with suitable boundary conditions [74]. This example also clearly demonstrates the essential role of *center manifolds* in the determination of normal forms.

We have attempted to present the results in this paper in a widely accessible form. Therefore, we have not tried to state the strongest possible structural stability results for codimension two bifurcations. Instead our goal has been to go as far as a few systematic methods will take us, indicating in §5 the issues whose resolution requires more sophisticated techniques. One might hope that these methods could also be extended to bifurcations of higher codimension, but problems with more degeneracy in the linear part at an equilibrium require substantial new insights into the calculations of the qualitative features of three-dimensional flows. Thus, we expect that analysis of codimension three bifurcations will require methods which go well beyond those used in this paper.

2. Codimension one. This section is a rapid review of some codimension one (or elementary) bifurcations of a dynamical system. It provides background for the discussion of codimension two bifurcations in the next three sections. Here we consider a vector field X_λ or system of ordinary differential equations $x=f_\lambda(x)$ defined on $\mathbb{R}^n$ with $\lambda\in\mathbb{R}$ a single parameter. We are interested in the variation of equilibria and periodic orbits of X_λ as functions of λ. In particular, we initially focus open changes in the equilibria and periodic orbits of X_λ. An *equilibrium* $p\in\mathbb{R}$ is a solution of $f_\lambda(p)=0$. A *periodic orbit* is a nonequilibrium trajectory with $x_\lambda(t)=x_\lambda(0)$ for some $t>0$. Bifurcations occur when there are changes in the number or stability of equilibria or periodic orbits of the vector field with varying parameter λ.

To obtain a picture of the simplest bifurcations, we must first recall some fundamental results about equilibria and periodic orbits which are not bifurcating. If p is an equilibrium of X_λ, then the flow of X_λ near p is studied initially by *linearizing* X_λ at p. One replaces the system of equations $\dot{x}=f_\lambda(x)$ by the linear system $\dot{\xi}=(Df_\lambda(p))\xi$. If the $n\times n$ matrix $Df_\lambda(x)$ has no zero or pure imaginary eigenvalues, then the equilibrium is called *hyperbolic*. At a hyperbolic equilibrium, one can write $\mathbb{R}^n$ as a direct sum of two invariant subspaces E^u, E^s such that the spectrum of $Df_\lambda(p)$ restricted to E^u is in the right half plane, and the spectrum of $Df_\lambda(p)$ restricted to E^s is in the left half plane. Solutions $\xi_\lambda(t)$ of the linear system which lie in E^u tend to the origin as $t\to-\infty$, and solutions which lie in E^s tend to the origin as $t\to+\infty$. Other solutions diverge both as $t\to-\infty$ and $t\to+\infty$. The *stable manifold theorem* [47] asserts that a similar property is true for the original vector field X_λ. There are submanifolds, the *stable manifold* W^s and *unstable manifold* W^u, invariant under the flow of X_λ with the property that $x_\lambda(t)\to p$ as $t\to+\infty$ if $x_\lambda(0)\in W^s$ and $x_\lambda(t)\to p$ as $t\to-\infty$ if $x_\lambda(0)\in W^u$. There is a neighborhood U of p such that all solutions not in W^s or W^u leave U both for times $t>0$ and times $t<0$. The tangent spaces of W^s and W^u at p are E^s and E^u. Hyperbolic equilibria vary smoothly with λ as do compact subsets of their stable and unstable manifolds.

Bifurcation at an equilibrium requires that the linearized vector field have zero or pure imaginary eigenvalues. In the simplest situations, these degeneracies of the linearization lead to the *saddle node* (or limit point) and *Hopf* bifurcations, respectively. In each case, the local structure of the flow near the bifurcating equilibrium p is controlled by specific nonlinear terms in the Taylor series of f_λ at p when these are not zero. This is a common theme for all the bifurcations we study. Another aspect of the analysis of all the bifurcations we study is that an extension of the stable manifold theorem allows one to consider dynamical behavior in a low dimensional submanifold of $\mathbb{R}^n$. The center manifold theorem [83] implies that hyperbolic behavior persists in the directions complementary to the eigenspaces for eigenvalues which lie on the imaginary axis [98]. Thus, for the saddle node and Hopf bifurcations, one need only study one- and two-dimensional vector fields to understand the dynamics near a general bifurcation of these types. We note that constraints or symmetries affect the analysis of the saddle node bifurcation.

Saddle node. The prototype (normal form) for the saddle node bifurcation is the one-parameter family of vector fields defined by $\dot{x}=\lambda-x^2$ on $\mathbb{R}^1$. Here the flow is trivial with all solutions $x_\lambda(t)\to-\infty$ when $\lambda<0$ and $t\to\pm\infty$. At $\lambda=0$, a single equilibrium appears at $x=0$. It is stable from the right and unstable from the left. When $\lambda>0$, there are two equilibria, one stable and the other unstable. The two equilibria separate from each other at a rate comparable to $\sqrt{\lambda}$. The saddle node is shown in Fig. 1. The extent to which these properties are satisfied by the general saddle node bifurcation is expressed by the following theorem.

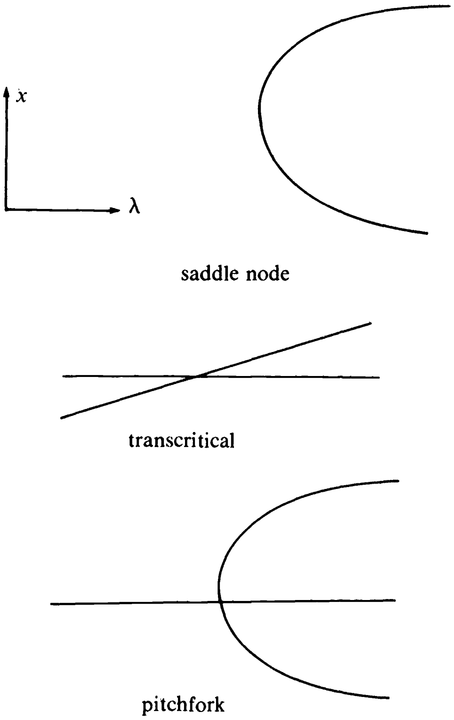
THEOREM 2.1. *Let X_λ be a one-parameter family of vector fields on $\mathbb{R}^n$. Let p be an equilibrium point for X_{λ_0} for which the following hypotheses are satisfied*:

(SN1) *The linearization of X_{λ_0} at p has a simple eigenvalue* 0 *with right eigenvector v and left eigenvector w. The remainder of the spectrum of X_{λ_0} lies off the imaginary axis.*

(SN2) $w((\partial/\partial\lambda)X\lambda(p;\lambda_0))\neq 0$.

(SN3) $w(D^2X_{\lambda_0}(p))(v,v)\neq 0$.

Then there is a smooth curve of equilibria for X_λ *in* $\mathbb{R}^n\times\mathbb{R}$ *which has a quadratic tangency with the hyperplane* $\mathbb{R}^n\times(\lambda_0)$. *If* (x_1,λ) *and* (x_2,λ) *are two equilibria of* X_λ *near* (p,λ_0), *then* x_1 *and* x_2 *are hyperbolic and* $|\dim W^s(x_1)-\dim W^s(x_2)|=1$. *The set of one-parameter families of vector fields satisfying* (SN1)–(SN2) *is an open set in the space* $\Xi=C^\infty(\mathbb{R}^n\times\mathbb{R},\mathbb{R}^n)$. *Here quadratic tangency of a curve with a hyperplane means that the second derivative of the curve does not lie in the hyperplane.*



saddle node

transcritical

pitchfork

FIG. 1. *Codimension* 1 *bifurcations of equilibria.*

This description of the saddle node bifurcation is inappropriate in settings for which hypotheses (SN2) and (SN3) cannot be satisfied. Two of these deserve mention as alternatives to the saddle node bifurcation. In classical bifurcation problems, one usually has the distinguished *trivial* equilibrium at the origin which is assumed to exist for all values of λ. Accordingly, hypothesis (SN2) cannot be satisfied at a bifurcation of the trivial equilibrium. The appropriate condition which replaces (SN2) is that $w((\partial/\partial\lambda)(DX_\lambda(0;\lambda_0))(v))\neq 0$ and the prototype for this new kind of bifurcation is the equation $\dot{x}=\lambda x-x^2$. This system describes a *transcritical bifurcation* in which two smooth curves of equilibria cross at $\lambda=0$ and *exchange stability* there. Within the class Ξ of systems which satisfy the constraint $f_\lambda(0)=0$ for all λ, transcritical bifurcation of the trivial equilibrium is typical behavior. If the constraint is dropped then the transcritical bifurcation can be perturbed to either a pair of saddle node bifurcations or to no bifurcation at all. Figure 1 illustrates transcritical bifurcation.

The second alternative to the saddle node involves systems with a simple symmetry. Examples often arise in which the systems $\dot{x}=f_\lambda(x)$ are equivariant with respect to a reflection. In a one-dimensional system, this means that $f_\lambda(x)=-f_\lambda(-x)$ or f_λ is odd for all λ. For such systems, both (SN2) and (SN3) will fail. The symmetry automatically implies that there is a trivial equilibrium, so (SN2) is once again replaced with the condition $w((\partial/\partial\lambda)(DX_\lambda(0))(v))\neq 0$. The hypothesis (SN3) is replaced by the assumption $w(D^3X_\lambda(0))(v,v,v)\neq 0$. The simplest example satisfying these assumptions and the symmetry condition is $\dot{x}=\lambda x-x^3$. The typical bifurcation behavior within the class of symmetric systems is a *pitchfork* or *symmetric* bifurcation in which a curve of nontrivial equilibria passes through the point of bifurcation $(0,\lambda_0)$ with a quadratic tangency to the plane $\mathbb{R}^n\times\{\lambda_0\}$. Pitchfork bifurcations can be perturbed to systems with either one saddle node or three saddle nodes if the symmetry condition is dropped. The theory of Golubitsky and Schaeffer describes these perturbations, though much of the theory in this example was understood previously.

The bifurcation diagrams in Fig. 1 show the loci of equilibria in $\mathbb{R}^n\times\mathbb{R}$ for these three different bifurcations involving one zero eigenvalue at an equilibrium. Each represents structurally stable behavior within different classes of systems. The form of the prototypical examples comes from the normal form analysis and transversality considerations discussed in §3.

Hopf bifurcation. The second kind of codimension one bifurcation which involves an equilibrium p occurs when the linearized vector field at p has a simple pair of pure imaginary eigenvalues and no other eigenvalues on the imaginary axis. This bifurcation is called *Hopf bifurcation* in recognition of E. Hopf's contribution to its study (see [86]). The simplest expression of a system with a nondegenerate Hopf bifurcation is given in polar coordinates by

$$\dot{r}=r(\lambda-r^2),\qquad \dot{\theta}=1. \tag{2.1}$$

From a static point of view, there is no bifurcation of equilibria at $\lambda=0$. However, the *stability* of the equilibrium solution at the origin changes as λ changes sign and this is accompanied by a change in the number of periodic orbits. The periodic solutions of (2.1) form the smooth surface defined by $\lambda=r^2$ with quadratic tangency to the plane $\mathbb{R}^2\times\{0\}$. See Fig. 2. Computationally, the Hopf bifurcation provides an important technique for locating periodic solutions of a system. Without having to integrate the differential equations, the change in stability at the equilibrium signals the occurrence of periodic orbits and their approximate location near the bifurcation. The following theorem gives a precise statement of these results.

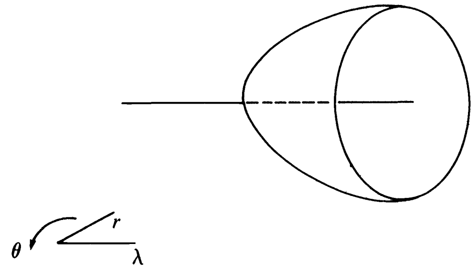


FIG. 2. *Hopf bifurcation.*

THEOREM 2.2. *Let X_λ be a one-parameter family of vector fields on $\mathbb{R}^2$. Let $p(\lambda)$ be an equilibrium for X_λ for which the following hypotheses are satisfied*:

(H1) *The linearization of X_λ at $p(\lambda)$ has a simple pair of complex eigenvalues $\alpha(\lambda)$,*

$\bar{\alpha}(\lambda)$ *with* $\mathrm{Re}\,\alpha(\lambda_0)=0$. *All other eigenvalues of the linearization of* X_λ *at* $p(\lambda)$ *have nonzero real parts.*

(H2) $(d/d\lambda)(\mathrm{Re}\,\alpha(\lambda))\neq 0$.

Hypotheses (H1) *and* (H2) *imply that in* $\mathbb{R}^n\times\mathbb{R}$ *there is a smooth two-dimensional surface* σ *tangent to the eigenspace of* $\alpha(\lambda_0)$ *and* $\bar{\alpha}(\lambda_0)$ *at* $p(\lambda_0)$ *which comprises a family of periodic orbits for the vector fields* X_λ. *Moreover* σ *is contained in a three-dimensional submanifold* M *of* $\mathbb{R}^n\times\mathbb{R}$ *in which* X_λ *can be expressed via polar coordinates on* $M\cap(\mathbb{R}^n\times\{\lambda\})$ *in the form*

$$\dot{r}=r(a\lambda+br^2)+\textit{higher order terms},\qquad \dot{\theta}=(c+dr^2)+\textit{higher order terms}.$$

Hypotheses (H1) *and* (H2) *imply that neither* a *nor* c *is* 0 *in this representation.*

(H3) $b\neq 0$.

Hypotheses (H1)–(H3) *imply that* σ *has quadratic tangency with the hyperplane* $\mathbb{R}^2\times\{\lambda_0\}$ *at* (p,λ_0). *Near* (p,λ_0), *the signs of* a *and* b *and the spectrum of the linearized equations determine the stability of the equilibria* $p(\lambda)$ *and the periodic orbits in* σ. *Systems satisfying* (H1)–(H3) *form an open set of one-parameter families of vector fields in* $C^\infty(\mathbb{R}^n\times\mathbb{R},\mathbb{R}^n)$.

The Hopf and saddle node bifurcations constitute a complete list of the codimension one bifurcations of equilibria in general one-parameter families of vector fields. More degenerate bifurcations can be perturbed to a succession of saddle nodes and Hopf bifurcations by small C^∞ changes in the family. It is of considerable interest to determine which successions can occur when a somewhat more degenerate bifurcation is perturbed. Results of this kind are readily obtained from the analysis of codimension two bifurcations in the next sections. Before proceeding, however, we need to discuss other kinds of codimension one bifurcations which do not involve qualitative changes in equilibria.

For bifurcations of periodic orbits, one has theorems analogous to those stated above which describe saddle node and (secondary) Hopf bifurcations for periodic orbits. In addition, there is a new type of bifurcation for periodic orbits, the *flip* (*periodic doubling* or *subharmonic*) bifurcation. The standard technique for investigating the stability and bifurcation of periodic orbits of a flow begins by choosing a *cross-section* and defining its (Poincaré) *return map*. The cross-section is a hypersurface $M^{n-1}\subset\mathbb{R}^n$ which is transverse to the vector field and intersects the periodic orbit in exactly one point p. The return map θ: $M\to M$ is defined by sending x to the first point on the trajectory through x which lies in M. The map θ has p for a fixed point and is defined in a neighborhood of p in M. The use of the return map reduces many questions about the periodic orbit to corresponding questions about fixed points of a system defined for *discrete* times. For discrete systems there is again a stable manifold theorem. Here *hyperbolicity* requires that $D\theta(p)$ has no eigenvalues of absolute value 1.

The stable manifold theory works as well for discrete time systems as for continuous time systems. Equilibria are replaced by fixed points, and pure imaginary eigenvalues are replaced by eigenvalues of modulus 1. We continue to use the notation $W^s(p)$ and $W^u(p)$ for the stable and unstable manifolds of a fixed point p for the discrete system obtained by iterating the map θ: $M\to M$. There are three generic codimension one bifurcations for fixed points in one-parameter families of discrete systems. These correspond to eigenvalues $+1$, -1, and pairs of complex eigenvalues of modulus 1. Complex eigenvalues which are third or fourth roots of unity are special.

THEOREM 2.3. *Let* θ_λ: $M\to M$ *be a one-parameter family of smooth maps. When* $\lambda=\lambda_0$ *assume that* θ_λ *has a fixed point* p *at which the following conditions are satisfied*:

(SN1) *The derivative* $D\theta_{\lambda_0}(p)$ *of* θ_{λ_0} *has a simple eigenvalue* 1 *with right eigenvector*

v and left eigenvector w. The remainder of the spectrum of $D\theta_{\lambda_0}(p)$ lies off the unit circle.

(SN2) $w((\partial\theta/\partial\lambda)(p))\neq 0$.

(SN3) $w(D^2\theta_{\lambda_0}(p)(v,v))\neq 0$.

Then there is a smooth curve of fixed points for θ_λ in $M\times\mathbb{R}$ which has quadratic tangency with the hyperplane $M\times\{\lambda_0\}$. If (x_1,λ) and (x_2,λ) are two distinct fixed points near (p,λ_0), then $|\dim W^s(x_1)-\dim W^s(x_2)|=1$.

THEOREM 2.4. *Let θ_λ: $M\to M$ be a one-parameter family of smooth maps. When $\lambda=\lambda_0$ assume that λ has a fixed point p at which the following conditions are satisfied*:

(SH1) *The derivative $D\theta_{\lambda_0}(p)$ has a simple eigenvalue -1 with left eigenvector v and right eigenvector w. The remainder of the spectrum of $D\theta_{\lambda_0}(p)$ lies off the unit circle.*

Hypothesis (SH1) *implies that there is a smooth curve $x(\lambda)$ of fixed points of θ_λ with $x(\lambda_0)=p$. Let $\alpha(\lambda)$ be the continuous function such that $\alpha(\lambda)$ is an eigenvalue of $D\theta_\lambda(x(\lambda))$ and $\alpha(\lambda_0)=-1$.*

(SH2) $(d/d\lambda)(\alpha(\lambda_0))\neq 0$.

(SH3) $w(D^3\theta^2_{\lambda_0}(p)(v,v,v))\neq 0$.

If hypotheses (SH1)–(SH3) *hold, then a curve of periodic orbits of period* 2 *bifurcates from (p,λ_0) in $M\times\mathbb{R}$. This curve has quadratic tangency with the hyperplane $M\times\{\lambda_0\}$.*

THEOREM 2.5. *Let θ_λ: $M\to M$ be a one-parameter family of maps which has a smooth curve of fixed points $x(\lambda)$. Assume that the derivatives $D\theta_\lambda(x(\lambda))$ have a continuous family of simple complex eigenvalues $\alpha(\lambda),\bar{\alpha}(\lambda)$ such that the following conditions hold*:

(H1)′ *At $x(\lambda_0)=p$, $|\alpha(\lambda_0)|=1$ and all other eigenvalues of $D\theta_{\lambda_0}(p)$ beside $\alpha(\lambda_0),\bar{\alpha}(\lambda_0)$ lie off the unit circle.*

(H2)′ $(d/d\lambda)|\alpha(\lambda_0)|\neq 0$.

If $\alpha^i(\lambda_0)\neq 1$ for $i=3$ or 4, *then hypotheses* (H1)′ *and* (H2)′ *imply that there is a smooth change of coordinates so that the expression for θ in polar coordinates in the plane spanned by the eigenvectors $\alpha(\lambda_0),\bar{\alpha}(\lambda_0)$ is*

$$\theta_\lambda(r,\phi)=\left(r\left(1+a(\lambda-\lambda_0)+br^2\right),\phi+c+dr^2\right)+\textit{higher order terms}.$$

Hypotheses (H1)′ *and* (H2)′ *imply that neither a nor c is zero.*

(H3)′ $b\neq 0$.

Hypotheses (H1)′–(H3)′ *imply that there is a two-dimensional surface σ (not infinitely differentiable!) having quadratic tangency with the hyperplane $M\times\{\lambda_0\}\subset M\times\mathbb{R}$ which is invariant for θ: $\theta(\sigma)=\sigma$. If $\sigma\cap(M\times\{\lambda\})$ is larger than a point, then it is a simple closed curve.*

Note here that the dynamics of θ on the invariant curves produced by this theorem remain to be determined. This involves questions of *resonance* and *small divisors* which we postpone until §5. In addition, the cases $\alpha^3(\lambda_0)=1$ and $\alpha^4(\lambda_0)=1$ have *strong resonance* for which additional terms enter the special polar coordinate representations of θ stated in the theorem [8]. These additional terms reflect more complicated behavior for the typical family having bifurcations with third or fourth roots of unity as eigenvalues.

The final type of codimension one bifurcation involves a lack of transversality between the stable and unstable manifolds of equilibria or periodic orbits. Since they involve trajectories far from these special orbits, they have a more global character than the bifurcations considered thus far. Nonetheless, saddle connections for two-dimensional flows are an important part of our analysis of codimension two bifurcations of equilibria. If a trajectory is asymptotic to a single hyperbolic equilibrium or periodic orbit for both $t \to +\infty$ and $t \to -\infty$, then it is called *homoclinic*. A nontransversal intersection of the stable manifold of an equilibrium or periodic orbit with the unstable manifold of another is a *heteroclinic* trajectory. The typical lack of transversality which occurs in codimension one depends upon whether equilibria or periodic orbits are involved because the dimension counts differ. (For a periodic orbit γ, $\dim W^s(\gamma) + W^u(\gamma) = n+1$, allowing the possibility of transversal intersections along a homoclinic trajectory.) A full discussion of the phenomena involved in the lack of transversality bifurcations is highly technical and would be out of place in this review. We restrict ourselves to the one result which will be used in the next section.

THEOREM 2.6 [114]. *Let X_λ be a one-parameter family of vector fields on $\mathbb{R}^2$ with a saddle point $x(\lambda)$. Assume that when $\lambda = \lambda_0$, there is a trajectory γ which tends to $p = x(\lambda_0)$ as $t \to \pm\infty$. If the trace τ of $DX_{\lambda_0}(p)$ is not zero, then there is an $\varepsilon > 0$ and a one-parameter family of periodic orbits $\pi(\lambda)$ for $\varepsilon + \lambda_0 > \lambda > \lambda_0$ or $\varepsilon - \lambda_0 < \lambda < \lambda$ such that $\pi(\lambda) \to \gamma$ as $\lambda \to \lambda_0$. The stability of the $\pi(\lambda)$ is determined by τ: $\pi(\lambda)$ is stable if $\tau < 0$ and unstable if $\tau > 0$.*

3. Codimension two bifurcations of equilibria I. Normal forms. Using the codimension one bifurcations described above as a basic dictionary, let us turn to the analysis of codimension two bifurcations of equilibria. We consider a two-parameter family X_λ of vector fields on $\mathbb{R}^n$ defined by the equations $\dot{x} = f_\lambda(x) = f(x, \lambda)$ with f_λ: $\mathbb{R}^n \times \mathbb{R}^n$ a smooth map. We shall assume that within a specified class of two-parameter families of vector fields Ξ there is a value λ_0 of λ for which X_{λ_0} has a degeneracy more complicated than those described in the last section. With this assumption, we want to describe as much as possible about the flows of X_λ for λ near λ_0. In particular we shall draw diagrams of the λ plane showing regions in which the X_λ have similar dynamic behavior and the curves bounding these regions along which various codimension one bifurcations take place. Experimentally based diagrams of this kind can be found in the literature of fluid dynamics [20], but little emphasis has been placed upon understanding the intersections of curves representing different bifurcations in that context.

THEOREM 3.1. *Let $\Xi = C^\infty(\mathbb{R}^n \times \mathbb{R}^2, \mathbb{R}^n)$ be the class of general two-parameter families of smooth vector fields. Any two-parameter family of vector fields in Ξ can be perturbed so that the only bifurcations of equilibria are saddle nodes, Hopf bifurcations, or one of the following five types*:

(i) *X_{λ_0} has an equilibrium p at which DX_{λ_0} has a simple zero eigenvalue and no other eigenvalues on the imaginary axis. Hypothesis* (SN3) *of Theorem* 2.1 *fails, but a corresponding cubic term is not zero.*

(ii) *X_{λ_0} has an equilibrium p at which DX_{λ_0} has a simple pair of pure imaginary eigenvalues and no other eigenvalues on the imaginary axis. Hypothesis* (H3) *of Theorem* 2.2 *fails, but a corresponding fifth degree term is not zero.*

(iii) *X_{λ_0} has an equilibrium p at which DX_{λ_0} has zero as an eigenvalue of multiplicity two and no other eigenvalues on the imaginary axis. DX_{λ_0} is nilpotent of rank* 1 *on the generalized eigenspace of zero* (*i.e., the Jordan canonical form on this subspace is* $(\begin{smallmatrix}0&1\\0&0\end{smallmatrix})$) *and higher order conditions specified below are satisfied.*

(iv) *X_{λ_0} has an equilibrium p at which DX_{λ_0} has zero as a simple eigenvalue and a simple pair of pure imaginary eigenvalues. No other eigenvalues of $DX_{\lambda_0}(p)$ lie on the imaginary axis, and higher order conditions specified below are satisfied.*

(v) *X_{λ_0} has an equilibrium p at which DX_{λ_0} has two simple pairs of pure imaginary eigenvalues and no other eigenvalues on the imaginary axis. Nonresonance conditions and higher order conditions specified below are satisfied.*

We shall summarize briefly some of the history of these different bifurcations. Case (i) is a dynamic version of Thom's cusp catastrophe [126], and its unfolding is the same in this context as it is in Thom's. The degenerate Hopf bifurcation (ii) has been studied by Takens [122] and Golubitsky and Langford [37]. Each gives a full picture of the unfolding of a persistent family. The double zero eigenvalue (iii) was analyzed independently by Takens [124] and Bogdanov [16]. Their work established the approach to codimension two bifurcation problems adopted here. The bifurcations (iv) have been studied in various contexts by Keener [69], [70], Langford [77], Guckenheimer [41], and Holmes [58]. The double Hopf bifurcation has been studied only recently and the work of a number of investigators will undoubtedly appear shortly after this is written.

In the remainder of this review we shall concentrate upon the analysis of cases (iii), (iv) and (v) from Theorem 3.1 which involve double degeneracies for the linearized problems at an equilibrium. For cases (iii) and (iv) we discuss alternatives which involve restricting our space of vector fields to satisfy a constraint or symmetry of the sort discussed in §2 with reference to the saddle node. In many cases, flows near a codimension two bifurcation are completely determined by considerations involving planar vector fields, but other cases involve resonance phenomena which are discussed in §5. To emphasize the common structure of the analysis employed in the different cases, we shall proceed by applying each step of the strategy outlined in the introduction to all cases simultaneously. The reader primarily interested in the results which pertain to a given case can find these presented succinctly in §4 in Figs. 3–9.

The first step involves making smooth coordinate changes which simplify the expression of our systems as much as possible. The practical meaning of this statement is that we try to transform to zero as many nonlinear terms as possible in the Taylor series of the vector field at the point of bifurcation. The procedure for doing so is inductive, working with terms of successively higher degree in the Taylor series. At each stage of the calculation one computes the image of a certain linear map that can be expressed in Lie algebraic terms. Terms in the Taylor series can be changed by addition of elements lying in the image of the linear map, so coordinate changes are possible in which the nonlinear terms of the vector field in new coordinates lie in specific complements to the images of the linear maps. The resulting expressions are called the *normal forms* of the vector fields.

Let us describe the procedure of calculating normal forms in more detail. Let X: $\mathbb{R}^n \to \mathbb{R}^n$ be a vector field and ϕ: $\mathbb{R}^n \to \mathbb{R}^n$ be a locally defined diffeomorphism which defines a change of coordinates. The expression of X in the new cooridnates is $Y(x) = D\phi_{\phi^{-1}x}(X(\phi^{-1}(x)))$. We are particularly interested in vector fields which have an equilibrium at the origin and coordinate changes which leave the origin fixed. The effect of such coordinate changes on the Taylor series of X at the origin can be computed by expanding both ϕ and X in their Taylor series:

$$X(x) = \sum_{i=1}^{k} A_i(x) + o(k), \qquad \phi(x) = \sum_{i=1}^{k} P_i(x) + o(k).$$

Here $A_i(x) = \sum_{j=1}^{n} A_{ij}(x)\,\partial/\partial x_j$ is a vector field whose coefficients $A_{ij}(x)$ are homogeneous polynomials of degree i and P_i is a vector valued homogeneous polynomial of degree i.

The linear terms of Y are the linear terms of $D\phi X \circ \phi^{-1}$. Since the Taylor series of ϕ^{-1} begins with P_1^{-1}, the linear part of Y is $P_1 A_1 P_1^{-1}$. We may choose P_1 so that the linear part of the vector field in new coordinates is in Jordan canonical form and then assume that all further coordinate changes have P_1 as the identity. To proceed inductively, we assume that $l>1$ is chosen so that $P_2=P_3=\cdots=P_{l-1}=0$ and P_1 is the identity. Then the expression of X and Y will have the same terms of degree smaller than l and the terms of degree l will differ in an easily computable way. If $\phi(x)=x+P_l(x)+o(l)$, then $\phi^{-1}(x)=x-P_l(x)+o(l)$ and $D\phi(x)=\mathrm{id}+DP_l(x)+o(l+1)$. Therefore, Y has Taylor expansion of degree l

$$(\mathrm{id}+DP_l(x))\cdot\sum_{i=1}^{l}A_i(x-P_l(x))=\sum_{i=1}^{l}A_i(x)+DP_lA_1(x)-A_1\circ P_l(x)+o(l).$$

Algebraically, this last formula can be expressed by introducing a linear map $\bar{A}$: $H_l\to H_l$ where H_l is the vector space of vector valued homogeneous polynomials of degree l. The map is defined by $\bar{A}_l(P_l)=DP_lA_1-A_1\circ P_l$. If A_1 and P_l are both interpreted as vector fields then this is the adjoint action of A_1 on H_l: $A_1(P_l)$ is the Lie bracket $[A_1,P_l]$. The nice aspect of this calculation is that the degree l terms of X have been changed in a way which depends linearly on the elements of H_l. This allows one to carry out normal form calculations quite effectively. One lets l increase, thereby inductively changing the terms of X of higher degrees to the desired form. However, one must be careful in carrying through the procedure with examples to remember that the coordinate change at stage l will affect higher degree terms of X in a more complicated nonlinear fashion. The end result of these computations is expressed by the following *normal form theorem*.

THEOREM 3.2 [132]. *Let X be a vector field on $\mathbb{R}^n$ with an equilibrium at* 0. *Denote $DX(0)$ by L and by H_l the space of vector fields on $\mathbb{R}^n$ whose components are homogeneous polynomials of degree l. Define the linear map* $\mathrm{ad}\,L$: *$H_l\to H_l$ by* $\mathrm{ad}\,L(P)=[L,P]$. *For each $l>1$, let B_l be the image of* $\mathrm{ad}\,L$ *and pick a complementary subspace G_l: $H_l=B_l+G_l$. Then, for any $k>1$, there is a polynomial change of coordinates in $\mathbb{R}^n$ leaving the origin fixed, so that the Taylor series of degree k of $D\phi\cdot X\circ\phi^{-1}$ is $\Sigma_{l=1}^{k}C_l(x)+o(k)$ with $C_l\in G_l$ for $1<l\leq k$.*

Before applying the normal form theorem to bifurcation problems, we recall some history of the *linearization problem* for hyperbolic equilibria. This is the problem of whether there exists an analytic or smooth change of coordinates near an equilibrium for which the vector field becomes linear. Formally, the solution to this problem requires that the map $\mathrm{ad}\,L$: $H_l\to H_l$ in the normal form theorem be surjective for all $l>1$. The eigenvalues of $\mathrm{ad}\,L$ acting on H_l have the form $\lambda_i-\Sigma_{j=1}^{n}\alpha_j\lambda_j$, where the λ_i,λ_j are eigenvalues of L and the α_j are nonnegative integers whose sum is l. Thus the formal solution of the linearization problem depends upon the eigenvalues of the linearized vector field. If all the real parts of the eigenvalues of L have the same sign and X is analytic, then Poincaré proved convergence of the power series defining the coordinate change. When L has eigenvalues in both the left and right half planes, then the convergence question involves *small divisors* because the sums $\lambda_i-\Sigma_{j=1}^{n}\alpha_j\lambda_j$ do not remain bounded away from zero. Nonetheless, Siegel [108] established convergence of the formal coordinate change which linearizes X for a set of L whose complement has Lebesgue measure zero on $\mathbb{R}^{n^2}$. Sternberg [118] considered a C^∞ version of this theorem. If smoothness conditions are dropped, then *Hartman's theorem* [47] shows that all vector fields are locally topologically equivalent to a linear vector field. (The kinds of geometric phenomena which prevent analytic linearization when there are solutions of an equation $\lambda_i - \Sigma_{j=1}^{n} \alpha_j \lambda_j = 0$ are evident when one interprets the vector field as a complex vector field [9].)

The linearization problem introduced above is never solvable at a nonhyperbolic equilibrium. There will always be $l > 1$ for which the maps $\operatorname{ad} L$ are not surjective. The nonlinear terms coming from the complementary subspaces G_l in the normal form theorem play an important role in determining the qualitative structure of the dynamics of these systems and their perturbations. Applying the normal form theorem to our bifurcation problems focuses attention on this issue. The relatively small number of remaining nonlinear terms in the normal forms makes feasible a study of the dynamics of the general equation with a given linear part. The computations which reduce the general system to its normal form can be lengthy, but they are straightforward.

For codimension one and two bifurcation problems, there are five choices of L for which we want to carry out the normal form computations:

(i) $L=0$ on $\mathbb{R}^1$, $L=(0)$.

(ii) L is skew symmetric on $\mathbb{R}^2$, $L=\begin{pmatrix} 0 & -w \\ w & 0 \end{pmatrix}$.

(iii) L is nilpotent of rank 1 on $\mathbb{R}^2$, $L=\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$.

(iv) L is an infinitesimal rotation on $\mathbb{R}^3$, $L=\begin{pmatrix} 0 & -w & 0 \\ w & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$.

(v) L acts on $\mathbb{R}^4$, being skew symmetric on two invariant orthogonal subspaces, $L=\begin{pmatrix} 0 & -w_1 & 0 & 0 \\ w_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -w_2 \\ 0 & 0 & w_2 & 0 \end{pmatrix}$.

The first two of these cases correspond to the saddle node and Hopf bifurcations, respectively. The last three correspond to the cases of codimension two bifurcations from Theorem 3.1.

Cases (ii), (iv) and (v) above involve pure imaginary eigenvalues. It simplifies matters to complexify the normal form calculations in these cases. Let us illustrate this process with case (ii), the Hopf theorem calculations. Begin with $\mathbb{R}^2$ and coordinates (x,y). If we allow each of x and y to become complex, then we introduce $(z,\bar{z})$ as the coordinates relative to the basis

$$\frac{1}{2}(1,-i)=\frac{\partial}{\partial z}, \qquad \frac{1}{2}(1,i)=\frac{\partial}{\partial \bar{z}}.$$

The matrix L becomes $\left(\begin{smallmatrix} -iw & 0 \\ 0 & w \end{smallmatrix}\right)$ in the new coordinates, or $L=iw(-z\,\partial/\partial z+\bar{z}\,\partial/\partial\bar{z})$. On H_l. the eigenvectors of $\operatorname{ad} L$ are the vector fields $z^j\bar{z}^{l-j}\,\partial/\partial z$ and $z^j\bar{z}^{l-j}\,\partial/\partial\bar{z}$ with eigenvalues $iw(l-2j\pm 1)$, respectively. Therefore $\operatorname{ad} L$ is surjective on H_l if l is even and has a two-dimensional kernel if l is odd. One choice of complementary subspace is the plane spanned by $(z\bar{z})^{(l-1)/2} z\,\partial/\partial z$ and $(z\bar{z})^{(l-1)/2}\bar{z}\,\partial/\partial\bar{z}$ or the real vectors $(x^2+y^2)^{(l-1)/2}(x\,\partial/\partial x+y\,\partial/\partial y)$ and $(x^2+y^2)^{(l-1)/2}(-y\,\partial/\partial x+x\,\partial/\partial y)$. In polar coordinates these vector fields are $r^l\,\partial/\partial r$ and $r^{l-1}\,\partial/\partial\theta$. For $l=3$, this computation provides the justification for the expressions introduced in the Hopf bifurcation theorem.

We make an additional remark about the normal form for the Hopf bifurcation before considering the codimension two problems. Note that the normal form expressed in polar coordinates has a right-hand side which is independent of the angular variable (apart from the error term). Thus when the error terms are ignored, the normal form equations are *equivariant* with respect to rotations γ of the plane. This means that $X(\gamma(x))=D\gamma X(x)$. This symmetry allows one to search for periodic orbits by finding equilibria of the equation for the radial coordinate. A similar reduction here can be accomplished also by the *method of averaging* [46] or a *Lyapunov–Schmidt* procedure. From all three of these approaches, the fact that equations have an approximate circular symmetry allows the effective use of the reduction. The reduced equation for the radial coordinate has an odd right-hand side, this special form being a remnant of the circular symmetry. The role of such "internal symmetries" will be important in our dynamical analysis of codimension two bifurcations, because the normal forms can be chosen to preserve the rotational symmetries of their linear parts.

Let us now compute the normal form equations for the codimension two bifurcation problems. The results are summarized in §7. For the case of a two-dimensional nilpotent space, we shall only need to compute the degree two terms of the normal form. In $\mathbb{R}^2, H_2$ is six-dimensional and $\operatorname{ad}L$: $H_2 \to H_2$ is computed routinely. With coordinates (x_1,x_2), $L=x_2\partial/\partial x_1$, and $\operatorname{ad}L(x_1^2\partial/\partial x_1)=2x_1x_2\partial/\partial x_1$, $\operatorname{ad}L(x_1x_2\partial/\partial x_1)=x_2^2\partial/\partial x_1$, $\operatorname{ad}L(x_2^2\partial/\partial x_1)=0$, $\operatorname{ad}L(x_1^2\partial/\partial x_2)=2x_1x_2\partial/\partial x_2 - x_1^2\partial/\partial x_2$, $\operatorname{ad}L(x_1x_2\partial/\partial x_2)=x_1x_2\partial/\partial x_1$, $\operatorname{ad}L(x_2^2\partial/\partial x_2)=x_2^2\partial/\partial x_1$. We conclude that the image of $\operatorname{ad}L$ is four-dimensional with a complementary subspace spanned by $x_1^2\partial/\partial x_2$ and $x_1x_2\partial/\partial x_2$. Any vector field X with linear part L can be transformed via smooth coordinate changes to $x_2\partial/\partial x_1+(ax_1^2+bx_1x_2)\partial/\partial x_2+O(|x|^3)$ for some constants a and b. Linear rescaling of (x_1,x_2,t) (which may reverse time) allows us to fix $a=b=1$ if each is nonzero. Thus the general bifurcation problem involving a two-dimensional nilpotent space is reduced to studying the system of equations

$$\dot{x}_1=x_2, \qquad \dot{x}_2=x_1^2+x_1x_2$$

and its perturbations.

The next case on our list is (iv): L is a linear vector field on $\mathbb{R}^3$ with a pair of pure imaginary eigenvalues and a zero eigenvalue. With coordinates (x_1,x_2,x_3) we may assume $L=w(x_1\partial/\partial x_2-x_2\partial/\partial x_1)$. By introducing complex coordinates in the (x_1,x_2) plane, we write $L=iw(-z\partial/\partial z+\bar{z}\partial/\partial\bar{z})$ as for the Hopf case described above. In terms of the coordinates $(z,\bar{z},x_3)$, the vector fields with monomial coefficients are eigenvectors of $\operatorname{ad}L$. We have

$$\operatorname{ad}L\left(z^j\bar{z}^k x_3^l\frac{\partial}{\partial z}\right)=iw(k-j+1)\left(z^j\bar{z}^k x_3^l\frac{\partial}{\partial z}\right),$$
$$\operatorname{ad}L\left(z^j\bar{z}^k x_3^l\frac{\partial}{\partial \bar{z}}\right)=iw(k-j-1)\left(z^j\bar{z}^k z^l\frac{\partial}{\partial \bar{z}}\right) \quad \text{and}$$
$$\operatorname{ad}L\left(z^j\bar{z}^k x_3^l\frac{\partial}{\partial x_3}\right)=iw(k-j)\left(z^j\bar{z}^k x_3^l\frac{\partial}{\partial x_3}\right).$$

Thus, the complementary subspaces to $\operatorname{ad}L$ in the spaces H_m are spanned by polynomials in $z\bar{z}$ and x_j multiplied by the vector fields $z\partial/\partial z$, $\bar{z}\partial/\partial\bar{z}$, and $\partial/\partial x_3$. In particular, the normal form of degree two expressed in cylindrical coordinates is

$$\dot{\theta}=w+o(1), \quad \dot{r}=arx_3+o(2), \quad \dot{x}_3=bx_3^2+cr^2+o(2).$$

Once again the right-hand sides have θ dependence only in the remainder terms. We note also that the cubic terms will play an important role in the analysis of this bifurcation. The normal form of degree three is given by

$$\begin{aligned}&\dot{\theta}=w+dr^2+o(2),\\&\dot{r}=arx_3+er^3+frx_3^2+o(3),\\&\dot{x}_3=bx_3^2+cr^2+gr^2x_3+hx_3^2+o(3).\end{aligned}$$

The final computation of normal form we do is for a vector field X having two pairs of pure imaginary eigenvalues $\pm iw_1$ and $\pm iw_2$ for its linearization at the origin. We shall use coordinates (x_1,y_1,x_2,y_2) for $\mathbb{R}^4$. If we complexify each (x_i,y_i) plane, then the linear part of X can be expressed as

$$iw_1\left(-z_1\frac{\partial}{\partial z_1}+\bar{z}_1\frac{\partial}{\partial \bar{z}_1}\right)+iw_2\left(-z_2\frac{\partial}{\partial z_2}+\bar{z}_2\frac{\partial}{\partial \bar{z}_2}\right)=L.$$

Once again the monomial vector fields are eigenvectors of $\operatorname{ad} L$ on the spaces H_m

$$\operatorname{ad} L\left(z_1^{j_1}\bar{z}_1^{k_1}z_2^{j_2}\bar{z}_2^{k_2}\frac{\partial}{\partial\xi}\right)=\left[iw_1(k_1-j_1)+iw_2(k_2-j_2)+\delta\right]z_1^{j_1}\bar{z}_1^{k_1}z_2^{j_2}\bar{z}_2^{k_2}\frac{\partial}{\partial\xi}$$

where $\delta=iw_l$ if $\xi=z_l$ and $\delta=-iw_l$ if $\xi=\bar{z}_l$. The complementary subspace to the image of $\operatorname{ad} L$ is spanned by polynomials in $(z_1\bar{z}_1)$ and $(z_2\bar{z}_2)$ times the vector fields $z_1\,\partial/\partial z_1$, $\bar{z}_1\,\partial/\partial\bar{z}_1$, $z_2\,\partial/\partial z_2$ and $\bar{z}_2\,\partial/\partial\bar{z}_2$, *provided that w_1/w_2 is irrational.* If w_1/w_2 is rational, then extra *resonance* terms appear in the normal form.

In the nonresonance cases, the normal forms without the error terms are again separable. With polar coordinates in the (x_1,y_1) and (x_1,y_2) planes, the normal forms of degree k (k odd) become

$$\dot{\theta}_i=w_i+B_i(r_1^2,r_2^2)+o\left((r_1^2+r_2^2)^{(k-1)/2}\right), \qquad \dot{r}_i=r_iA_i(r_1^2,r_2^2)+o\left((r_1^2+r_2^2)^{k/2}\right),$$

where A_i and B_i are polynomials of degree $(k-1)/2$ with no constant term. The absence of θ_i dependence in the truncated equations for r_i means that this two-dimensional system can be investigated initially and then information about the four-dimensional system inferred from this. An analysis of the resonance cases requires that one deal initially with equations that do not separate readily so that there is an invariant planar subsystem, and will not be attempted in this review.

Before passing to the next step in our treatment of codimension two bifurcations, the computation of unfoldings, we briefly indicate the changes necessary to deal with systems possessing simple symmetries (in addition to the "internal" symmetries of the normal forms themselves). If G is a group of symmetries acting on $\mathbb{R}^n$, then the vector fields which are equivariant with respect to G form a subspace of the space of all vector fields. Accordingly, we can perform the computation of normal forms within the class of equivariant vector fields by restricting $\operatorname{ad} L$ to act on the subspaces of the H_l which are equivariant. One simple case which arises in the applications we consider occurs when there is a zero eigenvalue and a pair of pure imaginary eigenvalues. If this system and its perturbations are equivariant with respect to a reflection symmetry in the direction of the 0 eigenvalue, then all of the quadratic terms in the normal form disappear. The normal form in the class of equivariant vector fields looks much like that in the case of the pure imaginary eigenvalues, except that one of the angular variables is missing.

The next step in our analysis involves inserting two-dimensional parameters into the problems so that we obtain a transversal family to the surface on which the codimension two bifurcation occurs. For those problems (iii)–(v) of Theorem 3.1 which have a double degeneracy of their linearizations, we can work in the context of families of linear vector fields (perhaps with a constant term). There are the three cases (iii)–(v) to consider, with alternatives for (iii) and (iv) depending upon whether 0 is constrained to be an equilibrium.

Let us consider first the cases in which 0 is forced to remain an equilibrium. Here one has the space of linear vector fields on $\mathbb{R}^n$, represented by $n\times n$ matrices M_n. In M_n there is the real algebraic variety V consisting of matrices with eigenvalues having zero real parts. We want to study L for which there is a submanifold M of V containing L which has codimension 2 in M_n. In the three cases (iii)–(v), M consists of matrices with (iii) a 2×2 nilpotent Jordan block, (iv) a zero eigenvalue and a pair of pure imaginary eigenvalues, or (v) two distinct pairs of pure imaginary eigenvalues. We may assume that L is in Jordan normal form in each case, and then easily compute a transversal T: $\mathbb{R}^2\to M^n$

$$\text{(iii)}\quad L=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix},\qquad T(\lambda_1,\lambda_2)=\begin{pmatrix}0 & 1\\ \lambda_1 & \lambda_2\end{pmatrix},$$

$$\text{(iv)}\quad L=\begin{pmatrix}0 & -w & 0\\ w & 0 & 0\\ 0 & 0 & 0\end{pmatrix},\qquad T(\lambda_1,\lambda_2)=\begin{pmatrix}\lambda_1 & -w & 0\\ w & \lambda_1 & 0\\ 0 & 0 & \lambda_2\end{pmatrix},$$

$$\text{(v)}\quad L=\begin{pmatrix}0 & -w_1 & 0 & 0\\ w_1 & 0 & 0 & 0\\ 0 & 0 & 0 & -w_2\\ 0 & 0 & w_2 & 0\end{pmatrix},\qquad T(\lambda_1,\lambda_2)=\begin{pmatrix}\lambda_1 & -w & 0 & 0\\ \lambda_1 & \lambda_1 & 0 & 0\\ 0 & 0 & \lambda_2 & -w_2\\ 0 & 0 & -w_2 & \lambda_2\end{pmatrix}.$$

In the remaining cases that do not necessarily preserve 0 as an equilibrium, we consider the space of affine vector fields and the possibility of perturbations without equilibria. The transversals to the submanifolds M in the space of affine vector fields are

$$\text{(vi)}\qquad (T(\lambda_1,\lambda_2))\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}x_2\\ \lambda_1+\lambda_2x_1\end{pmatrix},$$

$$\text{(vii)}\qquad (T(\lambda_1,\lambda_2))\begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix}=\begin{pmatrix}\lambda_1x_1-wx_2\\ wx_1+\lambda_1x_2\\ \lambda_2\end{pmatrix}.$$

These transversals are combined with the normal form computations to give us the two-dimensional families of vector fields whose dynamics we now study.

4. Codimension two bifurcations II. Dynamics. We now turn to the analysis of the dynamics of the normal form equations. In each of the cases which have a double degeneracy in their linear part, the normal form equations contain a pair which are separated from the rest (involving angular variables) when remainder terms are ignored. Our initial dynamical studies focus upon these planar systems. We begin by finding the equilibria of these systems as a function of the parameter and computing their stability. This portion of the analysis is straightforward and requires mainly that care be taken in enumerating the various possibilities which arise in the cases with pure imaginary eigenvalues. The bifurcation diagrams in Figs. 7–13 contain the results of these calculations in terms of phase portraits superimposed on the bifurcation diagrams in the λ parameter plane. In the λ plane there are curves along which saddle node or other codimension one bifurcations of equilibria take place. The phase diagrams show the stability of equilibria using a solid circle for sinks, a cross for saddle points and an open circle for sources. We have not distinguished the difference between sinks and sources with real eigenvalues (nodes) and complex eigenvalues (foci). The crosses representing saddles have arrows indicating the directions of stable and unstable manifolds. In some cases, there are changes from sinks to sources, indicating the presence of Hopf bifurcations and periodic orbits for the two dimensional flows.

The computations themselves which underlie these diagrams can be illustrated by the case of a double zero eigenvalue. One begins with the system

$$\dot{x}_1=x_2, \qquad \dot{x}_2=\lambda_1+\lambda_2 x_1+x_1^2+x_1 x_2.$$

The equilibria are found by setting $\dot{x}_1=\dot{x}_2=0$ or $x_2=0$, $\lambda_1+\lambda_2 x_1+x_1^2+x_1 x_2=0$. The linearization of the equation at an equilibrium is defined by the matrix

$$L=\begin{pmatrix} 0 & 1 \\ \lambda_2+2x_1+x_2 & x_1 \end{pmatrix}.$$

Saddle nodes occur when there is a zero eigenvalue at an equilibrium. Eliminating x_1 and x_2 from the three equations $\dot{x}_1=\dot{x}_2=\det L=0$ gives the equation $\lambda_1 2-\lambda_2^2/4=0$. This is the curve of the saddle nodes in Fig. 7. The condition for Hopf bifurcations to occur is that $\dot{x}_1=\dot{x}_2=\operatorname{Trace} L=0$ together with $\det L>0$. These equations yield $x_1=x_2=\lambda_1=0$ together with $\lambda_2<0$. This yields the Hopf curve H of the bifurcation diagram in Fig. 7.

Note here a distinction between the case of a double zero eigenvalue and the cases with pure imaginary eigenvalues. In the case of a double zero eigenvalue, the normal form theorem together with linear rescaling of variables produce a unique equation whose unfolding was to be studied. In the cases with pure imaginary eigenvalues, these steps leave one with equations that still contain undetermined coefficients for higher order terms. For the unfolding families to be persistent, these coefficients must satisfy a number of inequalities. Several subcases for the dynamics of the equations are present, and these are determined by the higher order coefficients. In the case of a zero plus pure imaginary eigenvalues, there are four different subcases at this stage (after allowing for time reversal). In the case of two pairs of pure imaginary eigenvalues, there are many more subcases. The diagrams illustrate four of these, leaving the reader the task of enumerating the entire list.

The second part of the dynamical analysis involves finding all the periodic and homoclinic orbits for the two-dimensional systems. These occur for only some subcases when there are pure imaginary eigenvalues. The procedure here is more subtle. One begins by introducing a small parameter δ and rescaling so that as $\delta \to 0$, one approaches an *integrable* system. This integrable system is interpreted as a *blown-up* version of the original codimension two bifurcation. If the integrable system has periodic orbits, then these can be studied for small δ using a variational argument. The result of this calculation is an estimate of which parameter values correspond to the appearance of a periodic orbit of given size and shape. When the estimated function is a Morse function with nondegenerate critical points, then the dependence of periodic orbits on parameter values in our two-dimensional problems will be qualitatively the same. When the function is not monotonic, then multiple limit cycles appear in these systems. Note, for instance, the pair of limit cycles which occur for a case with a double zero eigenvalue and a symmetry of rotation by π.

The rescalings appropriate to each of the three cases are indicated in §7. The limit systems obtained when $\delta=0$ are integrable along the curve in the parameter plane for which there is an equilibrium having pure imaginary eigenvalues. This means that there is a function H: $\mathbb{R}^2\to\mathbb{R}$ constant along the trajectories of the limit equation for these parameter values. The compact level curves of H which do not contain critical points form a continuous family of periodic orbits for the vector field. The functions H for the different cases are recorded in the summary in §7. In the limit $\delta=0$, the Hopf bifurcations have become totally degenerate. Their periodic orbits will approximate the limit cycles which occur for $\delta>0$ small, but a variational calculation involving δ is needed to determine how these limit cycles depend upon the parameters and their stability. If one carries through this procedure for the Hopf bifurcation itself, then the limiting vector field obtained for $\delta=0$ is linear.

The variational argument for finding limit cycles is based upon the formula expressing the rate of change of area of a plane region R as it moves with a two-dimensional flow ϕ_t. An elementary computation shows that

$$\frac{d}{dt}(\operatorname{area}\phi_t(R))\bigg|_{t=0}=\int_R \operatorname{div} X,$$

where X is the vector field of ϕ_t and $\operatorname{div} X$ is the divergence of X. If ϕ_t has a closed orbit γ and R is the interior of γ, then $\int_R \operatorname{div} X=0$ since the area of $\phi_t(R)$ is constant. Therefore, the vanishing of the divergence is a necessary condition for γ to be a closed orbit of ϕ_t. We want to apply this formula to the situation in which there is a one-parameter family of vector fields X_δ with X_0 integrable. If R is the interior of a periodic orbit γ of X_0, then a necessary condition for γ to be the limit of a family of periodic orbits γ_δ for X_δ, $\delta>0$, is that

$$\frac{\partial}{\partial\delta}\left(\int_R \operatorname{div} X_\delta\right)=0.$$

The following theorem says that simple zeros for the last integral give sufficient conditions for the existence of a family of periodic orbits γ_δ.

THEOREM 4.1 [3]. *Let X_δ be a one-parameter family of planar vector fields such that X_0 has a continuous family of periodic orbits γ_s. Let R_s be the interior of γ_s and define the function $g(s)=(\partial/\partial\delta)(\int_{R_s}\operatorname{div} X_\delta)$. If $g(s_0)=0$ and $dg(s_0)/ds\neq 0$, then there is a $\bar{\delta}>0$ and a continuous family of closed curves β_δ, $\delta\in[0,\delta_0]$ such that $\beta_0=\gamma_{s_0}$ and β_δ is a limit cycle of X_δ for $\delta>0$. If $dg(s_0)/ds<0$, then β_s is stable. If $dg(s_0)/ds>0$, then β_δ is unstable.*

We illustrate this last result in one of our bifurcation problems. For the case of a double zero eigenvalue with rotational symmetry, the divergence integral is given by $\delta\int_{R_c}(\Lambda_2+a_1X_1^1)$, where R_c is the interior of a compact component γ of a level curve of $H(X_1,X_2)=-X_2^2/2+\Lambda_1X_1^2/2+a_1X_1^4/4=c$. For $\delta>0$, setting this integral equal to 0, gives an equation for the approximate value of Λ_2 for which γ lies near a limit cycle of the rescaled system

$$\frac{dX_1}{dT}=X_2, \qquad \frac{dX_2}{dT}=\Lambda_1 X_1+a_1 X_1^3+\delta\left(\Lambda_2 X_2+a_2 X_1^2 X_2\right).$$

The values of Λ_2 for which the divergence integral vanishes can be computed explicitly in terms of complete elliptic integrals. Integrating with respect to X_2 gives

$$\Lambda_2=-\frac{a_1 \int X_1^2\left(2c+\Lambda_2 X_1^2+a_1 X_1^4/2\right)^{1/2} dX_1}{\int\left(2c+\Lambda_1 X_1^2+a_1 X_1^4/2\right)^{1/2} dX_1},$$

where the limits of integration are roots of the polynomial $2c+\Lambda_1 X_1^1+a_1 X_1^4/2$. This function $\Lambda_2(c)$ is not monotone in the case $\Lambda_1>0$, $a_1<0$, but rather has a single critical point. There are values of Λ_2 for which there are periodic orbits corresponding to two different values of c. At the critical point of $\Lambda_2(c)$, these periodic orbits coalesce with one another in a saddle node of periodic orbits.

For the bifurcations involving imaginary eigenvalues, the results are more complicated. To begin with, the rescaled limit equations are integrable but not divergence free. Therefore, the computations are simplified by multiplying the rescaled limit by an *integrating* factor which makes them divergence free. The second additional complication is that there are more terms of order δ in the normal forms, and these involve truncating the normal forms with higher degree than was necessary for our earlier analysis of equilibria. Thus cubic terms must be retained in the normal form for the zero–pure imaginary bifurcation, and fifth degree terms must be retained in the normal form for the double Hopf bifurcation.

In the first of these two cases, the divergence of the cubic terms contains two coefficients that can be varied. In the second case, the divergence of the fifth degree terms is a homogeneous quadratic polynomial in (r_1^2, r_2^2) and has three coefficients. One can arrange by the correct choice of coefficients that the function $g(s)$ in Theorem 3.3. is not monotone. When the contributions of these higher order terms are then balanced against the contribution obtained from a small variation of parameter values, one obtains multiple limit cycles in the unfolding of the planar normal form equations. Section 7 includes these divergence integrals. It seems that they cannot usually be evaluated in closed form. However, genericity arguments suggest that most values of coefficients will give a Morse function describing the parameter variations necessary to maintain the different closed level curves of H as periodic orbits. When this happens, the unfolding behavior of the planar systems will be structurally stable. Further study of these divergence integrals and the corresponding geometry seems to be of interest.

The final feature of the unfolding behavior of these planar systems involves the disappearance of the periodic orbits as the parameters are varied. The smallest periodic orbits are associated with Hopf bifurcations, while the largest are associated with *homoclinic* or *heteroclinic* trajectories. The limit of the periodic orbits as they grow larger is approximated by level curves of the function H which contain critical points. For the corresponding integrable vector fields one has closed curves composed of saddle point equilibria and portions of their stable and unstable manifolds. Thus the disappearance of the limit cycles in our unfoldings as the periodic orbits grow in size is associated with their periods becoming unbounded. The limit cycles terminate along a curve for which one has saddle loop bifurcations of the type described in Theorem 2.6. In the cases with pure imaginary eigenvalues, internal symmetry of the equations forces these loops to contain more than one saddle point.

The following theorem summarizes the two-dimensional dynamic information which we have obtained thus far.

THEOREM 4.2. *Let X_λ be one of the following two-parameter families of planar vector fields and Ξ be the class of vector fields satisfying the indicated constraint or symmetry*:

(1) $y\partial/\partial x+(\lambda_1+\lambda_2 x+x^2+xy)\partial/\partial y$;

(2) $y\partial/\partial x+(\lambda_1 y+\lambda_2 x+x^2+xy)\partial/\partial y$, 0 *remains equilibrium*;

(3) $y\partial/\partial x+(\lambda_1 y+\lambda_2 x\pm x^3+x^2 y)\partial/\partial y$, *rotational symmetry in* π;

(4) $x(\lambda_1+ay+bx^2+cy^2)\partial/\partial x+(\lambda_2+dx^2+ey^2+fx^2y+gy^3)\partial/\partial y$, *reflection symmetry in x-direction*;

(5) $x(\lambda_1+ay+bx^2+cy^2)\partial/\partial x+y(\lambda_2+dx^2+ey^2+fx^2y+gy^3)\partial/\partial y$, *reflection symmetry in x direction* $+0$ *remains equilibrium*;

(6) $x(\lambda_1+ax^2+by^2+cx^4+dx^2y^2+ey^4)\partial/\partial x+y(\lambda_2+fx^2+gy^2+hx^4+ix^2+jy^4)\partial/\partial y$, *reflection in the x and y directions.*

For almost all values of the coefficients $(a,b,c,\cdots,j)$ *the families* (1)–(6) *are structurally stable within the class of two-parameter families of vector fields in the indicated class* σ, *provided that the variational integrals listed in* §7 *define Morse functions. Diagrams for regions of structural stability in the* (λ_1,λ_2) *plane for cases* (1)–(5) *and selected cases of* (6) *are illustrated in Figs.* 3–9 (*apart from indicating the variation of limit cycles with parameters where limit cycles occur in a way that depends on two or three coefficients*).

The final results in this section involve the interpretation of cases (4)–(6) of Theorem 4.2 in terms of codimension two bifurcations which have pure imaginary eigenvalues. Cases (4) and (5) apply to bifurcations with zero and pure imaginary eigenvalues, with (5) pertaining to the constrained situation in which 0 always remains an equilibrium. Here x plays the role of the radial coordinate and y plays the role of the axial coordinate in a cylindrical coordinate system adapted to the problem. Case (6) applies both to the double Hopf bifurcation and to the bifurcation with zero and pure imaginary eigenvalues when there is a reflection symmetry in the direction of the zero eigenvector.

To draw pictures of the phase portraits corresponding to the two-dimensional system (4)–(6) of Theorem 4.2, when there are imaginary eigenvalues we must reintroduce the angular variables which have been ignored to this point. For the case of one zero and one pair of pure imaginary eigenvalues, each point in the interior of the right

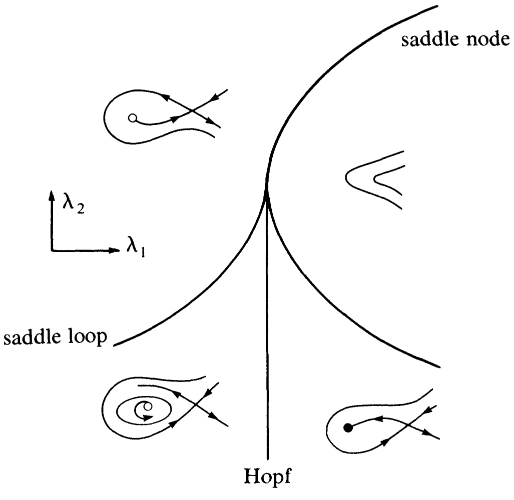


FIG. 3. *Stability diagram*; *double zero eigenvalue*. (*Theorem* 4.2(1).)

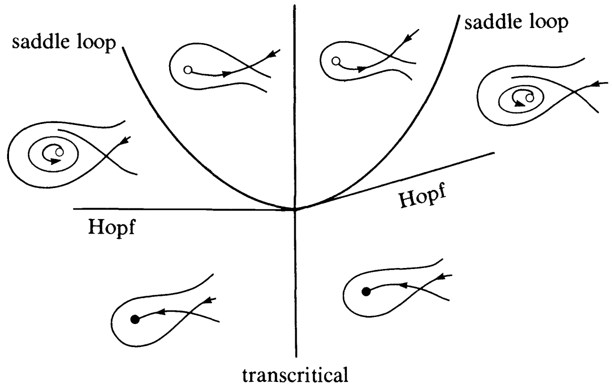


FIG. 4. *Stability diagram*; *double zero eigenvalue, trivial equilibrium.* (*Theorem* 4.2(2).)

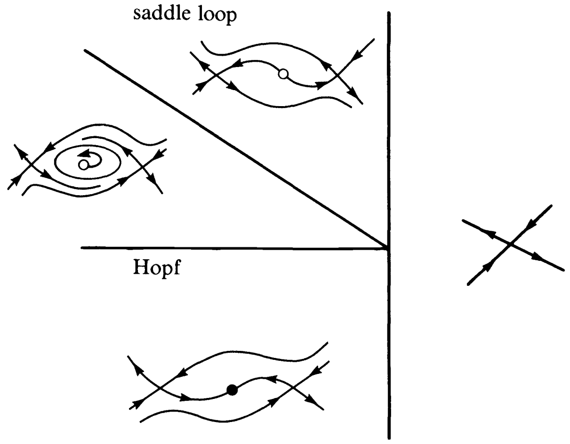


FIG. 5. *Stability diagram*; *double zero eigenvalue, symmetry.* (*Theorem* 4.2(3+).)

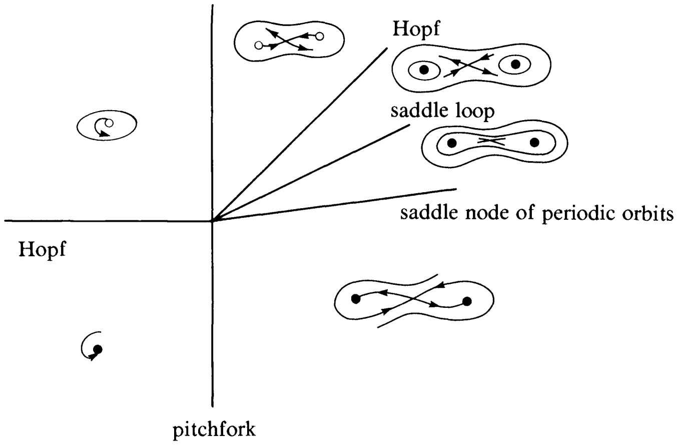


FIG. 6. *Stability diagram*; *double zero eigenvalue, symmetry.* (*Theorem* 4.2(3−).)

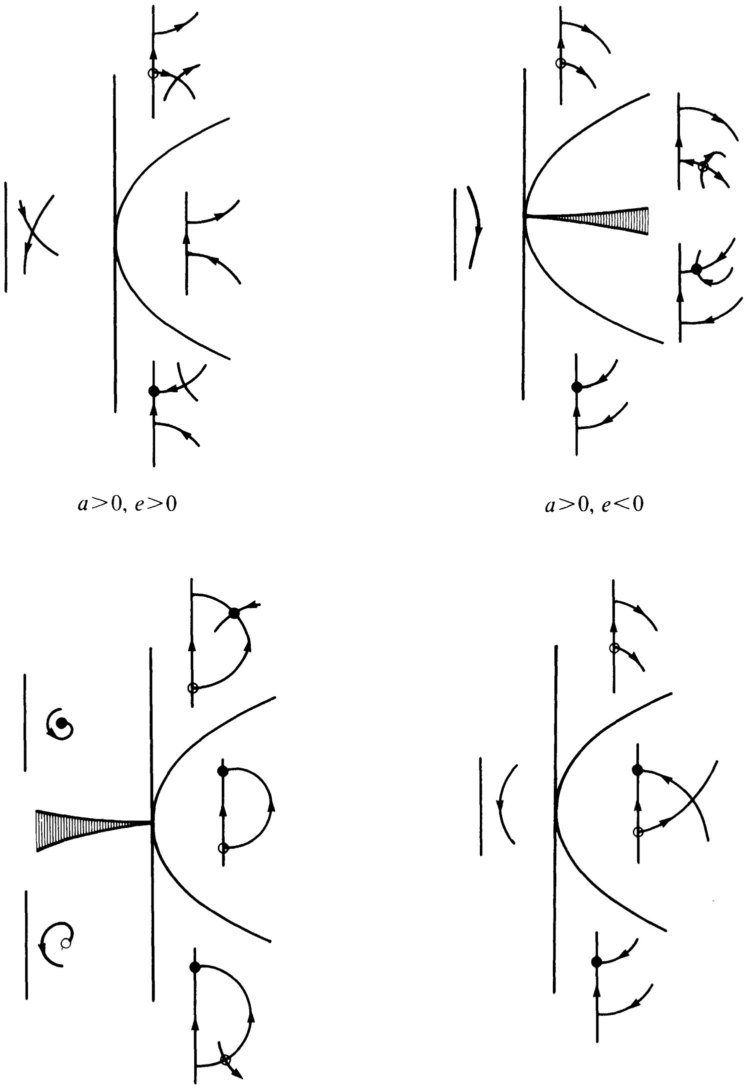


FIG. 7. *Stability diagrams*, 0+*pure imaginary eigenvalues*. (*Theorem* 4.2(4).)

(x,y) half plane corresponds to a circle in $\mathbb{R}^3$. Equilibria of the planar systems which lie off the y axis represent periodic orbits of the three-dimensional system, and periodic orbits of the planar system give rise to invariant two-dimensional tori in $\mathbb{R}^3$. Of special interest in the next section will be the three-dimensional flows which correspond to saddle loops for the two-dimensional flows. For the flow depicted in Fig. 1, one obtains three-dimensional flows with an invariant set consisting of the surface of a two-dimensional sphere together with a diameter joining two antipodal points. This invariant set is attracting from the interior of the sphere.

For the double Hopf flows there are two angular coordinates. In the flow of Theorem 4.2(6), nonzero points on the boundary of the positive quadrant correspond to circles in $\mathbb{R}^4$ while points in the interior of the quadrant represent two-dimensional

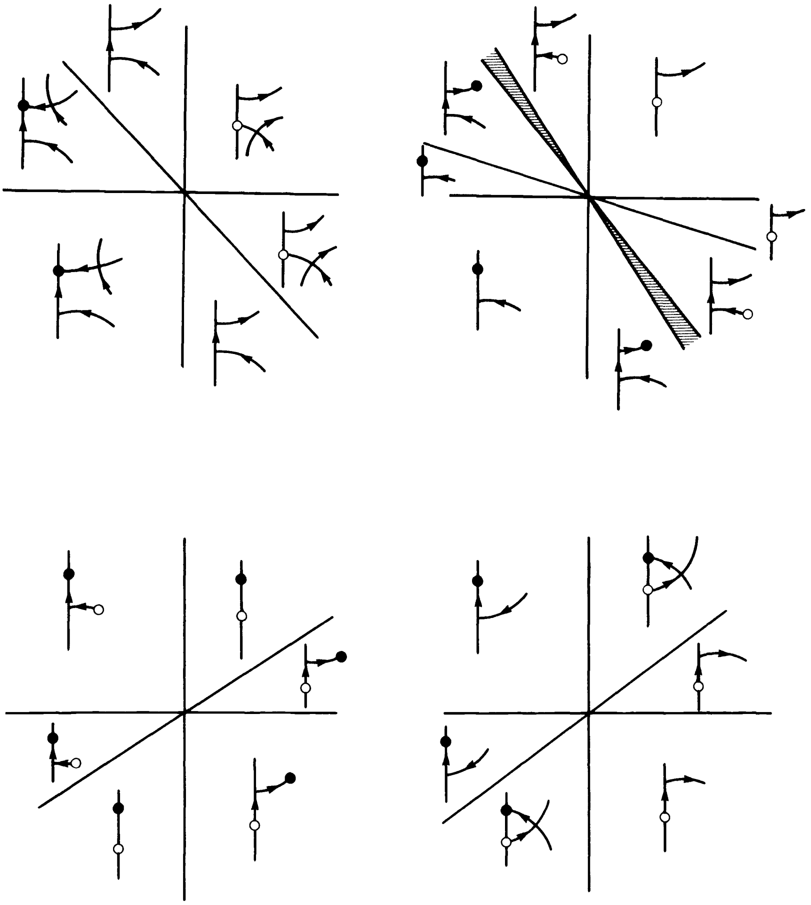
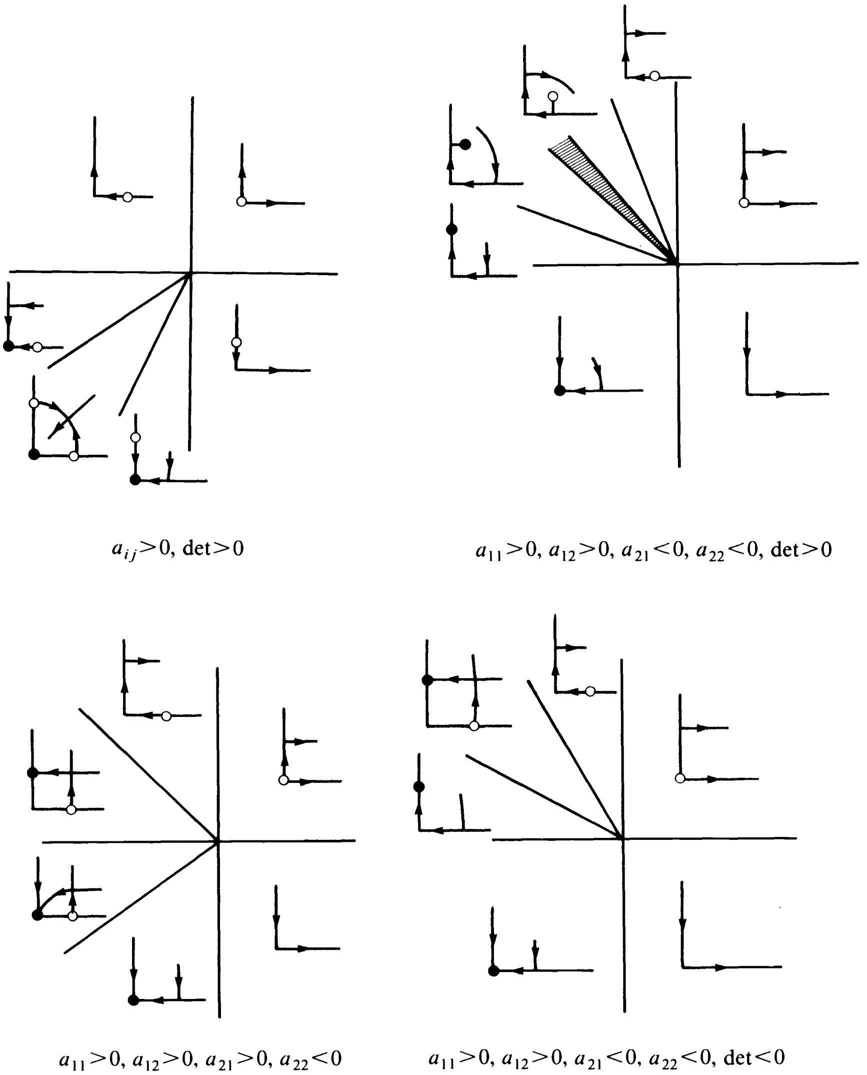


FIG. 8. *Stability diagrams,* 0+*pure imaginary eigenvalues trivial equilibrium.* (*Theorem* 4.2(6).)

tori in $\mathbb{R}^4$. These limit cycles of (6) correspond to three-dimensional invariant tori in $\mathbb{R}^4$.

We close this section with a brief discussion of the structural stability of the unfoldings of codimension two bifurcations deduced from the two-dimensional analysis described above. If the remainder terms of the normal forms are ignored, then the normal forms themselves have "internal symmetry". They are equivariant with respect to rotations in the plane of a pair of imaginary eigenvalues. If one restricts attention to classes Ξ of vector fields which possess these rotational symmetries, then the bifurcation diagrams represent persistent unfoldings of the corresponding codimension two bifurcations in the sense that perturbations of the family have homeomorphic bifurcation diagrams.

THEOREM 4.3. *Let Ξ be the class of vector fields in $\mathbb{R}^3$ which are equivariant with respect to rotations around the x_3 axis. Then there is a* $1-1$ *correspondence between the universal unfolding of codimension two bifurcations in Ξ of vector fields with a zero eigenvalue and a pair of pure imaginary eigenvalues and the persistent unfoldings of Theorem* 4.2(4). *If one further restricts Ξ to consist of vector fields with an equilibrium at the origin or an additional reflectional symmetry, then the correspondence is with Theorem* 4.2(5) *of Theorem* 4.2(6).



$a_{ij}>0$, det>0

$a_{11}>0$, $a_{12}>0$, $a_{21}<0$, $a_{22}<0$, det>0

$a_{11}>0$, $a_{12}>0$, $a_{21}>0$, $a_{22}<0$

$a_{11}>0$, $a_{12}>0$, $a_{21}<0$, $a_{22}<0$, det<0

FIG. 9. *Stability diagrams, 2 pairs pure imaginary eigenvalues.*

THEOREM 4.4. *Let Ξ be the class of vector fields on $\mathbb{R}^4$ which are equivariant with respect to rotations in two orthogonal planes. There is a* $1-1$ *correspondence between the persistent codimension two unfoldings in Ξ of vector fields with two pairs of pure imaginary eigenvalues and the persistent families in Theorem* 4.2(6).

When the equivariance assumptions in Theorems 4.3 and 4.4 are relaxed, the dynamic behavior of perturbations of the degenerate vector fields we have been studying can be much more complicated than the phenomena we have considered thus far. Equivariance forces the flows on invariant tori in the problems to be periodic or quasiperiodic depending upon the *rotation* numbers which measure the average rates of flow around different directions on the torus. Periodic flow on a torus is a highly unstable phenomenon in the absence of equivariance. Even though the nonequivariant portion is of arbitrarily high order in the Taylor series of the original vector field (of ∞ order in a C^∞ context!), the lack of equivariance causes substantial changes in the qualitative behavior of unfoldings. Another qualitative change which comes with the departure from equivariance is the occurrence of transversal homoclinic orbits in some unfoldings. These dynamic phenomena are the subject of the next section.

5. Codimension two bifurcations III. Resonance phenomena. In §4, we carried through an essentially complete analysis of persistent codimension two bifurcations of equilibria for classes of vector fields which possessed angular symmetries. Departures from angular symmetry are associated with more complex dynamical phenomena than those discussed in the previous sections. These new dynamical features are the subject of this section. A thorough discussion of these problems is beyond the scope of this review and strains the capabilities of the underlying mathematical theory. Therefore, we shall confine ourselves to a description of the nature of these phenomena without attempting to give an exhaustive treatment of their qualitative dynamics. Many details remain to be further elaborated before a fully developed mathematical picture can be presented.

There are two kinds of new phenomena that we describe. In some of the cases involving pure imaginary eigenvalues discussed in §4, there were flows which have invariant tori of two and three dimensions. The internal symmetry properties of the flows imply that the dynamics of the motion on these tori is either periodic or quasiperiodic. The basic issue for us is what kinds of new dynamics occur when these flows are perturbed. The technical questions can be split into ones which involve (1) the persistence of invariant tori, and (2) the kinds of new qualitative dynamics which occur. New dynamics may occur either on an invariant torus or may be associated with the destruction of an invariant torus. Considerable attention has been focussed upon questions of this sort in relation to "mild turbulence" of fluids and the instabilities of plasmas. For example, the Ruelle–Takens theory of turbulence [103] is based upon the observation that the instabilities of periodic or quasiperiodic motion on an invariant torus of dimension larger than two are incompatible with older theories of turbulence [76]. This review is an inappropriate place for an extended discussion of these issues of "chaotic" motion and turbulence, but the reader may wish to pursue their relationship to the bifurcation phenomena described here.

The persistence of invariant tori with nonsymmetric perturbations can be approached in two different ways. When the unfolding parameters are included as variables, the invariant tori in the symmetric unfoldings occur as two parameter families of invariant tori (of dimension two or three depending upon the problem). Some of these tori will typically have periodic flow and some will have quasiperiodic flow. Without symmetry, transversality arguments preclude the existence of nonisolated periodic orbits in any flow contained within a structurally stable finite dimensional family. Thus, it is reasonable to ask questions about the persistence of tori in one of the following two forms:

(a) Does the whole family of invariant tori persist with nonsymmetric perturbations (without regard to the dynamics of these tori)?

(b) Does an individual quasiperiodic invariant torus persist with nonsymmetric perturbations so that the flow on the perturbed torus remains quasiperiodic?

The analytic techniques for answering these two questions are quite different from one another and each involves its own complications.

The easiest problem in which question (a) above arises is the Hopf bifurcation of periodic orbits described in §2. There one is concerned with a family of two-dimensional invariant tori bifurcating from a periodic orbit as a pair of complex eigenvalues crosses the unit circle. This family exists, but with only a finite degree of smoothness which typically decreases with the distance from the bifurcation. The technique used for proving this uses the idea of a *graph transform*. The same approach can be used for at least some of the codimension two bifurcation problems considered here to deduce the persistence of parts of the families of invariant tori with nonsymmetric perturbations [64].

The graph transform method applies generally to prove the persistence of *normally hyperbolic* invariant submanifolds M of a flow. It proceeds by describing a Banach space E of maps on M whose graphs yield all perturbations of M (with the desired degree of smoothness and continuity). If ψ is a perturbation of the flow of ϕ and $t>0$, then ψ_t maps one perturbation M to another $\psi_t(\tilde{M})$. If care is used in selecting E and $\tilde{M}$ is the graph of a small $\xi\in E$, then $\psi_t(\tilde{M})$ will also be the graph of an element of E, denoted $\Gamma_\psi(\xi)$ and called the *graph transform* of ξ. Fixed points of Γ_ψ near $0\in E$ are maps whose graphs are invariant manifolds of ψ near M. To apply the method of graph transforms one tries to pick E so that the graph transform Γ_ϕ has a hyperbolic fixed point at $0\in E$. Then the implicit function theorem implies that this fixed point is isolated and varies continuously with perturbations of Γ_ϕ: $E\to E$. In particular, if ψ_t is near ϕ_t, then Γ_{ψ_t} will have a unique fixed point near zero.

The applicability of the graph transform method requires that assumptions be made which relate properties of the flow in M to properties of the flow in directions normal to M. Roughly speaking, constructing a space E of C^r maps for which the technique works requires a hypothesis guaranteeing that if there is an expansion or contraction of trajectories inside M at the rate $\exp(\lambda t)$, then all of the normal directions to M split into those for which the flow is expanding or contracting at a rate more extreme than $\exp(r\lambda t)$. In bifurcation problems having invariant tori, the normal hyperbolicity depends upon the parameters, and it becomes weaker as one approaches the collapse of these tori. However, at the same time that the normal hyperbolicity becomes weaker, the flow on the torus itself approaches periodic/quasiperiodic flow. The application of the graph transform method requires that one estimate the relative rates at which these two things are happening. When the normal hyperbolicity becomes weaker at a slower rate than the flow on the torus approaches periodic/quasiperiodic flow, then the method works. Iooss and Langford [64] have successfully carried through these calculations for some of the codimension two bifurcations. They introduce a number of small parameters and use successive rescalings to pinpoint those tori to which they apply the technique.

The graph transform method does not enable one to determine the dynamics of the nonsymmetric flow on the family of perturbed tori. Also it is limited to proving a finite degree of differentiability for the perturbed tori. At the expense of focussing upon individual tori and introducing still greater technical complexity, *small divisor* methods provide an alternate approach which surmounts these difficulties. This technique is based upon the work of Siegel, Kolmogorov, Arnold and Moser and is often presented in terms of *hard implicit function theorems*. There are a number of excellent references for this analysis [51], so we do little more than describe the relevant results.

The n-dimensional torus can be regarded as a set of points in $\mathbb{R}^n$ whose components all differ by integers: $T^n=\mathbb{R}^n/Z^n$. A flow on T^n defines a flow on $\mathbb{R}^n$ by means of this identification. If $x(t)$ is the lifted trajectory of such a flow on $\mathbb{R}^n$ with $x(0)=x$, then $\lim_{t\to\infty}(1/t)(x(t)-x(0))=\rho(x)$ exists and is called the *rotation vector* of $x(t)$. The rotation vector measures the average rate of increase of each angular component of the torus along the trajectory. If the flow on the torus is quasiperiodic or periodic, then ρ is independent of x. For periodic flows, $\rho=(\rho_1,\cdots,\rho_n)$ is a vector with ρ_i/ρ_j rational for all i and j. The small divisor methods begin with tori whose rotation vector is strongly irrational. One wants all of the ratios ρ_i/ρ_j to be irrational numbers which satisfy arithmetic conditions indicating that linear combinations of the ρ_i/ρ_j with integer coefficients are poorly approximated by rational numbers. For such a torus T of the original symmetric flow, one seeks an invariant torus $\tilde{T}$ on which the perturbed nonsymmetric flow is quasiperiodic and still has the same ratios ρ_i/ρ_j. The strategy of locating $\tilde{T}$ involves finding a smooth coordinate transformation from T to $\tilde{T}$ which carries the symmetric vector field on T to a multiple of the nonsymmetric vector field on $\tilde{T}$.

Formal expressions for the coordinate transformation from T to $\tilde{T}$ can be computed using Fourier series, but the convergence of these formal expressions is difficult to prove. They involve *small divisors*, linear combinations $\Sigma_{i=1}^n a_i\rho_i$, with integer coefficients a_i, which appear in the denominators of the Fourier coefficients of the coordinate transformation. Convergence requires hypotheses on how small these divisors can be in terms of the size of the coefficient vectors $(a_1,\cdots,a_n)$. In addition to these arithmetic conditions on rotation vectors, there can be difficulty in achieving the freedom necessary to solve the equations which give the constant terms in the Fourier series for the coordinate transformations. In our context, there are two necessary hypotheses. The first is that there be a whole family of tori in which the rotation numbers $\rho_2/\rho_1,\cdots,\rho_n/\rho_1$ vary in a nonsingular manner. Without this hypothesis perturbations of the symmetric family which contain no torus with the original rotation numbers might be possible. The second hypothesis requires that the tori be normally hyperbolic. Without this assumption or something which replaces it, perturbations of the symmetric family which destroy the whole set of invariant tori might be possible. These two sets of additional hypotheses are satisfied for most choices of higher order terms of the normal forms in the codimension two bifurcations discussed in this review. The arithmetic conditions are satisfied by almost all rotation vectors (in the sense of Lebesgue measure).

One expects those invariant tori for the symmetric problem consisting of periodic orbits to have their dynamics greatly altered by a nonsymmetric perturbation. If there is a continuous family of invariant tori, one expects an open dense set of these to have hyperbolic periodic orbits in the absence of symmetry. Thus quasiperiodic motion is only to be expected on a nowhere dense set of invariant tori. From this topological point of view, the typical parameter value in a nonsymmetric family will not yield a flow with a quasiperiodic invariant torus. Nevertheless, the set of parameter values which do yield a quasiperiodic invariant torus is likely to have positive Lebesgue measure in the parameter space. If one picks a parameter value at random (with respect to Lebesgue measure), then there is a positive probability that it will lie in the (nowhere dense) set of parameter values for which the corresponding flows have quasiperiodic invariant tori.

There are additional new dynamical phenomena which occur in nonsymmetric unfoldings of codimension two bifurcations besides invariant tori with hyperbolic periodic orbits. In particular, transversal homoclinic orbits appear. These orbits are generally associated with "chaotic" motion in dynamical systems and with "sensitive dependence to initial conditions." Here we shall emphasize the nature of transversal homoclinic orbits, describe some of the implications of their existence, and explore how they arise in nonsymmetric unfoldings of some codimension two equilibria. Questions about the full extent of the limit sets which contain the homoclinic phenomena we describe will not be considered. In particular, questions about the stability and structural stability of these sets are left aside. Once again, there is much scope for additional work in this area, and our current knowledge is fragmentary.

DEFINITION. Let γ be a hyperbolic periodic orbit for a flow in $\mathbb{R}^n$ with stable manifold $W^s(\gamma)$ and unstable manifold $W^u(\gamma)$. A *homoclinic* orbit for γ is a trajectory $\delta \subset W^u(\gamma) \cap W^s(\gamma)$ different from γ. If $W^u(\gamma)$ and $W^s(\gamma)$ intersect transversally along δ, then δ is a *transversal homoclinic* orbit.

Note that $\dim W^u(\gamma) + \dim W^s(\gamma) = n+1$, so that transversal homoclinic orbits are possible. Note also that the stable manifold theorem implies that the set of vector fields with transversal homoclinic orbits is an open set in the set of all smooth vector fields. A basic feature of transversal homoclinic orbits is that they imply the existence of larger sets of trajectories which have hyperbolicity and recurrence properties. Inside these sets, there are stable directions along which trajectories approach one another and unstable directions along which trajectories diverge. This instability within the set leads to sensitive dependence on initial values and an unpredictability about the long term behavior of individual trajectories. The prototypes of these hyperbolic invariant sets are *Smale's horseshoe* and the more abstract *subshifts of finite type*, both expressed in terms of discrete systems. As usual, one should interpret these models as return maps of a cross-section to a flow.

We shall describe subshifts of finite type in a manner suitable for application to our bifurcation problems. Consider a flow ϕ_t: $\mathbb{R}^n \to \mathbb{R}^n$ and a finite number of disjoint cross-sections $R_1, \cdots, R_m$ to ϕ_t. Each R_i will be called a *rectangle*, and the following hypotheses are made.

(M1) Each R_i has a continuous product structure $R_i = E_i^u \times E_i^s$ with E_i^u and E_i^s compact and homeomorphic to disks. Denote by $E_i^u(x)$ the set of $y \in R_i$ with the same E_i^s coordinate as x. $E_i^s(x)$ is defined similarly.

(M2) If A: $\cup_i R_i \to \cup_j R_j$ is the map which sends $x \in R_i$ to the first intersection of $\{x(t)|t>0\}$ with a rectangle (when this exists), then $A(x) \in R_j$ implies that $A(E_i^s(x)) \subset E_j^s(A(x))$ and $A(E_i^u(x)) \supset E_j^u(A(x))$.

(M3) There is a metric d on $\cup_i R_i$ and a constant $\lambda > 1$ with the property that $y \in E_i^u(x)$ implies $\lambda d(x,y) < d(A(x),A(y))$ and $y \in E_i^s(x)$ implies $\lambda d(A(x),A(y)) < d(x,y)$.

(M4) There is $l>0$ such that $A^l(R_i) \cap R_j \neq \varnothing$ for all i,j.

It is far from easy to verify that cross-sections satisfying (M1)–(M4) exist for a given flow, but the consequences are far reaching. Assuming that there is more than one rectangle in our collection, we want to examine the set $\Lambda = \{x | A^i(x)$ is defined for all $j \in Z\}$. This means that $x(t)$ intersects $\cup_j R_j$ an infinite number of times for $t \to +\infty$ and $t \to -\infty$. We will give the elements of Λ a convenient description as a *subshift or finite type*. This process is called *symbolic dynamics* and the sets R_i satisfying (M1)–(M4) constitute a *Markov partition* for Λ.

If $x,y \in \Lambda$ are distinct, then (M3) together with the disjointness and compactness of the R_j implies that there is i such that $A^i(x)$ and $A^i(y)$ lie in different rectangles. Therefore $x \in \Lambda$ is uniquely specified by the sequence $\{a^i\}_{i=-\infty}^{\infty} = a(x)$ defined by the property that $A^i(x) \in R_{a_i}$. Each a_i lies in the index set $\{1, \cdots, m\}$ for the collection of rectangles and is called the *ith address* or *ith symbol* of x. The symbol sequences $a(x)$ preserve much of the information about the dynamics of ϕ_t and A because applying A to x corresponds to shifting indices. More precisely, if $\mathbf{a} = \mathbf{a}(x)$, the symbol sequence of $A(x)$ is the sequence $\mathbf{b}$ with $b_i = a_{i+1}$. Thus we can use symbolic dynamics to obtain a qualitative characterization of the set Λ and the dynamics of Λ.

Define the *transition matrix* $T=(t_{jk})$ for $(A,R_1,\cdots,R_m)$ to be the $m\times m$ matrix with $t_{jk}=0$ if $R_j\cap A(R_k)=\varnothing$ and $t_{jk}=1$ if $R_j\cap A(R_k)\neq\varnothing$. We denote by σ the set of bi-infinite sequences $\{a_i\}_{i=-\infty}^{\infty}$ of the symbols $\{1,\cdots,m\}$ which satisfy $t_{a_ia_{i+1}}=1$ for all i. If we define a metric on the set of sequences by $a(a,b)=\Sigma_{i=-\infty}^{\infty}\delta_i2^{-|i|}$, $\delta_i=0$ or 1 as $a_i=b_i$ or $a_i\neq b_i$, then σ is a compact metric space. Together with the shift map σ: $\sigma\to\sigma$ which shifts indices one unit, σ is called the *subshift of finite type* with transition matrix T.

THEOREM 5.1. *Let* ϕ_t: $\mathbb{R}^n\to\mathbb{R}^n$ *be a flow which has cross-sections* $R_i,\cdots,R_m$ *with return map A satisfying* (M1)–(M4). *Define the transition matrix* $T=(t_{jk})$ *by* $t_{jk}=0$ *or* 1 *as* $R_j\cap A(R_k)$ *is empty or not. Denote by* Λ *the set of points whose trajectories intersect* $\bigcup R_i$ *an infinite number of times as* $t\to\infty$ *and* $t\to+\infty$. *Then the symbolic dynamics of A establish a homeomorphism h between* Λ *and the subshift of finite type* σ *with transition matrix T. The homeomorphism h carries A to the shift map* $\sigma=hAh^{-1}$.

The essence of this theorem is contained in Smale [112], where he treats the case $m=2$ and $T=(\begin{smallmatrix}1&1\\1&1\end{smallmatrix})$ (the *horseshoe*). Hypothesis (M2) plays an important role in establishing that the map h is onto, and, consequently, that the set Λ will be large. We note also that it is easy to establish a number of interesting dynamical properties for σ such as the existence of dense orbits, the density of periodic orbits, and sensitivity to initial conditions. These are carried back to the set Λ by the map h. Smale [112] also relates the concepts of subshifts of finite type and transversal homoclinic orbits.

THEOREM 5.2 [112]. *If f*: $M\to M$ *is a smooth invertible map defining a discrete dynamical system, and if p is a fixed point of f which has a transversal homoclinic orbit, then there is an iterate* f^n *of f and two sets* R_1 *and* R_2 *for which* (M1)–(M4) *are satisfied by the map* $A=f^n$.

COROLLARY 5.3. *A discrete dynamical system f*: $M\to M$ *has a transversal homoclinic orbit if and only if there is a set* $\Lambda\subset M$ *such that the symbolic dynamics of* $f|\Lambda$ *form a subshift of finite type.*

We make a few remarks about these results for discrete systems before returning to continuous flow. First, the sets Λ which they locate may be contained in larger invariant sets of the same type. It is unlikely that the construction we have outlined will determine a maximal invariant set Λ which is topologically transitive (has a dense orbit). Whether a maximal topologically transitive set Γ is an *attractor* is an important practical issue, but it is not easy to determine for examples that a set Γ is both an attractor and that is satisfies the hyperbolicity conditions implicit in (M1)–(M4) (Smale's Axiom A). The second remark is that the topology of the sets Λ identified above is relatively simple. The disjointness of the R_i required in (M1) forces the sets Λ to be homeomorphic to *Cantor* sets. A maximal topologically transitive set Γ may have a much more complicated topological structure, and the definition of Markov partition must be modified to allow for this possibility.

From the point of view of bifurcation theory for discrete systems, the transition from a system which has tranversal homoclinic orbits to one which does not is a complicated story. Our knowledge about this transition is woefully inadequate but steadily growing. Numerical studies such as those of Hénon [49] find large sets which behave as attractors, but the only theoretical evidence indicating their existence comes from the study of one-dimensional mappings. Newhouse [89] has proved some remarkable results which show that systems having an infinite number of stable periodic orbits are a persistent feature of the transition to transversal homoclinic behavior. Finally, we remark that "universality" properties have been found when homoclinic behavior first appears through an infinite sequence of flip bifurcations of stable periodic orbits. Feigenbaum [32] first observed these features in the context of one dimensional mappings, but a body of numerical, theoretical and experimental evidence indicates that they are widespread.

Thus far we have discussed transversal homoclinic behavior in the context of discrete systems or cross-sections to continuous flows. When a continuous flow admits a global cross-section (such as in forced oscillation problems), the translation from discrete geometry to that of the continuous system is relatively straightforward. The analogue of a subshift of finite type is a *special flow* and it has the topology of a *solenoid*. This is a one dimensional object which locally is the product of an interval and a Cantor set. As the much studied Lorenz system [80] illustrates, the geometry of homoclinic phenomena in continuous systems can be quite surprising when there are equilibria in the flows and global cross-sections do not exist.

The most immediate appearance of transversal homoclinic orbits in codimension two bifurcation problems does involve equilibria. Theorems 4.3 and 4.4 imply that it is also a resonance phenomemon in that the only homoclinic orbits in symmetric vector fields are not transversal. Thus transversal homoclinic orbits only appear in unfoldings which break symmetries involving the angular variables of the normal forms. We shall pick one case as an illustration of how the homoclinic behavior can be established. A systematic theory of the extent of transversal homoclinic behavior in this or other cases remains incomplete.

Consider the unfolding of a vector field with an equilibrium at which there is a zero eigenvalue and a pair of pure imaginary eigenvalues. Assume further that the quadratic terms of the normal form leave us with the following normal form equations after linear rescaling.

$$\dot{\theta}=w+o(1), \quad \dot{r}=r(\lambda_2+az)+o(2), \quad \dot{z}=\lambda_1-z^2-r^2+o(2)$$

with $a>0$. The normal form equations truncated at terms of order two and having parametric values along the curve $\lambda_2=0$, $\lambda_2>0$ have a flow which has a family of invariant tori which are level curves of the function $H(\theta,r,z)=ar^{2/a}/2(\lambda_1-r^2/(a+1)-z^2)$. The curve $H=0$ consists of the z axis together with the ellipsoid E defined by $\lambda_1-r^2/(a+1)-z^2=0$. The points $(0,0,\pm(\lambda_1)^{1/2})$ are hyperbolic equilibria $p_\pm$. Here $W^u(p_+)=E-\{p_-\}$ and $W^s(p_-)=E-\{p_+\}$. Both $W^s(p_+)$ and $W^u(p_-)$ are rays on the z axis, overlapping in the segment (p_+,p_-) interior to E; see Fig. 10. We shall

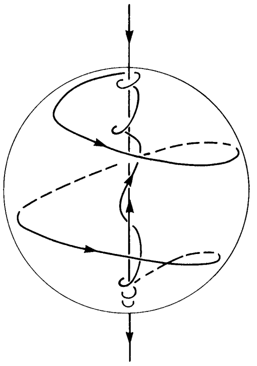


FIG. 10. *Flow of saddle loops in unfolding of* 0 + *pure imaginary eigenvalues vector field* ($a_2>0$, $a_3<0$, $a_4<0$ *in normal form*).

argue, using a theorem of Sil'nikov [111], that if $a<2$ in the normal form equation, then generic nonsymmetric perturbations of the family represented by the normal form will have transversal homoclinic orbits.

Consider trajectories near $W^u(p_+)\cap W^s(p_-)$ for the symmetric flows. If one fixes $\lambda_1>0$ and varies λ_2 near 0 then these trajectories pass through a transition in which they change from having bounded forward trajectories which remain inside E and approach a limit cycle to trajectories which become unbounded with $z\to-\infty$. This behavior persists for nonsymmetric perturbations of the family. In nonsymmetric perturbations, $W^u(p_+)\cap W^s(p_-)$ is usually empty. However, as the parameter λ_2 is varied $W^u(p_-)$ itself will undergo the transition described above, and it can only do so by lying in $W^s(p_-)$. In other words, there will be a curve in the parameter plane of the nonsymmetric systems for which p_- has a homoclinic trajectory.

THEOREM 5.4 [111]. *Let X be a three-dimensional vector field which has a hyperbolic equilibrium p_- for which the following hypotheses are satisfied*:

(1) *The linearization of X at p_- has two complex eigenvalues $\mu,\bar{\mu}$ and one real eigenvalue λ with $0<-\operatorname{Re}\mu<\lambda$.*

(2) *The equilibrium p_- has a homoclinic trajectory.*

Then within the space of C^1 vector fields satisfying (1) *and* (2), *there is a dense, open set which has transversal homoclinic orbits.*

This theorem of Sil'nikov can be visualized in terms of Fig. 11. Without leaving the set of vector fields with a homoclinic trajectory, we may linearize X at p_- by a perturbation, so we assume that X is linear in a neighborhood of p_-. If the eigenvalues of X at p_- are $\lambda,\gamma\pm iw$, the flow of X in cylindrical coordinates is given by $(r(t),\theta(t),z(t))=(r(0)e^{\gamma t},\ \theta(0)+tw, z(0)e^{\lambda t})$. Pick two cross-sections to the homoclinic orbit near p_-, the first M_1 contained in a cylinder $r=\rho$ and the second M_2 contained in a plane $z=\xi$. We want to compute the mapping g: $M_1\to M_2$ along trajectories. To do so, set $z(t)=\xi$ and $r(t)=\rho\exp(\gamma\lambda^{-1}\ln(\xi/z(0)))$ and $\theta(t)=\theta(0)+w\lambda^{-1}\ln(\xi/z(0))$. Hence g is defined on the set of points in M_1 for which $\xi/z(0)>0$ and $g(\theta,z)=(\rho\exp(\gamma\lambda^{-1}\ln(\xi/z)),\ \theta+w\lambda^{-1}\ln(\xi/z))$. Note that a curve parallel to the z axis in M_1 is mapped into a logarithmic spiral in M_2 and that circles of constant (r,z) are mapped to circles of constant (r,z). The condition that $0<\gamma<\lambda$ implies that $dg_{(r,z)}(0,1)$ is unbounded at $z\to 0$.

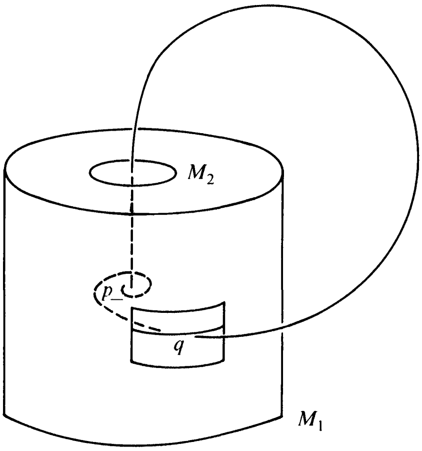


FIG. 11. *Geometry of Sil'nikov theorem.*

There is also a mapping h: $M_2 \to M_1$ along trajectories. The map h is defined in a neighborhood of $r=0$ and smooth. The composition $h \circ g$ is a return map for M_1 which is defined in a neighborhood of the boundary of the half space $\xi/z>0$. Now let us find the image of a curve parallel to the z axis under $h \circ g$. This will be a spiral converging to the point $q=W^u(p_-)\cap M_1$. If $x=(\rho,\theta,z)$, the distance from $h \circ g(x)$ to q will be of the order of $z^{-\gamma/\lambda}$. Thus as $x \to q$ along a curve parallel to the z axis, the distance from x to $h \circ g(z)$ will become larger than the distance from x to q. The rectangle R depicted in Fig. 12 has horizontal boundaries with $z=z_0, z_1$, where z_1/z_0 is slightly larger than $\exp(\pi\lambda w^{-1})$. A vertical segment in R is mapped by $h \circ g$ into a spiral segment in which θ varies by at least π. The values of z_0 and z_1 are determined so that this spiral intersects the annulus in M_1 defined by $z_0<z<z_1$ in two components. Choose the vertical boundaries $\theta=\theta_0,\theta_1$ of R so that they fall well outside the two components of the spiral arc which is the image of the segment $z_0<z<z_1$ in the line through q parallel to the z axis. The image of R will then overlap R in a figure with two components R_1, R_2 which looks like Smale's horseshoe. The hyperbolicity estimates required in property (M3) can be established provided that a certain constant defined by Sil'nikov does not vanish. Since this argument can be applied to a whole sequence of strips R which converge to the set $z=0$, one finds with it a countable collection of subshifts of finite type, each a Smale horseshoe. These occur in the unfolding of our condimension two bifurcation.

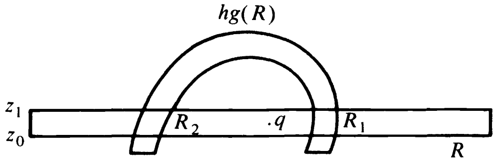


FIG. 12. *Smale's horseshoe in Sil'nikov theorem.*

6. Applications and examples. In this section we shall examine several situations in which the preceding theory yields a substantial amount of information about problems which have been considered previously by more classical techniques. We would have liked to include here a larger array, including newer and more interesting examples than those which are discussed, but fully developed applications of the theory described in earlier sections are still limited. Two of the examples point to different kinds of extensions to the theory which will be necessary for a full analysis of the bifurcations present within them.

Example 1. *Variational equation of Van der Pol.* The forced Van der Pol equation describes an oscillator with one degree of freedom and nonlinear resistance:

$$\ddot{x}=\varepsilon(1-x^2)x+x=b\cos(wt). \tag{6.1}$$

Experimental evidence during the late 1930's with electrical circuits led Cartwright and Littlewood to the first proofs of the existence of transversal homoclinic orbits in non-Hamiltonian systems of differential equations [22]. Their argument applies to the Van der Pol equation with large ε, where it corresponds to "relaxation oscillations." The dynamics of the Van der Pol equations are also of interest when ε is small, particularly when b and $(w-1)$ are of the same order of magnitude as ε. In this near resonance case, complicated dynamical phenomena occur.

The nearly resonant case of the Van der Pol equations can be studied by applying the *method of averaging* to the equation. After suitable rescaling, the average deviations of a trajectory from those of the simple harmonic oscillator (over one forcing period) are described by the systems of equations

$$(6.2) \qquad \dot{u} = -\sigma v + u\big(1-(u^2+v^2)\big), \qquad \dot{v} = F + \sigma u + v\big(1-(u^2+v^2)\big).$$

The properties of the solutions of these variational equations (6.2) as functions of the two parameters (σ, F) have been studied by Cartwright [21], Gillies [36], and Holmes and Rand [60]. Apart from minor uncertainties, they give a complete description of the dynamics of (6.2) for all values of (σ, F).

Figure 13 shows a picture of the (σ, F) plane with the bifurcation curves located. Of particular interest are the codimension two points $0, A$ and S. At the point $A = (1/\sqrt{3}\ 8/\sqrt{27})$ there is a cusp point at which the two curves of saddle node bifurcations (SN) terminate. At the point $0 = (\frac{1}{2}, \frac{1}{2})$, there is a codimension two bifurcation with a double zero eigenvalue. Holmes and Rand [60] use the computation of the normal form of (6.2) and Taken's analysis of this bifurcation to prove the existence of a curve of saddle loops (L) terminating at 0 in addition to the curve of directly calculable Hopf bifurcations (H). The point S is "known" only on the basis of numerical evidence and corresponds to a saddle node whose unstable manifold forms a loop which is part of the boundary of the stable manifold.

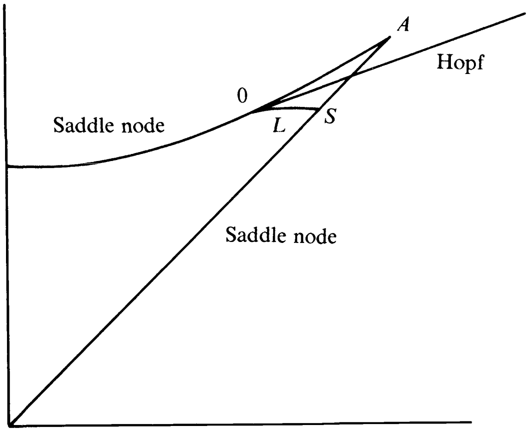


FIG. 13. *Stability diagram of Van der Pol variational equations.*

The features of (6.2) correspond to the features of the global return map for (6.1) obtained by integrating (6.1) for time $2\pi/w$. The equation (6.1) is not equivariant in any apparent way, so one does not expect that there will be resonant effects in relating the dynamics of (6.2) to those of (6.1). In particular in the region bounded by the curves H, L and the portion of SN joining 0 to S, (6.2) has a stable limit cycle, and one expects the corresponding (6.1) to have an invariant two-dimensional torus. The dynamics on this two-dimensional torus should exhibit *phase locking* and *entrainment*. Corresponding to the curve L of homoclinic loops for (6.2), one should find transversal homoclinic solutions for (6.1). Note here that the algebraic calculations which locate the point 0 in the bifurcation diagram for (6.2) indicate approximate parameter values for which (6.1) should have transversal homoclinic solutions. See [44] for a more thorough review of the dynamics of the Van der Pol equations.

Example 2. *Panel flutter*. The second example we discuss involves the oscillations of a thin elastic panel which is forced aerodynamically by a flow across the panel. We assume that all motions of the panel are normal deflections which are constant along lines parallel to the two ends of the panel. The ends of the panel are fixed at $z=0$ and $z=1$ and the lateral deflection is described by the function $v(z,t)$, v: $[0,1]\times\mathbb{R}^n\to\mathbb{R}$. The governing equation for the motions is

$$\alpha\dot{v}''''+v''''-\left\{\Gamma+k\int_0^1(v'(\xi))^2d\xi+\sigma\int_0^1(v'(\xi)\dot{v}'(\xi))\,d\xi\right\}v''+\rho v'+\sqrt{\rho\delta}\,\ddot{v}+\ddot{v}=0$$

with "•" representing differentiation with respect to t and "′" differentiation with respect to z. Of the various parameters, ρ represents the dynamical pressue of the flow across the plate and Γ represents a tensile load within the plate. The remaining parameters reflect structural characteristics of the plate which we assume to be fixed. Boundary conditions are $v=(\dot{v}+\alpha v)''=0$ (simply supported) or $v=v'=0$ (clamped). This equation is derived by Dowell [29] and has been studied by Holmes [55] and Holmes and Marsden [59] as an application of the theory of codimension two bifurcations.

The plate equation above is a rather formidable nonlinear partial differential equation. Even coping with the linearized equation is difficult. We shall use the example to illustrate the kinds of approximations which can be made to reduce a problem of this difficulty to manageable proportions without (apparently) throwing aside the qualitative features of its dynamics. The next example provides a simpler nonlinear partial differential equation for which a more complete analysis can be given, but in this one the answers from finite dimensional approximations must suffice for some calculations. Before describing this calculation, we briefly review the theory which does give one confidence that the reductions preserve the qualitative structure of the dynamic motions.

The approach which one adopts is that of trying to identify a suitable function space on which the plate equation defines a *smooth semiflow*, in the terminology of Marsden and McCracken [83]. This involves a set of technical hypotheses which allow one to prove a good existence theorem. It also provides the basis for an infinite dimensional *center manifold theorem*. Marsden [82] presents a recent review of these topics from the perspective of bifurcation theory. Briefly, the center manifold theorem applies to a system at an equilibrium p for which there is a $\delta>0$ such that all of its spectrum apart from a finite number of eigenvalues lies to the left of the line $\operatorname{Re}z=-\delta$ in the complex plane. In this situation, there will be an invariant finite dimensional manifold M passing through p such that the tangent space to M at p is spanned by the part of the spectrum which lies on the imaginary axis. If there are no eigenvalues in the right half plane, then this *center manifold* M will be *attracting* in the sense that p has a neighborhood U such that all solutions which remain inside U tend to M. Thus, one can study the dynamics of the equation on the center manifold of M where it defines a finite dimensional vector field.

There is no difficulty in including parameters in the center manifold theorem. If one begins with a system which has a "trivial" equilibrium ($v(z,t)\equiv0$ in the example here), then one can search for parameter values at which the equilibrium first loses its stability. By this we mean that all of the spectrum lies in the left half plane apart from a few eigenvalues on the imaginary axis. The corresponding center manifold M (including the parameters) then determines the bifurcation structure for the full set of equations regarded as a flow on the function space. If the normal form on M can be calculated and the parameterized family on M is persistent within the appropriate class, then the bifurcation analysis of this normal form can be applied to the example at hand. Apart from being able to give a complete calculation of the spectrum of the linearized problem, this is the program carried out by Holmes and Marsden [59] for the panel flutter problem.

In lieu of a full analysis of the spectrum, Holmes and Marsden employ a *Galerkin* approximation procedure which approximates the problem by a finite dimensional one. They then carry out calculations with these approximations, proving at the same time that as large approximations are used, the sequence of approximate solutions does converge in the function space to the solution of the original problem. Abstractly, the Galerkin procedure works in the following way. One chooses an orthonormal basis $\{w_1\}$ for the function space (which is now assumed to have an inner product). Let E_k be the finite dimensional space spanned by the functions $w_1,\cdots,w_k$ and π_k the projection of the function space onto E_k. If the original equation is expressed as an ordinary differential equation on function space $dx/dt=f(x)$, then $dx/dt=\pi_k f(x)$ defines a vector field on the finite dimensional space E_k. This last equation can be expressed as a system of k ordinary differential equations for the coefficients of the w_i along a trajectory.

To apply the Galerkin procedure to the panel problem, one uses the functions $\sin n\pi z$ as an orthonormal set. In terms of these, the panel equation can be described as a second order differential equation for the coefficients. Thus there is a two-dimensional space corresponding to each $\sin n\pi z$ in the Galerkin approximations. Provided that one retains at least four modes, numerical calculations indicate that the largest eigenvalues of the linearized problem remain almost unchanged by the addition of more modes. For reasonable parameter values $(\alpha,\delta,\Gamma,\rho)=(0.005,0.1,-2.29\pi^2,112.5)$, there is a codimension two bifurcation with a double zero eigenvalue. There is a symmetry to the panel equations which comes from replacing v with $-v$. This symmetry is present in the Galerkin approximations where it takes the form of rotation by π in the plane of each mode. Consequently, the normal forms will be those appropriate to the class of rotationally symmetric vector fields. Holmes computes the coefficients of the cubic terms of the normal form.

Example 3. *Brusselator*. The next example which we discuss is the *Brusselator* [12]. This is a model system of reaction diffusion equations representing the kind of dynamical behavior one suspects (hopes?) plays a role in regulating the formation of patterns in living organisms. One begins with the following reaction scheme:

$$A\to X,\quad B+X\to Y+D,\quad 2X+Y\to 3X,\quad X\to E.$$

In this scheme A,B,D and E are reactants whose concentrations are assumed to be fixed throughout the reaction. It is the dynamics of the intermediates X and Y which we want to examine with this assumption. In addition, we assume that the reaction is taking place in a one-dimensional medium and that X and Y diffuse with diffusion constants D_1 and D_2. This yields the following system of reaction-diffusion equations (X,Y):

$$\frac{\partial X}{\partial t}=D_1\frac{\partial^2 X}{\partial \xi^2}+X^2Y-(B+1)X+A,\qquad \frac{\partial Y}{\partial t}=D_2\frac{\partial Y}{\partial \xi^2}X^2Y+BX. \tag{6.3}$$

We assume further than the reaction is at equilibrium at the ends of the interval $[0,\pi]$, so that $X(0)=X(\pi)=A$ and $Y(0)=Y(\pi)=B/A$ for all $t\geq 0$.

The problem to be solved here is the initial value problem. In particular, we would like to know what kinds of dynamics are possible at $t\to\infty$ as a function of the parameters (A,B,D_1,D_2). The problem is very far from a complete solution, but we are able to give an argument for the existence of transversal homoclinic solutions based upon the theory described in this review. As with the forced Van der Pol example, this argument lacks completeness in that one is unable to prove that there is a hidden symmetry which would prevent the creation of transversal homoclinic solutions through resonance effects. Apart from this missing detail, this provides the first existence proof for transversal homoclinic solutions to a system of autonomous *partial differential equations* which does not admit a trivial reduction to a finite dimensional system. The Brusselator has served as a useful "model" system of partial differential equations, and its analysis is in the same spirit as that which one hopes to use for a variety of fluid dynamic problems.

The mathematical principles discussed for the panel problem allow us to apply the results of finite dimensional bifurcation theory to this infinite dimensional problem. There is an existence and uniqueness theory for solutions of (6.3) which guarantees that the equation defines a suitably smooth semiflow on the Banach space $C_0^2[0,\pi]$ of C^2 functions which satisfy the boundary conditions imposed by (6.3). Thus the center manifold theorem can be applied, and we shall adopt the attitude that the bifurcation structure of the problem is adequately described by the finite dimensional theory.

The Brusselator problem has the trivial equilibrium solution $x(\xi,t)\equiv A$, $y(\xi,t)=B/A$. We want to linearize the equations at this equilibrium and determine the spectrum of the linearized equations. Introducing

$$u=X-A, \qquad v=Y-\frac{B}{A},$$

(6.3) becomes

$$\frac{\partial u}{\partial t}=D_1\frac{\partial^2 u}{\partial \xi^2}+(B-1)u+A^2v+\left(\frac{B}{A}u^2+2Auv+u^2v\right), \tag{6.4}$$

$$\frac{\partial x}{\partial t}=D_2\frac{\partial^2 v}{\partial \xi^2}-Bu-A^2v\left(\frac{B}{A}u+2Auv+u^2v\right).$$

If $w=(u,v)$, we write $w_t=Lw+Nw$ where L is the linear part of (6.4). Representing $w=(u,v)$ as a Fourier series $w(t)=\Sigma_{n=0}^{\infty}w_n(t)\sin n\xi$, we find that the two-dimensional spaces spanned by the vector-valued functions $w_n\sin n\xi$ are invariant for L with spectrum given by the eigenvalues of

$$E_n=\begin{pmatrix} B-1-n^2D_1 & A^2 \\ B & -A^2-n^2D_2 \end{pmatrix}.$$

We want to find parameter values for which all of these eigenvalues have negative real parts (bounded away from 0) except for a finite number which lie on the imaginary axis.

We make two observations about the collection of eigenvalues of the matrices E_n as n varies. The first observation is that $\operatorname{Tr} E_n=B-1-A^2-n^2(D_1+D_2)$ is a decreasing function of n. Consequently, if E_n has pure imaginary eigenvalues for $n>1$, then $E_{(n-1)}$ has an eigenvalue with positive real part. The second observation is that $\det E_n$ is a quadratic function of n^2 with positive leading coefficient D_1D_2. Therefore, if E_k and E_l have zero eigenvalues and $|k-l|>1$, then there is an E_n which has negative determinant and an eigenvalue with positive real part. Thus, when no eigenvalues have positive real part, the maximum dimension of the eigenspace of the imaginary axis is 4. This situation results when E_k and E_{k+1} each have a zero eigenvalue for some $k>1$, and E_1

has pure imaginary eigenvalues. We compute the parameter values for which this maximal degeneracy occurs:

$$D_2=\frac{D_1^2k^2(k+1)^2+2D_1k(k+1)}{1+D_1k^2(k+1)^2},\quad A^2=D_1D_2k^2(k+1)^2,\quad B=1+A^2+D_1+D_2.$$

When these equations are satisfied, E_1 has pure imaginary eigenvalues, E_k and E_k+1 each have a zero eigenvalue, and all other eigenvalues of the E_n have negative real parts. To see this, note first that the third equation determines that $\operatorname{Tr} E_1=0$ and $\operatorname{Tr} E_n<0$ for $n>1$. Next observe that $\det E_n=A^2+n^2(A^2D_1+D_2-BD_2)+n^4D_1D_2$. Since this function is convex, $\det E_k=\det E_{k+1}=0$ implies that $\det E_n\geq 0$ for all n. The equation $A^2+k^2(A^2D_1+D_2-BD_2)+k^4D_1D_2=A^2+(k+1)^2(A^2D_1+D_2-BD_2)+(k+1)^4D_1D_2$ yields the second equation by eliminating the middle terms and solving for A^2. The first equation is then obtained by substituting the values of A^2 and B from the last two equations into the equation $\det E_k=0$ and solving for D_2.

This most degenerate equilibrium represents a bifurcation of codimension three. Its unfolding has not been calculated. In lieu of being able to calculate the unfolding for this codimension three bifurcation, we consider the easier problem of examining its behavior near an equilibrium in which there is one zero eigenvalue of E_k, pure imaginary eigenvalues for E_1 and all other eigenvalues have negative real parts. If we regard (A^2,B) as being experimental parameters with the diffusion rates (D_1,D_2) fixed, then the above conditions become

$$A^2=D_2k^2\left(\frac{D_1+D_2-D_2k^2}{1+D_1k^2-D_2k^2}\right),\qquad B=1+A^2+D+D \tag{6.5}$$

subject to the following inequalities on the diffusion rates:

$$D_2k^2\left(\frac{D_1+D_2-D_1k^2}{1+k^2(D_1-D_2)}\right)\left(1+(k\pm 1)^2(D_1-D_2)\right)$$
$$-(k\pm 1)^2D_2(D_1+D_2)+(k\pm 1)^4D_1D_2>0 \quad \text{for all } k\in z.$$

There are solutions to this system of equations and inequalities with A,B,D_1 and D_2 all positive.

Let E be the three-dimensional eigenspace of the imaginary axis for the linearized equations (6.4) and let P: $C_0^2[0,\pi]\to E$ be the projection onto E. We want to express in E the equations $(PW)_t=P(Lw+Nw)$ or $w_t=Lw+PNw$ for $w\in E$. These are the "truncated" equations of (6.3) which give an approximate description of the flow on its center manifold. To the extent that the Taylor expansions of degree two at the origin of the truncated equations and the full equations (6.4) agree, we can use the truncated equations to determine the unfolding of the codimension two equilibrium of (6.4). These computations are straightforward but somewhat lengthy. The details are not illuminating, so we merely outline the procedure. More detail can be found in [41].

Denote by X the vector field on E defined by the truncated equations. Recall that the normal form of X was $(\omega+a_1r^2)\partial/\partial\theta+a_2rz\,\partial/\partial r+(a_3z^2+a_4r^2)\partial/\partial z$ in appropriate cylindrical coordinates. The coefficients (a_2,a_3,a_4) determine the qualitative structure of the unfolding of this bifurcation. To compute these coefficients requires several steps:

(1) We find a basis for E so that the linearization of $DX(0)$ represents infinitesimal rotation about the z axis (with rate ω).

(2) For a vector field with the linear part of X, we identify those expressions which become the coefficients a_i in the normal form.

(3) We determine those terms in $N(w)$ which contribute to the expressions in step (2) and compute their projections onto the two-dimensional spaces of the form $(\sin\xi)w$ and $(\sin k\xi)w$.

(4) We compute the coordinates of the projections in step (3) relative to the basis which extends the basis for E found in step (1) by the second eigenvector of L in E_1. We read off the coefficients in the normal form.

We list the results of this computation:

$$z^2\frac{\partial}{\partial z}:\ \frac{\gamma B}{2A}\left(k^4D_2^2-A^4\right)\int_0^\pi \sin^3 k\xi\, d\xi,$$

$$r^2\frac{\partial}{\partial z}:\ \frac{\gamma}{4A}\left[B\left(1+d^2D_2^2\right)-2A^2\left(1+d^2D_2+d^2D_1D_2\right)\right]\int_0^\pi \sin^2\xi \sin k\xi\, d\xi,$$

$$rz\frac{\partial}{\partial r}:\ \frac{1}{A}\left[Bk^2D_2-A^2\left(A^2+k^2D_2\right)\right]\int_0^\pi \sin k\xi\, d\xi,$$

$$\gamma=\left(\left(A^2+k^2D_2\right)^2-A^2B\right)^{-1}\left(A^2+k^2D_2-B\right),$$

$$d^2=\left(A^2+A^2D_1+D_2+D_1D_2-BD_2\right)^{-1}.$$

There are some interesting aspects to these calculations. The trigonometric integrals above depend strongly on the parity of k. When k is even, they all vanish and the bifurcation is degenerate. This is due to the invariance of the functions $\sin 2\pi l$ on $[0,\pi]$ with respect to the symmetry $f(x)\to f(\pi-x)$. If we restrict attention to the class of functions which possess this symmetry, then the cubic terms in the Taylor expansion determine much of the qualitative behavior of the unfolding and the normal form is then for systems with a reflection geometry. To give a complete analysis of the bifurcation structure for varying boundary conditions, one needs to determine what occurs when one allows this symmetry to be broken.

When k is odd, the quadratic coefficients of the normal form do not vanish (for most allowable values of (D_1,D_2)) and the unfolding results from §§3–5 can be applied directly. One question of interest is whether or not there are values of (D_1,D_2k) which yield transversal homoclinic orbits in the unfolding (unless there are hidden constraints which prevent resonance effects). Such values of (D_1,D_2,k) do exist, indicating the presence of chaotic solutions to the Brusselator equations. We note that numerical solutions of the Brusselator have been computed which appear chaotic. These chaotic solutions are irregular both in time and space.

Example 4. *Double diffusive convection.* The final example we discuss is a "classical" fluid mechanics problem: thermohaline convection. Fluid motions exhibit a wide variety of dynamical phenomena, and fluid mechanics has been a fertile ground for applications of bifurcation theory. Bifurcation computations involving the Navier–Stokes equations are difficult unless they begin with steady flows of simple geometry. Consequently, most of the classical theory deals with the initial bifurcations in which an instability of a motion described by an explicit formula first occurs as a parameter is varied. One prospect for the use of more parameters and the computation of multiple bifurcations is that these provide a means for analytically coping with secondary (and higher in some cases) bifurcations without doing fluid calculations much more sophisticated than those which have been done in the past.

We have selected thermohaline convection as an example to illustrate these points because the linear calculations explicitly locate codimension two bifurcations and some work has been done to understand the behavior associated with these. Indeed, we shall see that bifurcations of both the double zero and double Hopf type occur with some symmetry in the system for the boundary conditions we employ. Knobloch and Proctor [74] give an analysis of this problem using perturbation theory methods.

Thermohaline convection is the problem of studying the fluid motions of salt water due to the buoyancy effects of opposing gradients of heat and salt (or of two other solutes which affect the fluid density). The different diffusivities of heat and salt lead to very different instabilities when hot salty water lies above colder fresher water and vice versa. There is a large body of experimental and numerical observation to compare with analytic results for this problem. The linear computation of stability as a function of the steepness of the salt and heat gradients leads to codimension two bifurcations as we now describe.

A horizontal layer of fluid of depth d has fixed temperatures and salt concentrations on its upper and lower boundaries. One assumes that the fluid is incompressible and that the buoyant force on the fluid depends linearly on the temperature and salt concentrations. The resulting equations are

$$\begin{aligned}
&\frac{\partial}{\partial t}v+v\cdot\nabla v=\frac{1}{\rho}\nabla p+g(\alpha T-\beta S)+\nu\nabla^2 v,\\
&\operatorname{div}(v)=0,\\
&\frac{\partial T}{\partial t}+v\cdot\nabla T-w\frac{\Delta T}{d}=k\nabla^2 T,\\
&\frac{\partial S}{\partial t}+v\cdot\nabla S-s\frac{\Delta S}{d}=k_s\nabla^2 S,
\end{aligned}$$

with v the fluid velocity vector, T and S the departures of the temperature and salt concentrations from their steady state distributions, ΔT and ΔS the imposed temperature and solute differences across the fluid layer and w the vertical velocity component. The remaining constants are the density ρ, the gravitational constant g, diffusivities k and k_s, kinematic viscosity ν and the buoyancy dependency on temperature and salinity given by α and β. The pressue p is eliminated from the system by taking the curl of the Navier–Stokes equation, thereby obtaining the vorticity equation. In the case of velocity fields which never have a component in the y direction, we can express the vorticity equation in terms of the stream function ψ. After rescaling, the system to be solved is now

$$\begin{aligned}
(6.6)\qquad &\sigma^{-1}\nabla^2\partial_t\psi-\sigma^{-1}J(\psi,\nabla^2\psi)=-R_T\partial_x T+R_S\partial_x S+\nabla^4\psi,\\
&\partial_t T+\partial_x\psi-J(\psi,T)=\nabla^2 T,\\
&\partial_t S+\partial_x\psi-J(\psi,S)=\tau\nabla^2 S,
\end{aligned}$$

where $J(f,g)$ is defined to be $\partial_x f\partial_z-\partial_z f\partial_x g$. The boundary conditions are $\psi=\partial_{zz}^2\psi=T=S=0$ when $z=0$ or 1.

The system above has trivial 0 solution, and the linearization at the trivial solution is obtained by dropping the Jacobian terms from each equation. For the linearized system, the normal modes (eigenvectors) are easily determined as functions of the form

$$\begin{pmatrix} \psi \\ T \\ S \end{pmatrix}(x,z,t)=e^{pt}\sin n\pi z\begin{pmatrix} \psi_0\sin\pi\alpha x \\ T_0\cos\pi\alpha x \\ S_0\cos\pi\alpha x \end{pmatrix},$$

where (p,n,α) satisfy the equation

$$\begin{aligned} &p^3+(\sigma+\tau+1)k^2p^2+\left[(\sigma+\tau+\sigma\tau)k^4-\pi^2\sigma\tau^2k^{-2}(R_T-R_S)\right]p \\ &\qquad\qquad +\sigma\tau k^6+\pi^2\sigma\alpha^2(R_S-\tau R_T)=0, \\ &k^2=\pi^2(n^2+\alpha^2). \end{aligned}$$

One seeks values of R_T and R_S for which no solutions p of this equation have positive real parts, but some have zero real parts. Choosing $\alpha^2=\frac{1}{2}$ and $n=1$, one obtains the values of p with the smallest maximum real parts. These can appear with either a pair of pure imaginary eigenvalues, as a zero eigenvalue, or as a zero eigenvalue of multiplicity two. The last possibility occurs for the special values of R_T and R_S for which

$$\begin{aligned} &(\sigma+\tau+\sigma\tau)k^4-\pi^2\sigma\alpha^2k^{-2}(R_T-R_S)=0, \\ &\sigma\tau k^6+\pi^2\sigma\alpha^2(R_S-\tau R_T)=0, \end{aligned}$$

or

$$\begin{pmatrix} R_S \\ R_T \end{pmatrix}=\frac{k^6}{\pi^2\alpha^2(1-\tau)}\begin{pmatrix} \tau^2+t^2/\sigma \\ 1+\tau/\sigma \end{pmatrix}.$$

These values of the Rayleigh numbers are a good candidate for the application of our codimension two bifurcation theory. At this point one should compute the normal form corresponding to this double zero eigenvalue and the transversality conditions for variations with respect to R_S and R_T, thereby determining the bifurcation structure. There is a symmetry in the system (6.6) which comes from simultaneously changing the signs of ψ, S, T and z. This symmetry forces the normal form of this problem to be one of type (iii) in Theorem 4.2. Following the perturbation calculations of Knobloch and Proctor [74], one projects the equations onto a five-dimensional space V which includes the two-dimensional zero eigenspace W. In the five-dimensional space V, the center manifold M has quadratic tangency to W. Restricting the equations to M we retain all of the terms that affect the cubic coefficients in the normal form and obtain an unfolding of the type indicated in Fig. 9.

We end by illustrating how a small change in the boundary conditions for the thermohaline problem produces a bifurcation of the double Hopf type, although the appropriate normal forms have not been computed. The computation of the spectrum of the linearized problem depends upon the horizontal wave number α. No restrictions were placed upon α corresponding to the (physically unrealistic) idealization of an infinite conducting layer. In addition to the difficulties in computing normal forms which we described above, there are problems in applying the center manifold theorem because the linearized operator has a continuous spectrum. The applied literature uniformly avoids this difficulty by examining only disturbances whose horizontal wavelength $\alpha=2^{1/2}$ corresponds to the eigenvalues with the largest real parts. One way of avoiding this second difficulty in applying the theory is to impose periodic boundary conditions in the x direction. For the normal modes of the linearized equation, this forces α to be a multiple of the imposed period$=m/D$ with D determining the scale of the imposed horizontal periodicity. When the function $(n^2+\alpha^2)^3/\alpha^2$ has equal values for $\alpha=m/D$, $\alpha=(m+1)/D$ with $m \in Z$ such that $m^2<2D^2<(m+1)^2$, then the corresponding modes of the linearized equation can both be pure imaginary and a double Hopf bifurcation results. The normal form for the example has not been computed, but it does illustrate that double Hopf bifurcations occur in fluid dynamics problems.

7. Summary.This summary will list normal forms, integrable limits from rescaling and variational integrals for the various types of bifurcations discussed in this paper.

Codimension one bifurcations.

Saddle node. Simple zero eigenvalue,

$$\dot{x}=\gamma+ax^2+o(2).$$

Transcritical. Simple zero eigenvalue, 0 constrained to be equilibrium,

$$\dot{x}=\gamma x+ax^2+o(2).$$

Pitchfork. Simple zero eigenvalue, reflection symmetry,

$$\dot{x}=\lambda x+ax^3+o(3).$$

Hopf. Simple pair pure imaginary eigenvalues,

$$\dot{\theta}=w+a_1r^2+o(2), \qquad \dot{r}=\lambda r-a_2r^3+o(3).$$

Codimension two bifurcations.

Two-dimensional nilpotent space.

$$\dot{x}_1=x_2+o(2), \qquad \dot{x}_2=\lambda_1+\lambda_2x_1+a_1x_1^2+a_2x_1x_2+o(2).$$

Rescaling.

$$x_1=\delta^2X_1, \quad x_2=\delta^3X_2, \quad \delta t=T, \quad \lambda_1=\delta^4\Lambda, \quad \lambda_2=\delta^2\Lambda_2.$$

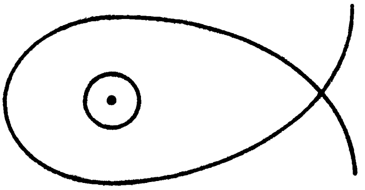
Integral.

$$H(X_1,X_2)=\frac{X_2^2}{2}+\Lambda_1X_2+\Lambda_2\frac{X_1^2}{2}+a_1\frac{X_1^3}{3}.$$

Variational calculation. Closed curve γ in set $H=c$ is approximation to closed orbit when

$$\int_{\text{interior}\,\gamma}X_1=0=\int X_1\left(-c+\Lambda_1X_1+\Lambda_2\frac{X_1^2}{2}+a_1\frac{X_1^3}{3}\right)^{1/2}dx_1.$$

This defines a surface in (Λ_1,Λ_2,c) space locating limit positions for periodic orbits.



Two-dimensional nilpotent space with 0 *constrained equilibrium.*

$$\dot{x}_1=x_2+o(2), \qquad \dot{x}_2=\lambda_1x_1+\lambda_2x_2+a_1x_1^2+a_2x_1x_2+o(2).$$

Rescaling.

$$x_1=\delta^2X_1, \quad x_2=\delta^3X_2, \quad \delta t=T, \quad \lambda_1=\delta^2\Lambda_1, \quad \lambda_2=\delta^2\Lambda_2.$$

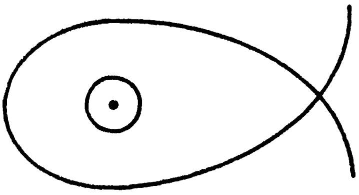
Integral.

$$H(X_1,X_2)=\frac{-X_2^2}{2}+\Lambda_1-\frac{X_1^2}{2}+a_1\frac{X_1^3}{3}.$$

Variational calculation. If D is compact with H constant on D,

$$\int_D(\Lambda_2+a_2X_1)=0 \quad \text{or} \quad \frac{\Lambda_2}{a_2}=\frac{\int_D X_1}{\int D^1}.$$

Integral curves.



Two-dimensional nilpotent space with symmetry of rotation by π.

$$\dot{x}_1=x_2+o(3), \qquad \dot{x}_2=\lambda_1x_1+\lambda_2x_2+a_1x_1^3+a_2x_1^2x_2+o(3).$$

Rescaling.

$$x_1=\delta X_1, \quad x_2=\delta^2X_2, \quad \delta t=T, \quad \lambda_1=\delta^2\Lambda_1, \quad \lambda_2=\delta^2\Lambda_2.$$

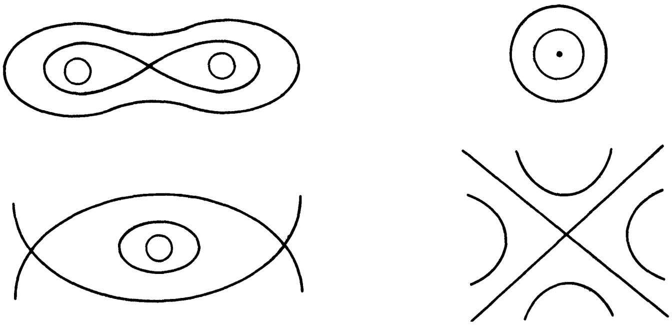
Integral.

$$H(X_1,X_2)=\frac{X_2^2}{2}+\Lambda_1\frac{X_1^2}{2}+a_1\frac{X_1^4}{4}.$$

Variational calculation. On the interior D of a compact component of $H=c$

$$\int_D(\Lambda_2+a_2X_1^2)=0.$$

Integral curves.



$0+$ *pure imaginary eigenvalue.*

$$\begin{aligned}
&\dot\theta=w+a_1r^2+o(2),\\
&\dot r=\lambda_2r+a_2rx_3+\left(b_1r^3+b_2rx_3^2\right)+o(3),\\
&\dot x_3=\lambda_2+a_3x_3^2+a_4r^2+\left(b_3r^2x_3+b_4x_3^3\right)+o(3).
\end{aligned}$$

Rescaling for equations in (r,x_3).

$$r=\delta R,\quad x_3=\delta X_3,\quad \delta t=T,\quad \lambda_1=\delta^2\Lambda_1,\quad \lambda_2=\delta^2\Lambda_2.$$

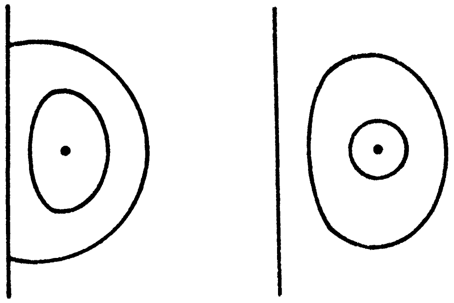
Integral.

$$H(R,X_3)=\frac{-a_2}{2a_3}R^{-2a_3/a_2}\left(\Lambda_2+a_3X_3^2+\frac{a_3a_4}{a_3-a_2}R^2\right).$$

Variational calculation. On the interior D of a compact component of $H=c$,

$$\int_D\left(\left(\Lambda_1+\left(\frac{a_3-a_2}{a_3}b_1-\frac{a_2b_3}{2a_3}\right)R^2+\left(b_2-\frac{3a_2}{2a_3}b_4X_3^2\right)\right)\right)R^{-2a_3/a_2-1}=0.$$

Integral curves.



$0+$ *pure imaginary eigenvalues + reflection symmetry in* x_3 *axis.*

$$\begin{aligned}
&\dot\theta=w+a_1r^2+o(2),\\
&\dot r=\lambda_1r+b_1rx_3^2+b_2r^3+rP_4+o(5),\\
&\dot x_3=\lambda_2x_3+b_3r^2x_3+b_4x_3^3+x_3Q_4+o(5).
\end{aligned}$$

two pairs of pure imaginary eigenvalues with no resonance.

$$\begin{aligned}
&\dot\theta_1=w_1+a_1r_1^2+a_2r_2^2+o(2),\\
&\dot\theta_2=w_2+a_3r_1^2+a_4r_2^2+o(2),\\
&\dot r_1=r_1\left(\lambda_1+b_1r_1^2+b_2r_2^2+P_4\right)+o(5),\\
&\dot r_2=r_2\left(\lambda_2+b_3r_1^2+b_4r_2^2+Q_4\right)+o(5).
\end{aligned}$$

(Two-dimensional systems for (r,x_3), (r_1,r_2) are the same.)
Rescaling for (r_1,r_2).

$$r_1=\delta R_1,\quad r_2=\delta R_2,\quad \delta^2t=T,\quad \lambda_1=\delta^2\Lambda_1,\quad \lambda_2=\delta^2\Lambda_2.$$

The integral exists when $(b_1-b_3)b_4\Lambda_1+(b_4-b_2)b_1\Lambda_2=0$,

$$H(R_1,R_2)=r_1^{\alpha_1}R_2^{\alpha_2}(\beta_0+\beta_1R_1^2+\beta_2R_2^2)$$

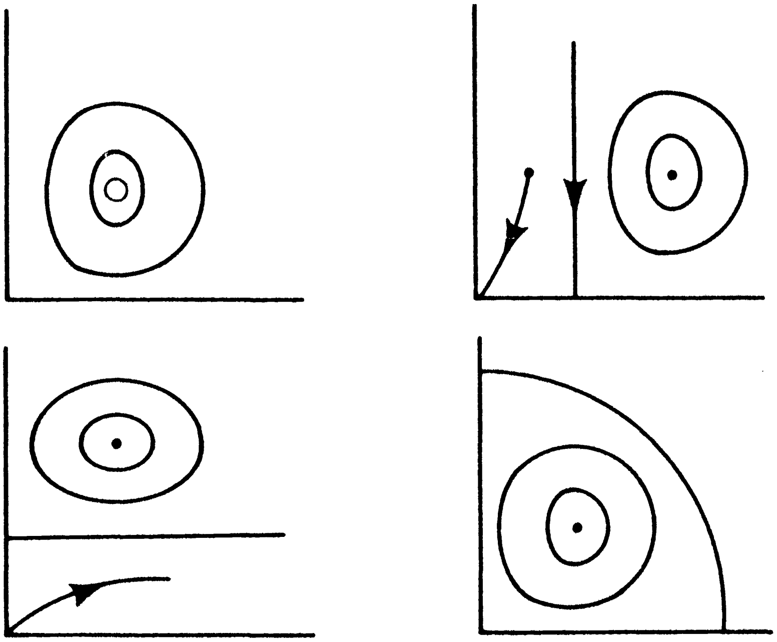
with

$$\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}=\begin{pmatrix}\dfrac{2b_1(b_2-b_4)}{b_1b_4-b_2b_3}\\[2ex] \dfrac{2b_4(b_3-b_1)}{b_1b_4-b_2b_3}\end{pmatrix}\quad\text{and}\quad \beta_0=\frac{\Lambda_2}{\alpha_1},\quad \beta_1=\frac{-a_{11}}{\alpha_2},\quad \beta_2=\frac{a_{22}}{\alpha_1}.$$

Variational calculation. On the interior D of a compact component of $H=c$ has the form

$$\int_D(\gamma_1\Lambda_1+\gamma_2\Gamma_2+\varepsilon_1R_1^4+\varepsilon_2R_1^2R_2^2+\varepsilon_3R_2^4)R^{\alpha_1-1}R^{\alpha_2-1}=0.$$

Integral curves.



Acknowledgments. I would like to thank warmly the editor and the reviewers for their helpful comments and the careful attention they have given to this manuscript.

REFERENCES

[1] V. S. Afraimovic and L. P. Sil'nikov, *On small periodic perturbations of autonomous systems*, Soviet Math. Dokl., 15 (1974), pp. 206–211.

[2] A. Andronov and E. Leontovich, *Sur la théorie de la variation de la structure qualitative de la division du plan en trajectories*, Dokl. Akad. Nauk SSSR, 21 (1938), pp. 427–430. (In Russian.)

[3] A. Andronov, E. Leontovich, I. Gordon and A. Maier, *The Theory of Bifurcation of Plane Dynamical Systems*, John Wiley, New York, 1973.

[4] P. Aris, *Chemical reactors and some bifurcation phenomena*, preprint, Minneapolis, 1977.

[5] V. I. Arnold, *Small denominators* I, *mappings of the circle onto itself*, Izv. Akad. Nauk. SSSR Ser. Mat., 25 (1961), pp. 21–86. (In Russian.)

[6] ______, *On matrices depending upon a parameter*, Russian Math Surveys, 26, no. 2 (1971), pp. 29–43.

[7] ______, *Lectures on bifurcations in versal families*, Russian Math Surveys, 27 (1972), pp. 54–123.

[8] ______, *Loss of stability of self oscillations close to resonance and versal deformations of equivariant vector fields*, Funct. Anal. Appl., 11 (1977), pp. 85–92.

[9] ______, *Bifurcations of invariant manifolds of differential equations and normal forms in neighborhoods of elliptic curves*, Funct. Anal. Appl., 10 (1976), pp. 249–258.

[10] D. G. Aronson, M. A. Chory, G. R. Hall and R. P. McGehee, *A discrete dynamical system with subtly wild behavior*, New Approaches to Nonlinear Problems in Dynamics, P. J. Holmes, ed., Society for Industrial and Applied Mathematics, Philadelphia, 1980, pp. 339–360.

[11] J. F. G. Auchmuty, *Bifurcating waves*, Ann. NY Acad. Sci., 316 (1979), pp. 263–278.

[12] J. F. G. Auchmuty and G. Nicolis, *Bifurcation analysis of non-linear reaction-diffusion equations*. I, Bull. Math. Biology, 37 (1975), pp. 323–365.

[13] ______, *Bifurcation analysis of reaction diffusion equations*. III. *Chemical oscillations*, Bull. Math. Biology, 38 (1976), pp. 325–349.

[14] P. G. Baines and A. E. Gill, *On thermohaline convection with linear gradients*, J. Fluid Mech., 37 (1969), pp. 289–306.

[15] L. Bauer, H. B. Keller and E. L. Reiss, *Multiple eigenvalues lead to secondary bifurcation*, SIAM Rev., 17 (1975), pp. 101–122.

[16] R. I. Bogdanov, *Versal deformations of a singular point of a vectorfield on a plane in the case of zero eigenvalues*, Proc. I. G. Petrovski Seminar, 2 (1976), pp. 37–65. (In Russian.)

[17] ______, *Orbital equivalence of singular points of vectorfields on the plane*, Funct. Anal. Appl., 10 (1976), pp. 316–317.

[18] N. N. Bogoliubov and Y. A. Mitropolsky, *Asymptotic Methods in the Theory of Nonlinear Oscillations*, Gordon and Breach, New York, 1961.

[19] P. Brunovsky, *On one parameter families of diffeomorphisms*, Comment. Math. Univ. Carolinae, 11 (1970), pp. 559–582.

[20] F. H. Busse and R. M. Clever, *Instabilities of convection rolls in a fluid of moderate Prandtl number*, J. Fluid Mech., 91 (1979), pp. 319–335.

[21] M. L. Cartwright, *Forced oscillations in nearly sinusoidal systems*, J. Inst. Elec. Eng., 95 (1948), pp. 88–96.

[22] M. L. Cartwright and J. E. Littlewood, *On non-linear differential equations of the second order*. I. *The equation* $y+k(1-y^2)\dot{y}+y=b\lambda k\cos(\lambda t+a)$, *k large*, J. London Math. Soc., 20 (1945), pp. 180–189.

[23] A. Chenciner and G. Iooss, *Bifurcations de tores invariants*, Arch. Rational Mech. Anal., 69 (1979), pp. 109–198.

[24] S. Chow, J. Hale and J. Mallet-Paret, *Applications of generic bifurcation*. II, Arch. Rational Mech. Anal., 62 (1976), pp. 209–235.

[25] J. D. Cole, *Perturbation Methods in Applied Mathematics*, Blaisdell, Waltham, MA, 1968.

[26] L. N. Da Costa, E. Knobloch and N. O. Weiss, *Oscillations in double-diffusive convection*, J. Fluid. Mech., to appear.

[27] A. Denjoy, *Sur les courbes definies par les équations differentielles à la surface du tore*, J. Math. Pures Appl., (9) 11 (1932), pp. 333–375.

[28] E. H. Dowell, *Nonlinear oscillations of a fluttering plate*, AIAA J., 4 (1966), pp. 1267–1275.

[29] ______, *Aeroelasticity of Plates and Shells*, Noordhoff, Leyden, 1975.

[30] ______, *Nonlinear elasticity*, New approaches to Nonlinear Problems in Dynamics, P. J. Holmes, ed., Society for Industrial and Applied Mathematics, Philadelphia, 1980, pp. 147–172.

[31] F. Dumortier, *Singularities of vector fields in the plane*, J. Differential Equations, 23 (1977), pp. 53–106.

[32] M. J. Feigenbaum, *Quantitative universitality for a class of nonlinear transformations*, J. Stat. Phys., 19 (1978), pp. 25–52.

[33] P. R. Fenstermacher, H. L. Swinney and J. P. Gollub, *Dynamical instabilities and the transition to chaotic Taylor vortex flow*, J. Fluid Mech., 94 (1979), pp. 103.

[34] J. E. Flaherty and F. C. Hoppensteadt, *Frequency entrainment of a forced Van der Pol oscillator*, Stud. Appl. Math., 85 (1978), pp. 5–15.

[35] N. K. Gavrilov and L. P. Sil'nikov, *On three dimensional dynamical systems close to systems with a structurally unstable homoclinic curve*, Math. USSR, Sb., 17 (1972), pp. 467–485; 19 (1973), pp. 139–156. (In Russian.)

[36] A. W. GILLIES, *On the transformation of singularities and limit cycles of the variational equations of Van der Pol*, Quart J. Mech. Appl. Math., VII(2) (1954), pp. 152–167.

[37] M. GOLUBITSKY AND W. F. LANGFORD, *Classification and unfoldings of degenerate Hopf bifurcations*, 41 (1981), pp. 375–415.

[38] M. GOLUBITSKY AND D. SCHAEFFER, *A theory for imperfect bifurcation via singularity theory*, Comm. Pure Appl. Math., 32 (1979), pp. 21–98.

[39] ______, *Imperfect bifurcation in the presence of symmetry*, Comm. Math. Phys., 67 (1979), pp. 205–232.

[40] J. GUCKENHEIMER, *Lectures on Bifurcation Theory, Dynamical Systems*, CIME Lectures 1978, Birkhauser, Basel, 1980.

[41] ______, *On a codimension two bifurcation*, Dynamical Systems and Turbulence, Warwick 1980, Lecture Notes in Mathematics 898, Springer-Verlag, New York, 1981, pp. 99–142.

[42] ______, *Patterns of bifurcation*, New Approaches to Nonlinear Problem in Dynamics, P. J. Holmes, ed., Society for Industrial and Applied Mathematics, Philadelphia, 1980, pp. 71–104.

[43] ______, *A strange, strange attractor*, The Hopf Bifurcation and Its Applications, J. Marsden and M. McCracken, eds., Springer-Verlag, New York, 1976.

[44] ______, *Dynamics of the Van der Pol equation*, IEEE Trans. Circuits and Systems, CS-11 (1980).

[45] J. GUCKENHEIMER AND R. W. WILLIAMS, *Structural stability of Lorenz attractors*, Pub. IHES, 50 (1979), pp. 59–72.

[46] J. K. HALE, *Ordinary Differential Equations*, John Wiley, New York, 1969.

[47] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[48] B. D. HASSARD, *Computation of invariant manifolds*, New Approaches to Nonlinear Problems in Dynamics, P. J. Holmes, ed., Society for Industrial and Applied Mathematics, Philadelphia, 1980, pp. 27–42.

[49] M. HÉNON, *A two dimensional mapping with a strange attactor*, Comm. Math. Phys., 50 (1976), pp. 69–78.

[50] M. HERMAN, *Mesure de Lebesgue et nombre de rotation*, Geometry and Topology, Lecture Notes in Mathematics, 597, Springer-Verlag, New York, 1977, pp. 271–293.

[51] ______, *Sur la conjugaison differentiable des diffeomorphismes du circle à des rotations*, Publ. IHES, 49 (1979), pp. 5–234.

[52] M. HERSCHKOWITZ-KAUFMAN, *Bifurcation analysis of reaction-diffusion equations*. II, Bull. Math. Biology, 37 (1975), pp. 589–636.

[53] M. HERSCHKOWITZ-KAUFMAN AND T. ERNEAUX, *The bifurcation diagram of model chemical reactions*, Ann. NY Acad. Sci., 316 (1976), pp. 296–313.

[54] M. HIRSCH, C. PUGH AND M. SHUB, *Invariant Manifolds*, Lecture Notes in Mathematics 583, Springer-Verlag, New York, 1977.

[55] P. J. HOLMES, *Bifurcations to divergence and flutter in flow-induced oscillations: a finite dimensional analysis*, J. Sound Vibration, 53 (1977), pp. 471–503.

[56] ______, *Domains of stability in a wind induced oscillation problem*, J. Appl. Math., 46 (1979), pp. 672–676.

[57] ______, *A nonlinear oscillator with a strange attractor*, Phil. Trans. Roy. Soc. London Ser. A, 292 (1979), pp. 419–448.

[58] ______, *Unfolding a degenerate nonlinear oscillation: a codimension two bifurcation*, Ann. NY Acad. Sci., to appear.

[59] P. J. HOLMES AND J. MARSDEN, *Bifurcation to divergence and flutter in flow-induced oscillations: an infinite dimensional analysis*, Automatica, 14 (1978), pp. 367–384.

[60] P. J. HOLMES AND D. A. RAND, *Bifurcations of the forced Van der Pol oscillator*, Quart. Appl. Math., (1978), pp. 495–509.

[61] H. E. HUPPERT AND D. R. MOORE, *Nonlinear double diffusive convection*, J. Fluid Mech., 78 (1976), pp. 821–854.

[62] G. IOOSS, *Bifurcation of Maps and Applications*, North-Holland, Amsterdam, 1979.

[63] G. IOOSS AND D. D. JOSEPH, *Elementary Stability and Bifurcation Theory*, Springer-Verlag, New York, 1980.

[64] G. IOOSS AND W. F. LANGFORD, *Conjectures on the routes to turbulence via bifurcations*, Ann NY. Acad. Sci., to appear.

[65] D. JOSEPH, *Stability of Fluid Motions*, 2 vols., Springer-Verlag, New York, 1976.

[66] R. JOST AND E. ZEHNDER, *A generalization of the Hopf bifurcation theorem*, Helvetica Physica Acta, 45 (1972), pp. 258–276.

[67] J. L. KAPLAN AND J. A. YORKE, *Preturbulence: a regime observed in a fluid flow of Lorenz*, preprint.

[68] J. P. KEENER, *Perturbed bifurcation theory at multiple eigenvalues*, Arch. Rational Mech. Anal., 56 (1974), pp. 348–366.

[69] ______, *Secondary bifurcation in non-linear diffusion reaction equations*, II, Stud. Appl. Math., 55 (1976), pp. 187–211.

[70] ______, *Infinite period bifurcation and global bifurcation branches*, preprint.

[71] J. P. KEENER AND H. B. KELLER, *Perturbed bifurcation theory*, Arch. Rational Mech. Anal., 50 (1973), pp. 159–175.

[72] H. B. KELLER AND W. F. LANGFORD, *Iterations, perturbations and multiplicities for non-linear bifurcation problems*, Arch. Rat. Mech. Anal. 48 (1972), pp. 83–108.

[73] A. KELLEY, *The stable, center-stable, center, center-unstable and unstable manifolds*, J. Differential Equations, (1967), pp. 546–570.

[74] E. KNOBLOCH AND M. R. PROCTOR, *Nonlinear periodic convection in double diffusive systems*, J. Fluid Mech., 108 (1981), pp. 291–316.

[75] Y. KURAMOTO, *Diffusion induced chaos in reacting systems*, Suppl. Prog. Theor. Phys., 64 (1978), pp. 346–367.

[76] L. LANDAU AND E. M. LIFSCHITZ, *Fluid Mechanics*, Pergamon, Oxford, 1959.

[77] W. LANGFORD, *Periodic and steady-state mode interactions lead to tori*, SIAM J. Appl. Math., 37 (1979), pp. 22–47.

[78] W. LANGFORD AND G. IOOSS, *Interactions of Hopf and pitchfork bifurcations*, Workshop on Bifurcation Problems, H. Mittelman, ed., Birkhauser Lecture Notes, Birkhauser, Basel, 1980.

[79] R. LEFEVER AND I. PRIGOGINE, *Symmetry breaking instabilities in dissipative systems*. II, J. Chem. Phys., 48 (1968), pp. 1695–1700.

[80] E. N. LORENZ, *Deterministic nonperiodic flow*, J. Atmospheric Sci., 20 (1963), pp. 130–141.

[81] T. MAHAR AND B. MATKOWSKY, *A model chemical reaction exhibiting secondary bifurcation*, SIAM J. Appl. Math., 32 (1977), pp. 394–404.

[82] J. E. MARSDEN, *Qualitative methods in bifurcation theory*, Bull. Amer. Math. Soc., 84 (1978), pp. 1125–1148.

[83] J. MARSDEN AND M. MCCRACKEN, *The Hopf Bifurcation and Its Applications*, Springer-Verlag, New York, 1976.

[84] J. N. MATHER, *Stability of C-mappings*, I, Ann. of Math, 87 (1968), pp. 89–104; II, Ann. of Math, 89 (1969), pp. 254–291; III, Publ. Math. IHES, 35 (1968), pp. 127–156; IV, Publ. Math. IHES, 37 (1969), pp. 223–248; V, Adv. in Math., 4 (1970), pp. 301–335; VI, Lecture Notes in Mathematics, 192, Springer-Verlag, New York, 1971, pp. 207–254.

[85] B. J. MATKOWSKY AND E. L. REISS, *Singular perturbations of bifurcations*, SIAM J. Appl. Math., 33 (1977), pp. 230–235.

[86] J. MARSDEN AND M. MCCRACKEN, *The Hopf Bifurcation Theorem and Its Applications*, Springer-Verlag, New York, 1976.

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

[88] D. W. MOORE AND E. A. SPIEGEL, *A thermally excited non-linear oscillator*, Astrophys. J., 143 (1966), pp. 871–887.

[89] S. NEWHOUSE, *Diffeomorphisms with infinitely many sinks*, Topology, 12 (1974), pp. 9–18.

[90] ______, *The abundance of wild hyperbolic sets and non-smooth stable sets for diffeomorphisms*, Publ. IHES, 50 (1979), pp. 101–151.

[91] ______, *On simple arcs between structurally stable flows*, Dynamical Systems-Warwick, 1974, Lecture Notes in Mathematics, Springer-Verlag, New York, 1975, pp. 209–233.

[92] S. NEWHOUSE AND J. PALIS, *Bifurcations of Morse-Smale dynamical systems*, Dynamical Systems, M. Peixoto, ed., Academic Press, New York, 1973, pp. 303–366.

[93] ______, *Cycles and bifurcation theory*, Asterisque, 31 (1976), pp. 43–140.

[94] S. NEWHOUSE, J. PALIS AND F. TAKENS, *Stable arcs of diffeomorphisms*, Bull. Amer. Math. Soc., 82 (1976), pp. 409–502.

[95] S. NEWHOUSE AND M. PEIXOTO, *There is a simple arc between any two Morse-Smale flows*, Asterisque, 31 (1976), pp. 15–42.

[96] J. PALIS, *Some developments on stability and bifurcations of dynamical systems*, Geometry and Topology, J. Palis and M. do Carmo, eds., Lecture Notes in Mathematics, 597, Springer-Verlag, New York, 1977, pp. 495–506.

[97] ______, *A differentiable invariant of toplogical conjugacies and moduli of stability*, Asterisque, 33, pp. 49–50.

[98] J. PALIS AND F. TAKENS, *Topological equivalence of normally hyperbolic dynamical systems*, Topology, 16 (1977), pp. 335–345.

[99] H. POINCARÉ, *Les methodes nouvelles de la mécanique celeste*, vols. I, II, III, Paris, 1892, 1893, 1899; reprint, Dover, New York, 1957.

[100] P. RABINOWITZ, ed., *Applications of Bifurcation Theory*, Academic Press, New York, 1977.

[101] L. A. RUBENFELD AND W. L. SIEGMAN, *Nonlinear dynamic theory for a double-diffusive convection model*, SIAM J. Appl. Math., 32 (1977), pp. 871–894.

[102] D. RUELLE, *Sensitive dependence on initial condition and turbulent behavior of dynamical systems*, Conference on Bifurcation Theory and Its Applications, NY Academy of Science, 1977.

[103] D. RUELLE AND F. TAKENS, *On the nature of turbulence*, Comm. Math. Phys., 20 (1971), pp. 167–192.

[104] D. H. SATTINGER, *Bifurcation and symmetry breaking in applied mathematics*, Bull. Amer. Math. Soc., 3 (1980), pp. 779–819.

[105] D. SCHAEFFER AND M. GOLUBITSKY, *Bifurcation analysis near a double eigenvalue of a model chemical reaction*, Arch. Nat. Mech. Anal., 67 (1981), pp. 315–347.

[106] ______, *Boundary conditions and mode jumping in the buckling of a rectangular plate*, Comm. Math. Phys., 69 (1979), pp. 209–236.

[107] M. SHEARER, *Secondary bifurcation near a double eigenvalue*, this Journal, 11 (1980), pp. 365–389.

[108] C. L. SIEGEL, *Über die normalform analytischer Differentialgleichungen in der Nähe einer Gleichgewichtslösung*, Gottinger Nachrichten der Akad. der Wissenschaften, (1952), pp. 21–30.

[109] C. L. SIEGEL AND J. MOSER, *Lectures on Celestial Mechanics*, Springer-Verlag, New York, 1971.

[110] L. SIL'NIKOV, *On a new type of bifurcation of multidimensional dynamical systems*, Sov. Math. Dokl., 10 (1969), pp. 1368–1371.

[111] L. P. SIL'NIKOV, *A contribution to the problem of the structure of an extended neighborhood of a structurally stable equilibrium of saddle focus type*, Math. USSR Sb., 10 (1970), pp. 91–102. (In Russian.)

[112] S. SMALE, *Diffeomorphisms with many periodic points*, Differential and Combinatorial Topology, Princeton, Univ. Press, Princeton, 1965, pp. 63–80.

[113] ______, *Differentiable dynamical systems*, Bull. Amer. Math. Soc., 73 (1967), pp. 747–817.

[114] J. SOTOMAYOR, *Bifurcations of vector fields on two dimensional manifolds*, Publ. IHES, 43 (1973), pp. 1–46.

[115] ______, *Generic bifurcations of dynamical systems*, Dynamical Systems, M. Peixoto, ed., Academic Press, New York, 1973, pp. 561–582.

[116] ______, *Structural stability and bifurcation theory*, Dynamical Systems, M. Peixoto, ed., Academic Press, New York, 1973, pp. 549–560.

[117] S. STERNBERG, *Local contractions and a theorem of Poincaré*, Amer. J. Math., 79 (1957), pp. 787–789.

[118] ______, *On the structure of local homeomorphisms of Euclidean n-space*, II, III, Amer. J. Math., 80 (1958), pp. 623–631, 81 (1959), pp. 578–604.

[119] J. J. STOKER, *Nonlinear Vibrations in Mechanical and Electrical Systems*, Wiley-Interscience, New York,

[120] F. TAKENS, *Partially hyperbolic fixed points*, Topology, 10 (1971), pp. 133–147.

[121] ______, *A nonstabilized jet of singularity of a vectorfield*, Dynamical Systems, M. Peixoto, ed., Academic Press, New York, 1973, pp. 583–598.

[122] ______, *Unfoldings of certain singularities of vectorfields: generalized Hopf bifurcations*, J. Differential Equations, 14 (1973), pp. 476–493.

[123] ______, *Integral curves near mildly degenerate singular points of vectorfields*, Dynamical Systems, M. Peixoto, ed., Academic Press, New York, 1973, pp. 599–617.

[124] ______, *Forced oscillations and bifurcations*, Applications of Global Analysis, Communications of Maths Institute, Rijksuniversiteit, Utrecht, 3 (1974), pp. 1–59.

[125] ______, *Singularities of vector fields*, Publ. IHES, 43 (1973), pp. 47–100.

[126] R. THOM, *Structural Stability and Morphogenesis*, W. A. Benjamin, Reading, MA., 1975.

[127] A. UPPAL, W. H. RAY AND A. B. POORE, *On the dynamic behavior of continuous stirred tank reactors*, Chemical Engrg. Science, 29 (1974), pp. 967–985.

[128] G. VERONIS, *On finite amplitude instability in thermohaline convection*, J. Marine Res., 23 (1965), pp. 1–17.

[129] ______, *Effect of a stabilizing gradient of solute on thermal convections*, J. Fluid Mech., 34 (1968), pp. 315–336.

[130] R. F. WILLIAMS, *The structure of Lorenz attractors*, Publ. IHES, 50 (1979).

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1501-0002 $01.25/0

CAUSTICS IN EXTENDED EUCLIDEAN SPACE*

J. W. BRUCE,[†] P. J. GIBLIN[‡] AND C. G. GIBSON[§]

Abstract. This paper is one of a series in which the authors investigate genericity properties of caustics by reflexion from smooth mirrors. Here, the mirror is a smooth surface in $\mathbb{R}^3$, and two related problems are considered: (1) what is the form of the caustic at infinity (the caustic in the "far field"), generically, with a finite light source; (2) what is the form of the caustic generically when the light source is at infinity? For (1) both "mirror genericity," where deformations of the mirror are allowed, and "source genericity," where only the source may be moved, are considered. With some assumptions on M it is shown that the caustic can be made generic in either of these two ways. (In the latter case, only a local result is proved.) For (2), only mirror genericity is considered, and it is shown that there is an obstruction to proving a result of this kind, caused by an inherent lack of genericity in wavefronts arising from parallel light reflected from a mirror in $\mathbb{R}^3$. It is proved, however, that for most mirrors in $\mathbb{R}^3$ with parallel incident light, the part of the caustic lying in a compact region is generic.

AMS-MOS subject classification (1980). Primary 58C27, 78A05

Key words. caustic, singularity, genericity, optics

Introduction. This paper is devoted to the study of light caustics obtained by reflexion of a light source in a smooth convex mirror M in $\mathbb{R}^3$. Here we consider the related questions:

(1) What is the form of the caustic at infinity, generically, with finite light source?

(2) What is the form of the caustic generically when the light source is at infinity?

Both questions are of some physical interest. In particular the study of question (1) is relevant for the appearance of the caustic in the far field, i.e. its appearance when cut by a distant screen. (Cutting the caustic by a screen is one way, of course, of actually seeing the caustic.) See §1.

The two questions are related because following [5] and [6] one can see that they can both be attacked using contact of paraboloids of revolution with the mirror M, without constructing an intermediate wavefront. (This observation is very useful since it simplifies the computations.) For question (1) the focus of the paraboloid is the light source and its axis is the direction of the reflected light. For question (2) the axis will be the direction of incident light and the focus the corresponding point on the caustic. However, the two situations are very different when it comes to the genericity questions. We will as before be considering two basic questions:

(A) Mirror genericity: Is it true that, fixing the light source, for almost any mirror the corresponding caustic is generic?

(B) Source genericity: Is it true that, fixing the mirror, almost any light source will give a generic caustic?

Concerning the more interesting question (B) we see that we have a far better chance of proving source genericity in (1) than in (2). For in question (1) we have a

* Received by the editors May 26, 1982 and in revised form November 27, 1982.

[†] Department of Mathematics, University College, Cork, Ireland.

[‡] Department of Pure Mathematics, University of Liverpool, Liverpool L69 3BX, England. This author was a visitor at the University of North Carolina at Chapel Hill while this article was written.

[§] Department of Pure Mathematics, University of Liverpool, Liverpool L69 3BX, England. This author was a visitor at Sonderforschungsbereich "Theoretische Mathematik" of Bonn University, West Germany, for part of the period during which this article was written.

three-parameter family of peturbations of the source and we are asking for a two-parameter family of functions (the height functions on the wavefront) to be generic. In question (2) we only have a two-parameter family of perturbations of the source (the different incident directions) and we are asking for a three-parameter family of functions (distance squared function on the wavefront) to be generic. (We have interchanged the roles of the axis and focus.)

In §§1 and 2 we obtain some positive results concerning source genericity for the caustic at infinity. The principal one (Theorem 2.12) states that if the mirror M is analytic then for almost any position of light source s inside M the caustic at infinity is locally generic. The corresponding assertion for mirror genericity (for a general smooth M) already follows from our work in [3] and the theorem of Looijenga which shows that generic wavefronts give generic caustics at infinity.

In §3 we begin by discussing some problems concerning mirror genericity with light source at infinity. (The fact that there *are* problems vindicates again our study of these questions; genericity results *do* need to be proved.) We do show (Theorem 3.3) that, given any compact region K of $\mathbb{R}^3$ and incident direction s for the light, for an open dense set of mirrors that part of the caustic in K is generic. The version of mirror genericity which asks for a generic compactified caustic in projective 3-space is shown to fail. Because of these problems, and the complications involved in getting results in the easier case of finite light source in [4], we do not consider source genericity for question (2). (Note that it already fails in the plane: see [6].)

1. Caustics and screens. In practice one usually observes light caustics by placing a screen in the ambient space; the caustic will then appear as a bright curve on the screen. Thus although for generic wavefronts in $\mathbb{R}^3$ the caustic has local forms of the type listed in [5], in practice one observes two-dimensional sections of these local forms. What form will these sections take? For a generic wavefront and generic screen one will observe the models associated with a one-dimensional wavefront as in Fig. 1. For if L is the screen and W the wavefront, then the proof of the genericity of wavefronts given by Looijenga in [10] (see also [12]) shows that a generic W gives a generic family of distance squared functions $W\times L\rightarrow\mathbb{R}$ and hence a generic Lagrangian or catastrophe map, with L as control space rather than the whole of $\mathbb{R}^3$.

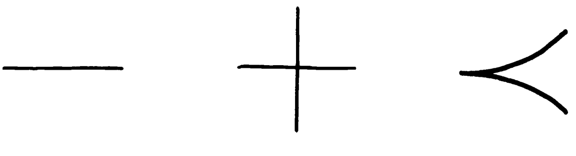


FIG. 1

In this paper we wish to prove a different type of result. We claim that choosing a screen sufficiently far away from the generic wavefront W one should obtain the generic forms of the diagram above, and the caustic curve on the screen should not change if we move the screen (i.e. the caustics should be isotopic). In physicists' terminology we are studying the *caustic in the far field*. To do this we need to introduce the family of height functions on W. So consider $S=\{(t,a)\in\mathbb{R}\times\mathbb{R}^3\colon t^2+\|a\|^2=1\}$ and define $G\colon W\times S\rightarrow\mathbb{R}$ by $G(x,t,a)=t\|x\|^2-2\langle a,x\rangle=t\|x-at^{-1}\|^2-t^{-1}\|a\|^2$. This is the compactification of the distance squared functions discussed in [10] and [12, p. 713]. For $t=0$ we have the height functions on W. As usual the (extended) caustic is the set of points (t,a) of S for which $G_{(t,a)}\colon W\rightarrow\mathbb{R}$ has an $A_{\geq 2}$ at some point. (We can think of this extended caustic as lying in the projective space $\mathbb{P}^3$, i.e. S modulo antipodal points, for $G_{(-t,-a)}=-G_{(t,a)}$.) If $t\neq 0$ a point of the extended caustic corresponds to a point of the usual caustic (i.e. at^{-1} is a centre of curvature of W.) However, the new points of the caustic in S are those on the sphere (at infinity) given by $\{t=0\}$. This is in fact the fold curve of the Gauss map of W (see [2]). The main idea is that *this caustic at infinity is essentially the caustic in the far field, i.e. the caustic cut with say a spherical screen of sufficiently large radius centered at the origin.* The key lemma we require is:

LEMMA 1.1. *For a generic wavefront the sphere $\{t=0\}$ cuts the natural stratification of the extended caustic transversally.*

Proof. By Looijenga's transversality result both G and its restriction to $W\times\{t=0\}$ are generic families of functions for generic W. Thus one generically has strata of type A_2,A_3 and A_2/A_2 of the extended caustic meeting the sphere $\{t=0\}$. To prove that $\{t=0\}$ meets the A_2,A_3 and A_2/A_2 strata of S *transversally* we need the following rather trivial extension of the standard Thom lemma. Let F: $X\times Y\rightarrow Z$ be smooth and $A\subset Y$, $B\subset Z$ be smooth submanifolds. If F is transverse to B then the restricted map F: $X\times A\rightarrow Z$ is transverse to B if and only if the projection π: $F^{-1}(B)\subset X\times Y\rightarrow Y$ is transverse to A.

(*Proof.* This is a local assertion, so near a point $a\in A$ write Y as $A\times A'$ with a now written $(a,a')\in A\times A'$. Write π_1 for projection of Y onto A' and consider F: $X\times A\times A'\rightarrow Z$. Now F is transverse to B, so by Thom's lemma $F_{a'}$: $X\times A\rightarrow Z$ is transverse to B at (x,a,a') if and only if the composite

$$F^{-1}(B)\xrightarrow{\pi}A\times A'\xrightarrow{\pi_1}A'$$

is a submersion at (x,a,a'). This is so if and only if π is transverse to A at (x,a,a') Q.E.D.)

We now take $X=W^{(r)}$, $Y=S$, $A=\{t=0\}$, $F={}_rj_1^kG$, $Z={}_rJ^k(W,\mathbb{R})$ and B the A_2 or A_3 set ($r=1$) or A_2/A_2 set (when $r=2$). For generic W we have ${}_rj_1^kG$ transverse to the relevant B as well as its restriction to $W^{(r)}\times\{t=0\}$. So the projection π: $({}_rj_1^kG)^{-1}(B)\rightarrow S$ is transverse to $\{t=0\}$, as required. Q.E.D.

The relevance of the lemma is as follows. We have a two-sheeted mapping $S-\{t=0\}\rightarrow\mathbb{R}^3$ defined by $(t,a)\rightarrow t^{-1}a$, with a (one-sided) inverse $\mathbb{R}^3\rightarrow S-\{t=0\}$ defined by $x\rightarrow((1+\|x\|^2)^{-1/2}, x(1+\|x\|^2)^{-1/2})$. Thus the spheres centred at $0\in\mathbb{R}^3$ correspond to the sets $t=\delta$, a constant, in S. Since $\{t=0\}$ meets the extended caustic transversally, so will $t=\delta$ for δ small, and we will have isotopic caustic curves on all sufficiently large spheres (with corresponding radius $\delta^{-2}-1$.)

One final crucial observation: in the proof of Lemma 1.1 above we assumed that G was a generic family. If however we only have G: $W\times\{t=0\}\rightarrow\mathbb{R}$ generic then since genericity is open we will have G: $W\times U\rightarrow\mathbb{R}$ generic for some suitably small neighbourhood U of $\{t=0\}$ provided W is compact. This in fact suffices to prove the lemma in this case, and make the deductions above. So for caustics in the far field we need only consider the family of height functions on compact wavefronts W.

Consequently let us review some facts concerning height functions on smooth surfaces in $\mathbb{R}^3$. (See [2] for details.) A height function in a given direction has an $A_{\geq 2}$ singularity at a point p of a surface W in $\mathbb{R}^3$ if and only if the normal to W at p is in the given direction and the Gaussian curvature of W at p is zero. Generically only A_1,A_2 and A_3 singularities occur. The A_3 singularities occur at points where the rib lines cut the parabolic curves, and give rise to cusps of the Gauss mapping. Contact of $A_{\leq 2}$ of a surface with its tangent plane is easily seen to be automatically transverse. For an A_3 point, however, there is a genuine condition to be satisfied.

As in [5] we shall study the caustic (at infinity this time) via contact between the mirror and quadrics (in fact paraboloids) of revolution rather than considering height functions on the wavefront. (We shall only consider local genericity, i.e., we ignore self-intersections of the caustic.) These two approaches yield the same result; the fact that contact is preserved on taking orthotomics is proved in [3, §4] and [1], and the proof that an A_3 arising from the paraboloids is transverse if and only if that arising from the corresponding height function is transverse is similar to the proof in [5, Appendix] for the A_4 and D_4 cases when considering distance squared functions.

Let M be the mirror and let R denote the space of quadratic forms on $\mathbb{R}^3$ whose zero-set is a paraboloid of revolution. In §2 we parametrize R by $\mathbb{R}\times\mathbb{R}^3\times\mathbb{R}^3$ where one factor $\mathbb{R}^3$ gives the focus s, which corresponds to the light source in our work. In what follows we take $s\in\operatorname{Int}M$, the open region of $\mathbb{R}^3$ inside M. There is a *global contact map*

$$\Gamma\colon M\times R\to\mathbb{R}$$

defined by $\Gamma(m,Q)=Q(m)$. What we actually do in §2 is to show that, for some closed set Ω of measure zero in $\operatorname{Int}M$, the map Γ with s restricted to $\operatorname{Int}M-\Omega$ has jet extension (as a family of functions on M) transverse to various strata in $J^k\,(M,\mathbb{R})$. This is done by using a local version γ of Γ and working case by case. (In fact, for strata of singularities of corank 2 we show directly that the contact map avoids these strata (Lemma 2.2).) We can now deduce from Thom's basic transversality lemma that, for all s off a set of measure zero in $\operatorname{Int}M-\Omega$ (or in $\operatorname{Int}M$), the contact mapping with s held *fixed* is transverse to these strata. This is equivalent to the corresponding result for height functions on the orthotomic (thus for $A_{\geq 4}$ and corank 2 the strata are actually avoided for almost all s), and establishes the genericity of the caustic at infinity for almost all source positions.

2. Transversality of the contact map. In this section we present detailed computations concerning the contact of the mirror M with paraboloids of revolution. We write A for the real vector space of all quadratic functions on $\mathbb{R}^3$ (i.e. with only constant, linear and quadratic terms) and R for the subset of those whose zero-sets are paraboloids of revolution. Note that R is of course closed under multiplication by nonzero scalars. We wish to show that R is actually a smooth submanifold of A, and to compute its tangent space at any point.

A paraboloid of revolution has a focus $s=(u,v,w)$ and a directrix plane. If t is in the direction of the axis ($t=(p,q,r)\neq 0$) then for some $k\in\mathbb{R}$ the directrix plane has the form $t\cdot\alpha=k$, where $\alpha=(x,y,z)$. Then the distances of any point of the paraboloid of revolution from the focus and the directrix are equal. We always assume that s does not lie on the directrix plane.

LEMMA 2.1. *The equation of the paraboloid of revolution with focus s and directrix* $t\cdot\alpha=k$ *is*

$$(*)\qquad (k-t\cdot\alpha)^2-t^2(s-\alpha)^2=0.$$

Here $t^2=t\cdot t$ is the square of the length of t, etc. The paraboloid can also be described as the antiorthotomic of the directrix plane relative to the focus. In our situation the directrix plane is the tangent plane to the wavefront W, and the focus is either the light source or the point on the caustic corresponding to light from infinity in the direction of the axis. Contact between W and its tangent plane is the same as contact between M and the paraboloid of revolution.

The left side of (∗) yields an explicit parametrization of R (more precisely of "half" of R: for every $Q \in R$ either Q or $-Q$, but not both, has the form given), the parameter space being the open subset of $\mathbb{R} \times \mathbb{R}^3 \times \mathbb{R}^3$ comprising those points (k,s,t) with $t \neq 0$ and $t \cdot s \neq k$. We claim that this parametrization has maximal rank 7. If we differentiate (∗) with respect to k,u,v,w respectively we see that the image of the differential contains $k - t \cdot \alpha$, $x-u$, $y-v$, $z-w$: the condition $t \cdot s \neq k$ ensures that these vectors are linearly independent, so the image of the differential contains $1,x,y,z$. If we now differentiate (∗) with respect to p,q,r respectively and disregard linear combinations of $1,x,y,z$ in the result, we see that

$$X = x(px+qy+rz) - p(x^2+y^2+z^2),$$
$$Y = y(px+qy+rz) - q(x^2+y^2+z^2),$$
$$Z = z(px+qy+rz) - r(x^2+y^2+z^2)$$

all lie in the image of the differential. One readily checks that X, Y, Z are linearly independent if and only if $t \neq 0$, which condition is automatically satisfied. Thus R is a smooth immersed submanifold of A of dimension 7 with tangent space spanned by $1,x,y,z,X,Y,Z$. In practice we shall be concerned more with the subset $R_{z=0}$ of R, comprising quadratic functions with zero-set paraboloids of revolution passing through the origin 0 in $\mathbb{R}^3$, and tangent there to the plane $z=0$. $R_{z=0}$ is likewise a smooth submanifold of A, of dimension 4, with tangent space spanned by z,X,Y,Z. In fact, by the reflexion property for paraboloids of revolution, the reflexion $\bar{s}=(u,v,-w)$ of the focus in the plane $z=0$ must be a scalar multiple of $t=(p,q,r)$, so we can write

$$X = x(ux+vy-wz) - u(x^2+y^2+z^2),$$
$$Y = y(ux+vy-wz) - v(x^2+y^2+z^2),$$
$$Z = z(ux+vy-wz) + w(x^2+y^2+z^2).$$

A trivial calculation based on Lemma 2.1 shows that the unique paraboloid of revolution

$$Q = ax^2+by^2+cz^2+2dxy+2eyz+2fzx+gz$$

through 0 in $\mathbb{R}^3$, tangent to the plane $z=0$, with focus $s=(u,v,w)$, where $w \neq 0$, has coefficients

$$\begin{array}{lll} a=v^2+w^2, & b=u^2+w^2, & c=u^2+v^2, \\ d=-uv, & e=vw, & f=uw, \end{array}$$
$$g=-4w(u^2+v^2+w^2).$$

Geometrically, it is more illuminating to write $Q = \Lambda S - P^2$ where $\Lambda = u^2+v^2+w^2$, $S = x^2+y^2+z^2-4wz$, $P = ux+vy-wz$. S represents the sphere through 0 centered at the point where the z-axis meets the axis of Q, and P represents the plane through the origin perpendicular to the axis of Q, so a translate of the directrix plane.

Contact of M with paraboloids of revolution. We need to make a few preliminary comments about the contact of M at a fixed point with paraboloids of revolution. By a rigid motion of $\mathbb{R}^3$ we can suppose the point in question is 0, and that (close to 0) the surface is given in the Monge form

$$z = \tfrac{1}{2}(\kappa_1 x^2 + \kappa_2 y^2) + C(x,y) + D(x,y) + \cdots$$

where $C, D, \cdots$ are binary forms of degrees $3, 4, \cdots$ and κ_1, κ_2 are the principal curvatures at 0. We are concerned only with surfaces M having positive Gaussian curvature, so we can suppose $\kappa_1>0$, $\kappa_2>0$ and $\kappa_1 \leq \kappa_2$. Now let N be an open neighborhood of 0 in $\mathbb{R}^2$ on which the function $z=z(x,y)$ is defined. We define the *contact map* $\gamma: N \times R \to \mathbb{R}$ by the formula

$$\gamma(x,y,Q)=Q(x,y,z(x,y)).$$

Of course for a fixed Q passing through 0 the germ of this function at 0 represents the contact [3, §4] of Q with M at 0, and is singular if and only if Q is tangent to $z=0$ at 0. For such a paraboloid of revolution

$$Q=ax^2+by^2+cz^2+2dxy+2eyz+2fzx+gz$$

with a, b, c, d, e, f, g expressed (as above) in terms of the coordinates u, v, w of the focus, the contact germ has initial Taylor expansion

$$\gamma(x,y)=\left(a+\tfrac{1}{2}\kappa_1 g\right)x^2+2dxy+\left(b+\tfrac{1}{2}\kappa_2 g\right)y^2+\cdots.$$

We wish to consider in detail the conditions for γ to be transverse to the various canonical strata in the jet-space: of course, this refers strictly not to γ, but to a corresponding jet-extension into a jet-space $J^k(2, 1)$.

Strata of corank 2. The condition for contact of corank 2 is that the quadratic part of the germ should be identically zero, i.e.,

$$\begin{aligned}
&0=2d=-2uv,\\
&0=a+\tfrac{1}{2}\kappa_1 g=v^2+w^2-2\kappa_1 w(u^2+v^2+w^2),\\
&0=b+\tfrac{1}{2}\kappa_2 g=u^2+w^2-2\kappa_2 w(u^2+v^2+w^2).
\end{aligned}$$

Note that since we are only interested in positions of the source s *inside* the ovaloid M, it cannot lie on any tangent plane to M, so we can always assume $w \neq 0$. When 0 is an umbilic of M, i.e. $\kappa_1=\kappa_2=\kappa$ (say), these equations have the unique solution $u=0$, $v=0$, $w=1/2\kappa$, i.e. the source is the midpoint of the line-segment joining 0 to the unique centre of principal curvature. And at a nonumbilical point the equations have two real solutions $w=1/2\kappa_1$, $u^2=(1/4\kappa_2)(1/\kappa_1-1/\kappa_2)$ in the principal plane $v=0$. These elementary deductions suggest that for almost all positions of the source we should be able to avoid contact of corank 2. Indeed at each point $m \in M$ we have at most two positions of the source giving rise to contact of corank 2, and as m moves over M we expect these points to sweep out a surface avoidable by arbitrarily small changes of the source. The corresponding mental picture on the orthotomic of M is that one is trying to force all the umbilics off the parabolic curve by small changes in the source.

LEMMA 2.2. *There exists a closed set $\Omega \subseteq \operatorname{Int} M$, of Lebesgue measure zero, such that for every position of the source s off Ω there is no paraboloid of revolution with focus s having contact of corank* 2 *with M.*

Proof. For $m \in M$ write Ω_m for the set of (at most two) positions of the source s for which a paraboloid of revolution with focus s has corank 2 contact with M at m. Take Ω' (resp. Ω'') to be the union of the sets Ω_m with m an umbilic (resp. not an umbilic). Using the local triviality of the normal bundle of M one sees easily that Ω' has Lebesgue measure zero. And that Ω'' likewise has Lebesgue measure zero follows exactly the initial part of [5, proof of Prop. (4.7)]. Thus $\Omega=\Omega' \cup \Omega''$ is of Lebesgue measure zero as well. Moreover Ω is closed. Indeed Ω is the image under the proper projection $M\times\mathbb{R}^3\rightarrow\mathbb{R}^3$ of the set V of pairs (m,p) with $m\in M$, $p\in\Omega_m$, so it suffices to show V is closed i.e. its complement is open in $M\times\mathbb{R}^3$. And that is a trivial consequence of the explicit description of Ω_m given above. Q.E.D.

Generic strata of corank 1. For any integer $k\geq 1$ the contact map $\gamma: N\times R\rightarrow\mathbb{R}$ induces a single jet-extension into the jet-space $J_0^k(2,1)$ without constants. Our next objective is to discuss the transversality of this map to the strata of corank 1, i.e. the A_n-strata. Note that the vector sum of the image of the differential and the tangent space to the orbit contains the vectors in the Jacobian ideal J_γ, modulo terms of degree $\geq k+1$: the remaining generators arise by taking the list of tangent vectors to R, and substituting $z=z(x,y)$ in each. In particular, the tangent vectors x,y produce the linear terms in the jet-space, so we need only consider terms of degree ≥ 2, and the list of tangent vectors to $R_{z=0}$. We start with the generic strata of corank 1.

LEMMA 2.3. *The contact map is always transverse to the A_n-strata for $n\leq 3$.*

Proof. For these strata the Jacobian ideal J_γ will contain all monomials of degree ≥ 3, so we need only consider x^2, xy, y^2. Modulo terms of degree ≥ 3 we have $Z\equiv w(x^2+y^2)$, and the tangent space to the orbit contains

$$\frac{1}{2}x\frac{\partial\gamma}{\partial x}\equiv\left(a+\frac{g}{2}\kappa_1\right)x^2+dxy,\qquad \frac{1}{2}x\frac{\partial\gamma}{\partial y}\equiv dx^2+\left(b+\frac{g}{2}\kappa_2\right)xy,$$

$$\frac{1}{2}y\frac{\partial\gamma}{\partial x}\equiv\left(a+\frac{g}{2}\kappa_1\right)xy+dy^2,\qquad \frac{1}{2}y\frac{\partial\gamma}{\partial y}\equiv dxy+\left(b+\frac{g}{2}\kappa_2\right)y^2.$$

These five vectors span the same space as x^2, xy, y^2 unless $a+g\kappa_1/2=0$, $d=0$, $b+g\kappa_2/2=0$: but these are precisely the conditions for contact of corank 2, and the result follows. Q.E.D.

The A_4-stratum. This stratum requires a rather detailed analysis. A necessary preliminary is to observe that the condition for γ to be of corank 1 is that at least one of $a+g\kappa_1/2$, $d, b+g\kappa_2/2$ should be nonzero, and that

$$d^2=\left(a+\tfrac{1}{2}\kappa_1 g\right)\left(b+\tfrac{1}{2}\kappa_2 g\right).$$

Patient manipulation, using the fact that we can always assume $w\neq 0$, reduces this to

$$(**)\qquad 4\kappa_1\kappa_2 w(u^2+v^2+w^2)-2\{\kappa_1 u^2+\kappa_2 v^2+(\kappa_1+\kappa_2)w^2\}+w=0$$

defining a circular cubic surface in $\mathbb{R}^3$, in fact precisely the first discriminant surface of [5, §3]. At an umbilic of M, where $\kappa_1=\kappa_2=\kappa$ (say), this reduces to the sphere $u^2+v^2+w^2=w/2\kappa$, and its tangent plane $w=1/2\kappa$. This simple fact allows us to make a rather direct attack on the question of transversality to the A_4-orbit at an umbilic of M, i.e., we can assume the source lies either on the sphere, or on the plane, and consider each possibility separately.

LEMMA 2.4. *Suppose* 0 *is an umbilic of M. There are only finitely many points on the plane $w=1/2\kappa$ where the contact map fails to be transverse to the A_4-orbit.*

Proof. For the A_4-orbit the Jacobian ideal J_γ contains all monomials of degree ≥ 4, so we can work modulo such terms. In fact we need only consider x^3, x^2y, xy^2, y^3 for then the argument of Lemma 2.3 will produce x^2, xy, y^2 as well. Modulo terms of degree ≥ 4 the tangent vectors to $R_{z=0}$ yield

$$z\equiv\frac{\kappa}{2}(x^2+y^2)+C(x,y),$$

$$X \equiv uy^2 - vxy + \frac{x}{4}(x^2+y^2),$$
$$Y \equiv vx^2 - uxy + \frac{y}{4}(x^2+y^2),$$
$$Z \equiv \frac{1}{\kappa}(x^2+y^2) + \kappa(ux+vy)(x^2+y^2).$$

Now when $w = 1/2\kappa$ we have $a + \frac{1}{2}\kappa g = -u^2$, $d = -uv$ and $b + \frac{1}{2}\kappa g = -v^2$. Using these relations we see that, modulo terms of degree ≥ 4, the tangent space to the A_4-orbit contains

$$\frac{1}{2}x^2\frac{\partial\gamma}{\partial x} \equiv ux^2(ux+vy), \quad \frac{1}{2}x^2\frac{\partial\gamma}{\partial y} \equiv vx^2(ux+vy),$$
$$\frac{1}{2}xy\frac{\partial\gamma}{\partial x} \equiv uxy(ux+vy), \quad \frac{1}{2}xy\frac{\partial\gamma}{\partial y} \equiv vxy(ux+vy),$$
$$\frac{1}{2}y^2\frac{\partial\gamma}{\partial x} \equiv uy^2(ux+vy), \quad \frac{1}{2}y^2\frac{\partial\gamma}{\partial y} \equiv vy^2(ux+vy).$$

We can assume $u \neq 0$ or $v \neq 0$, else the contact is of corank 2 so these vectors span the space of binary cubics with factor $ux+vy$, and we can work modulo such terms. Note now that from z, X, Y, Z we can construct essentially one linear combination without quadratic terms, namely

$$x - \kappa^2 Z \equiv C(x,y) - \frac{\kappa^3}{4}(ux+vy)(x^2+y^2).$$

Transversality can only fail if $ux+vy$ is a factor of this, i.e.

(i) $$C(-v,u) = 0.$$

We require a further condition on u, v which we obtain as follows. The tangent space to the A_4-orbit also contains $x\partial\gamma/\partial x$, $y\partial\gamma/\partial x, x\partial\gamma/\partial y$, $y\partial\gamma/\partial y$. By subtracting off appropriate multiples of X, Y, Z we obtain vectors without quadratic terms in the vector sum of the tangent space and the image of the differential. Explicitly, we consider (modulo terms of degree ≥ 4, and cubic terms with factor $ux+vy$)

$$x\frac{\partial\gamma}{\partial x} - uX + \kappa u^2 Z \equiv \frac{ux}{4}(x^2+y^2) + gx\frac{\partial C}{\partial x},$$
$$y\frac{\partial\gamma}{\partial x} - uY + \kappa uv Z \equiv \frac{uy}{4}(x^2+y^2) + gy\frac{\partial C}{\partial x},$$
$$x\frac{\partial\gamma}{\partial y} - vX + \kappa uv Z \equiv \frac{vx}{4}(x^2+y^2) + gx\frac{\partial C}{\partial y},$$
$$y\frac{\partial\gamma}{\partial y} - vY + \kappa v^2 Z \equiv \frac{vy}{4}(x^2+y^2) + gy\frac{\partial C}{\partial y}.$$

And transversality can only fail if $ux+vy$ is a factor of all four expressions, i.e.,

(ii) $$\frac{uv}{4}(u^2+v^2) + gv\frac{\partial C}{\partial x}(v, -u) = 0,$$

(iii) $$\frac{u^2}{4}(u^2+v^2) + gu\frac{\partial C}{\partial x}(v, -u) = 0,$$

$$\text{(iv)}\qquad \frac{v^2}{4}(u^2+v^2)+gv\frac{\partial C}{\partial y}(v,-u)=0,$$

$$\text{(v)}\qquad \frac{uv}{4}(u^2+v^2)+gu\frac{\partial C}{\partial y}(v,-u)=0.$$

When $u=0$, $v\neq 0$ or $u\neq 0$, $v=0$ a brief calculation verifies that the equations (i) to (v) have at most two solutions. When $u\neq 0$, $v\neq 0$ the equations (ii), (iv), (v) are redundant, and we need only consider (i), (iii). (i) represents three (possibly complex) lines through the origin in the (u,v)-plane. And, as $w=1/2\kappa$ implies $g=-(2/\kappa)(u^2+v^2+1/4\kappa^2)$, we see that (iii) represents a circular quartic curve in this plane with a singular point at the origin. Thus we obtain only finitely many points of intersection with $u\neq 0$, $v\neq 0$ (in fact ≤ 6) unless the quartic reduces to a line and a cubic, and the line lies in the set defined by (i). However, simple inspection of the equation of the quartic shows that this can only happen when the line is $u=0$, which case we have already dealt with. That completes the proof of Lemma 2.4. Q.E.D.

LEMMA 2.5. *Suppose* 0 *is an umbilic of* M. *There are no points on the sphere* $u^2+v^2+w^2=w/2\kappa$ *where the contact map fails to be transverse to the* A_4*-orbit.*

Proof. In principle this proceeds in the same style as that of Lemma 2.4. Again, we can work modulo terms of degree ≥ 4, and need only produce x^3, x^2y, xy^2, y^3. The tangent vectors to $R_{z=0}$ produce

$$z\equiv\frac{\kappa}{2}(x^2+y^2)+C(x,y),$$

$$X\equiv -y(vx-uy)+\frac{1}{2}w\kappa x(x^2+y^2),$$

$$Y\equiv x(vx-uy)+\frac{1}{2}w\kappa y(x^2+y^2),$$

$$Z\equiv w(x^2+y^2)+\frac{\kappa}{2}(ux+vy)(x^2+y^2).$$

When $u^2+v^2+w^2=w/2\kappa$ we have $a+\frac{1}{2}\kappa g=v^2$, $d=-uv$, $b+\frac{1}{2}\kappa g=u^2$, and the tangent space to the A_4-orbit contains

$$\frac{1}{2}x^2\frac{\partial\gamma}{\partial x}\equiv vx^2(vx-uy),\qquad \frac{1}{2}x^2\frac{\partial\gamma}{\partial y}\equiv -ux^2(vx-uy),$$

$$\frac{1}{2}xy\frac{\partial\gamma}{\partial x}\equiv vxy(vx-uy),\qquad \frac{1}{2}xy\frac{\partial\gamma}{\partial y}\equiv -uxy(vx-uy),$$

$$\frac{1}{2}y^2\frac{\partial\gamma}{\partial x}\equiv vy^2(vx-uy),\qquad \frac{1}{2}y^2\frac{\partial\gamma}{\partial y}\equiv -uy^2(vx-uy).$$

Either $u\neq 0$ or $v\neq 0$, or else we have contact of corank 2, so these vectors span the space of binary cubics with factor $vx-uy$, and we can work modulo such terms. From the vectors z, X, Y, Z we can produce a linear combination with no quadratic terms, namely

$$wz-\frac{\kappa}{2}Z\equiv wC(x,y)-\frac{\kappa^2}{4}(ux+vy)(x^2+y^2),$$

and transversality can only fail when $vx-uy$ is a factor of this, i.e.

$$\text{(i)}\qquad \frac{\kappa}{4}(u^2+v^2)^2-wC(u,v)=0.$$

We produce a second condition on u,v as follows. Using the Euler relation we see that the tangent space to the A_4-orbit contains the vector

$$x\frac{\partial\gamma}{\partial x}+y\frac{\partial\gamma}{\partial y}\equiv(vx-uy)^2+3\kappa w(ux+vy)(x^2+y^2)+3gC(x,y).$$

However, the image of the differential contains

$$uX+vY\equiv(vx-uy)^2+\frac{w\kappa}{2}(ux+vy)(x^2+y^2)$$

and subtraction produces the vector

$$\frac{5}{2}\kappa w(ux+vy)(x^2+y^2)+3gC(x,y).$$

Transversality can only fail when $vx-uy$ is a factor of this, i.e.,

$$\text{(ii)}\qquad \frac{5}{2}\kappa w(u^2+v^2)^2+3gC(u,v)=0.$$

Now (i), (ii) give two equations in $(u^2+v^2)^2, C(u,v)$, and the determinant of the coefficients is easily checked to be $\kappa w^2\neq 0$. It follows that $u=0$, $v=0$, establishing the result. Q.E.D.

We can sum up our discussion of what happens at an umbilic of M as follows:

LEMMA 2.6. *Suppose* 0 *is an umbilic of* M. *There are only finitely many positions of the source* $s=(u,v,w)$ *for which the contact mapping fails to be transverse to the* A_4 *orbit, all of which lie in the plane* $w=1/2\kappa$.

We turn our attention now to the question of what happens at a nonumbilical point of M. Our first step is provided by

LEMMA 2.7. *Suppose* 0 *is not an umbilic of* M. *There are no positions of the source* $s=(u,v,w)$ *off the principal planes* $u=0$, $v=0$ *for which the contact mapping fails to be transverse to the* A_4-*orbit.*

Proof. As in the two previous proofs, we work modulo terms of degree ≥ 4, and need only produce x^3, x^2y, xy^2, y^3. We suppose throughout that $u\neq 0$, $v\neq 0$. Following the philosophy so far adopted we need to construct a linear combination of z, X, Y, Z with no quadratic terms. Sheer calculation produces

$$\text{(1)}\qquad 4w(u^2+v^2)C(x,y)-(\kappa_1x^2+\kappa_2y^2)(Aux+Bvy)$$

where

$$A=\kappa_1u^2+\kappa_2v^2+(\kappa_2-\kappa_1)w^2,$$
$$B=\kappa_1u^2+\kappa_2v^2-(\kappa_2-\kappa_1)w^2.$$

The next step is to produce from $x\,\partial\gamma/\partial x$, $y\,\partial\gamma/\partial x$, $x\,\partial\gamma/\partial y$, $y\,\partial\gamma/\partial y$ a vector with zero quadratic part, by subtracting off an appropriate linear combination of X, Y, Z. Starting with $x\,\partial\gamma/\partial x+y\,\partial\gamma/\partial y$ this proceeds uniquely and produces the vector

$$\text{(2)}\qquad 3gC(x,y)+\tfrac{1}{2}(\kappa_1x^2+\kappa_2y^2)(A'x+B'y)$$

where

$$A'=6uw-pw-ru,\qquad B'=6vw-qw-rv.$$

The important practical consideration here is that by taking an appropriate linear combination of (1) and (2) we obtain a binary cubic with factor $\kappa_1x^2+\kappa_2y^2$, namely this position definite form times the linear form

$$(3) \qquad u(gA+4w^2c)x+v(gB+4w^2c)y.$$

We can now obtain a definite restriction on u,v,w. Looking at the tangent vectors to the A_4-orbit obtained by multiplying $\partial\gamma/\partial x$, $\partial\gamma/\partial y$ by x^2,xy,y^2 we see that we can work modulo cubic forms with factor

$$(4) \qquad u\{(a+\tfrac{1}{2}\kappa_1g)x+dy\}+v\{dx+(b+\tfrac{1}{2}\kappa_2g)y\}.$$

Transversality can only fail when the linear forms (3) and (4) are linearly dependent: a few lines of working reduces this to the condition

$$(5) \qquad \kappa_1u^2+\kappa_2v^2+(\kappa_1+\kappa_2)w^2=w.$$

However, the very fact that the contact is of corank 1 imposes a further condition on (u,v,w), namely that it satisfies the equation (**) of the first discriminant surface: in view of (5) this can be written as

$$(6) \qquad u^2+v^2+w^2=1/4\kappa_1\kappa_2.$$

Eliminating u,v from (5), (6) we obtain

$$(\kappa_2-\kappa_1)v^2+\kappa_2\left(w-\frac{1}{2\kappa_2}\right)^2=0,$$

$$(\kappa_1-\kappa_2)u^2+\kappa_1\left(w-\frac{1}{2\kappa_1}\right)^2=0,$$

and since κ_1,κ_2 are distinct and positive, we see that these equations can only be satisfied when one of u,v is zero. That concludes the proof. Q.E.D.

It remains to discover what happens in the principal planes at a nonumbilical point of M.

LEMMA 2.8. *Suppose 0 is not an umbilic of M. Then, for a position of the source in one of the principal planes, transversality to the A_4-orbit can only fail when 0 is an $A_{\geq 3}$-point of M. Moreover, in this case transversality will fail at not more than one point in each principal plane.*

Proof. By symmetry we can suppose $u=0$, $v\neq 0$. Following the initial steps in the proof of the preceding proposition we can produce a linear combination of X, Y, Z with no quadratic terms, namely

$$(1) \qquad 4vwC(x,y)-(\kappa_2v^2-\kappa_2w^2+\kappa_1w^2)y(\kappa_1x^2+\kappa_2y^2).$$

For contact of corank 1 exactly one of $a+\frac{1}{2}\kappa_1g, b+\frac{1}{2}\kappa_2g$ must be $\neq 0$. The former case is easily disposed of. Looking at the tangent vectors to the A_4-orbit obtained by multiplying $\partial\gamma/\partial x$, $\partial\gamma/\partial y$ by x^2, xy,y^2 (modulo terms of degree ≥ 4), we see that they span the space of binary cubics with factor x, and we can work modulo such terms. But then $vx\,\partial\gamma/\partial x-2(a+\frac{1}{2}\kappa_1g)Y$ has zero quadratic part, and the coefficient of y^3 is $\neq 0$, so we achieve transversality. It remains to discuss the case $b+\frac{1}{2}\kappa_2g\neq 0$. The first observation to make here is that then necessarily $a+\frac{1}{2}\kappa_1g=(v^2+w^2)(1-2\kappa_1w)=0$ so $w=1/2\kappa_1$, a line in the (v,w)-plane. Secondly, the tangent vectors to the A_4-orbit obtained by multiplying $\partial\gamma/\partial x$, $\partial\gamma/\partial y$ by x^2,xy,y^2 (modulo terms of degree ≥ 4) span the space of binary cubics with factor y, and we can work modulo such terms. In particular, transversality can only fail when y is a factor of (1), i.e. y is a factor of $C(x,y)$, which is precisely the condition for 0 to be an $A_{\geq 3}$-point of M. To obtain a definite restriction on v we consider the tangent vectors $x\partial\gamma/\partial x$, $y\partial\gamma/\partial x$, $x\partial\gamma/\partial y$, $y\partial\gamma/\partial y$ to the A_4-orbit. The first two are zero, modulo terms of degree ≥ 4, and cubic terms with factor y. For each of the latter two we can remove the quadratic terms by subtracting off appropriate multiples of X, Y, Z. The only positive information is provided by the vector $vx\partial\gamma/\partial y+2(b+\frac{1}{2}\kappa_2 g)$. The condition for y to be a factor of this is easily written down. If we agree to write $C(x,y)=y(c_1x^2+c_2xy+c_3y^2)$, and bear in mind that $w=1/2\kappa_1$, the condition reduces to

$$\left(4v^2+\frac{1}{\kappa_1^2}\right)\left\{vc_1-\frac{(\kappa_1-\kappa_2)}{4}\right\}=0$$

yielding *at most one* value of v for which transversality fails. Q.E.D.

We can now put the bits together to establish the main result we need concerning the A_4-stratum. Unfortunately at this point we need to assume M is analytic. We do not know whether this assumption can be avoided.

LEMMA 2.9. *Assume that M is analytic. Then there exists a closed set $\Omega\subseteq \mathrm{Int}\,M$, of Lebesgue zero, such that for every position of the source s off Ω the contact map is transverse to the A_4-orbit.*

Proof. For each point $m\in M$ there is, by the preceding propositions, a finite set $\Omega_m\subseteq\mathbb{R}^3$ of "bad" positions of the source, where the contact map fails to be transverse to the A_4-orbit. Write A for the set of points (m,s) in $M\times\mathbb{R}^3$ with $s\in\Omega_m$. Let us assume, temporarily, that A is subanalytic. For the properties of subanalytic sets used below, see [8], [9] or [4, §4]. Observe first that then A necessarily has dimension ≤ 2, since the fibres of the projection $M\times\mathbb{R}^3\to M$, restricted to A, are finite. Moreover, the image of A under the proper projection $M\times\mathbb{R}^3\to\mathbb{R}$ will likewise be subanalytic, of dimension ≤ 2, and hence its closure Ω will have the same properties. It follows that Ω has Lebesgue measure zero in $\mathbb{R}^3$, and has of course the properties required by the proposition. Thus it remains to check that A is subanalytic. Evidently, we need only consider points on M which are umbilics, or $A_{\geq 3}$-points, each condition defining a subanalytic subset of M. Then, for each type of point $m\in M$ we have simply to observe that the proofs of the preceding propositions yield the "bad" set Ω_m explicitly as the zero set of a finite system of analytic equations.

The A_n-strata with $n\geq 5$. Our final objective in this section is to prove that for almost all positions of the source s we can avoid nongeneric contact of corank 1.

LEMMA 2.10. *Assume that M is analytic. Then for almost all positions of the source s inside M there is no paraboloid of revolution with focus s having contact of type $A_{\geq 5}$ with M.*

Our strategy for proving this result is based on a refinement of the Thom basic transversality lemma [7] due to Mather [11].

LEMMA 2.11. *Let $F\colon A\times B\to N$ be a smooth mapping, and $P\subseteq N$ a smooth manifold. Suppose that for all $c=(a,b)$ in $A\times B$ either $F_a\colon B\to N$ is transverse to P, or the dimension of $\mathrm{Im}\,T_cF+T_{F(c)}P$ is strictly greater than that of $\mathrm{Im}\,T_bF_a+T_{F(c)}P$: then for almost all $a\in A$ in the sense of Lebesgue measure, $F_a\colon B\to N$ is transverse to P.*

We now set up the situation to which this result will be applied. Consider the "global" contact mapping

$$\Gamma\colon M\times\mathbb{R}^3\times\mathbb{R}\times\mathbb{R}^3\rightarrowtail\mathbb{R}$$

given by $\Gamma(m,t,c,s)=Q_\pi(m)$, where $\pi=(t,k,s)$ and Q_π is the left-hand side of (*) in Lemma 2.1. The flighted arrow indicates that the domain of Γ is the open subset of $M\times\mathbb{R}^3\times\mathbb{R}\times\mathbb{R}^3$ defined by $t\cdot s\neq k$ and $s\in \operatorname{Int}M$. We consider Γ as a family (Γ_s) of smooth mappings parametrized by the positions of the source s. Of course locally this is precisely the contact mapping we have studied in this section. Each Γ_s induces a smooth jet extension

$$(1) \qquad j_1^4\Gamma_s : M\times\mathbb{R}^3\times\mathbb{R} \rightarrowtail J^4(M,\mathbb{R})$$

and it is to this that we apply the Thom–Mather lemma. For this application we shall need to use the jet-bundle which includes constants in the fibres. The reason for this is that in essence k adjusts the constant term of Q_π and we want to use up the influence of k on this constant term so that it does not affect others.

Let us note at once that since multiplying k and t by $\lambda\in\mathbb{R}$, $\lambda\neq 0$, merely multiplies Q_π by λ^2, this cannot affect the contact singularity. It follows that the tangent vector corresponding to

$$p\frac{\partial}{\partial p}+q\frac{\partial}{\partial q}+r\frac{\partial}{\partial r}+k\frac{\partial}{\partial k}$$

always lies in the tangent space to (for example) an A stratum. This effectively reduces the tangent vectors provided by $\mathbb{R}^3\times\mathbb{R}$ from four to three.

Next, we claim that, if Q_π has contact of type $A_{\geq 5}$ with M at m, then the vector sum of the image of the differential at (m,t,k) of (1), and the tangent space to the $A_{\geq 5}$-stratum has codimension ≥ 2. Indeed if we write $\mathcal{E}$ for the algebra of germs at m of smooth functions on M, $\mathfrak{M}$ for its maximal ideal, and J_{Γ_s} for the Jacobian ideal generated by the partials of Γ_s with respect to some local coordinates on M, then $\mathfrak{M}^5+J_{\Gamma_s}$ has codimension 5 in $\mathcal{E}$. On the other hand, as we have noted, at most three extra independent vectors come from $\mathbb{R}^3\times\mathbb{R}$, so the claim is proved.

Consider now the corresponding mapping

$$(2) \qquad j_1^4\Gamma : M\times\mathbb{R}^3\times\mathbb{R}\times\mathbb{R}^3 \rightarrowtail J^4(M,\mathbb{R}).$$

We claim that here the corresponding vector sum has codimension ≤ 1. This needs only to be checked locally, using the contact map studied earlier in this section. First we note the crucial fact that $1, x$ and y are all in the image of the differential of the contact map, as they were when, using J_0, we ignored constant terms. To see this observe that differentiation with respect to k,p,q,r produces nonzero multiples of

$$k-px-qy-rz, \quad -u+x, \quad -v+y, \quad -w+z$$

respectively, where $z=z(x,y)$ as in the earlier parts of this section. A linear combination of these produces the *nonzero* constant

$$k-pu-qv-rw$$

without any higher terms. It is now clear that we obtain x and y too.

Next, the tangent vectors to the $A_{\geq 5}$ stratum obtained by multiplying $\partial\gamma/\partial x$ and $\partial\gamma/\partial y$ by monomials of degree 3 (and ignoring terms of degree ≥ 5) span a subspace of the space of binary quartics of codimension 1. Augmenting these vectors by a single vector, we obtain all terms of degree 4 in the jet space, and can work modulo terms of degree ≥ 4.

Let $\Omega\subset\operatorname{Int}M$ be the closed set of measure zero for which one or more of the transversality arguments of Lemmas 2.4, 2.5, 2.7, 2.8 fail. We shall assume now that

$s \notin \Omega$; clearly we can still apply the Thom–Mather lemma to Int$M-\Omega$ and still deduce that, for all s off a set of measure zero in Int$M$, the map $j_1^4\Gamma_s$ is transverse to $A_{\geq 5}$ (and hence does not intersect it). Then the arguments of the above four results yield terms of degree 3 modulo those of degree ≥ 4, and then terms of degree 2 modulo those of degree ≥ 3. We use here the fact that $1, x$ and y are automatically present, so that we need only, as in the proofs of the four results, consider quadratic and higher terms.

The end result is that Lemma 2.10 follows by an application of the Thom–Mather lemma, the appropriate submanifold of J^4 being the $A_{\geq 5}$-stratum. Q.E.D.

Putting together Lemmas 2.3, 2.6, 2.9 and 2.10 we obtain:

THEOREM 2.12. *Suppose that M is analytic. Then, for all $s \in$ Int$M$ off a set of measure zero, the caustic at infinity is locally generic.*

3. Light source at infinity. Here we consider only the question of mirror genericity. The general method used in [3] when the light source was finite was

(a) Given the mirror and source construct a smooth wavefront W.

(b) Deform W slightly to obtain a generic wavefront W'.

(c) Reconstruct a corresponding mirror M', a slight deformation of M.

When the light source is at infinity, i.e. there is given a direction s for parallel incident light, we have a choice of incident wavefronts. Namely, we can choose for "incident wavefront" any plane L given by $x \cdot a = \beta$, where a is a unit vector in the incident direction and β is some real number. (Note that the same L will serve for light in the directions a and $-a$.) We reconstruct the wavefront W in much the same way as we did for the case of a finite source (see Fig.2). Corresponding to each point m of the mirror there is a (unique) point x on L with the normal to L at x passing through m. We then reflect x in the tangent plane to M at m to obtain the corresponding point q of W. Of course W depends, as did L, on β. A short computation shows that W is the locus

$$(1) \qquad q = m + (\beta - m \cdot a)a - 2((\beta - m \cdot a)(a \cdot n))n$$

where n is the unit normal to M at m (note that n and $-n$ give same q). We suppose that a and β are selected so that $\beta - p \cdot a > 0$ for all m on M, so that M lies wholly on one side of L.

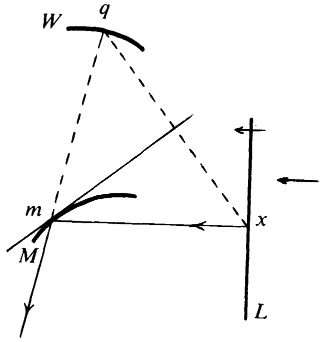


FIG. 2

LEMMA 3.1. (a) *If the tangent plane to M at m does not contain the incident direction s (i.e. if $a \cdot n \neq 0$) then W is immersed at q provided β is sufficiently large (depending on m).*

(b) *If the tangent plane to M at p does contain s then W is immersed at q.*

Proof. (a) Differentiating (1) with respect to coordinates u_1, u_2 on M we obtain

$$\frac{\partial q}{\partial u_i}=\frac{\partial m}{\partial u_i}-\left(\frac{\partial m}{\partial u_i}\cdot a\right)a-2\left\{\left(\frac{-\partial m}{\partial u_i}\cdot a\right)(a\cdot n)+(\beta-m\cdot a)\left(a\cdot\frac{\partial n}{\partial u_i}\right)\right\}n$$
$$-2\{(\beta-m\cdot a)(a\cdot n)\}\frac{\partial n}{\partial u_i}.$$

The condition for the $\partial q/\partial u_i$ to be linearly dependent is therefore a quadratic equation in β where the coefficient of β^2 is zero if and only if

$$\left(a\cdot\frac{\partial n}{\partial u_i}\right)n+(a\cdot n)\frac{\partial n}{\partial u_i}=0, \qquad i=1,2,$$

are linearly dependent. However, M has nonzero Gaussian curvature at m, so that n, $\partial n/\partial u_1$ and $\partial n/\partial u_2$ are linearly independent there, and by hypothesis $a\cdot n$ is nonzero. Hence the coefficient of β^2 cannot be zero, and for a given m the vectors $\partial q/\partial u_i$ will be linearly independent for all sufficiently large β.

(b) When $a\cdot n=0$ the expression for $\partial q/\partial u_i$ reduces to

$$\frac{\partial q}{\partial u_i}=\frac{\partial m}{\partial u_i}-\left(\frac{\partial m}{\partial u_i}\cdot a\right)a-2\left\{(\beta-m\cdot a)\left(a\cdot\frac{\partial n}{\partial u_i}\right)\right\}n.$$

Let b be a unit vector perpendicular to a and n. Then

$$\frac{\partial q}{\partial u_i}=\left(\frac{\partial m}{\partial u_i}\cdot b\right)b-2\left\{(\beta-m\cdot a)\left(a\cdot\frac{\partial n}{\partial u_i}\right)\right\}n.$$

Consequently W fails to be immersed precisely when

$$\left(\frac{\partial m}{\partial u_1}\cdot b\right)\left(\frac{\partial n}{\partial u_2}\cdot a\right)=\left(\frac{\partial m}{\partial u_2}\cdot b\right)\left(\frac{\partial n}{\partial u_1}\cdot a\right)$$

(recall that $\beta-m\cdot a$ is nonzero). Choose coordinates so that $b=\partial m/\partial u_1$, $a=\partial m/\partial u_2$ (both evaluated at m), so the above reduces to $(\partial n/\partial u_2)\cdot(\partial m/\partial u_2)=0$. Finally the Gaussian curvature of M at m is

$$\left(\frac{\partial n}{\partial u_1}\cdot\frac{\partial m}{\partial u_1}\right)\left(\frac{\partial n}{\partial u_2}\cdot\frac{\partial m}{\partial u_2}\right)-\left(\frac{\partial n}{\partial u_1}\cdot\frac{\partial m}{\partial u_2}\right)^2$$

which is therefore ≤ 0. This contradiction shows that W is always immersed at q. Q.E.D.

The problem with using this lemma to construct a wavefront W is that, given M and the incident direction s, we cannot choose a fixed β so large that the resulting W is immersed everywhere. No single β will work for all points of M away from the *profile curve* on M in the direction s (i.e. the set of points of M at which the tangent plane contains the direction s), since the sufficiently large value of β in Lemma 3.1(a) tends to infinity as m approaches the curve.

As a simple illustration take a circle in the plane and a line not meeting the circle (Fig. 3) Carrying out the wavefront construction in the plane, with the circle as mirror and the line as incident wavefront, we obtain the curve illustrated (Fig. 3), which inevitably has a node. Rotating the figure we obtain a mirror M in $\mathbb{R}^3$ and incident wavefront L for which the constructed wavefront W will have a cone-like point of nonimmersion, no matter how distant L is from M.

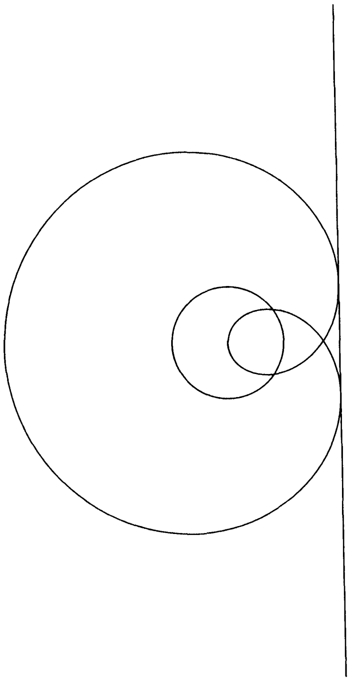


FIG. 3

We now show that there is a genuine obstruction to carrying out the proof of mirror genericity (see Fig. 4).

LEMMA 3.2. *Let m be a point on the profile of the mirror M for the incident direction s. If q is the corresponding point of W then there is one centre of curvature where the corresponding sphere has A_2 contact and one principal curvature zero, W having $A_{\geq 3}$ contact with its tangent plane.*

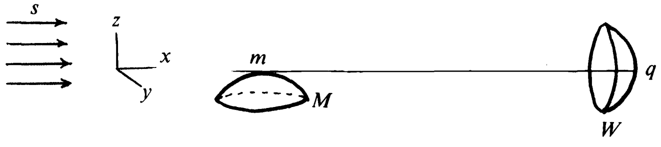


FIG. 4

Proof. Write M at $0=m$ locally as

$$z=f(x,y)=\lambda x^2+2\mu xy+\nu y^2+O(3).$$

Note that λ and ν are both nonzero, for otherwise the Gaussian curvature of M at 0 is ≤ 0. We can then take $a=(1,0,0)$. Using the parametrization of W given by (1) above we find

$$q_1=\beta-2(\beta-x)(f_x^2)(1+f_x^2+f_y^2)^{-1},$$

$$q_2 = y - 2(\beta - x) f_x f_y \left(1 + f_x^2 + f_y^2\right)^{-1},$$
$$q_3 = f + 2(\beta - x) f_x \left(1 + f_x^2 + f_y^2\right)^{-1}.$$

Hence

$$q_1 = \beta - 8\beta\lambda^2 x^2 - 16\beta\lambda\mu xy - 8\beta\mu^2 y^2 + O(3),$$
$$q_2 = y + O(2),$$
$$q_3 = 4\beta\lambda x + 4\beta\mu y + O(2).$$

Consequently $0 \in M$ corresponds to $(\beta, 0, 0)$ on W, and near this point W has equation $x = \beta - z^2/2\beta +$ higher terms in x, y, z. Thus the principal curvatures of W are $-1/\beta$ and 0 and one centre of curvature is at the origin, i.e. at the point $p \in M$.

The distance squared function from 0 to W close to $(\beta, 0, 0)$ is

$$q_1^2 + q_2^2 + q_3^2 = \beta^2 + y^2 - 8\beta\lambda^2 x^3 + \phi(x, y)$$

where ϕ involves only terms above the diagonal joining x^3 and y^2 in the Newton polygon of $y^2 - 8\beta\lambda^2 x^3$. Since $\lambda \neq 0$ the distance squared function does indeed have an A_2 at the point $(\beta, 0, 0)$ of W. The height function on W at this point in the incident direction is q_1, which has quadratic term $-8\beta(\lambda x + \mu y)^2$. Substituting $x = u - \mu\lambda^{-1}y$ we get quadratic term $-8\beta\lambda^2 u^2$ but no term in y^3. So, at $(\beta, 0, 0)$, W does have $A_{\geq 3}$ contact with its tangent plane. Q.E.D.

Thus one part of the caustic corresponding to points of the profile is well behaved. This is analogous to the plane case, where, with parallel incident light, a point m of M where the tangent is in the incident direction always gives a *nonsingular* point of the caustic, situated at the same point m.

However, generically the profile will be a smooth curve (in fact always since our mirrors have positive Gaussian curvature) and, since each point gives an $A_{\geq 3}$ for the height function on W, this part of the caustic at infinity is *never* generic. (Generically $A_{\geq 3}$ will occur only for isolated points.) Thus independently of M and the incident light direction the caustic is nongeneric at infinity, and the proof of mirror genericity sketched out at the beginning of §3 must fail. Putting it another way, we cannot expect to start with M, construct W, deform W to W' and reconstruct M' from W', since only wavefronts W with the *nongeneric* behaviour at infinity discussed above can arise from actual mirrors. There is a built-in lack of genericity for wavefronts arising from parallel light reflected from a mirror in $\mathbb{R}^3$.

Despite this we are able to prove a weaker result.

THEOREM 3.3. *Let K be a compact region in $\mathbb{R}^3$ and s an incident light direction. Then, for an open dense set of embeddings of the 2-sphere S^2 as a convex mirror in $\mathbb{R}^3$, that part of the caustic within K is generic.*

Proof. Openness is clear; we have only to prove density. Given M and s construct a wavefront W using a value of β large enough to ensure that M lies wholly on one side of the incident wavefront L. Now choose an open neighbourhood U of the profile of M relative to s, as follows. For each point m of the profile, one point of the caustic is at m and is an A_2 (by Lemma 3.2)—hence automatically transverse. So in some neighbourhood U of the profile each point m' will give one centre of curvature of W near to the profile, and it will be a transverse A_2. By shrinking U if necessary we can ensure that all of the centres of curvature of W inside K arise either from the transverse A_2's corresponding to points in U or from points of $M - U$.

Now increase β if necessary so that the construction for the wavefront gives a smooth point q for each p in $M-U$. (Recall that the bad values of β for each point p are the roots of a quadratic equation whose coefficients depend smoothly on m and whose β^2 coefficient is nonzero away from the profile.) This will not disturb the property of the previous paragraph; consider the resulting wavefront W. We now use Looijenga's results ([10], [12, p. 712]) to deform W slightly to get a generic wavefront W'—generic, that is, for distance squared functions from points in K. We can do this in such a way that W' coincides with W at points of W corresponding to $U \subset M$. (Strictly speaking, in order to make W' smooth we shall have to work with a pair of neighbourhoods $U_1 \subset U$ of the profile.)

We now wish to recover a new mirror M' from W'. Clearly we want M' to coincide with M on U; how do we construct the other part? Let q be a point on the wavefront W. The corresponding point m on M is given by $F(m,v_1,v_2)=0$ where

$$F(m,v_1,v_2)=m-q-(\beta-m\cdot a)N$$

where N is the unit normal *to the wavefront* at q (in Fig. 2 the distance from m to x is $\beta-m\cdot a$, and is equal to the distance from m to q). Here v_1,v_2 are local coordinates on W, so that q and N are functions of v_1,v_2. Now $F=0$ is a set of linear equations for m and there is a unique solution if and only if the three vectors e_i+a_iN are independent. (Here $e_i=(1,0,0)$, etc., and $a=(a_1,a_2,a_3)$.) If these vectors were dependent, then for some nonzero $\lambda=(\lambda_1,\lambda_2,\lambda_3)$ we would have $\Sigma\lambda_i(e_i+a_iN)=0$, which gives $\lambda=-(\lambda\cdot a)N$, so $\lambda\cdot a=-(\lambda\cdot a)(N\cdot a)$. But $\lambda\cdot a\neq 0$ so $N\cdot a=-1$ and since N and a are unit vectors $N=-a$, so that $m-q=(\beta-p\cdot a)a$.

When W *is* constructed from a mirror M then (writing as before n for the unit normal to M), comparing the last equation with (1) we have $(\beta-m\cdot a)(a\cdot n)=0$, so that $a\cdot n=0$, i.e., m is on the profile of M for the incident direction. Consequently we can smoothly reconstruct an M' from W' provided we have not perturbed W too much to obtain W'; away from $U\subset M$ we will have $a\cdot n\neq 0$.

Finally, will the parametrization $m(v_1,v_2)$ of the reconstructed mirror be an immersion? Differentiating $F=0$ with respect to v_i we obtain

$$\frac{\partial m}{\partial v_i}-\frac{\partial q}{\partial v_i}+\left(\frac{\partial m}{\partial v_i}\cdot a\right)N-(\beta-m\cdot a)\frac{\partial N}{\partial v_i}=0.$$

Assuming $\lambda\,\partial m/\partial v_1=\partial m/\partial v_2$ for some $\lambda\in\mathbb{R}$ we have

$$-\lambda\frac{dq}{\partial v_1}+\frac{\partial q}{\partial v_2}-(\beta-m\cdot a)\left(\lambda\frac{\partial N}{\partial v_1}-\frac{\partial N}{\partial v_2}\right)=0.$$

It is now a straightforward matter to show that m is actually a centre of curvature of W at q. (For example one can choose coordinates so that W is $\{(v_1,v_2,f(v_1,v_2))\}$ locally, with $f(v_1,v_2)=\frac{1}{2}\kappa_1v_1^2+\frac{1}{2}\kappa_2v_2^2+O(3)$. Then

$$\frac{\partial N}{\partial v_1}=(-\kappa_1,0,0),\qquad \frac{\partial N}{\partial v_2}=(0,-\kappa_2,0)$$

at $v_1=v_2=0$ and $m(0)=(0,0,m_3)$ where $m_3=\beta-m(0)\cdot a\neq 0$. The above equation then gives $\kappa_2=1/m_3$.)

Hence we are only in trouble if m is a centre of curvature of W at q, i.e. light rays reflected from M at m focus at m. But this is absurd, especially when the incident rays are parallel. (One easily obtains a formal contradiction by considering the distance squared function from m to W at q.)

Thus perturbing W slightly to W' we can recover M' as a smooth mirror, since the condition for m to be an immersion is necessarily open and is satisfied, as we have just seen, by M. This completes the proof. Q.E.D.

REFERENCES

[1] J. W. Bruce, *On contact of hypersurfaces*, Bull. London Math. Soc., 13 (1981), pp. 51–54.

[2] ______, *The duals of generic hypersurfaces*, Math. Scand., 49 (1981), pp. 36–80.

[3] J. W. Bruce, P. J. Giblin and C. G. Gibson, *On caustics by reflexion*, Topology, 21 (1982), pp. 179–199.

[4] ______, *Source genericity of caustics by reflexion in the plane*, Quarterly J. Math. (Oxford), 33 (1982), pp. 169–190.

[5] ______, *Source genericity of caustics by reflexion in* $\mathbf{R}^3$, Philos. Trans. Roy. Soc. London, A381 (1982), pp. 83–116.

[6] ______, *Caustics by reflexion in the extended plane*, Univ. of Liverpool, 1981, to appear in Geom. Dedicata.

[7] C. G. Gibson, *Singular points of smooth mappings*, Research Notes in Math., 25, Pitman, New York, 1979.

[8] H. Hironaka, *Subanalytic sets*, in Number Theory, Algebraic Geometry and Commutative Algebra (in honor of Y. Akizuki), Kinokuniya, Tokyo, 1973, pp. 453–493.

[9] ______, *Triangulation of algebraic sets*, in Algebraic Geometry, Arata 1974, (Proc. Symp. Pure Math A.M.S., 29 (1975)), pp. 165–185.

[10] E. J. N. Looijenga, *Structural stability of smooth families of* C^∞ *functions*, Thesis, Univ. of Amsterdam, 1974.

[11] J. N. Mather, *Generic projections*, Ann. Math., 98 (1973), pp. 226–245.

[12] C. T. C. Wall, *Geometric properties of generic differentiable manifolds*, Lecture Notes in Mathematics, 597, Springer, Berlin, 1977, pp. 707–774.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics

0036-1410/84/1501-0003 $01.25/0

REPEATED RESONANCE AND HOMOCLINIC BIFURCATION IN A PERIODICALLY FORCED FAMILY OF OSCILLATORS*

BERNIE GREENSPAN† AND PHILIP HOLMES‡

Abstract. We use global perturbation techniques originally due to Melnikov [1963] to study the bifurcation behavior exhibited by a family of nonlinear oscillators subject to periodic forcing. We concentrate on the case in which the unforced systems possess a one-parameter family of periodic orbits limiting on a homoclinic orbit.

1. Introduction. Many physical problems are modelled as single degree of freedom nonlinear oscillators subject to external periodic forcing. The books of Andronov, Vitt and Khaikin [1966] or Hayashi [1964], [1975] provide examples from mechanics and electrical circuit theory. Such oscillators without external time-dependent perturbations may be studied by phase plane techniques, and their typical behaviors are therefore fairly well understood (cf. Andronov et al. [1971], [1973]). However, the presence of external forcing greatly complicates the situation, and classical analyses (using, for example, averaging or perturbation methods) have generally been limited to the case of weak nonlinearity (cf. Nayfeh and Mook [1978]).

In earlier work (Holmes [1979], [1980], Moon and Holmes [1979], Greenspan and Holmes [1981]) we were able to overcome this limitation by studying small perturbations of strongly nonlinear, *integrable* systems. In the present paper we make use of these techniques to study a problem in which, as a parameter is varied, repeated resonances of successively higher and higher orders occur, culminating in "subharmonics of infinite order" and homoclinic orbits. Specifically, our main example is the nonlinear oscillator

$$\ddot{y}-y+y^3=\varepsilon y^2\dot{y}-\delta\dot{y}+\gamma\cos t, \tag{1.1}$$

where ε, δ and γ are (small) parameters. The corresponding unperturbed (Hamiltonian) system is

$$\ddot{y}-y+y^3=0, \tag{1.2}$$

with Hamiltonian

$$H(y,\dot{y})=\frac{\dot{y}^2}{2}-\frac{y^2}{2}+\frac{y^4}{4}, \tag{1.3}$$

which is completely integrable and whose solutions may be expressed in terms of elliptic functions and, in the homoclinic limit on $H(y,\dot{y})=0$, hyperbolic functions.

Equation (1.1) without periodic forcing ($\gamma=0$) was studied by Holmes and Rand [1980] and shown to exhibit planar homoclinic bifurcations as the parameters ε and δ are varied. Results of Takens [1974] involving a singular "blowing up" change of coordinates were used to do this. In the present paper these results are recovered more directly by Melnikov's method (Melnikov [1963], Greenspan and Holmes [1981]) and we are also able to treat the periodically forced case.

*Received by the editors May 7, 1982, and in revised form January 4, 1983.

† Department of Mathematics, Vanderbilt University, Nashville, Tennessee 37235.

‡ Department of Theoretical and Applied Mechanics and Center for Applied Mathematics, Cornell University, Ithaca, New York 14853.

Equation (1.1), without periodic forcing, occurs as a model for panel flutter in a steady supersonic air flow (Holmes [1977], Holmes and Marsden [1978], Holmes [1981]). Periodic perturbation of the pressure differential across the panel would give rise to an additional time-dependent term such as $\gamma \cos t$ (cf. Dowell [1966]). Our example is, therefore, not without physical interest.

The paper is arranged as follows: In §2 we review Melnikov's method and state the main results. In §3 we present a preliminary example to illustrate the main ideas and to point out a characteristic difficulty which often arises in such analyses. Moreover, this example displays all the typical behavior found in each resonance band of our main example. We prove a theorem (Theorem 3.1) giving a fairly complete description of an autonomous averaged system close to the full system, and from this obtain partial results on the Poincaré map of the latter. In §4, we turn to our main example. We show that, for fixed γ and $\varepsilon \ll 1$, as δ is increased a countable sequence of bifurcations occurs in which subharmonic motions of successively higher periods $2\pi m$ are created and destroyed until ultimately, for a critical value $\delta=\delta(\infty)$, for any $\gamma>0$ and ε sufficiently small we have countably many subharmonic orbits coexisting in a thickened "figure of eight" neighborhood of the level curve $H(y,\dot{y})=0$ of the unperturbed Hamiltonian system (1.2)–(1.3). The closure of the unstable manifolds of these orbits forms a complicated attracting set, which we briefly describe.

Related work on global bifurcations of two-dimensional diffeomorphisms with attracting invariant closed curves has been done by Takens [1974], Arnold [1977], Aronson et al. [1980], [1982], but in the former cases these authors concentrated on the resonances encountered in the neighborhood of a Hopf bifurcation. Here we are more concerned with passage through resonance and the analogue of the planar homoclinic bifurcation in which a periodic orbit vanishes as its period becomes infinite.

The papers of Aronson et al. are more directly relevant here and we shall see that the generic (time periodic) perturbations of our results on the averaged equation give rise to a Poincaré map displaying essentially the same features found by these authors in their numerical work.

2. Global perturbations on integrable systems: Melnikov's method. In this section we briefly review the analytical techniques to be used below. For more details, and proofs of the theorems, see Greenspan and Holmes [1983] or Guckenheimer and Holmes [1983]. We note that Chow, Hale and Mallet-Paret [1980] have obtained similar results by different methods.

We consider systems of the form

$$\dot{x}=f(x)+\varepsilon g(x,t), \qquad x=\binom{u}{v}\in\mathbb{R}^2, \tag{2.1}$$

where

$$f=\begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix}, \qquad g=\begin{pmatrix} g_1(x,t) \\ g_2(x,t) \end{pmatrix}$$

are sufficiently smooth $(C^r, r\geq 2)$ and bounded on bounded sets and g is T-periodic in t. For simplicity we assume that the unperturbed system is Hamiltonian: $f_1=\frac{\partial H}{\partial v}$, $f_2=-\frac{\partial H}{\partial u}$ The non-Hamiltonian case is considered by Melnikov [1963] and Holmes [1980a]. Specific assumptions on the unperturbed flow are (cf. Fig. 1):

A1. For $\varepsilon=0$, (2.1) possesses a homoclinic orbit $q^0(t)$ to a hyperbolic saddle point p_0.

A2. Let $\Gamma^0=\{q^0(t)|t\in\mathbb{R}\}\cup\{p_0\}$. The interior of Γ^0 is filled with a continuous family of periodic orbits $q^\alpha(t)$, $\alpha\in(-1,0)$. Letting $d(x,\Gamma^0)=\inf_{q\in\Gamma^0}|x-q|$ we have $\lim_{\alpha\to 0}\sup_{t\in\mathbb{R}} d(q^\alpha(t),\Gamma^0)=0$.

A3. Let $h\alpha=H(q^\alpha(t))$ and T_α be the period of $q^\alpha(t)$. Then T_α is a differentiable function of h_α and $dT_\alpha/dh_\alpha>0$ inside Γ^0.

We note that A2 and A3 imply that $T_\alpha\to\infty$ monotonically as $\alpha\to 0$. Many of the results to follow can be proved under less restrictive assumptions.

Since g is T-periodic, the extended (x,t) phase space of (2.1) is the product $\mathbb{R}^2\times S^1$, where S^1 is the circle of length T. Associated with (2.1) we have a Poincaré map $P_\varepsilon^{t_0}$ defined on a (global) cross section $\Sigma^{t_0}=\{(x,t)|t=t_0\}$. $P_\varepsilon^{t_0}$ is obtained by following solutions of (2.1) based on Σ^{t_0} to their next intersection with Σ^{t_0}, cf. Chillingworth [1976]. Thus the unperturbed Poincaré map $P_0^{t_0}$ is simply the time T map of the unperturbed flow of $\dot{x}=f(x)$. Fixed points and periodic cycles of period m of $P_\varepsilon^{t_0}$ correspond to T-periodic motions and mT-periodic subharmonics of (2.1) respectively, and stability types correspond. In what follows we are effectively using regular perturbation theory to approximate $P_\varepsilon^{t_0}$ based on our knowledge of $P_0^{t_0}$ from the integrable unperturbed problem. (The general theory tells us that any two Poincaré maps $P_\varepsilon^{t_1}$, $P_\varepsilon^{t_2}$ are diffeomorphic, and consequently we will sometimes drop the superscript t_0.)

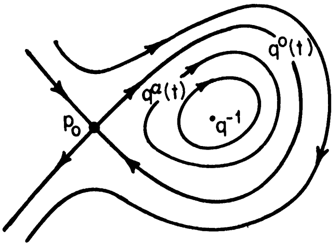


FIG. 1. *The unperturbed system.*

We first consider bifurcations from the homoclinic orbit $q^0(t)$ as ε increases. In this connection it is important to establish perturbation results for the fixed point p_0 of the Poincaré map and its invariant manifolds.

LEMMA 2.1. *Under the above assumptions, for ε sufficiently small* (2.1) *has a unique hyperbolic periodic orbit $\gamma_\varepsilon^0(t)=p_0+O(\varepsilon)$. Correspondingly, the Poincaré map $P_\varepsilon^{t_0}$ has a unique hyperbolic saddle point $p_\varepsilon^{t_0}=p_0+O(\varepsilon)$. Moreover, the local stable and unstable manifolds $W_{\mathrm{loc}}^s(\gamma_\varepsilon)$, $W_{\mathrm{loc}}^u(\gamma_\varepsilon)$ of the perturbed periodic orbit are C^r close to those of the unperturbed periodic orbit $p_0\times S^1$, and orbits $q_\varepsilon^s(t,t_0)$, $q_\varepsilon^u(t,t_0)$ lying in the global manifolds $W^s(\gamma_\varepsilon)$, $W^u(\gamma_\varepsilon)$ and based on Σ^{t_0} can be expressed as follows, with uniform validity in the indicated time intervals*:

$$(2.2)\qquad \begin{aligned} q_\varepsilon^s(t,t_0)&=q^0(t-t_0)+\varepsilon q_1^s(t,t_0)+O(\varepsilon^2), && t\in[t_0,\infty),\\ q_\varepsilon^u(t,t_0)&=q^0(t-t_0)+\varepsilon q_1^u(t,t_0)+O(\varepsilon^2), && t\in(-\infty,t_0].\end{aligned}$$

As described in Greenspan and Holmes [1983], the distance $d(t_0)$ between the manifolds $W^s(p_\varepsilon)$, $W^u(p_\varepsilon)$ of the perturbed fixed point $p_\varepsilon=\gamma_\varepsilon\cap\Sigma^{t_0}$ of the map $p_\varepsilon^{t_0}$ is well approximated by the *Melnikov function* $M(t_0)$:

$$(2.3)\qquad d(t_0)=\frac{\varepsilon M(t_0)}{|f(q^0(0))|}+O(\varepsilon^2).$$

Here $M(t_0)$ is given by the simple formula

$$M(t_0)=\int_{-\infty}^{\infty} f(q^0(t))\wedge g(q^0(t),t+t_0)\,dt, \tag{2.4}$$

where the wedge product is defined as $a\wedge b=a_1b_2-a_2b_1$. We then have the following:

THEOREM 2.2. *If $M(t_0)$ has simple zeros and maxima and minima of $O(1)$, then, for $\varepsilon>0$ sufficiently small, $W^u(p_\varepsilon^{t_0})$ and $W^s(p_\varepsilon^{t_0})$ intersect transversely. If $M(t_0)$ remains bounded away from zero then $W^u(p_\varepsilon^{t_0})\cap W^s(p_\varepsilon^{t_0})=\varnothing$.*

COROLLARY 2.3. *Consider the parameterized family $\dot{x}=f(x)+\varepsilon g(x,t;\mu)$, $\mu\in\mathbb{R}$ and let hypotheses* A1–A3 *hold. Suppose that the Melnikov function $M(t_0,\mu)$ has a quadratic zero $M(\tau,\mu_b)=(\partial M/\partial t_0)(\tau,\mu_b)=0$ but $(\partial^2M/\partial t_0^2)(\tau,\mu_b)\neq 0$ and $(\partial M/\partial\mu)(\tau,\mu_b)\neq 0$. Then $\mu_B=\mu_b+O(\varepsilon)$ is a bifurcation value for which quadratic homoclinic tangencies occur in the family of systems.*

We remark that, if $g=g(x)$ is not explicitly time-dependent, then we have, using Green's theorem,

$$\begin{aligned}\int_{-\infty}^{\infty} f(q^0(t))\wedge g(q^0(t))\,dt&=\int_{-\infty}^{\infty}(f_1g_2-f_2g_1)\,dt\\ &=\int\left(g_2(u^0,v^0)\dot{u}^0-g_1(u^0,v^0)\dot{v}^0\right)dt\\ &=\int_{\operatorname{int}\Gamma}\operatorname{trace}Dg(x)\,dx.\end{aligned} \tag{2.5}$$

Thus the formula obtained in Andronov et al. [1971] is a special (planar) case of the more general Melnikov function which describes the "splitting" of the separatrices.

We now turn to the periodic orbits $q^\alpha(t)$ within Γ^0. To study these we need the subharmonic Melnikov function. Letting $q^\alpha(t-t_0)$ be a periodic orbit of period mT/n, with m and n relatively prime, we set

$$M^{m/n}(t_0)=\int_0^{mT} f(q^\alpha(t))\wedge g(q^\alpha(t),t+t_0)\,dt. \tag{2.6}$$

THEOREM 2.4. *If $M^{m/n}(t_0)$ has simple zeros and maxima and minima of $O(1)$, and $dT_\alpha/dh_\alpha\neq 0$, then for $0<\varepsilon\leq\varepsilon(n)$,* (2.1) *has a subharmonic orbit of period mT. If $n=1$ then the result is uniformly valid in $0<\varepsilon\leq\varepsilon_0=\varepsilon(1)$.*

COROLLARY 2.5. *Consider the parametrized family $\dot{x}=f(x)+\varepsilon g(x,t;\mu)$, $\mu\in\mathbb{R}$, and let hypotheses* A1–A3 *hold. Suppose that $M^{m/n}(t_0,\mu)$ has a quadratic zero $M^{m/n}=\partial M^{m/n}/\partial t_0=0$, $\partial^2M^{m/n}/\partial t_0^2$, $\partial M^{m/n}/\partial\mu\neq 0$ at $\mu=\mu_b$. Then $\mu_{m/n}=\mu_b+O(\varepsilon)$ is a bifurcation value at which saddle-nodes occur.*

The final result is a generalization of one obtained by Chow, Hale and Mallet-Paret [1980]. It implies that the homoclinic bifurcation is the limit of a countable sequence of subharmonic saddle-node bifurcations.

THEOREM 2.6. *Let $M^{m/1}(t_0)=M^m(t_0)$. Then*

$$\lim_{m\to\infty} M^m(t_0)=M(t_0). \tag{2.7}$$

The existence and bifurcation results summarized above are supplemented by a perturbation method which enables us to compute the global structure of the perturbed Poincaré map $P_\varepsilon^{t_0}$, and to determine how the sets of subharmonics and homoclinic orbits are related. Our starting point is Melnikov [1963, §7], although we have somewhat modified his transformations.

Since the unperturbed system is Hamiltonian, a symplectic change of coordinates to action angle variables can be found in the interior of Γ^0:

$$I=I(u,v), \qquad \theta=\theta(u,v). \tag{2.8}$$

Under this change of coordinates (2.1) becomes

$$\begin{aligned} \dot{I}&=\varepsilon\left(\frac{\partial I}{\partial u}g_1+\frac{\partial I}{\partial v}g_2\right)\overset{\text{def}}{=}F(I,\theta,t),\\ \dot{\theta}&=\Omega(I)+\varepsilon\left(\frac{\partial \theta}{\partial u}g_1+\frac{\partial \theta}{\partial v}g_2\right)\overset{\text{def}}{=}G(I,\theta,t) \end{aligned} \tag{2.9}$$

where $\Omega(I^\alpha)=\frac{\partial H}{\partial I}(I^\alpha)=2\pi/T_\alpha$ is the angular frequency of the unperturbed orbit $q^\alpha(t)$ with action $I^\alpha=I(q^\alpha)$. We now consider small perturbations of a resonant orbit $T_\alpha=\frac{mT}{n}$. Letting

$$\begin{aligned} I&=I^\alpha+\sqrt{\varepsilon}\,h,\\ \theta&=\Omega(I^\alpha)t+\phi=\left(\frac{2\pi n}{mT}\right)t+\phi\overset{\text{def}}{=}\Omega^\alpha t+\phi, \end{aligned} \tag{2.10}$$

we obtain

$$\begin{aligned} \dot{h}&=\sqrt{\varepsilon}\,F(I^\alpha,\Omega^\alpha t+\phi,t)+\varepsilon F'(I^\alpha,\Omega^\alpha t+\phi,t)h+O(\varepsilon^{3/2}),\\ \dot{\phi}&=\sqrt{\varepsilon}\,\Omega'(I_\alpha)h+\varepsilon\left(G(I^\alpha,\Omega^\alpha t+\phi,t)+\frac{\Omega''(I^\alpha)h^2}{2}\right)+O(\varepsilon^{3/2}), \end{aligned} \tag{2.11}$$

where $'$ denotes $\frac{\partial}{\partial I}$. Here we have expanded in Taylor series and used the fact that $\Omega'\neq 0$, since $dT_\alpha/dh_\alpha\neq 0$. Since

$$\frac{\partial I}{\partial u}=\frac{\partial I}{\partial H}, \quad \frac{\partial H}{\partial u}=-\frac{1}{\Omega(I)}f_2 \quad\text{and}\quad \frac{\partial I}{\partial v}=\frac{1}{\Omega(I)}f_1, \tag{2.12}$$

(2.11) can be rewritten as

$$\begin{aligned} \dot{h}&=\sqrt{\varepsilon}\,\frac{1}{\Omega^\alpha}f\left(q^\alpha(t)\wedge g\left(q^\alpha(t),t+\frac{\phi}{\Omega^\alpha}\right)\right)+\varepsilon[F'(I^\alpha,\Omega^\alpha t+\phi,t)h]+O(\varepsilon^{3/2}),\\ \dot{\phi}&=\sqrt{\varepsilon}\,\Omega'(I^\alpha)h+\varepsilon\left[\frac{\Omega''(I^\alpha)h^2}{2}+G(I^\alpha,\Omega^\alpha t+\phi,t)\right]+O(\varepsilon^{3/2}). \end{aligned} \tag{2.13}$$

Provided that $\Omega'(I^\alpha)$ is bounded, for $\sqrt{\varepsilon}$ sufficiently small, the averaging theorem (cf. Hale [1963]) can be applied to the leading term of (2.13) to yield

$$\dot{\bar{h}}=\sqrt{\varepsilon}\,\frac{1}{\Omega^\alpha}\frac{1}{mT}\int_0^{mT}f(q^\alpha(t))\wedge g(q^\alpha(t),t+\bar{\phi}/\Omega^\alpha)\,dt$$

or

$$\dot{\bar{h}}=\sqrt{\varepsilon}\,\frac{1}{2\pi n}M^{m/n}\left(\frac{\bar{\phi}}{\Omega^\alpha}\right), \qquad \dot{\bar{\phi}}=\sqrt{\varepsilon}\,\Omega'(I^\alpha)\bar{h}. \tag{2.14}$$

Under the averaging theorem, the hyperbolic or elliptic fixed points of (2.14) correspond to small periodic motions of (2.11) and hence to subharmonics of order m/n of (2.1). It is, of course, no coincidence that a necessary and sufficient condition for the existence of such fixed points is that the Melnikov function $M^{m/n}$ have simple zeros and that $\Omega'(I^\alpha)\neq 0$ ($dT_\alpha/dh_\alpha\neq 0$). We note, however, that in approaching a homoclinic orbit, $\Omega'(I^\alpha)$ typically grows without bound, and, in contrast to the uniform validity of Theorems 2.2–2.6, the averaged results become invalid in this region.

We note that (2.14) is a structurally unstable Hamiltonian system with Hamiltonian

$$\mathcal{H}=\sqrt{\varepsilon}\left(\frac{\Omega' h^2}{2}-V(\bar{\phi})\right), \tag{2.15}$$

where $V(\phi)=(1/2\pi n)\int M^{m/n}(\bar{\phi}/\Omega^\alpha)\,d\phi$, and thus to determine the stability and the global behavior of orbits of the unperturbed system near the resonant orbit q^α, we must investigate the terms of $O(\varepsilon)$. Thus, second order averaging is necessary (cf. Holmes and Holmes [1981]). Letting $f\wedge g=(1/2\pi n)M^{m/n}(\phi/\Omega^\alpha)+\tilde{F}(\phi,t)$ where $\tilde{F}$ has period T and zero mean, the averaging transformation is

$$h=\bar{h}+\sqrt{\varepsilon}\int\tilde{F}(\bar{\phi},t)\,dt,\qquad \phi=\bar{\phi}, \tag{2.16}$$

where the antiderivative is defined up to a t-independent term generally taken to be zero. Using (2.16), (2.11) becomes

(2.17)

$$\begin{aligned}\dot{\bar{h}}&=\sqrt{\varepsilon}\,\frac{1}{2\pi n}M^{m/n}(\phi/\Omega^\alpha)+\varepsilon\left(F'(I^\alpha,\Omega^\alpha t+\bar{\phi},t)\bar{h}-\Omega'\frac{\partial}{\partial\phi}\int\tilde{F}\,dt\,\Omega'\bar{h}\right)+O(\varepsilon^{3/2}),\\ \dot{\bar{\phi}}&=\sqrt{\varepsilon}\,\Omega' h+\varepsilon\left(\frac{\Omega'' h^2}{2}+G(I^\alpha,\Omega^\alpha t+\bar{\phi},t)+\Omega'\int\tilde{F}\,dt\right)+O(\varepsilon^{3/2}).\end{aligned}$$

Since $\tilde{F}$ has zero mean (it is simply a sum of Fourier components), $\int\tilde{F}$ and $\frac{\partial}{\partial\phi}\int\tilde{F}$ also have zero mean and on a second application of averaging to the $O(\varepsilon)$ terms of (2.17) we obtain (dropping the bars)

$$\begin{aligned}\dot{h}&=\sqrt{\varepsilon}\,\frac{1}{2\pi n}M^{m/n}(\phi/\Omega^\alpha)+\overline{\varepsilon F'(\phi)}h+O(\varepsilon^{3/2}),\\ \dot{\phi}&=\sqrt{\varepsilon}\,\Omega' h+\varepsilon\left(\frac{\Omega'' h^2}{2}+\overline{G(\phi)}\right)+O(\varepsilon^{3/2}),\end{aligned} \tag{2.18}$$

where $\overline{F'}$, $\bar{G}$ are the averages of F' and G. As Morosov [1973] notes, this second order averaging generally suffices to determine the stability of the fixed points and hence of the bifurcating subharmonics, at least for $\Omega'<\infty$ and ε sufficiently small. However, as we shall see in our application in §§3 and 4, one can sometimes also obtain global information on the Poincaré map by considering the time T flow maps of the averaged systems (2.18) in the neighborhood of each resonant and nonresonant periodic orbit. These results on the full Poincaré map $P_\varepsilon^{t_0}$ follow from application of the averaging theorem (Hale [1969]). In other situations, the T-periodic terms in the $O(\varepsilon)$ components of (2.17) are of crucial importance in establishing the global structure of solutions of the system, and averaging leads to qualitatively incorrect results (cf. Holmes [1979], [1980]). We shall meet both situations in the examples which follow.

3. An example of a single passage through resonance: the nonlinear harmonic oscillator. The computations necessary for application of the Melnikov theory outlined above, while not conceptually difficult, are often tedious, and the main ideas tend to be obscured by lengthy computations with special functions. In this section, therefore, we present a simple example in which all functions are trigonometric and all transformations can be made explicitly. Moreover, this example exhibits all the behavior found in each resonance band in our main example, to follow in §4.

We consider the system

$$\begin{aligned}\dot{u}&=v(\Omega-(u^2+v^2))+\varepsilon(\delta u-u(u^2+v^2)+\gamma u\cos t),\\ \dot{v}&=-u(\Omega-(u^2+v^2))+\varepsilon(\delta v-v(u^2+v^2)),\end{aligned}\tag{3.1}$$

where Ω is a fixed parameter, δ and γ vary and $0<\varepsilon\ll 1$ is a (small) scaling parameter. We first make the transformation to standard action angle coordinates:

$$u=\sqrt{2I}\sin\theta,\qquad v=\sqrt{2I}\cos\theta,\tag{3.2}$$

to obtain

$$\begin{aligned}\dot{I}&=\varepsilon[2\delta I-4I^2+2\gamma I\sin^2\theta\cos t],\\ \dot{\theta}&=\Omega-2I+\varepsilon[\gamma\sin\theta\cos\theta\cos t].\end{aligned}\tag{3.3}$$

The period $T(I)$ of the orbits of the unperturbed system is given by

$$T=\frac{2\pi}{(\Omega-2I)},\tag{3.4}$$

and the unperturbed phase plane is filled with periodic orbits (apart from the circle $I=\Omega/2$ which is filled with degenerate fixed points); see Fig. 2.

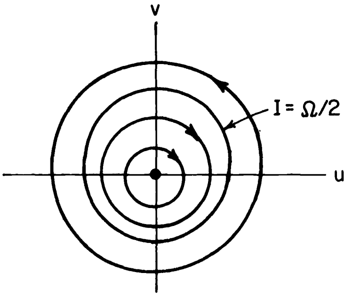


FIG. 2. *The unperturbed nonlinear harmonic oscillator.*

First consider the case $\gamma=0$ (no external forcing). It is easy to see that the dissipation parameter δ acts as follows. For $\delta\leq 0$ all the closed orbits of Fig. 2 are broken and there is a unique, globally stable sink at the origin, while for $\delta>0$ there is a unique stable, hyperbolic periodic orbit given by

$$I=\frac{\delta}{2},\tag{3.5a}$$

with period

$$T(\delta)=\frac{2\pi}{\Omega-\delta}.\tag{3.5b}$$

The Poincaré map associated with (3.3) therefore has an attracting invariant circle on which the rotation number is rational if $T(\delta)/2\pi = p/q$, $p, q \in \mathbb{Z}$, and otherwise irrational. (When $\delta = \Omega$, $T(\delta)$ is infinite and we have an attracting circle of degenerate fixed points.)

We now study the passage through resonance of subharmonics of order two; that is, we shall be concerned with the case in which $\delta \approx \Omega - \frac{1}{2}$, so that $T(\delta) \approx 4\pi$, and $\gamma \neq 0$. We therefore consider bifurcations from the resonant orbit given by

$$I = I^{\alpha} = \frac{\Omega}{2} - \frac{1}{4}. \tag{3.6}$$

The general theory leads us to expect the bifurcation of a finite set of points of period two in the Poincaré map. Here, where the perturbation calculations are straightforward, we are able to check this directly and also obtain more subtle, global information.

Following (2.10), we let

$$I = \frac{\Omega}{2} - \frac{1}{4} + \sqrt{\varepsilon}\, h, \qquad \theta = \frac{t}{2} + \phi, \tag{3.7}$$

so that (3.3) becomes, after some trignometrical expansion:

$$\begin{aligned} \dot{h} = {} & \sqrt{\varepsilon}\left[\delta\omega - \omega^2 + \frac{\gamma\omega}{2}\left(\cos t + \frac{\sin 2\phi}{2}\sin 2t - \frac{\cos 2\phi}{2}(1 + \cos 2t)\right)\right] \\ & + \varepsilon\left[2\delta - 4\omega + \gamma\left(\cos t + \frac{\sin 2\phi}{2}\sin 2t - \frac{\cos 2\phi}{2}(1 + \cos 2t)\right)\right] + 4\varepsilon^{3/2}h^2, \end{aligned} \tag{3.8a}$$

$$\dot{\phi} = -\sqrt{\varepsilon}\, 2h + \varepsilon\left[\frac{\gamma}{4}(\cos 2\phi \sin 2t + \sin 2\phi(1 + \cos 2t))\right], \tag{3.8b}$$

where $\omega = \Omega - \frac{1}{2}$.

To average the first order ($O(\sqrt{\varepsilon})$) terms we use the transformation (cf. (2.16))

$$h = \bar{h} + \sqrt{\varepsilon}\, \frac{\gamma\omega}{2} \int \left(\cos t + \frac{\sin 2\bar{\phi}}{2}\sin 2t - \frac{\cos 2\bar{\phi}}{2}\cos 2t\right) dt, \qquad \varphi = \bar{\varphi}, \tag{3.9}$$

where the bracketed term $\tilde{F}(\bar{\phi}, t)$ is the oscillating part of the leading term of (3.8a). We have the antiderivative

$$\int \tilde{F}(\bar{\phi}, t)\, dt = \left(\sin t - \frac{\sin 2\bar{\phi}}{4}\cos 2t - \frac{\cos 2\bar{\phi}}{4}\sin 2t\right), \tag{3.10}$$

and thus, applying the transformation (3.9), (3.8a, b) become (cf. (2.17)):

$$\begin{aligned} \dot{\bar{h}} = {} & \sqrt{\varepsilon}\,\omega\left[\delta - \omega - \frac{\gamma}{4}\cos 2\bar{\phi}\right] \\ & + \varepsilon\left[2\delta - 4\omega + \gamma\left(\cos t + \frac{\sin 2\bar{\phi}}{2}\sin 2t - \frac{\cos 2\bar{\phi}}{2}(1 - \cos 2t)\right)\right. \\ & \left. - \gamma\omega\left(\frac{\cos 2\bar{\phi}}{2}\cos 2t - \frac{\sin 2\bar{\phi}}{2}\sin 2t\right)\right] h + O(\varepsilon^{3/2}), \\ \dot{\bar{\phi}} = {} & -\sqrt{\varepsilon}\, 2h + \varepsilon\left[\frac{\gamma}{4}(\cos 2\bar{\phi}\sin 2t + \sin 2\bar{\phi}(1 + \cos 2t))\right. \\ & \left. - \gamma\omega\left(\sin t - \frac{\sin 2\bar{\phi}}{4}\cos 2t - \frac{\cos 2\bar{\phi}}{4}\sin 2t\right)\right] + O(\varepsilon^{3/2}). \end{aligned} \tag{3.11}$$

We note that a direct calculation of the Melnikov function from (3.1) yields $M(t_0)=2\pi\omega[\delta-\omega-\frac{\gamma}{4}\cos t_0]$, which, upon setting $t_0=\bar{\phi}/\Omega^{\alpha}=2\bar{\phi}$ and dividing by 2π, as in (2.14), yields the leading term of (3.11).

Averaging the $O(\varepsilon)$ terms, (implicitly) using a second transformation $(\bar{h},\bar{\phi})\to(\bar{\bar{h}},\bar{\bar{\phi}})$ and dropping the double bars, and rescaling time by a factor $\sqrt{\varepsilon}$ we obtain

$$(3.12)\qquad \begin{aligned}\dot{h}&=\omega\left[\delta-\omega-\frac{\gamma}{4}\cos 2\phi\right]+\sqrt{\varepsilon}\left[2\delta-4\omega-\frac{\gamma}{2}\cos 2\phi\right]h+O(\varepsilon),\\ \dot{\phi}&=-2h+\sqrt{\varepsilon}\left[\frac{\gamma}{4}\sin 2\phi\right]+O(\varepsilon).\end{aligned}$$

The remainder of this section is devoted to an analysis of the autonomous averaged system (3.12) and a discussion of the implications for the Poincaré map of the full system (3.8a, b) or (3.1)–(3.3). Throughout, ω is fixed, ε is a fixed and sufficiently small parameter, and δ and γ are allowed to vary.

THEOREM 3.1. *The bifurcations set and phase portraits of the autonomous averaged system* (3.12) *in the neighborhood of the resonant band* $h=0$ $(I=\Omega/2-\frac{1}{4})$ *are homeomorphic to those shown in Figs.* 3 *and* 4. *In particular*:

(i) *Two pairs of fixed points exist within the region bounded by the lines AO, DO given by* $\gamma=\pm4(\delta-\omega)+O(\sqrt{\varepsilon})$; *if* $\delta<2\omega$ *these are saddles and sinks* (*if* $\delta>2\omega$, *saddles and sources*). *These fixed points coalesce in saddle-node bifurcations on AO, DO.*

(ii) *There are two curves BE, CE lying within* $O(\sqrt{\varepsilon})$ *of the line* $\delta=\omega$ *and meeting AO, DO at the points E, F. A further curve GH connects BE, CF and curves EI, JF join this curve as shown. Outside the region BEIJFC the phase portrait has a smooth invariant closed curve, within EOF this curve contains the four fixed points. Within BEIJFC no such curve exists. The curve becomes nondifferentiable and vanishes in saddle-connection* (*homoclinic*) *bifurcations on BEI, JFC and on IJ the sinks change from nodes to spirals, also leading to a loss of differentiability. Approaching EIJF from below the curve successively loses degrees of differentiability.*

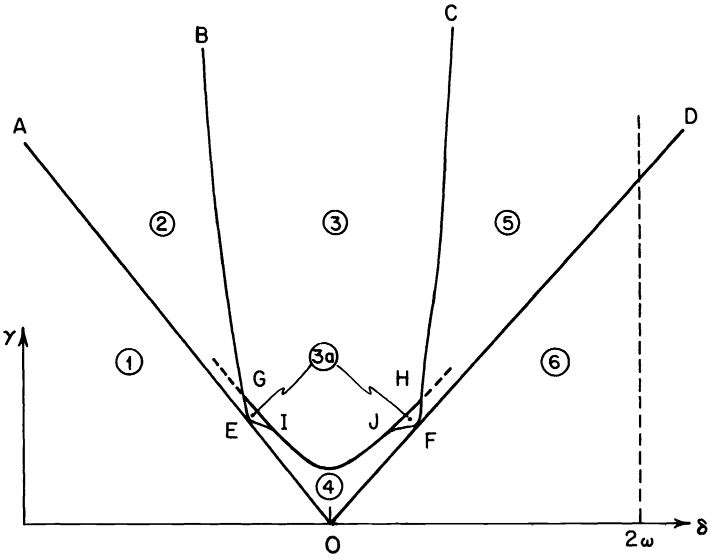


FIG. 3. *The bifurcation set for* (3.12). *For a detailed description of the region near E, F, see the proof of Lemma* 3.4.

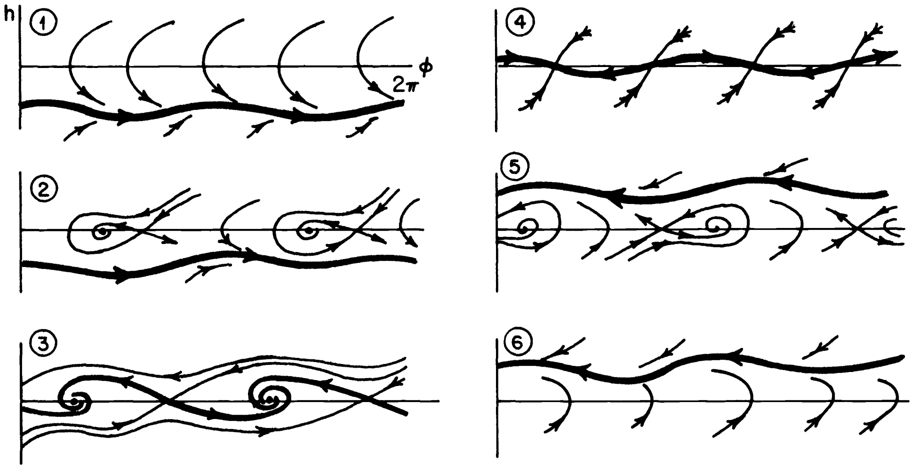


FIG. 4a. *Structurally stable phase portraits. Invariant closed curve shown as heavy line. Portraits in regions* 3a *are homeomorphic to those in* 3, *but the foci are nodes.*

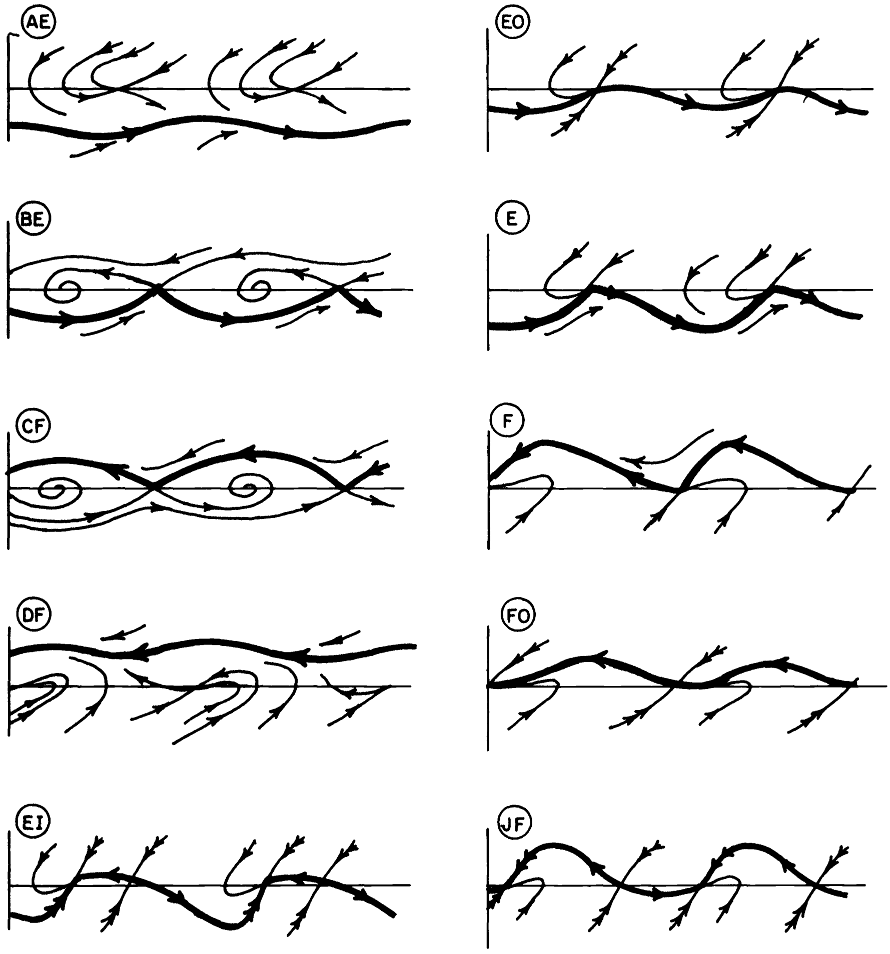


FIG. 4b. *Bifurcation phase portraits. Invariant closed curve shown as heavy line.*

Proof of Theorem 3.1.

Assertion (i). First consider the truncated system (3.12) with $O(\sqrt{\varepsilon})$ terms removed. This is a Hamiltonian system with Hamiltonian

$$\mathcal{H}(h,\phi)=h^2+\omega(\delta-\omega)\phi-\frac{\gamma\omega}{8}\sin 2\phi, \tag{3.13}$$

which has two saddles and two centers at $(h,\phi)=(0,\frac{1}{2}\cos^{-1}(4(\delta-\omega)/\gamma))$ if $\gamma>4|\delta-\omega|$. Thus we obtain the approximate saddle-node bifurcation curves $\gamma=\pm4(\delta-\omega)$. We next linearize the full system, including $O(\sqrt{\varepsilon})$ terms, to obtain the matrix

$$A=\begin{bmatrix}\sqrt{\varepsilon}\left(2\delta-4\omega-\frac{\gamma}{2}\cos 2\phi\right) & \frac{\gamma\omega}{2}\sin 2\phi+\sqrt{\varepsilon}\,\gamma h\sin 2\phi \\ -2 & \frac{\sqrt{\varepsilon}\,\gamma}{2}\cos 2\phi\end{bmatrix}. \tag{3.14}$$

Noting that trace $A=\sqrt{\varepsilon}(2\delta-4\omega)<0$ if $\delta<2\omega$, we obtain the stability results of assertion (i) in the theorem. Also, provided $\delta\neq2\omega$, by Bendixson's criterion no closed orbits exist in the *planar* flow of (3.12), although a unique closed curve does exist upon identification of $\phi=0$ with $\phi=2\pi$ (cf. assertion (ii), and see below). Finally, the phase portraits in regions ①, ②, ③, ⑤, ⑥ and on $\textcircled{\scriptsize AE}$ and $\textcircled{\scriptsize DF}$ follow from straightforward consideration of the level curves of $\mathcal{H}(h,\phi)$ and the perturbations due to the $O(\sqrt{\varepsilon})$ terms on solution curves, eigenvalues and eigenvectors.

Assertion (ii). This is proved in two stages. We first fix γ and $\delta\approx\omega$, taking $\delta=\omega+\sqrt{\varepsilon}\,\Delta$, and perturb the truncated Hamiltonian system (3.13) by adding the $O(\sqrt{\varepsilon})$ terms:

$$\begin{aligned}\dot{h}&=-\frac{\gamma\omega}{4}\cos 2\phi+\sqrt{\varepsilon}\left[\omega\Delta-\left(2\omega+\frac{\gamma}{2}\cos 2\phi\right)h\right]+O(\varepsilon),\\ \dot{\phi}&=-2h+\sqrt{\varepsilon}\,\frac{\gamma}{4}\sin 2\phi+O(\varepsilon).\end{aligned} \tag{3.15}$$

When $\sqrt{\varepsilon}=0$, (3.15) has a homoclinic cycle connecting the saddle points at $(h,\phi)=(0,3\pi/4)$, $(0,7\pi/4)$ and formed by the level curve $\mathcal{H}=\gamma\omega/8$ of the Hamiltonian

$$\mathcal{H}(h,\phi)=h^2-\frac{\gamma\omega}{8}\sin 2\phi. \tag{3.16}$$

The four branches of this cycle are conveniently given by solutions based at the points $(h,\phi)=(\pm\sqrt{\gamma\omega}/2,\pi/4)$, $(\pm\sqrt{\gamma\omega}/2,5\pi/4)$. Denoting one such branch by $(\hat{h}(t),\hat{\phi}(t))$ we can investigate whether it is broken or not for $\sqrt{\varepsilon}\neq0$ by computing the Melnikov function as in §2. Here the perturbation is time independent, and we have

$$M(\Delta)=\int_{-\infty}^{\infty}\left\{-\frac{\gamma^2\omega}{16}\cos 2\hat{\phi}\sin 2\hat{\phi}+2\hat{h}\left[\omega\Delta-\left(2\omega+\frac{\gamma}{2}\cos 2\hat{\phi}\right)\hat{h}\right]\right\}d\tau. \tag{3.17}$$

Noting that $\hat{h}(\tau)$ and $\sin 2\hat{\phi}(\tau)$ are even while $\cos 2\hat{\phi}(\tau)$ is odd, (3.17) may be simplified to

$$M(\Delta)=\int_{-\infty}^{\infty}[\omega\Delta-2\omega\hat{h}]2\hat{h}\,d\tau=\int_a^b[\omega\Delta-2\omega\hat{h}]\,d\phi, \tag{3.18}$$

where a and b take the values $3\pi/4$ and $7\pi/4$. Using $\hat{h}^2=(\gamma\omega/8)(1+\sin 2\hat{\phi})\equiv(\gamma\omega/4)\cos^2\psi$, where $\psi=\phi-\pi/4$, we obtain

$$M(\Delta)=\int_{\pi/2}^{3\pi/2}\left[\mp\omega\Delta+\omega\sqrt{\gamma\omega}\cos\psi\right]d\psi=\mp\pi\omega\Delta+2\omega\sqrt{\gamma\omega}\,, \tag{3.19}$$

for the upper and lower branches of $\mathcal{H}=\gamma\omega/8$ respectively.

A similar computation, involving integration from $\hat{\phi}=0$ to 2π or 2π to 0 along an unperturbed orbit with energy $\mathcal{H}>\gamma\omega/8$ shows that for each value of $\Delta<2\sqrt{\gamma\omega}/\pi$ precisely one such orbit is preserved below the resonance band, while for $\Delta>2\sqrt{\gamma\omega}/\pi$ one such orbit is preserved above it. In the original coordinates, these are smooth limit cycles lying in the annulus centered at $I=\Omega/2-\frac{1}{4}$ $(h=0)$. The topology of the (h,ϕ) phase space of (3.16) is important here, since, viewed as a planar system, there are no closed orbits! These limit cycles persist as δ moves away from $\delta=\omega$, as an examination of (3.12) shows, since for sufficiently large values of $|h|$ the term

$$\varepsilon\left[2\delta-4\omega-\frac{\gamma}{2}\cos 2\phi\right]h$$

becomes important, leading to a net upward trend of solutions for $h<0$ if $\delta<\omega$ and a net downward trend for $h>0$ if $\delta>\omega$. Finally, setting $\delta=\omega+\sqrt{\varepsilon}\Delta=\omega\pm2\sqrt{\varepsilon\gamma\omega}/\pi$ we obtain estimates of the saddle connection bifurcation curves BE, CF. Note, however, that this estimate is only valid for fixed γ, since as $\gamma\to0$ the term $\sqrt{\varepsilon}(2\delta-4\omega)h$ of (3.12) becomes greater than the "leading" term $(\gamma\omega/4)\cos 2\phi$. To complete the proof of assertion (ii) we now address this point.

We state several lemmas, which together complete the proof. These lemmas are proved at the end of the section.

LEMMA 3.2. *For* $\gamma=0$, $\delta<2\omega$ (3.12) *has a smooth normally hyperbolic attracting invariant closed curve* $h\approx\omega(\delta-\omega)/\sqrt{\varepsilon}(4\omega-2\delta)$.

By the persistence theory for such normally hyperbolic manifolds (*Hirsch, Pugh and Shub* [1977]), *this curve must persist for* γ *sufficiently small.*

LEMMA 3.3. *For* $\gamma>4|\delta-\omega|$ *and* γ *sufficiently small the invariant curve contains two saddles and two sinks and is composed of the union of these fixed points with the unstable manifolds of the saddles.*

LEMMA 3.4. *There are two unique points* $(\gamma^{\pm},\delta^{\pm})=(E,F)$ *near the curves* $\gamma=\pm4(\delta-\omega)$ *for which the invariant closed curve degenerates into two nondifferentiable saddle-node connections as shown in phase portraits* Ⓔ, Ⓕ.

LEMMA 3.5. *The sinks existing within the region* $\gamma>4|\delta-\omega|$ *are nodes below a curve given by* $\gamma\approx\sqrt{16(\delta-\omega)^2+\varepsilon^2\omega^4}$ *and foci above this curve.*

LEMMA 3.6. *Two curves EI, JF connect the points E, F with the curve* $\gamma\approx\sqrt{16(\delta-\omega)^2+\varepsilon^2\omega^4}$ *For parameter values on these curves the unstable manifolds of the saddle points make connections with the strong stable manifolds of the sinks, providing a nondifferentiable closed curve, as shown in the phase portraits* (EI), (JF).

These lemmas together complete the proof of assertion (ii), and their proofs, which follow, provide more details on the phase portraits of Figs. 4a and b. We note that results similar to the present ones were obtained by Levi, Hoppensteadt and Miranker [1978] in a study of bifurcations of the discrete sine-Gordon equation

$$\dot{\phi}=v,\qquad \dot{v}=-\sin\phi-\sigma v+I$$

under variation of dissipation, σ, and driving current, I.

Proof of Lemma 3.2. Setting $\gamma=0$ in (3.12) we find that $h=\omega(\delta-\omega)/\sqrt{\varepsilon}(4\omega-2\delta)$ is filled with degenerate saddle-nodes whose stable manifolds are a family of lines $h=\sqrt{\varepsilon}(2\omega-\delta)\phi+\text{const.}$ foliating the (h,ϕ) phase space. Moreover, solutions approach $h=\omega(\delta-\omega)/\sqrt{\varepsilon}(4\omega-2\delta)$ like $e^{-\sqrt{\varepsilon}(4\omega-2\delta)t}$. □

Proof of Lemma 3.3. Letting $\gamma=4\sqrt{\varepsilon}\,\Gamma$ and $\delta=\omega+\sqrt{\varepsilon}\,\Delta$, (3.12) becomes

$$\dot{h}=\sqrt{\varepsilon}\,(\Delta-\Gamma\cos 2\phi-2\omega h)+O(\varepsilon),\qquad \dot{\phi}=-2h+O(\varepsilon). \tag{3.20}$$

It is easy to check that, for $\Gamma>|\delta|$ $(\gamma>4|\delta-\omega|)$, (3.20) has sinks and saddles as specified, which coalesce pairwise in saddle-nodes when $\Delta=\pm\Gamma$. To prove that the unstable manifolds of the saddles form a smooth invariant curve, we consider the eigenvalues and eigenvectors of the linearized problem, with matrix

$$A=\begin{bmatrix} -2\sqrt{\varepsilon}\,\omega & 2\sqrt{\varepsilon}\sin 2\phi \\ -2 & 0 \end{bmatrix}. \tag{3.21}$$

Setting $\Gamma\sin 2\phi=\pm\sqrt{\Gamma^2-\Delta^2}$ $(\Gamma\cos 2\phi=\Delta)$ for sinks and saddles respectively, we obtain eigenvectors and eigenvalues as follows:

$$\text{(3.22)}\quad \text{sinks:}\qquad \lambda^s_{1,2}=-\sqrt{\varepsilon}\,\omega\pm\sqrt{\varepsilon\omega^2-4\sqrt{\varepsilon}(\Gamma^2-\Delta^2)}\ ,\qquad e^s_{1,2}=(-\lambda^s_{1,2},2);$$

$$\text{saddles:}\qquad \lambda^c_{1,2}=-\sqrt{\varepsilon}\,\omega\pm\sqrt{\varepsilon\omega^2+4\sqrt{\varepsilon}(\Gamma^2-\Delta^2)}\ ,\qquad e^c_{1,2}=(-\lambda^c_{1,2},2).$$

As $\Gamma,\Delta\to 0$, $\lambda^{s,c}_1\to 0$, $\lambda^{s,c}_2\to -2\sqrt{\varepsilon}\,\omega$ and the eigenvectors tend to $(0,1)$ and $(\sqrt{\varepsilon}\,\omega,1)$ respectively. Moreover, the lines $h=\sqrt{\varepsilon}\,\omega(\phi-c)$ are invariant stable manifolds for the fixed points $(h,\phi)=(0,c)$. For sufficiently small Γ,Δ, the stable manifolds of the surviving fixed points $(\hat{h},\hat{\phi})=(0,\frac{1}{2}\cos^{-1}(\Delta/\Gamma))$ must lie close to $h=\sqrt{\varepsilon}\,\omega(\phi-\hat{\phi})$ and hence cannot recross $h=0$ but must lie as in phase portrait ④ of Fig. 4a. This implies that the unstable manifold of each saddle must limit in the two sinks, and, moreover, must do so tangent to the slow eigenvector $(-\lambda^s_1,2)$, providing the required smooth curve. □

Proof of Lemma 3.4. Set $\gamma\approx 4(\delta-\omega)$, so that (3.12) has a pair of saddle-nodes at $(h,\phi)\approx(0,0)$, $(0,\pi)$. From the proof of assertion (i), and in particular consideration of the truncated Hamiltonian system, for sufficiently large γ solutions in both branches of the stable manifold of each saddle node lie in $h<0$ as $t\to-\infty$, while solutions in the center manifolds lie in $h>0$ as $t\to+\infty$. Thus, the left-hand branch of each center manifold lies above the right-hand branch of each stable manifold (phase portrait $\textcircled{\scriptsize DF}$, Fig. 4b.). Conversely, for sufficiently small γ we will show that the center manifold lies below the stable manifold, so that there must be at least one point (γ^+,δ^+) where the manifolds coincide as specified. Finally, we will show that this point is unique, thus establishing the bifurcation structures on *DFO*. Those on *AEO* $(\gamma\approx -4(\delta-\omega))$ follow in an analogous manner.

We set $\Delta=\Gamma$ in (3.20):

$$\dot{h}=\sqrt{\varepsilon}\,(\Gamma(1-\cos 2\phi)-2\omega h)+O(\varepsilon), \qquad (3.23)$$
$$\dot{\phi}=-2h+O(\varepsilon).$$

Since the vector field is π-periodic in ϕ we need only consider the interval $\phi\in[0,\pi]$. From (3.23), for $\Gamma=0$ the stable manifold of $(0,0)$ in the interval is given by the graph $h=\sqrt{\varepsilon}\,\omega\phi$, and thus, for Γ sufficiently small this manifold intersects the line $\phi=\pi$ above the fixed point $(0,\pi)$. A straighforward consideration of the vector field of (3.23) shows that solutions leaving $(0,\pi)$ in the left-hand branch of the center manifold of $(0,\pi)$ must remain in the interval $[0,\pi]$ below the stable manifold of $(0,0)$. All such solutions must therefore limit in the point $(0,0)$. Moreover, they must do so along one of the (nonunique) right-hand branches of the center manifold of this point. Since any center manifold for this analytic system is C^k for all k, any union of such center manifolds, joined at $(0,\pi)$ and $(0,0)$, must also be C^k for all k. Such a union provides the required smooth attracting invariant curve. See Fig. 5.

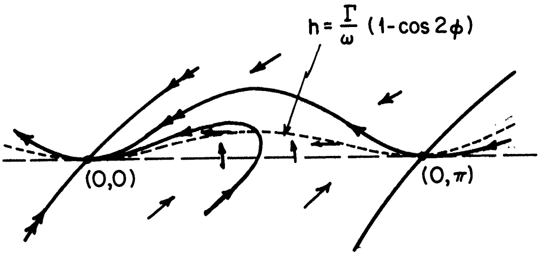


FIG. 5. *The center manifold for* $\Delta=\Gamma$.

To complete the proof of Lemma 3.4 it suffices to show that the upper (right-hand) branch of the stable manifold of $(0,0)$ moves down monotonically as Γ increases, so that there is precisely one value, Γ^+, for which it connects $(0,0)$ and $(0,\pi)$, as in portrait Ⓕ of Fig. 4b.

Let this manifold on the interval $(0,\pi]$ be given by the graph $h=h_\Gamma(\phi)$. Choose Γ_1 small, so that $h_{\Gamma_1}(\phi)>0$ on $(0,\pi]$ as in the proof of Lemma 3.3. Let $\Gamma_2>\Gamma_1$. Then, since everywhere the Γ_2 vector field (3.23) has a greater vertical component than the Γ_1 vector field, any solution based on h_{Γ_1} and integrated *backward* for Γ_2 must enter the region below h_{Γ_1} and continue to lie in it. Thus the curve h_{Γ_2}, forward asymptotic to $(0,0)$, lies below h_{Γ_1}. A similar argument shows that if h_{Γ_1} intersects $h=0$ at $\phi=\phi_1\in(0,\pi)$, then h_{Γ_2} intersects $h=0$ at some $\phi=\phi_2<\phi_1$; see Fig. 6. □

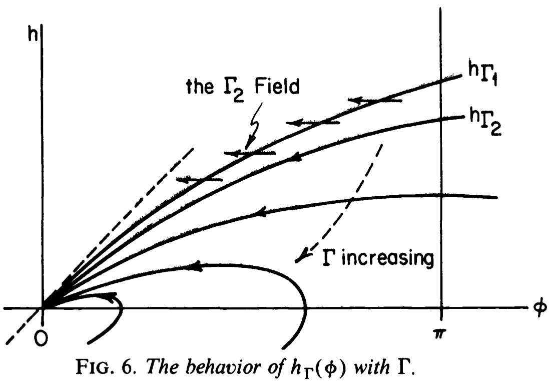


FIG. 6. *The behavior of* $h_\Gamma(\phi)$ *with* Γ.

Proof of Lemma 3.5. The eigenvalues $\lambda_{1,2}^s$ of the sinks given in (3.22) become complex on the curve

$$\varepsilon^2\omega^4 = 16\varepsilon(\Gamma^2-\Delta^2). \tag{3.24}$$

Setting $\gamma=4\sqrt{\varepsilon}\,\Gamma$, $\delta=\omega+\sqrt{\varepsilon}\,\Delta$ in (3.30) we obtain

$$\gamma=\sqrt{16(\delta-\omega)^2+\varepsilon^2\omega^4}, \tag{3.25}$$

as claimed. Crossing this parameter curve, the invariant closed curve changes from C^1 to C^0 (see Remark 3.1). □

Proof of Lemma 3.6. The proof of this result is essentially the same as that of Lemma 3.4. □

Remark. The precise degree of differentiability of the smooth closed curve which exists below the bifurcation curve *EIJF* is determined by the ratio of the eigenvalues of the sinks:

$$\frac{\lambda_2^s}{\lambda_1^s}=\frac{\sqrt{\varepsilon}\,\omega+\sqrt{\varepsilon\omega^2-4\sqrt{\varepsilon(\Gamma^2-\Delta^2)}}}{\sqrt{\varepsilon}\,\omega-\sqrt{\varepsilon\omega^2-4\sqrt{\varepsilon(\Gamma^2-\Delta^2)}}}\equiv\alpha.$$

Using linear theory near either of the sinks, it is easy to see that the two pieces of the closed curve meeting at a sink are approximated, in the canonical local coordinate system, by $y=c_1|x|^\alpha$, $x\le 0$, and $y=c_2|x|^\alpha$, $x\ge 0$, for some constants c_1, c_2. Letting $n(\alpha)$ denote the integer part of α, we see that all derivatives up to and including the nth are continuous at the sink, and hence that the curve changes from C^n to C^{n-1} near the parameter value

$$\gamma=\sqrt{16(\delta-\omega)^2+\left(\frac{4n}{(n+1)^2}\right)^2\varepsilon^2\omega^4}\,.$$

Note that, for $n=1$, we recover the C^1 to C^0 value of (3.25).

We now consider the implications of Theorem 3.1 for the full, time-dependent system (3.8) or, equivalently (3.1), (3.3). Under the conditions of the averaging theorem (Hale [1969]), hyperbolic fixed points of (3.12) correspond to points of period two for the Poincaré map of (3.1), (3.3) in the neighborhood of the unperturbed resonant orbit $I=\Omega/2-\frac{1}{4}$. Similarly the hyperbolic limit cycles of (3.12) correspond to smooth hyperbolic invariant closed curves of the map. Such closed curves may contain higher order subharmonic or dense orbits, depending on the rotation number, but this more delicate behavior is not revealed by our $O(\varepsilon)$ analysis.

Except on the curves *BEI* and *JFC* the phase portraits are all either structurally stable or exhibit saddle-node bifurcations of codimension one. Therefore, since the Poincaré map of the full system is close to the flow map of the autonomous system for $\varepsilon\neq 0$ and sufficiently small, this behavior persists for the full system in the sense that its Poincaré map is diffeomorphic to the time $2\sqrt{\varepsilon}\,\pi$ flow map of (3.12), via the change of coordinates (3.7). In particular, invariant curves of this map are diffeomorphic to solution curves of (3.12) and the saddle separatrices of the latter are diffeomorphic to the stable and unstable manifolds of the map. However, on *BEI* and *JFC* pairs of separatrices coincide, and this behavior is nongeneric for one-parameter families of maps, in which we expect, at worst, quadratic tangencies of manifolds (cf. Newhouse [1980]).

To check this in a specific case (on BE, CF) we restore the terms to (3.12) which were removed in the second averaging process and consider the periodically perturbed system (3.11), with $\delta = \omega + \sqrt{\varepsilon}\,\Delta$, as above. After rescaling time as in (3.12), we have

$$\begin{aligned}(3.26)\quad \dot{h} &= -\frac{\gamma\omega}{4}\cos 2\phi + \sqrt{\varepsilon}\left[\omega\Delta - \left\{2\omega - \gamma\left(c + \frac{\sin 2\phi}{2}s - \frac{\cos 2\phi}{2}(1+c)\right)\right.\right.\\ &\qquad\qquad \left.\left. - \gamma\omega(\cos 2\phi c - \sin 2\phi s)\right\}h\right],\\ \dot{\phi} &= -2h + \sqrt{\varepsilon}\left[\frac{\gamma}{4}(\cos 2\phi s + \sin 2\phi(1+c)) - \gamma\omega\left(s - \frac{\sin 2\phi}{4}c - \frac{\cos 2\phi}{4}s\right)\right],\end{aligned}$$

where c denotes $\cos(2t/\sqrt{\varepsilon})$ and s denotes $\sin(2t/\sqrt{\varepsilon})$. Computing the (time-dependent) Melnikov function by integrating along the unperturbed heteroclinic branches $(\hat{h}(t), \hat{\phi}(t))$ as before, we obtain

$$(3.27)\quad M(t_0,\Delta) \sim \pm\pi\omega\Delta + 2\omega\sqrt{\gamma\omega} - \left[K(\gamma,\omega)\varepsilon^{-3/2}e^{-\beta/\sqrt{\varepsilon}}\right]\sin\left(2t_0/\sqrt{\varepsilon}\right) + O\left(\frac{1}{\varepsilon}e^{-\beta/\sqrt{\varepsilon}}\right),$$

as $\varepsilon \to 0$, where $\beta = \pi\sqrt{\gamma\omega}/4$ and K is $O(1)$.

Since the oscillating part of M is exponentially small in $\sqrt{\varepsilon}$, it does not immediately follow that, if M has simple zeros, then the true distance function $d(t_0) = \sqrt{\varepsilon}M(t_0,\Delta) + O(\varepsilon)$ also has simple zeros (cf. (2.3)). However, the constant part of $d(t_0)$ certainly vanishes near $\Delta = \pm 2\sqrt{\gamma\omega}/\pi$, since $d(t_0)$ depends continuously on Δ, and the leading $O(\sqrt{\varepsilon})$ term has a simple zero with respect to Δ at $\pm 2\sqrt{\gamma\omega}/\pi$. Thus, choosing $\delta \approx \omega \pm 2\sqrt{\varepsilon\gamma\omega}/\pi$ such that $d(t_0)$ has zero mean, and assuming that $d(t_0)$ is analytic in γ, ω and ε (cf. Melnikov [1963]), we can conclude that, since at least one term of the oscillating part of $d(t_0)$ in (3.33) has simple zeros, there is an open set of $\gamma, \omega, \varepsilon$ values for which $d(t_0)$ also has simple zeros. It follows that the stable and unstable manifolds of the period two saddles in the Poincaré map intersect transversally with exponentially small oscillations in exponentially small neighborhoods of the "averaged" heteroclinic bifurcation points $\delta = \omega \pm \pi\sqrt{\varepsilon\gamma\omega}/2$. Moreover, on the boundaries of these regions the manifolds have quadratic tangencies, in view of Corollary 2.3. We illustrate this in Fig. 7a for the case $\delta \approx \omega - 2\sqrt{\varepsilon\gamma\omega}/\pi$. Similar splitting of the coincident manifolds of phase portraits E, F, EI and JF can also be expected to occur. These results agree with the generic case.

For more information and general results on exponentially small Melnikov functions and their implications, see Marsden and Holmes [1983].

We note that the behavior proven to occur for the system within the resonance region AOD, together with the partially conjectural results on homoclinic tangencies and transverse homoclinic orbits near BEI and JFC, is in agreement with the detailed computer observations of Aronson et al. [1980], [1982]. In particular, our computations suggest that, on the curves AO and DO, away from the points E, F, simple saddle-node bifurcations occur as in the portraits $\textcircled{AE}$, $\textcircled{EO}$, $\textcircled{DF}$, $\textcircled{FO}$ of Fig. 4b, and the rich homoclinic behavior detected by Aronson et al., if it all occurs in the present case, must be confined to a narrow band near the curves BEI and JFC. In Fig. 7b we conjecture a generic bifurcation diagram for the full Poincaré map, the letters and roman and arabic numerals on this figure correspond to the notation of Aronson et al. [1981], [1982], and we note that our results agree with their analysis and numerical computations. (cf. Aronson et al. [1982, Figs. 9.1–9.4]).

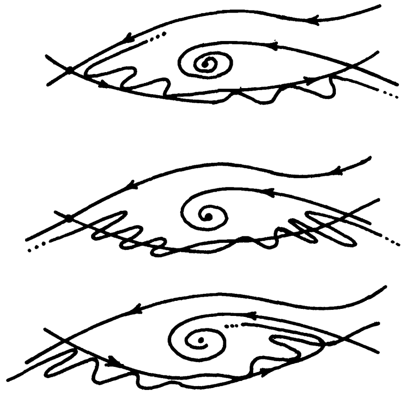


FIG. 7a. *Passage through homoclinic bifurcations for the Poincaré map near the curve BE.* (*cf. Aronson's cases* 6, *f and* 5 *respectively*).

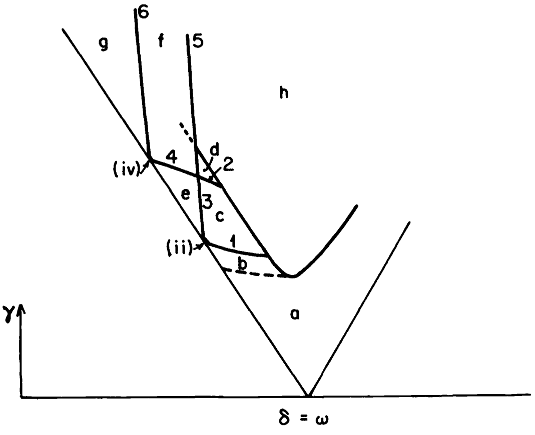


FIG. 7b. *A generic bifurcation set for the Poincaré map; only one side is shown. The letters and roman and arabic numerals refer to the cases classified by Aronson et al.* [1980], [1982].

Summarizing our results, we have

THEOREM 3.7. *The bifurcation set and associated invariant manifolds for the Poincaré map of* (3.1)–(3.3), *in the neighborhood of* $I=\Omega/2-\frac{1}{4}$, *are diffeomorphic to those described in Theorem* 3.1 *for the autonomous averaged system* (3.12), *with the following exception*: *Exponentially close* (*with respect to* $\sqrt{\varepsilon}$) *to the curve BEIJFC of Fig.* 3 *more complex global behavior involving transverse homoclinic orbits and quadratic tangencies will occur in the generic case.*

Using the Smale–Birkhoff homoclinic theorem (Smale [1963], [1967]), and Newhouse's [1979], [1980] results, we can therefore conclude that, in the generic case in a sufficiently small neighborhood of *BEIJFC*, the Poincaré map has countably many unstable periodic orbits of arbitrarily long periods, uncountably many bounded nonperiodic motions, and for a residual subset of parameter values, countably many stable periodic orbits. However, note that the stable period-two sinks will probably be the only "observable" attractors throughout the region bounded by the curve *AOD*.

4. An example of repeated resonance and homoclinic bifurcation. We now return to the example outlined in §1. Letting $\delta=\varepsilon\bar{\delta}$, $\gamma=\varepsilon\bar{\gamma}$, we have the system

$$\dot{u}=v, \qquad \dot{v}=u-u^3+\varepsilon\left(u^2v-\bar{\delta}v+\bar{\gamma}\cos t\right). \tag{4.1}$$

The unperturbed system ($\varepsilon=0$) has the phase portrait sketched in Fig. 8. We shall study perturbations from the family of periodic orbits (u^k,v^k) given by the elliptic functions with modulus $k\in(0,1)$:

$$u^k(t)=\sqrt{\frac{2}{2-k^2}}\,\mathrm{dn}\left(\frac{t}{\sqrt{2-k^2}},k\right), \tag{4.2}$$

$$v^k=-\left(\frac{\sqrt{2}k^2}{2-k^2}\right)\mathrm{sn}\left(\frac{t}{\sqrt{2-k^2}},k\right)\mathrm{cn}\left(\frac{t}{\sqrt{2-k^2}},k\right),$$

which limit on the center $(1,0)$ as $k\to 0$, and on the homoclinic orbit

$$u^1(t)=\sqrt{2}\,\mathrm{sech}\,t, \qquad v^1(t)=-\sqrt{2}\,\mathrm{sech}\,t\tanh t, \tag{4.3}$$

as $k\to 1$. These orbits are based at points $(\sqrt{2/(2-k^2)},0)$. A similar family exists within the left-hand half plane and periodic orbits encircling all three fixed points can also be given in terms of elliptic functions. For more details, see Greenspan [1981], Greenspan and Holmes [1983], and for general information on elliptic functions, see Byrd and Friedman [1971]. The period of the orbit (u^k,v^k) is

$$T(k)=2\int_{u^-(k)}^{u^+(k)}\frac{du}{\sqrt{2h+u^2-u^4/2}}=2\sqrt{2-k^2}\,K(k), \tag{4.4}$$

where $u^{\pm}=(1\pm\sqrt{1+4h})^{1/2}$, where $h=H(k)$ is the Hamiltonian energy defined below, and where $K(k)$ is the complete elliptic integral of the first kind. The unperturbed Hamiltonian of (4.1) can also be expressed as a monotonically increasing function of k for $k\in(0,1)$:

$$H(u^k,v^k)=\frac{v^{k2}}{2}-\frac{u^{k2}}{2}+\frac{u^{k4}}{4}=\frac{k^2-1}{(2-k^2)^2}\equiv H(k). \tag{4.5}$$

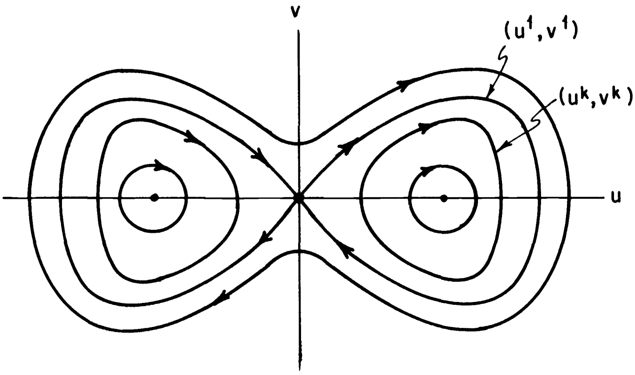


FIG. 8. *The unperturbed Duffing equation.*

Note that $H(k)\in(-\frac{1}{4},0)$ for the two families of orbits (u^k,v^k). Also we have

$$dT/dH=\frac{dT/dk}{dH/dk}>0 \quad \text{for } k\in(0,1), \tag{4.6}$$

and $dT/dH\to\infty$ as $k\to 1$. Hence assumptions A1–A3 of §2 are all satisfied, with the homoclinic orbit (u^1,v^1) playing the role of q^0.

We first compute the Melnikov function $M(t_0)$ for the homoclinic orbit:

$$\begin{aligned} M(t_0)&=\int_{-\infty}^{\infty} v^1(t)\left[(u^1(t))^2 v^1(t)-\bar{\delta}v^1(t)+\bar{\gamma}\cos(t+t_0)\right]dt \\ &=4\int_{-\infty}^{\infty}\operatorname{sech}^4 t\tanh^2 t\,dt-2\bar{\delta}\int_{-\infty}^{\infty}\operatorname{sech}^2 t\tanh^2 t\,dt \\ &\quad+\sqrt{2}\,\bar{\gamma}\int_{-\infty}^{\infty}\operatorname{sech} t\tanh t\cos(t+t_0)\,dt \\ &=\frac{16}{15}-\frac{4\bar{\delta}}{3}-\sqrt{2}\,\pi\bar{\gamma}\operatorname{sech}\left(\frac{\pi}{2}\right)\sin t_0. \end{aligned} \tag{4.7}$$

Here the final integral is computed by the method of residues. Thus, by Theorem 2.2 and Corollary 2.3, if

$$\bar{\gamma}>\left|\left(\frac{16-20\bar{\delta}}{15\sqrt{2}\,\pi}\right)\cosh\left(\frac{\pi}{2}\right)\right|, \tag{4.8}$$

transverse homoclinic orbits exist near the level curve $H(u,v)=0$, and quadratic homoclinic tangencies occur on curves near

$$\bar{\gamma}=\pm\left(\frac{16-20\bar{\delta}}{15\sqrt{2}\,\pi}\right)\cosh\left(\frac{\pi}{2}\right). \tag{4.9}$$

We next compute the subharmonic Melnikov function $M^{m/n}(t_0)$ of (2.6), selecting the unique resonant orbit within the right-hand homoclinic loop with period

$$T(k(m,n))=2\sqrt{2-k(m,n)^2}\,K(k(m,n))=\frac{2\pi m}{n}. \tag{4.10}$$

We obtain

$$M^{m/n}(\delta,\gamma,t_0)=J_1(m,n)-\delta J_2(m,n)-\gamma J_3(m,n)\sin t_0,$$

where

$$\begin{aligned} J_1(m,n)&=2\int_{u^-(k)}^{u^+(k)} u^2\sqrt{2H(k)+u^2-u^4/2}\,du \\ &=\frac{8}{15}\left[2(k^4+k'^2)E(k)+k'^2(k^2-2)K(k)\right](2-k^2)^{-5/2}, \\ J_2(m,n)&=2\int_{u^-(k)}^{u^+(k)}\sqrt{2H(k)+u^2-u^4/2}\,du \\ &=\frac{4}{3}\left[(2-k^2)E(k)-2k'^2K(k)\right](2-k^2)^{-3/2}, \end{aligned} \tag{4.11}$$

with

$$u^{\pm}(k)=\left[1\pm\sqrt{1+4H(k)}\right]^{1/2},$$

and

$$J_3(m,n)=\begin{cases}0, & n\neq 1,\\ \sqrt{2}\,\pi\,\mathrm{sech}(m\pi K'(k)/K(k)), & n=1.\end{cases}$$

Here $k=k(m,n)$, E is the complete elliptic integral of the second kind, $K'(k)=K(k')$ is the complementary complete elliptic integral of the first kind and $k'=\sqrt{1-k^2}$ is the complementary elliptic modulus. These integrals were evaluated using formulae and recursion relations found in Byrd and Friedman [1971].

Now, while J_3 is only defined for $k=k(m,n)$, J_1 and J_2 are defined for all $k\in(0,1)$. Replacing $J_i(m,n)$ by $J_i(k)$ for all $k\in(0,1)$ we can compute the following limits:

$$\lim_{k\to 1} J_1(k)=\tfrac{16}{15},\quad \lim_{k\to 0} J_1(0)=0,\quad \lim_{k\to 1} J_2(k)=\tfrac{4}{3},\quad \lim_{k\to 0} J_2(0)=0;$$

while

$$\lim_{k\to 0}\left(\frac{J_1(k)}{J_2(k)}\right)=1,$$

and

$$\lim_{m\to\infty} J_3(m,1)=\sqrt{2}\,\pi\,\mathrm{sech}\left(\frac{\pi}{2}\right). \tag{4.12}$$

Thus we verify that $M^{m/1}\to M$, as expected from Theorem 2.6. Moreover, we find that $J_1(k)$ and $J_2(k)$ increase monotonically from 0 to their respective limits as k increases from 0 to 1 (cf. Carr [1981, Chap. 4]).

We now define the *resonance ratio* $\hat{\delta}(k)$ as

$$\hat{\delta}(k)=\frac{J_1(k)}{J_2(k)}, \tag{4.13}$$

and, when $k=k(m,1)$, we write $\hat{\delta}(k(m,1))$ as

$$\bar{\delta}(m)=\frac{J_1(m,1)}{J_2(m,1)}. \tag{4.14}$$

LEMMA 4.1. *There exists $0\leq\tilde{k}<1$ such that $\hat{\delta}(k)$ decreases monotonically for $\tilde{k}<k<1$. Hence for m sufficiently large $\bar{\delta}(m)$ decreases monotonically with increasing m, limiting on $\bar{\delta}(\infty)=\frac{4}{5}$, so that a countable sequence of resonance ratios accumlates on this point from above.*

Proof. A direct (if tedious) computation using (4.11) shows that

$$\begin{aligned} &J_1'(k)J_2(k)-J_1(k)J_2'(k) \\ &\quad=\frac{1}{15}(2-k^2)^{-1}\big[80(2-k^2)E^2(k)-160k'^2K(k)E(k) \\ &\qquad -32(k^4+k'^2)K(k)E(k)+16k'^2(2-k^2)K^2(k)\big]. \end{aligned} \tag{4.15}$$

As $k \to 1$, we have the asymptotic behavior $k'=\sqrt{1-k^2} \to 0$, $E(k) \to 1$ and $K(k) \sim \ln(4/k') \to \infty$. Hence all terms in this expression remain bounded except $-32(k^4+k'^2)K(k)$, which approaches $-\infty$. Thus, for k sufficiently close to 1 (m sufficiently large) we have $\hat{\delta}'(k)=(J_1'J_2-J_1J_2')/J_2^2<0$. But from (4.4)–(4.6) we see that the period $T(k)$ of the unperturbed orbits increases monotonically with k, or, conversely, that $k(m,1)$ increases monotonically with m. It follows that $\bar{\delta}(m)=\hat{\delta}(k(m,1))$ decreases monotonically with m, as claimed. In fact, a numerical evaluation reveals that the expression (4.15) is negative for *all* $k \in (0,1)$, and so we can take $\bar{k} \equiv 0$. □

Remark 4.1. Carr [1981] obtained this lemma along with other results without computing J_1 and J_2 directly.

These results imply that the sequence of approximate saddle-node subharmonic bifurcation values

$$\bar{\gamma}=\pm\frac{J_1(m,1)-\bar{\delta}J_2(m,1)}{J_3(m,1)} \tag{4.16a}$$

accumulate on the homoclinic bifurcation curves (4.9). Each pair of lines (4.16a) forms the boundary of a resonance sector like that of §3, meeting the $\bar{\delta}$ axis at the point $\bar{\delta}(m)=J_1(m,1)/J_2(m,1)$. (Since $J_3(m,n)=0$ for $n \neq 1$ we are only concerned with resonances of order $m/1$.) As in the proof of Lemma 4.1, it can be checked that $J_2(m,1)/J_3(m,1)$ increases for sufficiently large m as $m \to \infty$ ($J_2'J_3-J_2J_3'>0$), so that the higher order resonance sectors become progressively narrower (cf. Greenspan [1981]).

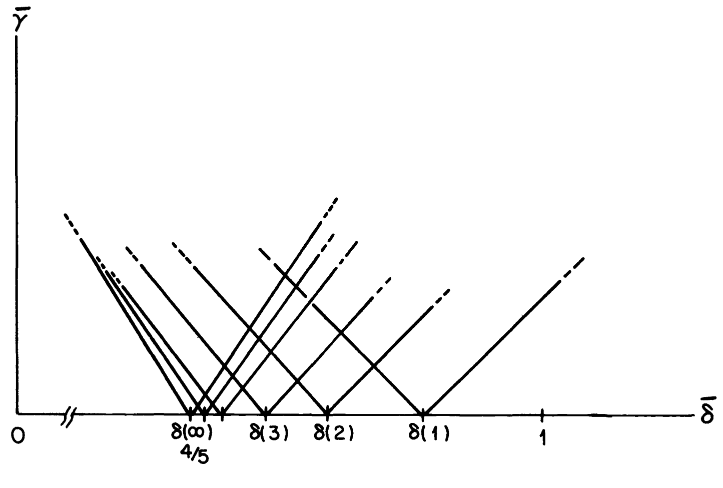


FIG. 9. *The countable sequence of resonance regions (not to scale). Here* $\delta=\varepsilon\bar{\delta}$, $\gamma=\varepsilon\bar{\gamma}$.

We illustrate these bifurcation curves in Fig. 9. However, we note that, while the existence results for subharmonics are uniformly valid for $0<\varepsilon\le\varepsilon_0$ (ε_0 is independent of m, cf. Theorem 2.4), the bifurcation curves generally vary with ε. In fact, from Corollary 2.5 and recalling that $\varepsilon\bar{\gamma}=\gamma$, $\varepsilon\bar{\delta}=\delta$, the approximate condition of (4.16a) should be replaced by

$$\gamma=\pm\frac{\varepsilon J_1(m,1)-\delta J_2(m,1)}{J_3(m,1)}+\varepsilon^2 C(\delta,m,\varepsilon), \tag{4.16b}$$

where $c_1 \leq C(\delta, m, \varepsilon) \leq c_2$ is uniformly bounded. (Here $\bar{\gamma}$ plays the role of μ in that corollary, and $\bar{\delta}$ is regarded as fixed.) Thus we cannot guarantee that the true bifurcation curves accumulate uniformly in ε as $m \to \infty$, as illustrated in Fig. 9. However, since the actual resonance regions are ε-close to the approximate regions illustrated, we *can* conclude that there are values of $\gamma, \delta, \varepsilon > 0$ for which countably many subharmonic coexist. (For example, pick $\gamma = \varepsilon$, $\delta = 4\varepsilon/5$ and $\varepsilon > 0$.) In particular, we note that the resonance sectors do not shrink to zero as $m \to \infty$.

To study the interior structure of each resonance sector we must compute the $O(\varepsilon)$ terms of (2.18). In the present case, rescaling time as in (3.11) we have

$$\begin{aligned} \dot{h} &= \frac{1}{2\pi} M^m(m\phi) + \sqrt{\varepsilon}\,\bar{F}'(\phi) h, \\ \dot{\phi} &= \Omega'(I^m) h + \sqrt{\varepsilon}\left(\frac{\Omega''(I^m) h^2}{2} + \overline{G(\phi)}\right), \end{aligned} \tag{4.17}$$

where $M^m(m\phi) = J_1(k) - \bar{\delta} J_2(k) - \bar{\gamma} J_3(k) \sin m\phi$, $k = k(m, 1)$ and $\Omega'(I^m)$ can be calculated as

$$\Omega'(I^m) = \frac{-\pi^2(2-k^2)[(2-k^2)E(k) - 2k'^2 K(k)]}{2k^4 k'^2 K^3(k)}. \tag{4.18}$$

We note that, as $m \to \infty$, $k \to 1$ and $\Omega'(I^m) \to \infty$; in fact

$$\Omega'(I^m) \sim \frac{e^{2\pi m}}{m^3}, \tag{4.19}$$

so that the averaged equations (4.17) are *not* uniformly valid in m, since $\sqrt{\varepsilon}$ must be taken successively smaller as $\Omega'(I^m)$ increases with m.

Now for $m < \infty$ the second order term $\Omega''(I^m)$ is a (negative) constant which need not be computed explicitly (in the example of §3, $\Omega'' \equiv 0$), but we do need the averaged functions

(4.20)

$$\overline{F'(\phi)} = \frac{\partial}{\partial I} \frac{1}{2\pi m} \int_0^{2\pi m} \frac{1}{\Omega(I)} V(I,\theta)[U^2(I,\theta)V(I,\theta) - \bar{\delta} V(I,\theta) + \bar{\gamma} \cos t]\, dt \,\Big|_{I=I^m},$$

and

$$\overline{G(\phi)} = \frac{1}{2\pi m} \int_0^{2\pi m} \frac{\partial \theta}{\partial V} [U^2(I^m,\theta)V(I^m,\theta) - \bar{\delta} V(I^m,\theta) + \bar{\gamma} \cos t]\, dt \Big|_{k=k(m,1)},$$

where $\theta = \Omega(I^m)t + \phi = t/m + \phi$ and U, V are determined by the action angle transformation. Here some results due to Greenspan [1981] which we now summarize are useful.

Derivatives of the symplectic action angle transformation $(U, V)\to(I(u,v),\theta(u,v))$ and its inverse $(I,\theta)\to(U(I,\theta), V(I,\theta))$ are related via their matrices of partial derivatives:

$$\begin{bmatrix} I_u & I_v \\ \theta_u & \theta_v \end{bmatrix}^{-1}=\begin{bmatrix} -\theta_v & I_v \\ \theta_u & -I_u \end{bmatrix}=\begin{bmatrix} U_I & U_\theta \\ V_I & V_\theta \end{bmatrix}, \tag{4.21}$$

since $\det\left[\begin{smallmatrix} I_u & I_v \\ \theta_u & \theta_v \end{smallmatrix}\right]=-1$. Thus we have $\theta_v=-U_I$ and using this, we may rewrite (4.20) as

(4.22)

$$\overline{F'(\phi)}=\frac{1}{2\pi m}\int_0^{2\pi m}(F_\gamma(\phi)+F_\delta(\phi))\,dt,\quad \overline{G(\phi)}=\frac{1}{2\pi m}\int_0^{2\pi m}(G_\gamma(\phi)+G_\delta(\phi))\,dt$$

where

$$\begin{aligned}
&F_\gamma(\phi)=\frac{\partial}{\partial I}\left(\frac{\bar\gamma}{\Omega(I)}V(I,\theta)\cos t\right)\Bigg|_{I=I^m},\\
&F_\delta(\phi)=\frac{\partial}{\partial I}\left(\frac{1}{\Omega(I)}\left[U^2(I,\theta)V^2(I,\theta)-\bar\delta V^2(I,\theta)\right]\right)\Bigg|_{I=I^m},\\
&F_\delta(\phi)=\frac{\partial}{\partial I}\left(\frac{1}{\Omega(I)}\left[U^2(I,\theta)V^2(I,\theta)-\bar\delta V^2(I,\theta)\right]\right)\Bigg|_{I=I^m},\\
&G_\gamma(\phi)=-\frac{\partial U}{\partial I}(I,\theta)\bar\gamma\cos t\Bigg|_{I=I^m},\\
&G_\delta(\phi)=-\frac{\partial U}{\partial I}(I,\theta)\left[U^2(I,\theta)V(I,\theta)-\bar\delta V(I,\theta)\right]\Bigg|_{I=I^m},
\end{aligned} \tag{4.23}$$

where $\theta=t/m+\phi$.

We first claim that the averaging transformations (cf. §2) can be chosen so that

$$\overline{F'_\gamma(\phi)}+\frac{\partial}{\partial\phi}\overline{(G_\gamma(\phi))}\equiv 0. \tag{4.24}$$

This follows from the fact that, when the perturbation term $\varepsilon(u^2v-\bar\delta v)$ is absent in (4.1), both the perturbed system and the truncated averaged system (4.17) are Hamiltonian. For details, see Greenspan [1981].

We next claim that $\overline{G_\delta(\phi)}$ is in fact independent of ϕ. To see this, note that in

$$\begin{aligned}
\overline{G_\delta(\phi)}=\frac{1}{2\pi m}\int_0^{2\pi m}&-\frac{\partial U}{\partial I}\left(I^m,\frac{t}{m}+\phi\right)\\
&\cdot\left[U^2\left(I^m,\frac{t}{m}+\phi\right)V\left(I^m,\frac{t}{m}+\phi\right)-\bar\delta V\left(I^m,\frac{t}{m}+\phi\right)\right]dt,
\end{aligned} \tag{4.25}$$

the change of variables $s=t/m+\phi$ removes explicit ϕ dependence in the integrand, while leaving the limits unchanged, due to the fact that all functions are $2\pi m$-periodic in t. Similarly, $\overline{F_\delta(\phi)}$ is independent of ϕ. Thus $\overline{G_\delta}$ and $\overline{F_\delta}$ are constants in any specific resonance order calculation.

As in the example of §3, the fixed points of (4.17) lie near $(h,\phi)=(0,\hat{\phi})$, where $\hat{\phi}$ is a root of $M^m(m\phi)=0$, and their stability types are determined by the matrix A of the linearization of (4.17). In particular, from the discussion above, we have

$$
\begin{aligned}
&\text{(4.26)}\\
&\operatorname{trace} A=\sqrt{\varepsilon}\,\overline{F_{\delta}'(\phi)}\\
&=\frac{\sqrt{\varepsilon}}{2\pi m}\frac{\partial}{\partial I}\left\{\frac{1}{\Omega(I)}\int_0^{2\pi m}V\left(I,\frac{t}{m}+\phi\right)\right.\\
&\qquad\qquad\left.\left.\cdot\left[U^2\left(I,\frac{t}{m}+\phi\right)V\left(I,\frac{t}{m}+\phi\right)-V\left(I,\frac{t}{m}+\phi\right)\right]dt\right\}\right|_{I=I^m}\\
&=\frac{\sqrt{\varepsilon}}{2\pi m}\left.\left\{\frac{\partial}{\partial I}\left(\frac{1}{\Omega(I)}\right)[J_1(k)-\bar{\delta}J_2(k)]+\left(\frac{1}{\Omega(I)}\right)\frac{\partial}{\partial k}[J_1(k)-\bar{\delta}J_2(k)]\frac{\partial k}{\partial H}\frac{\partial H}{\partial I}\right\}\right|_{I=I^m}\\
&=\frac{\sqrt{\varepsilon}}{2\pi m}\left\{-\frac{\Omega'(I^m)}{\Omega^2(I^m)}[J_1(k)-\bar{\delta}J_2(k)]+[J_1'(k)-\bar{\delta}J_2'(k)]\left.\frac{\partial k}{\partial H}\right|_{I=I^m}\right\}\\
&=\frac{\sqrt{\varepsilon}}{2\pi}\left\{-m\Omega'(I^m)[J_1(k)-\bar{\delta}J_2(k)]+\frac{1}{2m}\left(\frac{2-k^2}{k}\right)^3[J_1'(k)-\bar{\delta}J_2'(k)]\right\},\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k=k(m,1).
\end{aligned}
$$

At resonance $\bar{\delta}=\bar{\delta}(m)=J_1(k(m,1))/J_2(k(m,1))$, and this reduces to

$$
\frac{\sqrt{\varepsilon}}{4\pi m}\left(\frac{2-k^2}{k}\right)^3\left[J_1'(k)-\frac{J_1(k)J_2'(k)}{J_2(k)}\right],
$$

which, in view of (4.15), is a negative constant. Thus the fixed points are saddles or sinks, as in §3. Away from resonance, we have $J_1-\bar{\delta}J_2<0$ below resonance ($\bar{\delta}>\bar{\delta}(m)$) and $J_1-\bar{\delta}J_2>0$ above resonance ($\bar{\delta}<\bar{\delta}(m)$). Thus, since $\Omega'<0$, we have $\operatorname{trace}A<0$ for all $\bar{\delta}>\bar{\delta}(m)$, but $\operatorname{trace}A$ can change sign for $\bar{\delta}<\bar{\delta}(m)$. This is precisely as in §3, where we found that the fixed points are saddles and sinks for $\bar{\delta}<2\omega$ but saddles and sources for $\bar{\delta}>2\omega$. (In the present example, $\bar{\delta}$ *decreases* as we pass to successively longer period resonant orbits, rather than increasing as in the example of §3).

From the above discussion, and writing $J_i(m,1)$ as $J_i(m)$, (4.17) becomes

$$
\begin{aligned}
\text{(4.27)}\quad \dot{h}&=\frac{1}{2\pi}[J_1(m)-\bar{\delta}J_2(m)-\bar{\gamma}J_3(m)\sin m\phi]\\
&\quad+\sqrt{\varepsilon}\left[\omega(m)(J_1(m)-\bar{\delta}J_2(m)-\bar{\gamma}J_3(m)\sin m\phi)\right.\\
&\qquad\left.+\frac{1}{4\pi m}\left(\frac{2-k^2}{k}\right)^3(J_1'(m)-\bar{\delta}J_2'(m)-\bar{\gamma}J_3'(m)\sin m\phi)\right]h,\\
\dot{\phi}&=\Omega'(m)h+\sqrt{\varepsilon}\left[\frac{\Omega''(m)}{2}h^2-\bar{\gamma}\left(\omega(m)J_3(m)+\frac{1}{4\pi m}\left(\frac{2-k^2}{k}\right)^3J_3'(m)\right)\right.\\
&\qquad\left.\cdot m\cos m\phi+K_1(m)-\bar{\delta}K_2(m)\right],
\end{aligned}
$$

where $\omega(m)=-(m/2\pi)\Omega'(m)>0$, $\Omega'(m)=\Omega'(I^m)<0$ and Ω'', K_1, K_2 are constants, whose precise values we shall not require. Apart from the presence of these constants and the term $\Omega''h^2/2$ in the second component, (4.27) is, term by term, equivalent to (3.12). We are therefore able to conclude that the system governing the averaged behavior in a single resonance band in the present problem behaves just as does the model problem of §3. In particular, the delicate analysis necessary near each bifurcation point $(\delta,\gamma)=(\delta(m),0)$ is effected by writing

$$\bar{\delta}=\bar{\delta}\,(m)+\sqrt{\varepsilon}\,\Delta=J_1(m)/J_2(m)+\sqrt{\varepsilon}\,\Delta, \qquad \bar{\gamma}=\sqrt{\varepsilon}\,\Gamma, \tag{4.28}$$

so that (4.27) becomes

$$\begin{aligned} h'=\sqrt{\varepsilon}\Bigg[&-\frac{1}{2\pi}(J_2(m)\Delta+\Gamma J_3(m)\sin m\phi) \\ &+\frac{1}{4\pi m}\left(\frac{2-k^2}{k}\right)\frac{1}{J_2(m)}(J_1'(m)J_2(m)-J_1(m)J_2'(m))h\Bigg]+O(\varepsilon), \\ \dot{\phi}=\Omega'(m)h+\sqrt{\varepsilon}&\left[\frac{\Omega''(m)}{2}h^2+K_1(m)-\bar{\delta}K_2(m)\right]+O(\varepsilon). \end{aligned} \tag{4.29}$$

Equation (4.29) is the analogue of (3.20), and an analysis of its phase portraits yields analogues of Lemmas 3.2–3.6:

LEMMA 4.2. *For $\bar{\gamma}=0$, (4.27) has a smooth normally hyperbolic attracting invariant closed curve given by*

$$h\approx\frac{J_1(m)-\delta J_2(m)}{2\pi\sqrt{\varepsilon}\left[\omega(m)(J_1(m)-\bar{\delta}J_2(m))+\dfrac{1}{4\pi m}\left(\dfrac{2-k^2}{k}\right)^3(J_1'(m)-J_2'(m))\right]}. \tag{4.30}$$

This curve persists for γ sufficiently small.

LEMMA 4.3. *For $\bar{\gamma}>|(J_1(m)-\bar{\delta}J_2(m))/J_3(m)|$ and $\bar{\gamma}$ sufficiently small the invariant curve contains m saddles and m sinks and is composed of the union of these fixed points with the unstable manifolds of the saddles.*

LEMMA 4.4. *There are unique points $(\bar{\gamma}_m^{\pm},\bar{\delta}_m^{\pm})$ on the curves $\bar{\gamma}=\pm((J_1(m)-\bar{\delta}J_2(m))/J_3(m))$ at which the invariant curve degenerates into a set of m nondifferentiable saddle-node connections.*

LEMMA 4.5. *There is a curve $\mathcal{C}_m$ within the sector $\bar{\gamma}=|(J_1(m)-\bar{\delta}J_2(m))/J_3(m)|$ and within $O(\sqrt{\varepsilon})$ of its boundary below which the sinks are nodes and above which they are foci.*

LEMMA 4.6. *Two curves $\mathcal{C}_m^{\pm}$ connect the points $(\bar{\gamma}_m^{\pm},\bar{\delta}_m^{\pm})$ with $\mathcal{C}_m$. For parameter values on these curves the unstable manifolds of the saddle points make connections with the strong stable manifolds of the sinks, providing a nondifferentiable closed curve.*

The proofs of these results proceed just as do those of Lemmas 3.2–3.6. The presence of the additional terms in (4.27) and (4.29) in comparison with (3.12) and (3.20), respectively, introduces no new qualitative features.

As in §3, we therefore obtain a theorem on bifurcations of the autonomous averaged system for the order m resonance band entirely analogous to Theorem 3.1. Rather than stating this result, we recall that just as in the former case, the "gross" behavior carries over the Poincaré map for the full system, and we have the following final results:

THEOREM 4.7 (existence of subharmonics). *For* $0<\varepsilon\leq\varepsilon_0$ *and for parameter values within each resonance sector bounded by the curves* $\bar{\gamma}=\pm((J_1(m)-\bar{\delta}J_2(m))/J_3(m))$ *the Poincaré map of* (4.1) *has precisely* $2m$ *periodic points of period* m, m *of which are saddles. This result is uniform in the sense that* ε_0 *is a (small) constant independent of* m.

We remark that, for values of $\bar{\delta}$ sufficiently close to $\delta(\infty)=\frac{4}{5}$, and $\bar{\gamma}>0$, this implies that countably many pairs of subharmonics of arbitrarily high period coexist, since the resonance sectors overlap and accumulate on the homoclinic bifurcation curves (4.9).

The stability results are more delicate:

THEOREM 4.8 (stability and global behavior). *For* $0<\varepsilon\leq\varepsilon(m)$, *where* $\varepsilon(m)\to 0$ *as* $m\to\infty$, *the global structure of the bifurcation set and Poincaré map for* (4.1) *are diffeomorphic within each resonance band to those of the model problem of* §3 (*Theorems* 3.1, 3.8 *and Figs.* 3, 4), *with the following changes*:

(i) *There are* $2m$ *points of period* m, *rather than* 4 *of period* 2;

(ii) m *of these points are saddles and* m *sinks for* $\bar{\delta}>\bar{\delta}_s(m)$ *and* m *are saddles and* m *sources for* $\bar{\delta}<\bar{\delta}_s(m)$, *where*

$$\bar{\delta}_s(m)=\frac{\frac{1}{2}\left(\frac{2-k^2}{k}\right)^3 J_1(m)-m^2\Omega'(m)J_1(m)}{\frac{1}{2}\left(\frac{2-k^2}{k}\right)^3 J_2'(m)-m^2\Omega'(m)J_2(m)}. \tag{4.31}$$

Also, as in Theorem 3.8,

(iii) *Within each sector there are curves analogous to BEIJFC of Fig.* 3. *Transverse homoclinic orbits and quadratic tangencies will occur for parameters exponentially close (with respect to* $\sqrt{\varepsilon}$*) to these curves in the generic case.*

We recall that the nonuniform validity of the results of Theorem 4.8 are due to the fact that the derivative $\Omega'(I^m)$ in (4.17) grows without limit as $m\to\infty$.

We remark that the results of Greenspan [1981] and Greenspan and Holmes [1983] on perturbations of periodic motions outside the level curve $H(u,v)=0$, together with the results of Carr [1981], demonstrate that a second sequence of resonance sectors bounded by lines of the form

$$\gamma=\pm\left(\frac{\varepsilon\hat{J}_1(m)-\delta\hat{J}_2(m)}{\hat{J}_3(m)}\right)+O(\varepsilon^2) \tag{4.32}$$

accumulate on the homoclinic bifurcation sector from below. These sectors meet the δ axis at points $\hat{\delta}(m)\to 4/5^-$ as $m\to\infty$. Result analogous to Theorems 4.7–4.8 can be stated for these subharmonics and their bifurcations.

We close with some comments on the attracting set for the case $\bar{\delta}\approx\bar{\delta}(\infty)=\frac{4}{5}$. It is not difficult to check that, for $\gamma=0$ and $\bar{\delta}\approx\frac{4}{5}$ and $0<\varepsilon\leq\varepsilon_0$ sufficiently small, (4.1) possesses a double (figure of eight) attracting homoclinic orbit. This follows directly from the Melnikov theory, which shows that for $\bar{\delta}$ near $\bar{\delta}(\infty)=(J_1(k)/J_2(k))|_{k=1}=\frac{4}{5}$ the figure of eight level curve $H(u,v)=0$ is preserved, and a calculation of the trace of the linearized vectorfield at the saddle points $(u,v)=(0,0)$: yielding $-\varepsilon\delta$. A theorem of Andronov et al. [1966] on planar systems then implies that the homoclinic orbits attract nearby solutions, so that the Poincaré map of (4.1) with $\gamma=0$ takes a "thickened figure of eight": U, bounded, say, by the level curves $H=\pm\alpha$, into its interior:

$$\mathrm{Cl}(P(U))\subset U.$$

The attracting set is then defined as

$$A=\mathrm{Cl}\left(\bigcap_{n\geq 0}P^n(U)\right),$$

where Cl denotes closure. We note that, since the vector field is dissipative near the unperturbed saddle loop, the Poincaré map contracts areas and A therefore has zero Lesbesgue measure. For $\bar{\gamma}=0$, A is simply the union of the two homoclinic orbits and the saddle point. We now have

THEOREM 4.9. *For $0<\varepsilon\leq\varepsilon_0$ sufficiently small and $\bar{\delta}=\frac{4}{5}+O(\varepsilon)$, one may select $\bar{\gamma}>0$ such that the attracting set A of* (4.1) *contains horseshoes and hence contains a countable set of saddle type periodic points of arbitrarily high period and an uncountable set of bounded nonperiodic orbits. Moreover, while A may contain finite or countable sets of stable periodic orbits, none of their periods are less than some integer $N(\bar{\gamma})$, and $N(\bar{\gamma})\to\infty$ as $\bar{\gamma}\to 0$.*

Remark 4.2. The attracting set does not qualify as a nice attractor (or a strange attractor) since it may not contain a dense orbit. However, since the stable orbits can be made to have arbitrarily high period (for small $\bar{\gamma}$) they will be effectively unobservable in any numerical study and one will see "pseudochaos".

Proof. The Melnikov computations given above, together with Theorem 2.2, show that for all $\bar{\gamma}>0$ and $\bar{\delta}=\frac{4}{5}$, there exists an ε_0 such that for $\varepsilon<\varepsilon_0$ the Poincaré map P_ε has transverse homoclinic orbits. It follows by the standard arguments of the Smale–Birkhoff homoclinic theorem that some iterate P_ε^M of P_ε has horseshoes, i.e. P_ε^M has an invariant cantor set Λ^M on which P_ε^M is conjugate to a shift on two symbols. See Smale [1963], [1967]; Moser [1972], or Guckenheimer and Holmes [1983] for details. This proves the first part of the theorem.

Now as Newhouse [1974], [1979], [1980] pointed out, transversal homoclinic orbits can coexist with homoclinic tangencies, wild hyperbolic sets and their attendant stable periodic orbits. In fact the stable sinks we find in each resonance sector for finite m correspond to (some of) Newhouse's sinks. However, we *can* guarantee that as $\bar{\gamma}\to 0$ for $\bar{\delta}=\frac{4}{5}+O(\varepsilon)$ the periods of any such sinks $N(\bar{\gamma})\to\infty$. This is proved as follows.

First set $\bar{\gamma}=0$ in (4.1) to obtain an autonomous planar system. From the computations of (4.11) or the theorems of Carr [1981, Chap. 4], we see that the homoclinic figure of eight loop is preserved if $\bar{\delta}$ lies on a curve given by

$$\bar{\delta}=\frac{4}{5}+O(\varepsilon) \tag{4.33}$$

(cf. Corollary (2.3)). Choose $\bar{\delta}$ accordingly and now let $\bar{\gamma}$ vary. Our computations show that, for any finite $\bar{\gamma}$ and sufficiently small ε, countably many subharmonics coexist (cf. (4.16a-b), (4.32), and Theorem 2.6). However, as $\bar{\gamma}\to 0$ the values of m for which (4.16) and (4.32) are satisfied approach infinity. Hence saddle-node bifurcations occur in which orbits of successively higher periods $N(\bar{\gamma})$ coalesce and vanish (cf. Fig. 9). It follows that, for any specified integer N, we can choose values of $\bar{\gamma}$, $\bar{\delta}$ and $\varepsilon>0$ such that no periodic orbits of period $m<N$ exist. □

It is a reasonable conjecture that $A=\mathrm{Cl}(W^u(p))$, i.e. A is the closure of the unstable manifold of the unique saddle point $(u,v)=(0,0)+O(\varepsilon\gamma)$ for the perturbed Poincaré map. More details on attractors of this type appear in Holmes and Whitley [1983a, b].

REFERENCES

A. A. ANDRONOV, E. A. VITT AND S. E. KHAIKEN [1966], *Theory of Oscillators*, trans. F. Immirzi, Pergamon Press, Oxford.

A. A. ANDRONOV, E. A. LEONTOVICH, I. I. GORDON AND A. G. MAIER [1971], *Theory of Bifurcations of Dynamic Systems on a Plane*, Israel Program of Scientific Translation, Jerusalem.

______ [1973], *Theory of Dynamic Systems on a Plane*, Israel Program of Scientific Translation, Jerusalem.

V. I. ARNOLD [1976], *Loss of stability of self oscillations close to resonances and versal deformations of equivariant vector fields*, Functional Anal. Appl., 11, pp. 1–10.

D. G. ARONSON, M. A. CHORY, G. R. HALL AND R. P. MCGEEHEE [1980], *A discrete dynamical system with subtly wild behavior*, in New Approaches to Nonlinear Problems in Dynamics, P. J. Holmes, ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 339–359.

______ [1982], *Bifurcations from an invariant circle for two-parameter families of maps of the plane: A computer assisted study*, Comm. Math. Physics, 83, pp. 303–354.

P. F. BYRD AND M. D. FRIEDMAN [1971], *Handbook of Elliptic Integrals for Scientists and Engineers*, Springer-Verlag, Berlin.

J. CARR [1979], *Applications of Centre Manifold Theory*, Lecture Notes, Lefschetz Center for Dynamical Systems, Brown Univ., Providence, RI.

D. R. J. CHILLINGWORTH [1976], *Differentiable Topology with a View to Applications*, Pitman, London.

S. N. CHOW, J. K. HALE AND J. MALLET-PARET [1980], *An example of bifurcation to homoclinic orbits*, J. Differential Equations, 37, pp. 351–373.

E. H. DOWELL [1966], *Nonlinear oscillations of a fluttering plate*, AIAA J., 4, pp. 1267–1275.

B. D. GREENSPAN [1981], *Bifurcations in periodically forced oscillations: subharmonics and homoclinic orbits*, Ph. D. Thesis, Center for Applied Mathematics, Cornell Univ., Ithaca, NY.

B. D. GREENSPAN AND P. J. HOLMES [1983], *Homoclinic orbits, subharmonics and global bifurcations in forced oscillations*, Chapter 10 in Nonlinear Dynamics and Turbulence, G. Barenblatt, G. Iooss and D. D. Joseph, eds., Pitman, London, pp. 172–214.

J. GUCKENHEIMER AND P. J. HOLMES [1983], *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer-Verlag, New York.

J. K. HALE [1969], *Ordinary Differential Equations*, John Wiley, New York.

C. HAYASHI [1964], *Nonlinear Oscillations in Physical Systems*, McGraw-Hill, New York.

______ [1975], *Selected Papers on Nonlinear Oscillations*, Nippon Publ. Co., Osaka.

C. HOLMES AND P. J. HOLMES [1981], *Second order averaging and bifurcations to subharmonics in Duffing's equation*, J. Sound and Vibration, 78, pp. 161–174.

P. J. HOLMES [1977], *Bifurcations to divergence and flutter in flow-induced oscillations: A finite dimensional analysis*, J. Sound Vibration, 53, pp. 471–503.

______ [1979], *A nonlinear oscillator with a strange attractor*, Philos. Trans. Roy Soc. London Ser. A, 292, pp. 419–448.

______ [1980], *Averaging and chaotic motions in forced oscillations*, SIAM J. Appl. Math., 38, pp. 65–80.

______ [1981], *Center manifolds, normal forms and bifurcations of vector fields with application to coupling between periodic and steady motions*, Physica, 2D, pp. 449–481.

P. J. HOLMES AND J. E. MARSDEN [1978], *Bifurcations to divergence and flutter in flow-induced oscillations: an infinite dimensional analysis*, Automatica, 14, pp. 367–384.

P. J. HOLMES AND R. A. RAND [1980], *Phase portraits and bifurcations of the nonlinear oscillator* $\ddot{x}+(\alpha+\gamma x^2)\dot{x}+\beta x+\delta x^3=0$, Internat. J. Non-Linear Mech., 15, pp. 449–458.

P. J. HOLMES AND D. C. WHITLEY [1982], *A geometrically defined attractor for Duffing's equation*, in preparation.

______ [1983a], *On the attracting set of Duffing's equation*, I: *Analytical methods for small force and damping*, in Proc. Univ. of Houston Year of Concentration in Partial Differential Equations and Dynamical Systems, to appear.

______ [1983b], *On the attracting set for Duffing's equation* II: *A geometrical model for large force and damping*, Proc. Order in Chaos, Los Alamos National Lab., to appear in Physica D.

M. LEVI, F. HOPPENSTEADT AND W. MIRANKER [1978], *Dynamics of the Josephson Junction*, Quart. Appl. Math., 36, pp. 167–198.

J. E. MARSDEN AND P. J. HOLMES [1983], *On averaging and exponentially small Melnikov functions*, in preparation.

V. K. MELNIKOV [1963], *On the stability of the center for time periodic perturbations*, Trans. Moscow Math. Soc., 12, pp. 1–57.

F. C. MOON AND P. J. HOLMES [1979], *A magnetoelastic strange attractor*, J. Sound Vibration, 65, pp. 275–296.

A. D. MOROSOV [1973], *Approach to a complete qualitative study of Duffing's equation*, USSR Computational Math. and Math. Phys., 13, pp. 1134–1152.

______ [1976], *A complete qualitative investigation of Duffing's equation*, Differential Equations, 12, pp. 164–174.

J. MOSER [1973], *Stable and Random Motions in Dynamical Systems*, Princeton Univ. Pr., Princeton, NJ.

A. H. NAYFEH AND D. T. MOOK [1978], *Nonlinear Oscillations*, John Wiley, New York.

S. E. NEWHOUSE [1974], *Diffeomorphisms with infinitely many sinks*, Topology, 13, pp. 9–18.

______ [1979], *The abundance of wild hyperbolic sets and non-smooth stable sets for diffeomorphisms*, Publ. IHES, 50, pp. 101–151.

______ [1980], *Lectures on dynamical systems*, in Dynamical Systems, C.I.M.E. Lectures Bressanone, Italy, June 1978, Progress in Mathematics, 8, Birkhauser, Boston.

S. SMALE [1963], *Diffeomorphisms with many periodic points*, in Differential and Combinatorial Topology, S. S. Cairns, ed., pp. 63–80, Princeton Univ. Pr., Princeton, NJ.

______ [1967], *Differentiable dynamical systems*, Bull. Amer. Math. Soc., 73, pp. 747–817.

F. TAKENS [1974], *Forced oscillations and bifurcations*, in Applications of Global Analysis, Communications of Maths. Institute, Rijksuniversiteit, Utrecht, pp. 1–59.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics

0036-1410/84/1501-0004 $01.25/0

ON THE GLOBAL CONVERGENCE OF THE TODA LATTICE FOR REAL NORMAL MATRICES AND ITS APPLICATIONS TO THE EIGENVALUE PROBLEM*

MOODY T. CHU†

Abstract. The asymptotic behavior of the Toda lattice, when acting on real normal matrices, is studied. It is shown that the solution flow eventually converges to a diagonal block form where for a real eigenvalue the associated block is of size 1×1 with that eigenvalue as its element and for complex-conjugate pairs of eigenvalues the associated block is of size 2×2 with the real part as its diagonal elements and the (negative) imaginary part as its off-diagonal elements. This result generalizes the well-known asymptotic behavior of Jacobi matrices and is consistent with that from the QR-algorithm.

1. Introduction. Recently the dynamic flow of a special system of differential equations, known as the Toda lattice, has been found to be closely related to the important QR-algorithm [1], [2], [4], [7]. Roughly speaking, the QR-algorithm can be shown to be the time-1 mapping of the solution to the Toda lattice. Specifically, if we consider the following dynamic system for matrices in $\mathbb{R}^{n\times n}$:

$$\dot{X}=[X,\Pi_0X]=X\cdot\Pi_0X-\Pi_0X\cdot X \tag{1.1}$$

where $\Pi_0X=X^- -X^{-T}$ and X^- is the strictly lower triangular part of X, then the following properties concerning the solution flow $X(t)$ with initial data X_0 at $t=0$ can be derived from the general results presented in the previous paper [1].

LEMMA 1.1. *The solution $X(t)$ is given by*

$$X(t)=Q^*(t)X_0Q(t), \tag{1.2}$$

where $Q(t)$ solves the initial value problem

$$\dot{Q}=Q\cdot\Pi_0X,\qquad Q(0)=I. \tag{1.3}$$

Indeed $Q(t)$ is exactly the unitary matrix involved in the QR-decomposition [3], [6] of the matrix e^{tX_0}, namely

$$e^{tX_0}=Q(t)R(t) \tag{1.4}$$

where $R(t)$ is an upper triangular matrix with real nonnegative diagonal elements.

LEMMA 1.2. *For $k=0,\pm 1,\pm 2,\cdots$, suppose the matrix $e^{X(k)}$ has the QR-decomposition*

$$e^{X(k)}=Q^{(k)}R^{(k)}. \tag{1.5}$$

Then

$$e^{X(k+1)}=R^{(k)}Q^{(k)}. \tag{1.6}$$

Observe that, by (1.2), the trajectory $X(t)$ is bounded in $\mathbb{R}^{n\times n}$, so its ω-limit set is nonempty, compact and connected. We are interested in finding this set. A special case, when X_0 is a Jacobi matrix (and hence when X_0 is a real symmetric matrix by a standard tridiagonalization algorithm), has been studied extensively by a number of authors [2], [4], [7]. In fact, based on the continuous dependence of the initial data for

*Received by the editors December 7, 1982.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27650.

the system (1.1) and a well-known theorem [5], [6] in the numerical analysis concerning the convergence of the QR-algorithm, we have the following generalization [1].

THEOREM 1. *If the matrix* $X_0 \in \mathbb{R}^{n\times n}$ *has real distinct eigenvalues* $\{\lambda_1 > \lambda_2 > \cdots > \lambda_n\}$, *then the Toda flow* $X(t)$ *converges to an upper triangular matrix with the eigenvalues appearing on the diagonal* ı∵ *the descending order.*

In this paper we want to study the behavior of this flow when complex-conjugate pairs of eigenvalues occur. As is shown in [1], for an arbitrary (nonnormal) 2×2 matrix, the appearance of such a pair of eigenvalues will result in a periodic (in fact, a circular) portrait in the phase plane and thus $X(t)$ has no convergence at all. It is natural, therefore, to restrict ourselves in the study of the normal matrices first.

We begin in the next section with some preliminary facts. Especially, we point out the differential system which governs the dynamics of the corresponding eigenvectors of the flow $X(t)$. It turns out this system is much easier to handle than the system (1.1) itself. In §3 we discuss how eigenvalues affect eigenvectors and, hence, the entire flow $X(t)$ by the inverse algorithm. Although we only analyze two situations there, they seem to be generic enough to get general conclusions.

2. Preliminary facts. It is obvious, from Lemma 1.1, that normality is preserved along the flow provided that X_0 is a normal matrix. It is also known that there exists a unitary matrix U_0 such that

$$X_0 = U_0^* T U_0, \tag{2.1}$$

where T is a diagonal matrix with eigenvalues as its elements. Without loss of generality we shall assume these elements are arranged in such a way that

$$\operatorname{Re}\lambda_1 \geq \operatorname{Re}\lambda_2 \geq \cdots \geq \operatorname{Re}\lambda_n, \tag{2.2}$$

and that whenever there are complex-conjugate pairs, they are adjacent to each other. By (1.2), it follows that

$$X(t) = U^*(t) T U(t) \tag{2.3}$$

where

$$U(t) = U_0 Q(t). \tag{2.4}$$

Notice that, by (1.3), $U(t)$ satisfies the differential system

$$\dot{U} = U \cdot \Pi_0 X. \tag{2.5}$$

We shall assume X_0 is an upper Hessenberg matrix. Then the following lemma [1] guarantees the preservation of this structure along the entire flow. Recall that this useful property is also enjoyed by the classical QR-algorithm.

LEMMA 2.1. *If X is an upper Hessenberg matrix, so is* $\dot{X} = [X, \Pi_0 X]$.

Let us denote the matrix $U(t)$ in (2.4) by $U(t) = [u_1(t), \cdots, u_n(t)]$ where $u_i(t)$ is the ith column of $U(t)$. Then by (2.3) we have

$$[u_1, \cdots, u_n] \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \ddots & \\ & x_{32} & \ddots & \vdots \\ & & \ddots & \\ 0 & & x_{n,n-1} & x_{nn} \end{bmatrix} = T[u_1, \cdots, u_n]. \tag{2.6}$$

So the following equality holds for each $k=1,\cdots,n$.

$$\sum_{i=1}^{k+1} x_{ik}u_i = Tu_k, \tag{2.7}$$

where it is understood that $u_{n+1}=0$. Since all the vectors u_i are mutually orthogonal, we know that for all $1\leq i\leq n$ and $1\leq j\leq n$

$$x_{ij}=\langle u_i, Tu_j\rangle, \tag{2.8}$$

where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathbf{C}^n$.

From (2.5), (2.6) and (2.8), it is not hard to see now that

LEMMA 2.2. *For $i=1,\cdots,n$, the vector $u_i(t)$ satisfies the differential system*

$$\dot{u}_i = Tu_i - \sum_{j=1}^{i} \langle u_j, Tu_i\rangle u_j - \langle u_i, Tu_{i-1}\rangle u_{i-1}. \tag{2.9}$$

In particular, the first column $u_1(t)$ of $U(t)$ satisfies the equation

$$\dot{u}_1 = Tu_1 - \langle u_1, Tu_1\rangle u_1. \tag{2.10}$$

Direct substitution also shows that

LEMMA 2.3. *The solution to* (2.9) *is given explicitly by*

$$u_1(t)=\frac{e^{Tt}u_1(0)}{\|e^{Tt}u_1(0)\|_2}. \tag{2.11}$$

We note that the ith component $u_{i1}(t)$ of u_1 is given by

$$u_{il}(t)=\frac{e^{\lambda_i t}u_{i0}}{\left\{\sum_{j=1}^{n}\left|e^{\lambda_j t}u_{j0}\right|^2\right\}^{1/2}} \tag{2.12}$$

where u_{i0} is the complex conjugate of the first component of the ith eigenvector of X_0. The following useful inverse algorithm [5] turns out to be very important.

THEOREM 2.1. *Suppose B is an unreduced upper Hessenberg matrix with positive subdiagonal elements and Q is a unitary matrix, then Q and B are uniquely determined by the first column of Q, provided A is given and $B=Q^*AQ$.*

For our application, observe that the subdiagonal elements of $X(t)$ can never change signs along the positive orbit. If we assume, without loss, that X_0 not only is an upper Hessenberg matrix but also is unreduced to begin with, then from (2.6), (2.10) and the above theorem, we know that $X(t)$ and $U(t)$ are completely determined. The detailed analysis is presented in the next section.

3. Convergence of $X(t)$. First of all we should explain the meaning of convergence used in our context. Strictly speaking, convergence would be taken to mean the convergence of the flow $X(t)$ to some limit matrix. In our context, however, we mean convergence under deflations, i.e. we are concerned about the convergence of a submatrix obtained by deflation, as soon as the subdiagonal element is negligible, to another submatrix. The precise meaning will become clear later and indeed, as will be seen also, these two notions of convergence are essentially the same when the Toda lattice is acting on normal matrices.

For the simplicity of discussion, we shall make one more generic assumption, namely $u_{10} \neq 0$ whenever we need it and that X_0 is nonsingular. We shall also use the notation " → " to mean "converges to."

LEMMA 3.1. *If the eigenvalues in* (2.2) *are such that*

$$\operatorname{Re}\lambda_1 = \lambda_1 > \operatorname{Re}\lambda_2 \geq \cdots \geq \operatorname{Re}\lambda_n, \tag{3.1}$$

then

$$x_{11}(t) \to \lambda_1, \quad x_{21}(t) \to 0 \quad \textit{and} \quad x_{ik}(t) \to 0 \tag{3.2}$$

for every $2 \leq k \leq n$ *as* $t \to \infty$.

Proof. It is clear from (2.12) that as $t \to \infty$,

$$u_{11} \to \frac{u_{10}}{|u_{10}|} \quad \text{and} \quad u_{i1}(t) \to 0 \tag{3.3}$$

for all $i \geq 2$. Let us adopt the following notation in its intuitive sense:

$$\lim_{t \to \infty} u_i(t) = \hat{u}_i. \tag{3.4}$$

Then we have, from (2.8),

$$x_{11}(t) = \langle u_1, Tu_1 \rangle \to \langle \hat{u}_1, T\hat{u}_1 \rangle = \lambda_1 \tag{3.5}$$

and, from (2.7),

$$|x_{21}(t)| = \|Tu_1 - x_{11}u_1\|_2 \to \|T\hat{u}_1 - \lambda_1 \hat{u}_1\|_2 = 0. \tag{3.6}$$

Observe that, by (2.8) and (3.6),

$$x_{21}(t) = \langle u_2, Tu_1 \rangle \to \bar{u}_{12}\lambda_1 \hat{u}_{11} \to 0 \tag{3.7}$$

implies

$$u_{12}(t) \to 0 \tag{3.8}$$

where $^-$ means the complex conjugate. Therefore,

$$x_{12}(t) = \langle u_1, Tu_2 \rangle = \langle T^* u_1, u_2 \rangle \to \lambda_1 \bar{\hat{u}}_{11} u_{12} \to 0. \tag{3.9}$$

Indeed, for every $k > 2$, it is always true that

$$x_{k1}(t) = \langle u_k, Tu_1 \rangle \equiv 0 \to \bar{u}_{1k}\lambda_1 \hat{u}_{11} \tag{3.10}$$

implies

$$u_{1k}(t) \to 0. \tag{3.11}$$

Therefore,

$$x_{1k}(t) = \langle u_1, Tu_k \rangle = \langle T^* u_1, u_k \rangle \to \lambda_1 \bar{\hat{u}}_{11} u_{1k} \to 0. \tag{3.12}$$

In other words, if condition (3.1) is satisfied, then as $t \to \infty$

$$X(t) \to \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} \end{bmatrix}$$

where "x" represents either a nonzero element or an uncertain position.

Apparently when this convergence phenomenon happens, one is tempted to perform the deflation and to proceed the computation on the submatrix. We would like to point out, however, that those uncertain positions are really not entirely uncertain (they are uncertain simply because we don't care to include the analysis in Lemma 3.1). As a matter of fact, from (2.9), we know that for each $k\geq 2$, the eigenvector u_k is governed by

$$u_k = Tu_k = \sum_{i=1}^{k} \langle u_i, Tu_k \rangle u_i - \langle u_k, Tu_{k-1} \rangle u_{k-1}, \tag{3.13}$$

whereas, from (3.8), (3.9), (3.11) and (3.12), we see that the vector $\tilde{u}_k \in \mathbb{C}^{n-1}$, governed by

$$\dot{\tilde{u}}_k = \tilde{T}\tilde{u}_k - \sum_{i=2}^{k} \langle \tilde{u}_i, \tilde{T}\tilde{u}_k \rangle \tilde{u}_i - \langle \tilde{u}_k, \tilde{T}\tilde{u}_{k-1} \rangle \tilde{u}_{k-1},$$

where $\tilde{T}$ is obtained from T by deleting the first row and column, would describe the behavior of u_k as well when t is large enough. Therefore, those uncertain positions are actually converging according to either Lemma 3.1, with λ_1 being replaced by λ_2, or the next lemma, with λ_1 and λ_2 being replaced by λ_2 and λ_3. It is in this sense that we mean convergence.

LEMMA 3.2. *If the eigenvalues in* (2.2) *are such that*

$$\operatorname{Re}\lambda_1 = \operatorname{Re}\lambda_2 > \operatorname{Re}\lambda_3 \geq \cdots \geq \operatorname{Re}\lambda_n \tag{3.14}$$

and if $\lambda_1 = a + ib$ *with* $b \neq 0$, *then as* $t \to \infty$, *we have*

$$\begin{aligned} &x_{11}(t) \to a, \quad x_{22}(t) \to a, \quad x_{32}(t) \to 0, \\ &x_{21}(t) \to (\operatorname{sgn} x_{21}(0))|b|, \qquad x_{12}(t) \to -(\operatorname{sgn} x_{21}(0))|b|, \end{aligned} \tag{3.15}$$

and for all $k \geq 3$

$$x_{1k}(t) \to 0, \qquad x_{2k}(t) \to 0. \tag{3.16}$$

Proof. It is clear again from (2.12) that as $t \to \infty$,

$$u_{11}(t) \to \frac{e^{ibt}u_{10}}{\left\{|u_{10}|^2 + |u_{20}|^2\right\}^{1/2}}, \qquad u_{21}(t) \to \frac{e^{-ibt}u_{20}}{\left\{|u_{10}|^2 + |u_{20}|^2\right\}^{1/2}} \tag{3.17}$$

and for all $i \geq 3$,

$$u_{i1}(t) \to 0. \tag{3.18}$$

Notice that $u_{11}(t)$ and $u_{22}(t)$ do not converge at all. But we still use the notation (3.17) to indicate how they behave when t becomes large. Since X_0 is a real matrix, it must be that $u_{10} = \bar{u}_{20}$. Therefore

$$x_{11}(t) = \langle u_1, Tu_1 \rangle = \sum_{i=1}^{n} \lambda_1 |u_{i1}|^2 \to (a+ib)|\hat{u}_{11}|^2 + (a-ib)|\hat{u}_{21}|^2 = a. \tag{3.19}$$

Thus

$$|x_{21}(t)| = \|Tu_1 - x_{11}u_1\|_2 \to \|T\hat{u}_1 - a\hat{u}_1\| = b \tag{3.20}$$

implies that

$$x_{21}(t)\to\pm b \tag{3.21}$$

where the sign of this limit is the same as that of $x_{21}(0)$ since $x_{21}(t)$ can never change signs. Since $b\neq 0$, it follows, assuming $x_{21}(t)\to b$, from the fact

$$u_2=\frac{Tu_1-x_{11}u_1}{x_{21}} \tag{3.22}$$

that

$$u_{12}(t)\to iu_{11}(t),\quad u_{22}(t)\to iu_{21}(t),\quad u_{i2}(t)\to 0 \tag{3.23}$$

for all $i\geq 3$. So by (2.8), we know

$$x_{22}(t)=\langle u_2,Tu_2\rangle\to a \tag{3.24}$$

and

$$x_{12}(t)=\langle u_1,Tu_2\rangle\to -b. \tag{3.25}$$

By (2.7), simple calculation also shows

$$|x_{32}(t)|=\|Tu_2-x_{12}u_1-x_{22}u_2\|_2\to\|T\hat{u}_2+b\hat{u}_1-a\hat{u}_2\|=0. \tag{3.26}$$

We now claim for all $k\geq 3$, as $t\to\infty$

$$u_{1k}(t)\to 0,\qquad u_{2k}(t)\to 0. \tag{3.27}$$

Indeed this fact follows from solving the following system of equations

$$\langle u_k,Tu_1\rangle=0,\qquad\langle u_k,Tu_2\rangle=0, \tag{3.28}$$

or equivalently

$$\begin{aligned}&\bar{\hat{u}}_{1k}(a+ib)\hat{u}_{11}+\bar{\hat{u}}_{2k}(a-ib)\hat{u}_{21}=0,\\ &\bar{\hat{u}}_{1k}(a+ib)i\hat{u}_{11}-\bar{\hat{u}}_{2k}(a-ib)i\hat{u}_{21}=0.\end{aligned} \tag{3.29}$$

Therefore, for all $k\geq 3$,

$$x_{1k}(t)=\langle u_1,Tu_k\rangle=\langle T^*u_1,u_k\rangle\to\langle T^*\hat{u}_1,\hat{u}_k\rangle=0, \tag{3.30}$$

$$x_{2k}(t)=\langle u_2,Tu_k\rangle=\langle T^*u_2,u_k\rangle\to\langle T^*u_2,u_k\rangle=0. \tag{3.31}$$

In summary, this lemma states that if condition (3.14) holds, then

$$X(t)\to\begin{bmatrix}a & -b & 0 & 0 & 0\\ b & a & 0 & 0 & 0\\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x}\\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x}\\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x}\end{bmatrix}$$

where again "x" represents uncertain positions.

Finally we note that for the case $b=0$ (multiple eigenvalues), similar results (a 2×2 diagonal block) still can be obtained. Even for the nongeneric case when $\mathrm{Re}\lambda_1=\mathrm{Re}\lambda_2=\mathrm{Re}\lambda_3=\mathrm{Re}\lambda_4$, an argument analogous to Lemma 3.2 can still show the convergence. It is interesting to see the asymptotic behavior of the general flow [1]

$$\dot{X}=[X,\Pi_0(G(x))] \tag{3.32}$$

where $G(z)$ is an analytic function defined on an open set containing the spectrum on X_0. The analysis, nevertheless, is much harder than (1.1) since we don't have a system as nice as (2.9) and we are still working on it.

REFERENCES

[1] M. T. CHU, *The generalized Toda lattice, the QR-algorithm and the centre manifold theory*, SIAM J. Alg. Discrete Meth., 5 (1984), to appear.

[2] P. DEIFT, T. NANDA, AND C. TOMEI, *Differential equations for the symmetric eigenvalue problem*, SIAM J. Numer. Anal., 20 (1983), pp. 1–22.

[3] J. G. F. FRANCIS, *The QR transformation, a unitary analogue to the LR transformation*, Comput. J., 4 (1961), pp. 265–281.

[4] J. MOSER, *Finitely many mass points on the line under the influence of an exponential potential—an integrable system*, Dynamical Systems, Theory and Applications, J. Moser, ed., Lecture Notes in Physics 38, Springer-Verlag, Berlin, 1975, pp. 467–497.

[5] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[6] A. RALSTON AND P. RABINOWITZ, *A First Course in Numerical Analysis*, McGraw-Hill, New York, 1978.

[7] W. W. SYMES, *The QR-algorithm and scattering for the finite nonperiodic Toda lattice*, Physica, 40 (1982), pp. 275–280.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1501-0005 $01.25/0

ON THE WIDTH OF THE INSTABILITY INTERVALS OF THE MATHIEU EQUATION*

HARRY HOCHSTADT†

Abstract. It is shown that the widths of the instability intervals of the Mathieu equation are asymptotically given by

$$\frac{8h^{2m}}{4^m[(m-1)!]^2}\left[1+O\left(\frac{h^4}{m^2}\right)\right].$$

Recently the asymptotic widths of the instability intervals of the Mathieu equation have been determined by Avran and Simon [1] as well as Harrell [2]. The purpose of this note is to provide a rather simple method for the calculation of these widths. To do so, some formulas developed in the book by Meixner and Schaefke [3] will be used, and for convenience their notation will be used.

The Mathieu equation is

$$y''+(\lambda-2h^2\cos 2z)y=0. \tag{1}$$

The eigenvalues of the periodic spectrum fall into four classes, as follows:

$$\begin{array}{lll} \text{(I)} & y'(0)=y'(\pi/2)=0, & \{a_{2n}\}, \\ \text{(II)} & y'(0)=y(\pi/2)=0, & \{a_{2n+1}\}, \\ \text{(III)} & y(0)=y'(\pi/2)=0, & \{b_{2n+1}\}, \\ \text{(IV)} & y(0)=y(\pi/2)=0, & \{b_{2n}\}. \end{array}$$

These eigenvalues can be ordered as follows ([3], p. 119): $a_0<b_1<a_1<b_2<a_2<b_3<a_3<\cdots$ provided $h^2>0$. One can easily show that for large n, $a_{2n}\cong(2n)^2$, $b_{2n}\cong(2n)^2$, $a_{2n+1}\cong(2n+1)^2$, $b_{2n+1}\cong(2n+1)^2$. The widths of the instability intervals are given by b_k-a_k, and we shall demonstrate that for large k we have

$$b_k-a_k\cong\frac{8h^{2k}}{4^k[(k-1)!]^2}. \tag{2}$$

To derive (2) we shall make use of the continued fractions which the eigenvalues satisfy. Corresponding to the four cases we have [3, p. 118]

$$\text{(I)}\qquad \lambda-(2n)^2-\frac{h^4}{\lambda-(2n-2)^2-}\cdots\frac{h^4}{-\lambda-4-}\frac{2h^4}{\lambda}$$
$$=-\frac{h^4}{(2n+2)^2-\lambda-}\frac{h^4}{(2n+4)^2-\lambda-\cdots},$$

*Received by the editors April 19, 1982. This work was supported by the National Science Foundation under grant MCS-8103373.

†Polytechnic Institute of New York, Brooklyn, New York 11201.

$$\text{(II)}\qquad \lambda-(2n+1)^2-\frac{h^4}{\lambda-(2n-1)^2-}\cdots-\frac{h^4}{\lambda-1-h^2}$$
$$=\frac{-h^4}{(2n+3)^2-\lambda-}\,\frac{h^4}{(2n+5)^2-\lambda-\cdots},$$

$$\text{(III)}\qquad \lambda-(2n+1)^2-\frac{h^4}{\lambda-(2n-1)^2-}\cdots-\frac{h^4}{\lambda-1+h^2}$$
$$=\frac{-h^4}{(2n+3)^2-\lambda-}\,\frac{h^4}{(2n+5)^2-\lambda-\cdots},$$

$$\text{(IV)}\qquad \lambda-(2n)^2-\frac{h^4}{\lambda-(2n-2)^2-}\cdots-\frac{h^4}{\lambda-4}$$
$$=-\frac{h^4}{(2n+2)^2-\lambda-}\,\frac{h^4}{(2n+4)^2-\lambda-\cdots}.$$

(I)–(IV) are transcendental equations, whose solutions are a_{2n}, a_{2n+1}, b_{2n+1}, b_{2n}, respectively.

To estimate the above continued fractions we recall the following facts. Suppose we consider the following continued fraction:

$$\text{(3)}\qquad \frac{A_1}{B_1+}\,\frac{A_2}{B_2+}\,\frac{A_3}{B_3+}\cdots$$

and let p_k/q_k denote the kth convergent, where

$$\text{(4)}\qquad p_1=A_1,\quad p_2=A_1B_2,\quad q_1=B_1,\quad q_2=B_1B_2+A_2.$$

Then p_k and q_k satisfy the following recurrence formulas,

$$\text{(5)}\qquad \begin{aligned} p_{n+1}&=B_{n+1}p_n+A_{n+1}p_{n-1},\\ q_{n+1}&=B_{n+1}q_n+A_{n+1}q_{n-1}, \end{aligned}$$

with the initial condition given in (4). From (5) one can easily deduce that

$$\text{(6)}\qquad \frac{p_k}{q_k}-\frac{p_{k-1}}{p_{k-1}}=\frac{(-1)^{k+1}\Pi_1^k A_i}{q_kq_{k-1}}.$$

An explicit solution for the q_n can be found, as follows. Let

$$\text{(7)}\qquad q_n=B_1B_2\cdots B_nU_n.$$

Then

$$\text{(8)}\qquad \begin{pmatrix} U_{n-1}\\ U_n \end{pmatrix}=\prod_{k=1}^{n-1}\left[\begin{pmatrix} 0 & 1\\ 0 & 1 \end{pmatrix}+\frac{A_{k+1}}{B_kB_{k+1}}\begin{pmatrix} 0 & 0\\ 1 & 0 \end{pmatrix}\right]\begin{pmatrix} 1\\ 1 \end{pmatrix},\qquad n\geq 1.$$

To calculate b_{2r} we return to the continued fraction (IV) and let

$$\text{(9)}\qquad \left.\begin{aligned} A_k&=-h^4\\ B_k&=\lambda-(2(r-k))^2 \end{aligned}\right\}k=1,2,\cdots r-1,\qquad \left.\begin{aligned} \tilde{A}_k&=-h^4\\ \tilde{B}_k&=(2(r+k))^2-\lambda \end{aligned}\right\}k=1,2,\cdots.$$

Then we can rewrite the continued fraction (IV) in the form

$$(10)\qquad \lambda-(2r)^2+\frac{p_{r-1}}{q_{r-1}}=\frac{\tilde{p}_k}{\tilde{q}_k}+\sum_{m=0}^{\infty}\left(\frac{\tilde{p}_{k+m+1}}{\tilde{q}_{k+m+1}}-\frac{\tilde{p}_{k+m}}{\tilde{q}_{k+m}}\right).$$

To estimate the sum in (10) we note that $\lambda\cong(2r)^2$ and $\tilde{B}_k\cong 4k(k+r)$. Use of (6), (7), (8) shows that

(11)

$$\frac{\tilde{p}_{k+m+1}}{\tilde{q}_{k+m+1}}-\frac{\tilde{p}_{k+m}}{\tilde{q}_{k+m}}=\frac{-h^{4(k+m+1)}}{\tilde{q}_{k+m+1}\tilde{q}_{k+m}}\cong\frac{-h^{4(k+m+1)}}{\left[\dfrac{4^{k+m}(k+m)!(k+m+2r)!}{r!}\right]^2[4(k+m+1+2r)]}.$$

Equation (10) can therefore be estimated by

$$(12)\qquad \lambda-(2r)^2+\frac{p_{r-1}}{q_{r-1}}=\frac{\tilde{p}_r}{\tilde{q}_r}+O\left(\frac{h^{4(r+1)}}{r^{6r}}\right).$$

Similarly a_{2r} can be estimated from (I), with the A_k, B_k, $\tilde{A}_k$, $\tilde{B}_k$ as defined in (9) and also

$$(13)\qquad A_r=-2h^4,\qquad B_r=\lambda.$$

Then we find

$$(14)\qquad \lambda-(2r)^2+\frac{p_{r-1}}{q_{r-1}}-\frac{2h^{4r}}{q_rq_{r-1}}=\frac{\tilde{p}_r}{\tilde{q}_r}+O\left(\frac{h^{4(r+1)}}{r^{6r}}\right).$$

b_{2r} is a solution of (12) and a_{2r} a solution of (14). A comparison shows that the asymptotic developments of a_{2r} and b_{2r} will agree up to terms of $O(h^{4r}/q_rq_{r-1})$. Then it follows that, with

$$B_k=\lambda-(2(r-k))^2\cong 4k(2r-k),$$

$$(15)\qquad b_{2r}-a_{2r}\cong\frac{2h^{4r}}{q_rq_{r-1}}\cong\frac{2h^{4r}}{4^{2r-1}[(2r-1)!]^2}\left[1+O\left(\frac{h^4}{r^2}\right)\right].$$

A similar analysis using the continued fractions for a_{2r+1} and b_{2r+1} finally shows that

$$(16)\qquad b_m-a_m=\frac{8h^{2m}}{4^m[(m-1)]^2}\left[1+O\left(\frac{h^4}{m^2}\right)\right].$$

REFERENCES

[1] JOSEPH AVRAN AND BARRY SIMON, *The asymptotics of the gap in the Mathieu equation*, Ann. Phys., 134 (1981), pp. 76–84.

[2] M. EVANS HARRELL II, *On the effect of the boundary conditions on the eigenvalues of ordinary differential equations*, Amer. J. Math. supplement (1981), pp. 139–150.

[3] J. MEIXNER AND F. W. SCHAEFKE, *Mathieuische Funktionen und Sphäroidfunktionen*, Springer, Berlin-Göttingen-Heidelberg, 1954.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics

0036-1410/84/1501-0006 $01.25/0

FACTORED PRODUCT EXPANSIONS OF SOLUTIONS OF NONLINEAR DIFFERENTIAL EQUATIONS*

STANLY STEINBERG†

Abstract. Lie transformations are used to represent solutions of initial value problems for systems of nonlinear ordinary differential equations as exponentials of first order linear partial differential operators. These exponentials are then expanded using an analog of the usual exponential identities. This expansion is called the *factored product expansion*. Such expansions have been found useful in the study of magnetic and optical lenses.

1. Introduction. In this paper we will discuss *factored product expansions* of solutions of initial value problems for systems of nonlinear ordinary differential equations, that is, for problems of the form

$$\frac{d}{dt}y(t)=f(y(t)), \qquad y(0)=y.$$

Here

$$\begin{aligned}&y=(y_1,\cdots,y_n),\\&y(t)=(y_1(t),y_2(t),\cdots,y_n(t)),\\&f(y)=(f_1(y),\cdots,f_n(y)),\end{aligned}$$

where n is a positive integer, t and y_j are real parameters and $y_j(t)$ and $f_j(y)$ are real analytic functions with

$$f(0)=0.$$

We associate with the initial value problem the first order linear partial differential operator

$$L=\sum_{j=1}^{n} f_j(y)\frac{\partial}{\partial y_j}$$

and then use Lie transformations (which are sometimes called Lie series) to write the solution of the initial value problem in the form

$$y(t)=e^{tL}y.$$

Because $f(y)$ is analytic we can use a power series expansion to write

$$L=\sum_{k=1}^{\infty} L_k$$

where each L_k is a linear first order partial differential operator with homogeneous polynomial coefficients, and where the degree of homogeneity of the coefficients of L_k is k.

*Received by the editors April 16, 1982, and in revised form September 23, 1982. This work was partially supported by the National Science Foundation under grant MCS-8102683.

†Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131.

A natural generalization of the exponential identities to this situation gives an infinite product expansion

$$e^{t(L_1+L_2+L_3+\cdots)}=e^{F_1}e^{F_2}e^{F_3}\ldots$$

where, again, F_k is a linear first order partial differential operator with degree k homogeneous polynomial coefficients. It is this last formula that we call the *factored product expansion.* This formula is closely related to the noncommuting exponential identities frequently called Baker–Campbell–Hausdorff formulas.

The Baker–Campbell–Hausdorff formulas have a long history and have important applications in the study of Lie groups [7]. The applications we have in mind are quite different from Lie group applications and, to our knowledge, first appeared in the papers of Dragt et al. [1]–[5]. These papers use factored product expansions to study magnetic and optical lens systems in a Hamiltonian mechanics setting.

Our main result, Theorem 1 of §4, gives a procedure for calculating all of the exponents in a factored product expansion. This result is a generalization of Dragt and Finn [1, Thm. 2 and Lemma 6] to a non-Hamiltonian setting. More importantly, our method of proof is different from that of Dragt and Finn. This new method of proof allows us to obtain some new results on the degree of approximation given by truncated factored product expansions that should be useful in applications. We also note that our results are easily specialized to the Hamiltonian setting.

This paper is organized as follows. In §2 we give a brief summary of the properties of Lie transformations in a non-Hamiltonian setting. In this paper we use the name *Lie transformation* to describe the transformation generated by the exponential of a first order linear partial differential operator and the name *Lie series* to denote the power series definition of the Lie transformation. Other works sometimes use this terminology differently. We do not include any proofs because, with the exception of the noncommuting exponential identities, the proofs are elementary and can be found in the literature [1]–[6], [8], [9]. The results on noncommuting exponential identities follow from our results in §3. An expository account, with proofs, of these results on Lie transformations and series along with extensive references to literature on the theory and applications of Lie transformations can be found in Steinberg [8].

In §3 we give some preliminary results and in §4 we give the main result on factored product expansions along with some new corollaries on the degree of approximation given by truncated expansions.

2. Lie transformations. Here, for the convenience of the reader, we state the basic properties of Lie transformations. As we noted in the introduction, other accounts can be found in the literature. A Lie transformation is an exponential of a first order linear analytic partial differential operator in n variables.

$$L=\sum_{j=0}^{n} f_j(y)D_j,\qquad D_j=\frac{\partial}{\partial y_j},$$

where $f(y)$ is an analytic function near $y=0$. A Lie transformation is then given by

$$e^{tL}=\sum_{n=0}^{\infty}\frac{t^nL^n}{n!}.$$

The right-hand side of the last equation is called a *Lie series*. The action of the Lie transformation on a function $g(y)$, analytic near $y=0$, is given by

$$e^{tL}g(y)=\sum_{n=0}^{\infty}\frac{t^nL^n}{n!}g(y)=\sum_{n=0}^{\infty}\frac{t^n}{n!}(f(y)\cdot D)^n g(y).$$

Properties. We assume that $f(y)$, $g(y)$ and $h(y)$ are analytic functions near $y=0$, that a and b are real constants, that $c(t)$ is a smooth real valued function and that L is as above.

1) *Convergence.*

$$e^{tL}g(y)$$

is a well defined analytic function of y and t for y and t small enough.

2) *Time derivative.*

$$\frac{d}{dt}e^{c(t)L}=c'(t)Le^{c(t)L}=c^{c(t)L}c'(t)L.$$

3) *Linearity.*

$$e^{tL}(ag+bh)=ae^{tL}g+be^{tL}h.$$

4) *Product preservation.*

$$e^{tL}(gh)=(e^{tL}g)(e^{tL}h).$$

5) *Composition.*

$$e^{tL}g(y)=g(e^{tL}y).$$

We now suppose that P is another first order differential operator and define successive commutators by

$$\begin{aligned}&[L,\cdot\,]^0P=P,\\&[L,\cdot\,]^1P=LP-PL,\\&[L,\cdot\,]^nP=[L,\cdot\,]^{n-1}[L,P],\qquad n\geq 1.\end{aligned}$$

6) *Similarity.*

$$e^{tL}Pe^{-tL}=e^{t[L,\cdot\,]}P=\sum_{n=0}^{\infty}\frac{t^n}{n!}[L,\cdot\,]^nP.$$

7) *Function multiplier.*

$$e^{tL}ge^{-tL}h=(e^{tL}g)h.$$

8) *Noncommuting exponential identities.*

$$\begin{aligned}e^{t(L+P)}&=e^{tL}e^{tP}e^{t^2L^{(2)}}e^{t^3L^{(3)}}e^{t^4L^{(4)}}e^{t^5L^{(5)}}\cdots\\&=\cdots e^{t^5L^{(5)}}e^{-t^4L^{(4)}}e^{t^3L^{(3)}}e^{-t^2L^{(2)}}e^{tP}e^{tL},\end{aligned}$$

$$e^{tL}e^{tP}=e^{tL+tP+t^2W^{(2)}+t^3W^{(3)}}+\cdots,$$

where

$$L^{(2)}=-\tfrac{1}{2}[L,P], \qquad W^{(2)}=\tfrac{1}{2}[L,P],$$

and so forth. Here each $L^{(k)}$ and $W^{(k)}$ are k-fold commutators of L and P.

9) *Differential equation property.* If

$$y(t)=e^{tL}y,$$

then

$$y'(t)=f(y(t)), \qquad y(0)=y.$$

3. Preliminary results. In this section we will derive a formula for the derivative of an exponential of a time-dependent operator. Both this formula and the applications of this formula use operators that are defined as analytic functions of first order differential operators or as analytic functions of commutator operators. We define these operators using infinite series. Because we use only formal series properties of these expressions, we will not worry about the convergence of the series. We note that the formula we derive in this section can be used to derive the noncommuting exponential identities of the previous section.

Let

$$A(z)=\sum_{k=0}^{\infty} a_k z^k$$

be analytic near $z=0$ and assume that L is an operator. Then

$$A(L)=\sum_{k=0}^{\infty} a_k L^k.$$

As we said before, we consider the series as a formal expression. However, if we know all of the eigenvectors and eigenvalues of L, then this information can be used to calculate $A(L)$ in terms of A applied to the eigenvalues of L. If the coefficients a_k decrease like $\frac{1}{k!}$, then it is possible to show that $A(L)$ is well defined for first order differential operators of the type we are discussing. In addition, if L is a bounded operator and the power series of $A(z)$ has a finite radius of convergence, then it can be shown that $A(tL)$ is well defined for sufficiently small scaler t.

We have already met one example:

$$A(z)=e^{z}.$$

In the next proposition we will use

$$A(z)=\frac{e^{z}-1}{z}.$$

In the next section we will use

$$A(z)=\frac{z}{e^{z}-1}.$$

It is easy to see that the last $A(z)$ is singular at $i2\pi$, and consequently the power series of this function has a radius of convergence equal to 2π. Thus it seems unlikely that the series for $A(L)$ will converge for operators of the type we are considering. We also note that we will apply our formulas when the commutator operator $[L, \cdot]$ is used in place of the operator L, and thus compute $A([L, \cdot])$.

The next well-known result is key to our computations.

PROPOSITION 1. *If $L(t)$ is a linear first order differential operator with coefficients that are analytic in both t and y, then*

$$\frac{d}{dt}e^{L(t)}=\frac{e^{[L(t),\cdot]}-1}{[L(t),\cdot]}L'(t)e^{L(t)}=e^{L(t)}\frac{1-e^{-[L(t),\cdot]}}{[L(t),\cdot]}L'(t).$$

Proof. Set $L=L(t)$, $L'=dL(t)/dt$ and then compute:

$$\frac{d}{dt}e^{L}=\frac{d}{dt}\sum_{k=0}^{\infty}\frac{L^k}{k!}=\sum_{k=1}^{\infty}\frac{1}{k!}\sum_{m=0}^{k-1}L^mL'L^{k-m-1}$$
$$=\sum_{m=0}^{\infty}\sum_{k=m+1}^{\infty}\frac{1}{k!}L^mL'L^{k-m-1}$$
$$=\sum_{m=0}^{\infty}\sum_{j=0}^{\infty}\frac{m!j!}{(m+j+1)!}\frac{L^m}{m!}L'\frac{L^j}{j!}.$$

However,

$$\frac{m!j!}{(m+j+1)!}=\int_0^1\tau^m(1-\tau)^j d\tau,$$

so

$$\frac{d}{dt}e^{L}=\int_0^1 e^{\tau L}L'e^{(1-\tau)L}d\tau=\int_0^1 e^{\tau[L,\cdot]}L'\,d\tau e^{L}=\frac{e^{[L,\cdot]}-1}{[L,\cdot]}L'e^{L}.$$

Replacing τ by $1-\tau$ in the above integrals gives the second form of the result.

4. Main results. Our main result is the following:

THEOREM 1. *If*

$$L=\sum_{k=1}^{\infty}L_k,$$

where L_k is a first order differential operator with homogeneous polynomial coefficients of degree k, then

$$e^{tL}=e^{F_1(t)}e^{F_2(t)}e^{F_3(t)}\cdots$$

where $F_k(t)$ is a first order differential operator with homogeneous polynomial coefficients of degree k. The equality is meant in the sense of formal series.

Proof. Before we begin the proof we note that if L_j and L_k are first order partial differential operators with homogeneous polynomial coefficients of degree given by their subscripts, then

$$[L_j, L_k]=L_{j+k-1}.$$

Here we mean that L_{j+k-1} is another first order partial differential operator with homogeneous polynomial coefficients of degree given by the subscript $j+k-1$. Note that if $j=1$ then commutation by L_j does not increase the degree of homogeneity of the coefficients of L_k. This relationship between the commutator and the degree of homogeneity of the coefficients is a central aspect in our calculations.

Differentiate the proposed identity with respects to t:

$$\begin{aligned} Le^{tL}=&\frac{e^{[F_1,\cdot]}-1}{[F_1,\cdot]}F_1'e^{F_1}e^{F_2}e^{F_3}\dots\\ &+e^{F_1}\frac{e^{[F_2,\cdot]}-1}{[F_2,\cdot]}F_2'e^{F_2}e^{F_3}\dots\\ &+e^{F_1}e^{F_2}\frac{e^{[F_3,\cdot]}-1}{[F_3,\cdot]}F_3'e^{F_3}\dots\\ &+\cdots. \end{aligned}$$

The inverse of the proposed identity is given by

$$e^{-tL}=\cdots e^{-F_3}e^{-F_2}e^{-F_1}.$$

Multiply the derivative of the proposed identity on the right by this to obtain

$$\begin{aligned} L=&\frac{e^{[F_1,\cdot]}-1}{[F_1,\cdot]}F_1'+e^{F_1}\frac{e^{[F_2,\cdot]}-1}{[F_2,\cdot]}F_2'e^{-F_1}\\ &+e^{F_1}e^{F_2}\frac{e^{[F_3,\cdot]}-1}{[F_3,\cdot]}F_3'e^{-F_2}e^{-F_3}\\ &+\cdots\\ =&\frac{e^{[F_1,\cdot]}-1}{[F_1,\cdot]}F_1'+e^{[F_1,\cdot]}\frac{e^{[F_2,\cdot]}-1}{[F_2,\cdot]}F_2'\\ &+e^{[F_1,\cdot]}e^{[F_2,\cdot]}\frac{e^{[F_3,\cdot]}-1}{[F_3,\cdot]}F_3'\\ &+\cdots. \end{aligned}$$

To make the coefficient of the linear expressions in y zero, we need

$$F_1'(t)=\frac{[F_1(t),\cdot]}{e^{[F_1(t),\cdot]}-1}L_1.$$

An obvious solution for this equation is

$$F_1(t)=tL_1.$$

We now proceed to set the coefficient of higher powers of y equal to zero. Set $R^{(1)}=L_1$ and then for $k\geq 2$ define

$$\begin{aligned} R^{(k)}&=e^{-[F_{k-1},\cdot]}\left(R^{(k-1)}-\frac{[F_{k-1},\cdot]}{e^{[F_{k-1},\cdot]}-1}F_{k-1}'\right),\\ F_k'&=R_k^{(k)}. \end{aligned}$$

Here $R_j^{(k)}$ means the terms of degree j in $R^{(k)}$, and F_k is determined by integration. Some care must be used in evaluating the formula for R_k. When $k=1$, the formulas involve $[L_1,\cdot\,]$. In this case we note that this commutator is a linear mapping on the space of first order linear partial differential operators with degree k homogeneous coefficients, and that this space is finite dimensional. Consequently functions of $[L_1,\cdot\,]$ can be computed using spectral theory on finite dimensional spaces. When $k>1$, the powers series definitions of e^z and $z/(e^z-1)$ are used to compute functions of $[L_k,\cdot\,]$.

Things are set up so that

$$R^{(k)}=\frac{e^{[F_k,\cdot]}-1}{[F_k,\cdot\,]}F_k'+e^{[F_k,\cdot]}\frac{e^{[F_{k+1},\cdot]}-1}{[F_{k-1},\cdot\,]}F_{k+1}'+\cdots.$$

and then our choice for F_k gives

$$R_j^{(k)}=0,\qquad 1\leq j<k,$$

which completes the proof.

Remark. If in Theorem 1 we replace t by $-t$ and take the inverse of the resulting identity, then we obtain

$$e^{tL}=\cdots e^{-F_3(-t)}e^{-F_2(-t)}e^{-F_1(-t)}.$$

We note that, in general, it will not be possible to compute the expressions $\exp(F_k)$ or $\exp([F_k,\cdot\,])$ in closed form or to do the necessary integrals in closed form. In such a case F_k may be determined using truncated power series in t, say through terms of order m. The only thing that changes in the above proof is that all equations then mean that the first $m+1$ terms of the power series in t agree.

The next result estimates the error made when factored product expansions are used to approximate solutions of ordinary differential equations.

COROLLARY 1. *If*

$$y(t)=e^{tL}y$$

and

$$y_k(t)=e^{F_1(t)}e^{F_2(t)}\cdots e^{F_k(t)}y,$$

then

$$|y(t)-y_k(t)|\leq C_k|t||y|^{k+1}$$

for some constant C_k.

Proof. Because Lie transformations yield well defined analytic functions, both $y(t)$ and $y_k(t)$ are analytic functions of t and y. Thus we need to show that all terms in the power series of $y(t)-y_k(t)$ of order lower than given in the estimate are zero. This is exactly what the previous theorem does.

In another paper [9], we have given examples in one space dimension that show the above estimates are the best possible. We note that in this elementary setting it is possible to compute all of the needed expressions in closed form. As we stated above, it is not possible, in general, to compute $\exp(L_k)$ or $\exp([L_k,\cdot\,])$ in closed form. In view of this, the next result is useful in applications.

COROLLARY 2. *Given $m>0$, it is possible, using only power series techniques, to determine $y_k(t)$ so that*

$$|y(t)-y_k(t)|\leq K_m|y||t|^{m+1}+C_k|t||y|^{k+1}$$

for some constants K_m and C_k.

Proof. In the previous theorem use power series techniques to determine F_j, $1\leq j\leq k$ through terms of order t^m and then use the power series to determine $\exp(F_j)$ through terms of order t^m and then write

$$y_k(t)=e^{F_1}e^{F_2}\cdots e^{F_k}.$$

5. Comments. It is our belief that the factored product expansions are generically divergent. Another way to say this is that the constants C_k in Corollary 1 will, in general, grow as k becomes large. The growth of the constants C_k was confirmed by some of the numerical experiments done for [9]. As stated before, the error estimate given in Corollary 1 is best possible for the class of equations being considered [9]. However, in some special circumstances the errors may oscillate, giving better results than our estimates indicate.

REFERENCES

[1] ALEX J. DRAGT AND JOHN M. FINN, *Lie series and invariant functions for analytic symplectic maps*, J. Math. Phys., 17 (1976), pp. 2215–2227.

[2] ALEX J. DRAGT, *A method of transfer maps for linear and nonlinear beam elements*, IEEE Trans. Nuclear Sci., NS-26 (1979), pp. 3601–3603.

[3] ALEX J. DRAGT AND JOHN M. FINN, *Normal form for mirror machine Hamiltonians*, J. Math. Phys., 20 (1979), pp. 2649–2660.

[4] ALEX J. DRAGT AND D. R. DOUGLAS, *Charged particle beam transport using Lie algebraic methods*, IEEE Trans. Nuclear Sci., NS-28 (1981), pp. 2522–2524.

[5] ALEX J. DRAGT, *A Lie algebraic theory of geometrical optics and optical aberrations*, J. Opt. Soc. Amer., 72 (1982), pp. 372–379.

[6] W. GROBNER AND H. KNOPP, *Contributions to the Method of Lie Series*, Bibliographisches Institut. Mannheim, 1967.

[7] WILFRIED SHMID, *Poincaré and Lie groups*, Bull. Amer. Math. Soc., 6 (1982), pp. 175–186.

[8] STANLY STEINBERG, *Lie series and their applications*, in preparation.

[9] ______, *Lie series and nonlinear differential equations*, J. Math. Anal. Appl., to appear.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1501-0007 $01.25/0

ASYMPTOTIC INTEGRATION OF PERTURBED LINEAR DIFFERENTIAL EQUATIONS UNDER CONDITIONS INVOLVING ORDINARY INTEGRAL CONVERGENCE*

JAROMIR ŠIMŠA

Abstract. In this paper, we consider asymptotic integration of nth order linear differential equations with constant coefficients, modified by the addition of small functions. The integral smallness of the perturbation functions is expressed in terms of ordinary convergence instead of the classic conditions which require absolute integrability. The proof of our result is based on the Banach contraction principle.

1. Introduction. Statement of the result. We consider the scalar linear differential equation

$$(1.1)\quad x^{(n)}+[a_1+p_1(t)]x^{(n-1)}+\cdots+[a_{n-1}+p_{n-1}(t)]x'+[a_n+p_n(t)]x=0,$$

where a_k are complex numbers and $p_k(t)$ are continuous complex-valued functions defined on the half-line $0\leq t<\infty$.

A classical asymptotic theorem (see Hartman [2, Thm. 17.2]) gives asymptotic estimates for a fundamental system of solutions of (1.1): if the functions $|p_k(t)|t^q$ are integrable on $[0,\infty)$ for some $q\geq 0$ and if the real parts of roots λ_j of the equation

$$(1.2)\qquad \lambda^n+a_1\lambda^{n-1}+\cdots+a_{n-1}\lambda+a_n=0$$

are distinct, then there exist n solutions $x_j(t)$ of (1.1) such that

$$(1.3)\qquad x_j^{(k)}(t)=\left(\lambda_j^k+o(t^{-q})\right)\exp(\lambda_j t),\qquad 0\leq k\leq n-1\quad \text{as } t\to\infty.$$

In this paper we shall obtain a similar result under the weaker assumptions that the integrals

$$(1.4)\qquad \int^\infty p_k(t)t^q\,dt$$

converge (possibly not absolutely) and that the roots of (1.2) are distinct. The case $\lambda_1=\lambda_2=\cdots=\lambda_n=0$ has been discussed by Trench [3].

Furthermore, instead of (1.4) we shall consider more general integrals

$$\int^\infty p_k(t)\,e^{\rho t}t^q\,dt$$

with nonnegative constants ρ and q.

We now state our result.

THEOREM 1. *Suppose* (1.2) *has distinct roots* $\lambda_1,\lambda_2,\cdots,\lambda_n$. *Let the complex-valued functions* $p_k(t)$ *be continuous on* $[0,\infty)$ *and satisfy the following conditions*:

(i) $\int^\infty|p_1(t)|\,dt<\infty$;

(ii) *the integrals*

$$\int^\infty p_k(t)\exp\{(\rho+i\beta)t\}t^q\,dt,\qquad 1\leq k\leq n,$$

where ρ *and* q *are nonnegative constants, converge* (*perhaps conditionally*) *for* $\beta=0$ *and also for all* $\beta=\beta_{jm}=\operatorname{Im}(\lambda_j-\lambda_m)$ *satisfying* $\operatorname{Re}(\lambda_j-\lambda_m)=\rho$;

* Received by the editors July 24, 1981, and in revised form July 7, 1982.

† Department of Mathematical Analysis, J. E. Purkyně University, Janáčkovo nám. 2a, 662 95 Brno, Czechoslovakia.

(iii) *if* $\rho=0$ *and* $0\le q<1$ *in* (ii), *then*

$$\int^{\infty} t^{-q}\left|\int_t^{\infty} p_k(s)s^q\,ds\right|dt<\infty,\qquad 2\le k\le n.$$

Then (1.1) *has n solutions* $x_1(t),\cdots,x_n(t)$ *such that*

$$x_j^{(k)}(t)=\left(\lambda_j^k+o(e^{-\rho t}t^{-q})\right)\exp(\lambda_j t),\qquad 0\le k\le n-1. \tag{1.5}$$

This theorem contains a result concerning perturbations of the nonoscillatory equation $x''-x=0$ (Trench [3]) as a special case.

2. Preparatory estimates. In the proof of Theorem 1 we shall use the following lemma.

LEMMA 1. *Let* $h(t)$ *be a complex-valued continuous function integrable (perhaps conditionally) on the half-line* $[t_0,\infty)$, *where* $t_0>0$. *Denote*

$$H(t)=\sup_{t_1\ge t}\left|\int_{t_1}^{\infty} h(s)\,ds\right|,\qquad t\ge t_0.$$

Then the function $h(t)t^{-q}$ *is integrable for all* $q\ge0$ *and*

$$\left|\int_t^{\infty} h(s)s^{-q}\,ds\right|\le 2H(t)t^{-q},\qquad t\ge t_0. \tag{2.1}$$

Further, let $K(t)$ *be a continuously differentiable complex-valued function satisfying*

$$|K(t)|\le K_0e^{\alpha t}t^{-q} \tag{2.2}$$

and

$$|K'(t)|\le K_1e^{\alpha t}t^{-q} \tag{2.3}$$

on $[t_0,\infty)$, *where* K_0, K_1, $\alpha\ne0$ *and* $q\ge0$ *are real constants.*

(i) *If* $\alpha<0$, *then the integral of* $K(t)h(t)$ *converges and*

$$\left|\int_t^{\infty} K(s)h(s)\,ds\right|\le\left(K_0+|\alpha|^{-1}K_1\right)H(t)e^{\alpha t}t^{-q},\qquad t_0\le t<\infty. \tag{2.4}$$

(ii) *If* $\alpha>0$, *then*

(2.5)
$$\left|\int_{t_0}^{t} K(s)h(s)\,ds\right|\le\Bigg[\left(K_0+\alpha^{-1}K_1\right)t_0^{-q}H(t_0)t^q\exp\left\{\frac{\alpha(t_0-t)}{2}\right\}$$
$$+2^q\alpha^{-1}K_1H\left(\frac{t_0+t}{2}\right)+K_0H(t)\Bigg]e^{\alpha t}t^{-q}\quad \textit{for } t_0\le t<\infty.$$

Proof. Denote

$$H_1(t)=\int_t^{\infty} h(s)\,ds.$$

Integrating by parts yields

$$\int_t^{t_1} h(s)s^{-q}\,ds=-s^{-q}H_1(s)\Big|_t^{t_1}-\int_t^{t_1} qs^{-q-1}H_1(s)\,ds.$$

Since $H_1(t)\to0$ as $t\to\infty$ and

$$\int_t^{\infty}\left|qs^{-q-1}H_1(s)\right|ds\le H(t)\int_t^{\infty} qs^{-q-1}\,ds,$$

the integral of $h(t)t^{-q}$ converges and satisfies (2.1).

Assuming now a function $K(t)$ to have the properties stated in the hypotheses, we can write

$$(2.6)\qquad \int_t^{t_1} K(s)h(s)\,ds = -K(s)H_1(s)\Big|_t^{t_1} + \int_t^{t_1} K'(s)H_1(s)\,ds$$

and

$$(2.7)\qquad |K^{(j)}(t)H_1(t)| \leq K_j H(t) e^{\alpha t} t^{-q}, \qquad t \geq t_0,\ \ j=0,1.$$

If $\alpha<0$, then (2.7) implies that

$$(2.8)\qquad \int_t^{\infty} |K'(s)H_1(s)|\,ds \leq K_1 H(t) t^{-q} \int_t^{\infty} e^{\alpha s}\,ds = K_1 |\alpha|^{-1} H(t) e^{\alpha t} t^{-q}.$$

By (2.6)–(2.8), $\int^{\infty} K(t)h(t)\,dt$ converges and (2.4) holds.

If $\alpha>0$, let $T=(t_0+t)/2$. From (2.7),

$$\begin{aligned}(2.9)\qquad \int_{t_0}^{t} |K'(s)H_1(s)|\,ds &= \int_{t_0}^{T} |K'(s)H_1(s)|\,ds + \int_{T}^{t} |K'(s)H_1(s)|\,ds \\ &\leq K_1 H(t_0) t_0^{-q} \int_{-\infty}^{T} e^{\alpha s}\,ds + K_1 H(T) T^{-q} \int_{-\infty}^{t} e^{\alpha s}\,ds \\ &= \alpha^{-1} K_1 \left[t_0^{-q} H(t_0) e^{\alpha T} + H(T) T^{-q} e^{\alpha t} \right].\end{aligned}$$

By (2.6) with $t=t_0$ and $t_1=t$, (2.7) and (2.9) the inequality (2.5) is valid for $t_0 \leq t < \infty$. This completes the proof of Lemma 1.

Remark 1. If $K(t)$ satisfies (2.2) and (2.3) with $\alpha=0$, then, in general, the inequality (2.4) is useless. In this case,

$$(2.10)\qquad \left|\int_t^{\infty} K(s)h(s)\,ds\right| \leq |K(t)| \left|\int_t^{\infty} h(s)\,ds\right| + \int_t^{\infty} |K'(s)| \left|\int_s^{\infty} h(v)\,dv\right| ds,$$

because of (2.6). However, it is now necessary to show that the integral on the right of (2.10) converges.

Remark 2. The right-hand sides of (2.4) and (2.5) can be written as

$$(K_0+K_1) m(t_0,t,\alpha,q) e^{\alpha t} t^{-q},$$

where the function $m(t_0,t,\alpha,q)$ is independent of $K(t)$,

$$(2.11)\qquad \lim_{t\to\infty} m(t_0,t,\alpha,q)=0, \qquad t_0>0$$

and

$$(2.12)\qquad \lim_{t_0\to\infty}\ \sup_{t\geq t_0} m(t_0,t,\alpha,q)=0.$$

These properties of $m(t_0,t,\alpha,q)$ are the only ones which will be used in the proof of Theorem 1.

The proof of (2.11) and (2.12) is easy. Indeed, if $\alpha<0$, then $m(t_0,t,\alpha,q) \leq (1+|\alpha|^{-1})H(t)$, and the relations are valid. If $\alpha>0$, then

$$\begin{aligned} m(t_0,t,\alpha,q) &\leq (1+\alpha^{-1}) t_0^{-q} H(t_0) t^{q} \exp\left\{ \frac{\alpha(t_0-t)}{2} \right\} \\ &\quad + 2^{q} \alpha^{-1} H\left(\frac{t_0+t}{2}\right) + H(t), \qquad t \geq t_0.\end{aligned}$$

Obviously, it is sufficient to verify (2.12) for the function

$$m(t_0,t)=t_0^{-q}H(t_0)t^q\exp\{\alpha(t_0-t)/2\},\qquad \alpha>0.$$

We shall show that

$$\sup_{t\geq t_0} m(t_0,t)\leq 2^qH(t_0)(1+2^qt_0^{-q}M),\qquad t_0>0, \tag{2.13}$$

where M is an upper bound of $e^{-\alpha t}t^q$ on $[0,\infty)$. Observe that $m(t_0,t)\leq 2^qH(t_0)$ for $t_0\leq t\leq 2t_0$. Further, $t(t-t_0)^{-1}\leq 2$ for $t\geq 2t_0$. For such a t,

$$\begin{aligned} m(t_0,t)&=2^qt_0^{-q}H(t_0)\cdot t^q(t-t_0)^{-q}\cdot e^{-\alpha s}s^q\big|_{s=(t-t_0)/2}\\ &\leq 2^qt_0^{-q}H(t_0)\cdot 2^q\cdot M. \end{aligned}$$

Consequently, (2.13) is valid.

3. Proof of Theorem 1. To avoid unnecessary subscripts, we let r be a fixed integer $(1\leq r\leq n)$ throughout the proof. We shall show that under the conditions of Theorem 1, there is a solution $x=x_r$ of (1.1) satisfying (1.5) with $j=r$.

If $t_0>0$, let $U[t_0,\infty)$ be the space of all functions $u(t)$ in $C^{n-1}[t_0,\infty)$ satisfying

$$u^{(k)}(t)\exp(-\lambda_rt)=O(e^{-\rho t}t^{-q}),\qquad 0\leq k\leq n-1,\quad \text{as } t\to\infty. \tag{3.1}$$

Then $U[t_0,\infty)$ is a Banach space with respect to the norm

$$\|u\|=\sup_{t\geq t_0}\sum_{k=0}^{n-1}\left|u^{(k)}(t)\exp(-\lambda_rt)\right|e^{\rho t}t^q. \tag{3.2}$$

In the following, assume that $t\geq t_0$. From (3.2) and the identity

$$\left(u^{(k)}(t)\exp(-\lambda_rt)\right)'=\left(u^{(k+1)}(t)-\lambda_ru^{(k)}(t)\right)\exp(-\lambda_rt),$$

we find that

$$\left|u^{(k)}(t)\exp(-\lambda_rt)\right|\leq\|u\|e^{-\rho t}t^{-q}\qquad (0\leq k\leq n-1), \tag{3.3}$$

and

$$\left|\left(u^{(k)}(t)\exp(-\lambda_rt)\right)'\right|\leq\left(1+|\lambda_r|\right)\|u\|e^{-\rho t}t^{-q}\qquad (0\leq k\leq n-2) \tag{3.4}$$

if $u\in U[t_0,\infty)$.

For convenience, let

$$L_j(t)=\int_{t_{0j}}^t\sum_{k=1}^n\lambda_r^{n-k}p_k(s)\exp(\lambda_r-\lambda_j)s\,ds\qquad (1\leq j\leq n), \tag{3.5}$$

and

$$L_{jk}[u](t)=\int_{t_{0j}}^tp_k(s)\exp(-\lambda_js)u^{(n-k)}(s)\,ds\qquad (1\leq j,k\leq n), \tag{3.6}$$

where

$$t_{0j}=\begin{cases} t_0 & \text{if } \operatorname{Re}\lambda_j<\operatorname{Re}\lambda_r-\rho,\\ \infty & \text{if } \operatorname{Re}\lambda_j\geq\operatorname{Re}\lambda_r-\rho.\end{cases}$$

LEMMA 2. *Under the hypotheses of Theorem* 1, *the functions* $L_j(t)$, $L_{jk}[u](t)$ *are defined on* $[t_0,\infty)$ *and satisfy the inequalities*

$$\left|L_j(t)\right|\leq m_{j0}(t_0,t)t^{-q}\exp\left\{\left(\operatorname{Re}\lambda_r-\operatorname{Re}\lambda_j-\rho\right)t\right\}, \tag{3.7}$$

and

$$|L_{jk}[u](t)|\le\|u\|m_{jk}(t_0,t)t^{-q}\exp\{(\mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-\rho)t\} \tag{3.8}$$

for every u in $U[t_0,\infty)$ and $1\le j$, $k\le n$. Here the functions $m_{jk}(t_0,t)$ are independent of u and satisfy

$$\lim_{t\to\infty} m_{jk}(t_0,t)=0, \qquad t_0>0, \tag{3.9}$$

and

$$\lim_{t_0\to\infty}\ \sup_{t\ge t_0} m_{jk}(t_0,t)=0. \tag{3.10}$$

Proof. If $\mathrm{Re}\lambda_j\ne\mathrm{Re}\lambda_r-\rho$, let

$$h(t)=\sum_{k=1}^{n}\lambda_r^{n-k}p_k(t)e^{\rho t}t^q,$$

and

$$K_j(t)=t^{-q}\exp\{(\lambda_r-\lambda_j-\rho)t\}.$$

By condition (ii) of Theorem 1, $\int^\infty h(t)\,dt$ converges. Obviously the function $K_j(t)$ satisfies (2.2) and (2.3) on $[t_0,\infty)$, with $\alpha=\mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-\rho\ne0$, $K_0=1$ and $K_1=|\lambda_r-\lambda_j-\rho|+qt_0^{-1}$. From Lemma 1 and Remark 2, the integral

$$L_j(t)=\int_{t_{0j}}^{t}K_j(s)h(s)\,ds$$

converges and satisfies (3.7), with $m_{j0}(t_0,t)$ satisfying (3.9) and (3.10). If $\mathrm{Re}\lambda_j=\mathrm{Re}\lambda_r-\rho$, we can apply Lemma 1 to the integral

$$L_j(t)=\int_{\infty}^{t}h_j(s)s^{-q}\,ds,$$

where

$$h_j(t)=\sum_{k=1}^{n}\lambda_r^{n-k}p_k(t)\exp\{(\lambda_r-\lambda_j)t\}t^q,$$

because, by condition (ii), $\int^\infty h_j(t)\,dt$ converges. This proves (3.7).

Now (3.3) with $k=n-1$ implies that

$$\int_t^\infty\left|p_1(s)\exp(-\lambda_js)u^{(n-1)}(s)\right|ds\le\|u\|\int_t^\infty|p_1(s)|s^{-q}\exp\{(\mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-\rho)s\}\,ds$$
$$\le\|u\|t^{-q}\exp\{(\mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-\rho)t\}\int_t^\infty|p_1(s)|\,ds$$

if $\mathrm{Re}\lambda_j\ge-\rho+\mathrm{Re}\lambda_r$, which shows that $L_{j1}[u](t)$ is defined and (3.8) holds for $k=1$. (Here we need assumption (i) of Theorem 1.) If $\mathrm{Re}\lambda_j<-\rho+\mathrm{Re}\lambda_r$, let $T=(t_0+t)/2$. From (3.3) with $k=n-1$,

$$\int_{t_0}^{t}\left|p_1(s)\exp(-\lambda_js)u^{(n-1)}(s)\right|ds$$
$$=\int_{t_0}^{T}\left|p_1(s)\exp(-\lambda_js)u^{(n-1)}(s)\right|ds+\int_{T}^{t}\left|p_1(s)\exp(-\lambda_js)u^{(n-1)}(s)\right|ds$$
$$\le\|u\|t_0^{-q}\exp\{(\mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-\rho)T\}\int_{t_0}^{T}|p_1(s)|\,ds$$
$$+\|u\|T^{-q}\exp\{(\mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-\rho)t\}\int_T^t|p_1(s)|\,ds, \qquad u\in U[t_0,\infty).$$

Therefore (3.8) with $k=1$ holds if

$$m_{j1}(t_0,t)=t_0^{-q}t^q\exp\left\{(\mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-\rho)\frac{t_0-t}{2}\right\}\int_{t_0}^{\infty}|p_1(s)|\,ds+2^q\int_{(t_0+t)/2}^{\infty}|p_1(s)|\,ds.$$

(The properties (3.9) and (3.10) of such a function $m_{j1}(t_0,t)$ were proved at the end of §2.)

Now, let k be a fixed integer, $2\leq k\leq n$. Let $h(t)=p_k(t)t^qe^{\rho t}$ and $K_j(t)=t^{-q}\exp\{-(\lambda_j+\rho)t\}u^{(n-k)}(t)$. Then $K_j(t)\in C^1[t_0,\infty)$ and, by (3.3) and (3.4),

$$(3.11)\qquad |K_j(t)|\leq\|u\|\exp(\alpha_1 t)t^{-2q}$$

and

$$(3.12)\qquad |K_j'(t)|\leq M\|u\|\exp(\alpha_1 t)t^{-2q},$$

where $\alpha_1=\mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-2\rho\leq\mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-\rho$ and $M=1+|\lambda_r|+|\lambda_r-\lambda_j-\rho|+qt_0^{-1}$. To prove (3.8), we can also apply Lemma 1 to the integral

$$L_{jk}[u](t)=\int_{t_{0j}}^{t}K_j(s)h(s)\,ds,$$

the number $\alpha\neq 0$ in (2.2) and (2.3) is given by

$$\alpha=\begin{cases}\mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-\rho & \text{if } \mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-\rho\neq 0,\\ -\rho, & \text{if } \mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-\rho=0 \text{ and } \rho>0.\end{cases}$$

In fact, from (3.11) and (3.12), (2.2) and (2.3) are valid for $K_0=\|u\|t_0^{-q}$ and $K_1=M\|u\|t_0^{-q}$, since $\alpha_1\leq\alpha$.

It remains to prove (3.8) in the case where $\mathrm{Re}\lambda_j=\mathrm{Re}\lambda_r$ and $\rho=0$. Using (2.10), we obtain

$$(3.13)\qquad \begin{aligned}|L_{jk}[u](t)|&=\left|\int_t^{\infty}K_j(s)h(s)\,ds\right|\\ &\leq\|u\|t^{-2q}|H_1(t)|+M\|u\|\int_t^{\infty}s^{-2q}|H_1(s)|\,ds,\end{aligned}$$

where

$$H_1(t)=\int_t^{\infty}h(s)\,ds=\int_t^{\infty}p_k(s)s^q\,ds.$$

Inequality (3.13) shows that (3.8) holds if the integral on the right-hand side (3.13) converges and

$$t^q\int_t^{\infty}s^{-2q}|H_1(s)|\,ds\to 0\quad \text{as } t\to\infty.$$

If $0\leq q<1$, the last assertion follows from condition (iii) of Theorem 1. If $q\geq 1$, we may apply Hospital's rule:

$$\lim\frac{\int_t^{\infty}s^{-2q}|H_1(s)|\,ds}{t^{-q}}=\lim\frac{t^{-2q}|H_1(t)|}{(qt^{-q-1})}=0.$$

This completes the proof of Lemma 2.

Returning to the proof of Theorem 1, we define the operator T by

$$T[u](t)=-\sum_{j=1}^{n}c_j\exp(\lambda_j t)I_j[u](t), \tag{3.14}$$

where

$$I_j[u](t)=L_j(t)+\sum_{k=1}^{n}L_{jk}[u](t), \qquad 1\le j\le n. \tag{3.15}$$

Here $L_j(t)$, $L_{jk}[u](t)$ are the same as before (see (3.5) and (3.6)) and the numbers c_j satisfy the system

$$\sum_{j=1}^{n}c_j\lambda_j^k=\begin{cases}0 & \text{if } 0\le k\le n-2,\\ 1 & \text{if } k=n-1.\end{cases} \tag{3.16}$$

By Lemma 2, the functions $I_j[u](t)$ are defined on $[t_0,\infty)$ for every $u\in U[t_0,\infty)$ and

$$|I_j[u](t)|\le(1+\|u\|)m_j(t_0,t)t^{-q}\exp\{(\mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-\rho)t\}, \qquad 1\le j\le n. \tag{3.17}$$

For any elements $\tilde{u}$, $\tilde{\tilde{u}}$ in $U[t_0,\infty)$,

$$\begin{aligned}|I_j[\tilde{u}](t)-I_j[\tilde{\tilde{u}}](t)|&=\left|\sum_{k=1}^{n}L_{jk}[\tilde{u}-\tilde{\tilde{u}}](t)\right|\\ &\le\|\tilde{u}-\tilde{\tilde{u}}\|m_j(t_0,t)t^{-q}\exp\{(\mathrm{Re}\lambda_r-\mathrm{Re}\lambda_j-\rho)t\}, \qquad 1\le j\le n,\end{aligned} \tag{3.18}$$

where the functions $m_j(t_0,t)$ in (3.17) and (3.18) are independent of u and

$$\lim_{t\to\infty}m_j(t_0,t)=0, \qquad t_0>0, \tag{3.19}$$

and

$$\lim_{t_0\to\infty}\sup_{t\ge t_0}m_j(t_0,t)=0. \tag{3.20}$$

Consequently, $I_j[u]\in C^1[t_0,\infty)$ and, by (3.5) and (3.6),

$$I_j'[u](t)=L_j'(t)+\sum_{k=1}^{n}L_{jk}'[u](t)=\exp(-\lambda_j t)\sum_{k=1}^{n}p_k(t)x^{(n-k)}(t), \tag{3.21}$$

where $x(t)=\exp(\lambda_r t)+u(t)$.

From (3.14), (3.16), (3.21) and induction,

$$T^{(k)}[u](t)=-\sum_{j=1}^{n}c_j\lambda_j^k\exp(\lambda_j t)I_j[u](t) \qquad (0\le k\le n-1) \tag{3.22}$$

and

$$T^{(n)}[u](t)+\sum_{k=1}^{n}p_k(t)x^{(n-k)}(t)=-\sum_{j=1}^{n}c_j\lambda_j^n\exp(\lambda_j t)I_j[u](t). \tag{3.23}$$

From (3.17) and (3.22),

$$(3.24)\qquad |T^{(k)}[u](t)\exp(-\lambda_r t)| \le (1+\|u\|)\sum_{j=1}^{n}|c_j\lambda_j^k|m_j(t_0,t)e^{-\rho t}t^{-q},$$
$$0\le k\le n-1,\quad u\in U[t_0,\infty).$$

Thus $T[u]$ is in $U[t_0,\infty)$ for every $u\in U[t_0,\infty)$ (see the definition (3.1) and property (3.20) of $m_j(t_0,t)$). From (3.18) and (3.22),

$$|(T^{(k)}[\tilde{u}](t)-T^{(k)}[\tilde{\tilde{u}}](t))\exp(-\lambda_r t)| \le \|\tilde{u}-\tilde{\tilde{u}}\| \sum_{j=1}^{n}|c_j\lambda_j^k|m_j(t_0,t)e^{-\rho t}t^{-q},$$

and so, by the definition (3.2) of the norm in $U[t_0,\infty)$,

$$\|T[\tilde{u}]-T[\tilde{\tilde{u}}]\| \le \|\tilde{u}-\tilde{\tilde{u}}\| \sup_{t\ge t_0}\sum_{k=0}^{n-1}\sum_{j=1}^{n}|c_j\lambda_j^k|m_j(t_0,t)$$

for arbitrary functions $\tilde{u},\tilde{\tilde{u}}$ in $U[t_0,\infty)$. By (3.20), there is a $t_0>0$ such that

$$\sup_{t\ge t_0}\sum_{k=0}^{n-1}\sum_{j=1}^{n}|c_j\lambda_j^k|m_j(t_0,t)<1.$$

For such a t_0 the operator T is a contraction mapping of the space $U[t_0,\infty)$ into itself. According to the Banach contraction principle ([1, p. 11]), there exists u_r such that $T[u_r]=u_r$, i.e., $T[u_r](t)=u_r(t)$ for $t_0\le t<\infty$. From this identity and (3.21)–(3.23) it follows that $x_r(t)=\exp(\lambda_r t)+u_r(t)$ is a solution of (1.1) on $[t_0,\infty)$. This solution can be extended to $[0,\infty)$. Using (3.19) and (3.24) with $u=u_r=T[u_r]$ we find that this solution satisfies (1.5) with $j=r$. This completes the proof of Theorem 1.

REFERENCES

[1] W. A. Coppel, *Stability and Asymptotic Behavior of Differential Equations*, D.C. Heath, Boston, 1965.

[2] P. Hartman, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[3] W. F. Trench, *Asymptotic integration of linear differential equations subject to integral smallness conditions involving ordinary convergence*, this Journal, 7 (1976), pp. 213–221.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1501-0008 $01.25/0

BOUNDARY VALUE PROBLEMS FOR AN nTH ORDER LINEAR DIFFERENCE EQUATION*

ALLAN C. PETERSON†

Abstract. We are concerned with boundary value problems for the nth order linear difference equation $Pu(m)=\sum_{j=0}^{n}\alpha_j(m)u(m+j)=0$, where $\alpha_n(m)=1$ and $\alpha_0(m)\neq 0$. We give necessary conditions for the coefficients $\alpha_j(m)$ for $Pu(m)=0$ to be $(l,n-l)$-disconjugate. The Green's functions for $(l,n-l)$-boundary value problems, $1\le l\le n-1$, are also considered.

Key words. linear difference equation, boundary value problem, disconjugate, Green's function

We are concerned with the nth order linear difference equation

$$Pu(m)=\sum_{j=0}^{n}\alpha_j(m)u(m+j)=0,\qquad m\in I, \tag{1}$$

where the coefficients are defined on either the finite "interval" $I=[a,b]\equiv\{a,a+1,\cdots,b\}$, a and b integers, or the infinite "interval" $I=[a,\infty)\equiv\{a,a+1,\cdots\}$, a an integer. We assume $\alpha_n(m)\equiv 1$ and $\alpha_0(m)\neq 0$ for $m\in I$. Solutions for (1) are defined on I^n, where $I^n=[a,b+n]$ when $I=[a,b]$, and $I^n=I$ when $I=[a,\infty)$. A lot of notation used in this paper is the same as used by Hartman in [1].

We now introduce an adjoint difference equation [3, p. 289] of (1). To this end we first define quasi differences $D_kz(m)$ as follows. If $z(m)$ is defined on I^n, then

$$D_0z(m)=z(m),\qquad m\in I^n.$$

$$D_kz(m)=D_{k-1}z(m+1)+\frac{\alpha_k(m)}{\alpha_0(m)}z(m),\qquad m\in I^{n-k} \tag{2}$$

for $1\le k\le n$. It can easily be shown that

$$D_kz(m)=\sum_{j=0}^{k}\frac{\alpha_{k-j}(m+j)}{\alpha_0(m+j)}z(m+j) \tag{3}$$

for $0\le k\le n$. We then define the adjoint operator P^* and adjoint difference equation by the equation

$$P^*z(m)=D_nz(m)=0. \tag{4}$$

By use of (3) with $k=n$ we could also write this as (see [3, p. 289])

$$P^*z(m)=\sum_{j=0}^{n}\frac{\alpha_{n-j}(m+j)}{\alpha_0(m+j)}z(m+j)=0. \tag{5}$$

If $u(m)$ and $z(m)$ are defined on I^n, then as usual, the Lagrange bracket of $z(m)$ and $u(m)$ is defined by

$$\{z;u\}=\sum_{k=0}^{n-1}D_kz(m)u(m+k),\qquad m\in I^1. \tag{6}$$

*Received by the editors February 16, 1982, and in revised form August 10, 1982.

†Department of Mathematics and Statistics, University of Nebraska, Lincoln, Nebraska 68588.

If $u(m)$ is defined on I^n, then we define the usual difference operator Δ by

$$\Delta u(m)=u(m+1)-u(m), \qquad m\in I^{n-1}.$$

If $u(m)$ and $z(m)$ are defined on I^n, then by operating on both sides of (6) by Δ, it is easy to obtain the Lagrange identity

$$(7) \qquad -\frac{z(m)}{\alpha_0(m)}Pu(m)+u(m+n)P^*z(m)=\Delta\{z;u\}, \qquad m\in I.$$

For each $p, 0\leq p\leq n-1$, define $u_p(m,t)$ to be, for each fixed $t\in I$, the solution of (1) satisfying the initial conditions

$$u_p(t+k,t)=\delta_{pk}, \qquad k=0,\cdots,n-1,$$

where δ_{pk} is the Kronecker delta. Similarly, let $z_p(m,t)$, for each fixed $t\in I$, be the solution of (4) satisfying the initial conditions

$$D_k z_p(t,t)=\delta_{pk}, \qquad k=0,\cdots,n-1.$$

By the Lagrange identity (7), we get that for fixed $s,t\in I$ $\{z_p(m,s);u_q(m,t)\}$ is constant in I^1. Hence

$$\{z_p(m,s);u_q(m,t)\}|_{m=s}=\{z_p(m,s);u_q(m,t)\}|_{m=t}.$$

It follows easily from this that

$$(8) \qquad u_q(s+p,t)=D_q z_p(t,s), \qquad 0\leq p,q\leq n-1, \quad s,t\in I.$$

These are very important formulas. (See [4(6)] for the analogous results.)

Let $u_1(m),\cdots,u_n(m)$ be functions defined on I^n. Then as in [1] we define

$$W[u_1,\cdots,u_k](m)=\begin{vmatrix} u_1(m) & \dots & u_k(m) \\ u_1(m+1) & \dots & u_k(m+1) \\ \dots & & \dots \\ u_1(m+k-1) & \dots & u_k(m+k-1) \end{vmatrix}$$

for $m\in I^{n+1-k}$, $1\leq k\leq n$.

We can assume that our difference equation (1) is defined on $(-\infty,\infty)=\{\text{integers}\}$ by defining $\alpha_i(m)=\alpha_i(a)$, $m\leq a$ and $\alpha_i(b)$ for $m\geq b$. This will be assumed whenever necessary in the remainder of this paper.

It is easy to use (8) to derive

THEOREM 1.

$$W[u_0(m,t),\cdots,u_{k-1}(m,t)]|_{m=s+n-k}=W[z_{n-k}(m,s),\cdots,z_{n-1}(m,s)]|_{m=t}$$

for $k=1,\cdots,n$.

COROLLARY 2. *Assume* $t+n-1<s$. *The difference equation* $Pu(m)=0$ *has a nontrivial solution* $u(m)$ *with*

$$u(t+j)=0, \qquad j=k,\cdots,n-1,$$
$$u(s+j)=0, \qquad j=n-k,\cdots,n-1,$$

iff the adjoint equation $P^*z(m)=0$ *has a nontrivial solution* $z(m)$ *with*

$$\begin{aligned} z(t+j)&=0, \quad j=0,\cdots,k-1,\\ z(s+j)&=0, \quad j=0,\cdots,n-k-1. \end{aligned}$$

A solution $u(m)$ of (1) is said to have a zero at $t\in I^n$ provided $u(t)=0$. We say that $u(m)$ has a generalized zero (see [1]) at t, provided $u(t)=0$ when $t=a$ and, when $t>a$, either $u(t)=0$ or there is an integer k, $1\le k\le t-a$, such that $(-1)^k u(t-k)u(t)>0$, and, if $k>1$, $u(t-k+1)=\cdots=u(t-1)=0$. A solution $u(m)$ will be said to have a $(k,n-k)$-pair of zeros in I^n provided u has zeros at $t,t+1,\cdots,t+k-1$, followed by $n-k$ generalized zeros at $s,s+1,\cdots,s+n-k-1$, where $a\le t<t+k-1<s<s+n-k-1\le b+n$. We will say that (1) is $(k,n-k)$-disconjugate on I^n, provided no nontrivial solution has a $(k,n-k)$-pair of zeros in I^n. We say that (1) is disconjugate on I^n if no nontrivial solution of (1) has n generalized zeros on I^n.

In the following result we do not assume (1) is disconjugate on I^n. With this in mind compare the following result with [1, Thm. 5.2].

THEOREM 3. *Assume* $I=[a,b]$ *and that* (1) *is* $(l,n-l)$*-disconjugate on* I^n *for a fixed* $l\in\{1,\cdots,n-1\}$. *Then*

$$(9)\quad (-1)^{k(n+l)}\begin{vmatrix} \alpha_l(m) & \alpha_{l+1}(m) & \cdots & \alpha_{l+k-1}(m)\\ \alpha_{l-1}(m+1) & \alpha_l(m+1) & \cdots & \alpha_{l+k-2}(m)\\ \cdots & & & \cdots\\ \alpha_{l-k+1}(m+k-1) & \alpha_{l-k+2}(m+k-1) & \cdots & \alpha_l(m+k-1) \end{vmatrix}>0$$

for $m\in I^{1-k}$ $(I^0=I)$ *for* $k=1,\cdots,$ card I. (*Here* $\alpha_j(m)\equiv 0$ *for* $j>n$ *or* $j<0$.)

Proof. We prove (9) by induction on k. For $k=1$ we want to show that

$$(-1)^{n+l}\alpha_l(m)>0, \qquad m\in I.$$

If we assume not, then there is an $m_0\in I$ such that

$$(-1)^{n+l}\alpha_l(m_0)\le 0.$$

Let u be the solution of (1) such that

$$\begin{aligned} u(m_0+j)&=0, \quad j\in\{0,\cdots,n\}-\{l\},\\ u(m_0+l)&=1. \end{aligned}$$

By use of (1) evaluated at m_0, we get that

$$u(m_0+n)=-\alpha_l(m_0)u(m_0+l)=-\alpha_l(m_0).$$

Since $u(m_0+l+1)=\cdots=u(m_0+n-1)=0$, and

$$(-1)^{n-l}u(m_0+l)u(m_0+n)=(-1)^{n+l+1}\alpha_l(m_0)\ge 0,$$

$u(m)$ has a generalized zero at m_0+n. Hence u has an $(l,n-l)$ pair of zeros at $m_0<m_0+l+1$, which is a contradiction. Hence $(-1)^{n+l}\alpha_l(m)>0$ for $m\in I$.

Assume $k>1$ and the inequalities (9) are true with k replaced by $1,2,\cdots,k-1$.

Assume

$$D \equiv \begin{vmatrix} \alpha_l(m_0) & \dots & \alpha_{l+k-1}(m_0) \\ \alpha_{l-1}(m_0+1) & \dots & \alpha_{l+k-2}(m_0+1) \\ \dots & & \dots \\ \alpha_{l-k+1}(m_0+k-1) & \dots & \alpha_l(m_0+k-1) \end{vmatrix} = 0.$$

Then there are constants $A_1, \cdots, A_k$, not all zero, such that

$$\tag{10} \begin{array}{lcl} \alpha_l(m_0)A_1+\alpha_{l+1}(m_0)A_2 & +\cdots+ & \alpha_{l+k-1}(m_0)A_k=0, \\ \alpha_{l-1}(m_0+1)A_1+\alpha_l(m_0+1)A_2 & +\cdots+ & \alpha_{l+k-2}(m_0+1)A_k=0, \\ \dots & & \dots \\ \alpha_{l-k+1}(m_0+k-1)A_1 & +\cdots+ & \alpha_l(m_0+k-1)A_k=0. \end{array}$$

We consider the two cases $l+k \leq n$ and $l+k > n$. We will show that both cases lead to a contradiction.

First assume $l+k \leq n$. In this case let $u(m)$ be the solution of (1) such that

$$\tag{11} u(m_0+j)=0, \qquad 0 \leq j \leq l-1,$$

$$\tag{12} u(m_0+l+j)=A_{j+1}, \qquad 0 \leq j \leq k-1,$$

and if $l+k<n$,

$$u(m_0+l+k-1+j)=0, \qquad 0 \leq j \leq n-l-k.$$

Note that $u(m)$ is a nontrivial solution, as not all $A_1, \cdots, A_k$ are zero.

Using the first equation in (10), $Pu(m_0)=0$, (11) and (12) yields

$$u(m_0+n)=0.$$

Proceeding in this fashion we finally get, using the kth equation in (10), $Pu(m_0+k-1)=0$, (11) and (12), that

$$u(m_0+n+k-1)=0.$$

But then $u(m)$ is a nontrivial solution of (1) with an $(l, n-l)$-pair of zeros at $m_0 < m_0+l+k$, which is a contradiction.

Now assume $l+k>n$. In this case let $u(m)$ be the solution of (1) satisfying

$$\tag{13} u(m_0+j)=0, \qquad 0 \leq j \leq l-1,$$

$$\tag{14} u(m_0+l+j)=A_{j+1}, \qquad 0 \leq j \leq n-l.$$

In this case (10) becomes

(15)

$$\begin{array}{lcl} \alpha_l(m_0)A_1 & +\cdots+ & \alpha_n(m_0)A_{n-l+1}=0, \\ \alpha_{l-1}(m_0+1)A_1 & +\cdots+ & \alpha_{n-1}(m_0+1)A_{n-l+1}+\alpha_n(m_0+1)A_{n-l+2}=0, \\ \dots & \dots & \dots \\ \alpha_{n+1-k}(m_0+l+k-n-1)A_1 & +\cdots+ & \alpha_n(m_0+l+k-n-1)A_k=0, \\ \alpha_{n-k}(m_0+l+k-n)A_1 & +\cdots+ & \alpha_{n-1}(m_0+l+k-n)A_k=0, \\ \dots & \dots & \dots \\ \alpha_{l-k-1}(m_0+k-1)A_1 & +\cdots+ & \alpha_l(m_0+k-1)A_k=0. \end{array}$$

Using $Pu(m_0)=0$, (13) and (14), we get the equation

$$\alpha_l(m_0)A_1+\cdots+\alpha_{n-1}(m_0)A_{n-l+1}+\alpha_n(m_0)u(m_0+n)=0.$$

Combining this with the first equation in (15), we get that

$$\alpha_n(m_0)[u(m_0+n)-A_{n-l+1}]=0.$$

Hence

$$u(m_0+n)=A_{n-l+1}.$$

Similarly, using $Pu(m_0+1)=0$, (13), (14) and the second equation in (15), we get that

$$u(m_0+n+1)=A_{n-l+2}.$$

Proceeding in this manner we finally get, using $Pu(m_0+l+k-n-1)=0$, (13), (14) and the $(l+k-n)$th equation in (15), that

$$u(m_0+l+k-1)=A_k.$$

Since at least one of $A_1,\cdots,A_k$ is nonzero, we now know that $u(m)$ is a nontrivial solution of (1).

Using $Pu(m_0+l+k-n)=0$, (13), (14) and the $(l+k-n+1)$st equation in (15), we get that

$$u(m_0+l+k)=0.$$

Finally, using $Pu(m_0+k-1)=0$, (13), (14) and the kth equation in (15), we get that

$$u(m_0+n+k-1)=0.$$

But then $u(m)$ is a nontrivial solution of (1) with an $(l,n-l)$-pair of zeros at $m_0<m_0+l+k$, which is a contradiction. Hence $D\neq 0$.

If we assume (9) is not valid, then there is a $m_0\in I^{1-k}$ such that

$$(-1)^{k(n+l)}D$$
$$\equiv(-1)^{k(n+l)}\begin{vmatrix} \alpha_l(m_0) & \alpha_{l+1}(m_0) & \cdots & \alpha_{l+k-1}(m_0) \\ \alpha_{l-1}(m_0+1) & \alpha_l(m_0+1) & \cdots & \alpha_{l+k-2}(m_0+1) \\ \cdots & \cdots & & \cdots \\ \alpha_{l-k+1}(m_0+k-1) & \alpha_{l-k+2}(m_0+k-1) & \cdots & \alpha_l(m_0+k-1) \end{vmatrix}<0.$$

Let $u(m)$ be the solution of (1) satisfying the boundary conditions

$$\begin{aligned} &u(m_0+j)=0, && 0\le j\le l-1,\\ &u(m_0+l+k+j)=0, && 0\le j\le n-l-2,\\ &u(m_0+n+k-1)=1. \end{aligned}$$

Evaluating (1) at m_0, $m_0+1,\cdots,m_0+k-1$ respectively, we are lead to the equations

$$\begin{aligned}
&\alpha_l(m_0)u(m_0+l)+\alpha_{l+1}(m_0)u(m_0+l+1)\\
&\qquad\qquad+\cdots+\alpha_{l+k-1}(m_0)u(m_0+l+k-1)=0,\\
&\alpha_{l-1}(m_0+1)u(m_0+l)+\alpha_l(m_0+1)u(m_0+l+1)\\
&\qquad\qquad+\cdots+\alpha_{l+k-2}(m_0+1)u(m_0+l+k-1)=0,\\
&\cdots\qquad\cdots\\
&\alpha_{l-k+1}(m_0+k-1)u(m_0+l)+\alpha_{l-k}(m_0+k-1)u(m_0+l+1)\\
&\qquad\qquad+\cdots+\alpha_l(m_0+k-1)u(m_0+l+k-1)=-1.
\end{aligned}$$

Solving for $u(m_0+l+k-1)$, we get that

$$u(m_0+l+k-1)=\frac{-1}{D}\begin{vmatrix}\alpha_l(m_0) & \cdots & \alpha_{l+k-2}(m_0)\\ \cdots & & \cdots\\ \alpha_{l-k+2}(m_0+k-2) & \cdots & \alpha_l(m_0+k-2)\end{vmatrix}.$$

Hence

$$\begin{aligned}
&(-1)^{n-l}u(m_0+l+k-1)\\
&\quad=\frac{(-1)^{(k-1)(n+l)}}{-(-1)^{k(n+l)}D}\begin{vmatrix}\alpha_l(m_0) & \cdots & \alpha_{l+k-2}(m_0)\\ \cdots & & \cdots\\ \alpha_{l-k+2}(m_0+k-2) & \cdots & \alpha_l(m_0+k-2)\end{vmatrix}>0,
\end{aligned}$$

so u has a generalized zero at m_0+k-1. Hence $u(m)$ is a nontrivial solution of (1) which has an $(l,n-l)$-pair of zeros at $m_0<m_0+l+k$, which contradicts the $(l,n-l)$-disconjugacy of (1) on I^n. Hence

$$(-1)^{k(n+l)}D>0,$$

and the proof is complete. □

Let $U(m,\nu)$ be the Cauchy function (see [1]) for (1). That is, for each fixed $\nu\in[a,b]$, $U(m,\nu)$ is the solution of (1) satisfying

$$\begin{aligned}
&U(\nu+j,\nu)=0, \qquad j=1,\cdots,n-1,\\
&U(\nu+n,\nu)=1.
\end{aligned}$$

THEOREM 4. *If we assume* (1) *is* $(k,n-k)$*-disconjugate on* $I^n=[a,b+n]$ *for a fixed* k, $1\le k\le n-1$, *then the Green's function* $G_k(m,\nu)$ *for the problem*

$$\begin{aligned}
&Pu(m)=f(m),\\
&u(a+j)=0, \qquad 0\le j\le k-1,\\
&u(b+n-j)=0\,, \quad 0\le j\le n-k-1
\end{aligned}$$

exists. It is defined on $I^n \times I$ *and can be expressed in the form*:

$$(16)\quad G_k(m,\nu)=\frac{1}{D}\begin{vmatrix} 0 & u_k(m,a) & \cdots & u_{n-1}(m,a) \\ U(b+k+1,\nu) & u_k(b+k+1,a) & \cdots & u_{n-1}(b+k-1,a) \\ \cdots & & & \cdots \\ U(b+n,\nu) & u_k(b+n,a) & \cdots & u_{n-1}(b+n,a) \end{vmatrix},$$

for $m \leq \nu$, *and for* $m > \nu$

$$(17)\quad G_k(m,\nu)=\frac{1}{D}\begin{vmatrix} U(m,\nu) & u_k(m,a) & \cdots & u_{n-1}(m,a) \\ U(b+k+1,\nu) & u_k(b+k+1,a) & \cdots & u_{n-1}(b+k+1,a) \\ \cdots & & & \cdots \\ U(b+n,\nu) & u_k(b+n,a) & \cdots & u_{n-1}(b+n,a) \end{vmatrix},$$

where

$$D=\begin{vmatrix} u_k(b+k-1,a) & \cdots & u_{n-1}(b+k-1,a) \\ \cdots & & \cdots \\ u_k(b+n,a) & \cdots & u_{n-1}(b+n,a) \end{vmatrix},$$

and $u_j(m,a)$ *are defined as before* (8).

Proof. The proof of the existence is elementary and will be omitted. We will show that the Green's function $G_k(m,\nu)$ is given by (16) and (17). The Green's function $G_k(m,\nu)$ is characterized (see [1,p. 20]) by $PG_k(m,\nu)=\delta_{m\nu}$ for $(m,\nu)\in I^n \times I$, where $\delta_{m\nu}$ is the Kronecker delta, and where $v(m)=G_k(m,\nu)$ satisfies the boundary conditions

$$(18)\qquad v(a+j)=0, \qquad 0\leq j\leq k-1,$$

$$(19)\qquad v(b+n-j)=0, \qquad 0\leq j\leq n-k-1.$$

Define $G(m,\nu)$ on $I^n \times I$ by the right-hand side of (16) for $m\leq \nu$, and by the right-hand side of (17) when $\nu<m$. It suffices to show $G(m,\nu)$ satisfies the above properties that characterize $G(m,\nu)$.

Assume throughout this paragraph that $m\in[a,a+k-1]$. We now show that $v(m)=G(m,\nu)$ satisfies the boundary conditions (18). If $m\leq\nu$, then by (16) we get that $G(m,\nu)=0$. If $\nu<m$, then by (17)

$$G(m,\nu)=\frac{1}{D}\begin{vmatrix} U(m,\nu) & 0 & \cdots & 0 \\ \cdots & & & \cdots \\ U(b+n,\nu) & u_k(b+n,a) & \cdots & u_{n-1}(b+n,\nu) \end{vmatrix}.$$

But $a\leq\nu<m\leq a+k-1$, so $U(m,\nu)=0$ in the above determinant. So again $G(m,\nu)=0$. Hence $v(m)=G(m,\nu)$ satisfies (18).

Now assume $m\in[b+k+1,b+n]$. Then $m>\nu$, and it follows from (17) that $G(m,\nu)=0$. Hence $v(m)=G(m,\nu)$ satisfies (19).

It remains to be shown that $PG(m,\nu)=\delta_{m\nu}$ for $(m\nu)\in I^n\times I$. If $m>\nu$, then $PG(m,\nu)=0$ follows easily from (17). Similarly, if $a\leq m<\nu-n+1$, then $PG(m,\nu)=0$ follows easily from (16).

Assume $a \leq \nu - n + 1 \leq m < \nu$. Using (16) and (17), we get that

$$PG(m,\nu) = \frac{1}{D} \sum_{j=0}^{\nu-m} \begin{vmatrix} 0 & \alpha_j(m)u_k(m+j,a) & \cdots & \alpha_j(m)u_n(m+j,a) \\ \cdots & & & \cdots \\ U(b+n,\nu) & u_k(b+n,a) & \cdots & u_n(b+n,a) \end{vmatrix}$$

$$+\frac{1}{D} \sum_{j=\nu-m+1}^{n} \begin{vmatrix} \alpha_j(m)U(m+j,\nu) & \cdots & \alpha_j(m)u_{n-1}(m+j,a) \\ \cdots & & \cdots \\ U(b+n,\nu) & \cdots & u_{n-1}(b+n,a) \end{vmatrix}.$$

Using $U(m+j,\nu) = 0, j = \nu - m + 1, \cdots, n$, we get that

$$PG(m\nu) = \begin{vmatrix} 0 & Pu_k(m,a) & \cdots & Pu_{n-1}(m,a) \\ \cdots & & & \cdots \\ U(b+n,\nu) & u_k(b+n,a) & \cdots & u_{n-1}(b+n,a) \end{vmatrix} = 0.$$

Finally, consider the case $m = \nu$. Using (16) and (17) it is easy to see that

$$PG(\nu,\nu) = \frac{1}{D} \begin{vmatrix} 0 & \alpha_0(\nu)u_k(\nu,a) & \cdots & \alpha_0(\nu)u_{n-1}(\nu,a) \\ \cdots & & & \cdots \\ U(b+n,\nu) & u_k(b+n,a) & \cdots & u_{n-1}(b+n,a) \end{vmatrix}$$

$$-\frac{1}{D} \begin{vmatrix} \alpha_0(\nu)U(\nu,\nu) & \cdots & \alpha_0(\nu)u_{n-1}(\nu,a) \\ \cdots & & \cdots \\ U(b+n,\nu) & \cdots & u_{n-1}(b+n,a) \end{vmatrix}$$

$$+\begin{vmatrix} PU(\nu,\nu) & \cdots & Pu_{n-1}(\nu,a) \\ \cdots & & \cdots \\ U(b+n,\nu) & \cdots & u_{n-1}(b+n,a) \end{vmatrix}.$$

Since the last determinant is zero, and $U(\nu,\nu) = -(1/\alpha_0(\nu))$,

$$PG(\nu,\nu) = 1.$$

DEFINITION. Let $1 \leq p \leq n-1$. We say that $Pu(m) = 0$ is ρ_p-disconjugate on I^n, provided there is no nontrivial solution of $Pu(m) = 0$ such that $u(a+j) = 0$, $0 \leq j \leq p-1$, and $u(m)$ has $n-p$ generalized zeros in $[a+p, b+n]$.

For results concerning ρ_p-disconjugacy for differential equations, see [5].

Notation. Assume $u_1, \cdots, u_k$ are functions defined on some interval J, and $\mu(j) \in J$, $1 \leq j \leq k$. Then set

$$D_k(\mu(1), \cdots, \mu(k)) = \begin{vmatrix} u_1(\mu(1)) & \cdots & u_k(\mu(1)) \\ u_1(\mu(2)) & \cdots & u_k(\mu(2)) \\ \cdots & & \cdots \\ u_1(\mu(k)) & \cdots & u_k(\mu(k)) \end{vmatrix}.$$

THEOREM 5. *Suppose that* (1) *is* ρ_{n-k}*-disconjugate on* I^n. *Let* $u_j(m)$, $1 \leq j \leq k$, *be a solution of* (1) *such that* $u_j(a+l)=0$, $0 \leq l \leq n-j-1$ *and* $(-1)^{j-1}u_j(a+n-j)>0$. *Then*

$$D_k(\mu(1),\cdots,\mu(k))>0$$

for $a+n-k \leq \mu(1)< \cdots <\mu(k) \leq b+n$. *In particular*

$$\omega_k(m) \equiv W(u_1,\cdots,u_k)(m)>0$$

for $a+n-k \leq m \leq b+n-k+1$.

If one reads the proof of [1, Prop. 5.2], it is easy to see how to prove this result. In the proof of Prop. 5.2 it need not be true that $c_k \neq 0$ as claimed, because it is possible that $\mu_0(1)=a+n-k+1$. This is easy to correct.

Acknowledgment. The author would like to thank L. Jackson and the referee for their help with this paper.

REFERENCES

[1] P. HARTMAN, *Difference equations: disconjugacy, principal solutions, Green's functions, complete monotonicity*, Trans. Amer. Math. Soc., 246 (1978), pp. 1–30.

[2] K. MILLER, *Linear Difference Equations*, W. A. Benjamin, New York, 1968.

[3] N. E. NORLUND, *Vorlesungen über Differenzenrechnung*, Lecture Notes 13, Springer-Verlag, Berlin, 1924.

[4] A. C. PETERSON, *On the sign of Green's functions*, J. Differential Equations, 21 (1976), pp. 167–178.

[5] ———, *Existence-uniqueness for ordinary differential equations*, J. Math. Anal. Appl., 64 (1978), pp. 166–171.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1501-0009 $01.25/0

EIGENVALUES OF ANALYTIC KERNELS*

G. LITTLE[†] AND J. B. READE[†]

Abstract. It is shown that the eigenvalues of an analytic kernel on a finite interval go to zero at least as fast as R^{-n} for some fixed $R<1$. The best possible value of R is related to the domain of analyticity of the kernel. The method is to apply the Weyl–Courant minimax principle to the tail of the Chebyshev expansion for the kernel. An example involving Legendre polynomials is given for which R is critical.

Key words. eigenvalue, integral equation

Introduction. Let E_R denote the ellipse with foci at ±1 and semi-axis sum $R>1$. We prove the following theorem.

If $K(x,t)=K(t,x)\in C[-1,1]^2$, and for each $t\in[-1,1]$ there is an analytic continuation to $K(z,t)$ for z inside E_R, which is uniformly bounded in z,t in this range, and if the operator

$$Tf(x)=\int_{-1}^{1}K(x,t)f(t)\,dt$$

has eigenvalues

$$|\lambda_1|\geq|\lambda_2|\geq\cdots\geq|\lambda_n|\geq\cdots,$$

then $\lambda_n=O(R^{-n})$.

This improves on the estimate $O(R^{-n/4})$ obtained by Hille and Tamarkin in 1931 using infinite determinants. Our method is to use Chebyshev polynomials to approximate $K(x,t)$ on $[-1,1]^2$ by a kernel of finite rank, and to relate the operator norm of the difference kernel to the nth eigenvalue of $K(x,t)$ by means of the Weyl–Courant minimax principle. We give an example to show our estimate is best possible by nth powers.

1. The Weyl–Courant minimax principle.

LEMMA 1. *If T is any compact symmetric operator on a Hilbert space H with eigenvalues*

$$|\lambda_1|\geq|\lambda_2|\geq\cdots\geq|\lambda_n|\geq\cdots,$$

and if S is any operator of rank $\leq n$, then

$$\|T-S\|\geq|\lambda_{n+1}|.$$

Proof. Let (ϕ_n) be orthonormal eigenfunctions corresponding to (λ_n). Then we can choose

$$\phi=\alpha_1\phi_1+\alpha_2\phi_2+\cdots+\alpha_{n+1}\phi_{n+1}$$

*Received by the editors June 2, 1982.

[†] Department of Mathematics, The University, Manchester, England M13 9PL.

with $\|\phi\|=1$ such that $S\phi=0$. Therefore

$$\begin{aligned}\|(T-S)\phi\|^2=\|T\phi\|^2&=\|\lambda_1\alpha_1\phi_1+\lambda_2\alpha_2\phi_2+\cdots+\lambda_{n+1}\alpha_{n+1}\phi_{n+1}\|^2\\&=|\lambda_1\alpha_1|^2+|\lambda_2\alpha_2|^2+\cdots+|\lambda_{n+1}\alpha_{n+1}|^2\\&\geq|\lambda_{n+1}|^2\left(|\alpha_1|^2+|\alpha_2|^2+\cdots+|\alpha_{n+1}|^2\right)\\&=|\lambda_{n+1}|^2.\end{aligned}$$

The lemma follows.

2. Chebyshev expansions. Let $T_n(\cos\theta)=\cos n\theta$ denote the nth Chebyshev polynomial.

LEMMA 2. *If $f(z)$ is analytic inside E_R, then $f(z)$ has an expansion in Chebyshev polynomials*

$$f(z)=\frac{1}{2}a_0+\sum_1^\infty a_nT_n(z)$$

valid for z inside E_R. If $f(z)$ is bounded inside E_R, then $a_n=O(R^{-n})$.

Proof. For z inside E_R we have $z=\frac{1}{2}(w+w^{-1})$ where $R^{-1}<|w|<R$. Therefore $2f(\frac{1}{2}(w+w^{-1}))$ is analytic for all w satisfying $R^{-1}<|w|<R$, and so has a Laurent expansion

$$2f\left(\frac{1}{2}(w+w^{-1})\right)=\sum_{-\infty}^{\infty}a_nw^n$$

valid in this range, where

$$a_n=\frac{1}{\pi i}\int_C f\left(\frac{1}{2}(w+w^{-1})\right)w^{-n-1}dw,$$

C being any contour lying in $R^{-1}<|w|<R$ which circulates the origin once positively. Clearly $a_{-n}=a_n$, and so

$$2f\left(\frac{1}{2}(w+w^{-1})\right)=a_0+\sum_1^\infty a_n(w^n+w^{-n})=a_0+2\sum_1^\infty a_nT_n\left(\frac{1}{2}(w+w^{-1})\right)$$

for all $R^{-1}<|w|<R$. Hence

$$f(z)=\frac{1}{2}a_0+\sum_1^\infty a_nT_n(z)$$

for all z inside E_R. Taking C to be the circle with center at the origin and radius r satisfying $R^{-1}<r<R$, we have

$$|a_n|\leq 2Mr^{-n},$$

where $M=\sup|f(z)|$ over z inside E_R. Hence, if we let $r\to R$, we obtain

$$|a_n|\leq 2MR^{-n}.$$

3. Proof of the theorem. By Lemma 2 we have

$$K(x,t)=\frac{1}{2}a_0(t)+\sum_1^\infty a_n(t)T_n(x)$$

for all $x,t\in[-1,1]$, where $a_n(t)=O(R^{-n})$ uniformly in t. Taking the contour C of the proof of Lemma 2 to be the unit circle, we obtain

$$a_n(t)=\frac{1}{\pi}\int_{-\pi}^{\pi}K(\cos\theta,t)e^{-in\theta}d\theta,$$

which shows that $a_n(t)\in C[-1,1]$. Therefore, if we define

$$S_n(x,t)=\frac{1}{2}a_0(t)+\sum_1^n a_k(t)T_k(x),$$

we have a continuous kernel of rank $\leq n+1$. Also

$$|K(x,t)-S_n(x,t)|=\left|\sum_{n+1}^\infty a_k(t)T_k(x)\right|\leq\sum_{n+1}^\infty|a_k(t)|=\sum_{n+1}^\infty O(R^{-k})=O(R^{-n}).$$

Hence by Lemma 1, we have $\lambda_{n+2}=O(R^{-n})$, which gives the result.

4. Legendre polynomials. Let

$$P_n(z)=\frac{1}{2^n n!}\frac{d^n}{dz^n}(z^2-1)^n$$

denote the nth Legendre polynomial.

LEMMA 3. *$|P_n(z)|\leq R^n$ for $z\in E_R$.*

Proof. For $z\in E_R$ we have $z=\frac{1}{2}(w+w^{-1})$ where $|w|=R$. Therefore

$$P_n(z)=P_n\left(\frac{1}{2}(w+w^{-1})\right)=\frac{1}{\pi}\int_0^\pi\left(\frac{1}{2}(w+w^{-1})+\frac{1}{2}(w-w^{-1})\cos\phi\right)^n d\phi,$$

by Laplace's integral. (See [2, p. 312].) Now

$$\begin{aligned}\left|\frac{1}{2}(w+w^{-1})+\frac{1}{2}(w-w^{-1})\cos\phi\right|&=\left|w\cos^2\frac{1}{2}\phi+w^{-1}\sin^2\frac{1}{2}\phi\right|\\&\leq R\cos^2\frac{1}{2}\phi+R^{-1}\sin^2\frac{1}{2}\phi\\&<R\left(\cos^2\frac{1}{2}\phi+\sin^2\frac{1}{2}\phi\right)=R.\end{aligned}$$

Hence

$$|P_n(z)|\leq\frac{1}{\pi}\int_0^\pi R^n d\phi=R^n.$$

COROLLARY. *The estimate of the theorem cannot be improved to $O(R^{-(1+\varepsilon)n})$ for any $\varepsilon>0$.*

Proof. Consider

$$K(x,t)=\sum_1^\infty n^{-2}R^{-n}P_n(x)P_n(t).$$

$K(x,t)$ satisfies the hypotheses of the theorem and has eigenvalues

$$\lambda_n=\left(n+\frac{1}{2}\right)^{-1}n^{-2}R^{-n}.$$

REFERENCES

[1] E. Hille and J. D. Tamarkin, *On the characteristic values of linear integral equations*, Acta. Math., 57 (1931), pp. 1–76.

[2] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed., Cambridge Univ. Press, Cambridge, 1927.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1501-0010 $01.25/0

EIGENVALUES OF POSITIVE DEFINITE KERNELS II*

J. B. READE†

Abstract. We prove what we conjectured in our earlier paper of the same title [SIAM J. Math. Anal., 14 (1983), pp. 152–157], that the eigenvalues of any p times continuously differentiable positive definite kernel are $o(1/n^{p+1})$. The method is the same as we used to prove the case $p=1$ except that we now approximate the kernel by trigonometric polynomials obtained from certain combinations of Jackson kernels.

1. Introduction. Suppose that the real kernel $K(x,t)$ has continuous pth order partial derivatives and is 2π-periodic in x,t. Suppose also that $K(x,t)$ is symmetric and positive definite, so that the operator

$$Tf(x)=\frac{1}{2\pi}\int_{-\pi}^{\pi}K(x,t)f(t)\,dt$$

on the Hilbert space $L^2[-\pi,\pi]$ has positive eigenvalues (λ_n) which can be arranged in a decreasing sequence converging to zero. We show

$$\lambda_n=o\left(\frac{1}{n^{p+1}}\right)$$

as $n\to\infty$.

We give the details for the case $p=2$. The generalisation to $p>2$ involves no essentially new ideas. The method can also be applied when $p=1$, though the proof in [1] is considerably simpler.

2. C^2 functions. We say the function of one variable $f(x)\in C^2$ if $f(x)$ has a continuous second derivative.

LEMMA 1. *If $f(x)\in C^2$ is 2π-periodic, then*

$$\frac{1}{6}(f(x+2h)+f(x-2h))-\frac{2}{3}(f(x+h)+f(x-h))+f(x)=o(h^2)$$

uniformly in x as $h\to 0$.

Proof. Given $\varepsilon>0$, choose $\delta>0$ such that

$$|f''(x)-f''(y)|<\varepsilon$$

whenever $|x-y|<\delta$. Then, by the second mean value theorem, we have

$$\left|\frac{1}{6}(f(x+2h)+f(x-2h))-\frac{2}{3}(f(x+h)+f(x-h))+f(x)\right|$$

$$=\frac{1}{3}h^2|f''(x+2h\theta_1)+f''(x-2h\theta_2)-f''(x+h\theta_3)-f''(x-h\theta_4)|,$$

for some $0<\theta_1,\theta_2,\theta_3,\theta_4<1$,

$$<\frac{2}{3}\varepsilon h^2$$

for all $|h|<\delta/2$.

*Received by the editors July 15, 1982 and in final form September 24, 1982.

†Department of Mathematics, The University, Manchester, England M13 9PL.

3. Jackson kernels. Let

$$J_n(t)=\frac{3}{n(2n^2+1)}\frac{\sin^4(nt/2)}{\sin^4(t/2)}$$

be the nth Jackson kernel. The fact that

$$\frac{\sin^2(nt/2)}{\sin^2(t/2)}=\sum_{k=-n+1}^{n-1}(n-|k|)e^{ikt}$$

shows that $J_n(t)$ is a trigonometric polynomial of degree $2n-2$, and also that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}J_n(t)\,dt=1.$$

LEMMA 2.

$$\int_{-\pi}^{\pi}t^2J_n(t)\,dt=O\left(\frac{1}{n^2}\right).$$

Proof.

$$\int_{-\pi}^{\pi}t^2\frac{\sin^4(nt/2)}{\sin^4(t/2)}dt\le\pi^4\int_{-\pi}^{\pi}\frac{\sin^4(nt/2)}{t^2}dt,$$

since $\sin t>2t/\pi$ for all $0<t<\pi/2$,

$$=\pi^4n\int_{-n\pi}^{n\pi}\frac{\sin^4(u/2)}{u^2}du,$$

putting $u=nt$,

$$=O(n).$$

LEMMA 3. *For any symmetric 2π-periodic continuous kernel $K(x,t)$,*

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}K(x,2x-t)J_n(x-t)\,dx\,dt=\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}K(x,t)J_n(x-t)\,dx\,dt,$$

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}K(x,3x-2t)J_n(x-t)\,dx\,dt=\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}K(x,2t-x)J_n(x-t)\,dx\,dt$$

$$=\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}K(x,t)H_n(x-t)\,dx\,dt,$$

where

$$H_n(t)=\sum_k a_{2k}e^{ikt}$$

if

$$J_n(t)=\sum_k a_ke^{ikt}.$$

Proof.

$$\int_{-\pi}^{\pi}K(x,2x-t)J_n(x-t)\,dt=\int_{2x-\pi}^{2x+\pi}K(x,u)J_n(u-x)\,du,$$

putting $u=2x-t$,

$$=\int_{-\pi}^{\pi} K(x,t)J_n(x-t)\,dt,$$

since $K(x,t)$ is 2π-periodic, which gives the first identity. Using the same substitution we have

$$\int_{-\pi}^{\pi} K(x,3x-2t)J_n(x-t)\,dt=\int_{2x-\pi}^{2x+\pi} K(x,2u-x)J_n(u-x)\,du$$
$$=\int_{-\pi}^{\pi} K(x,2t-x)J_n(x-t)\,dt,$$

which gives the first half of the second identity. To prove the second half of the second identity we observe firstly that, if k is odd,

$$\int_{-\pi}^{\pi} K(x,2t-x)e^{ikt}\,dt=-\int_{0}^{2\pi} K(x,2u-x)e^{iku}\,du,$$

putting $u=t+\pi$,

$$=-\int_{-\pi}^{\pi} K(x,2t-x)e^{ikt}\,dt$$
$$=0,$$

whilst, if k is even,

$$\int_{-\pi}^{\pi} K(x,2t-x)e^{ikt}\,dt=\frac{1}{2}\int_{-2\pi-x}^{2\pi-x} K(x,u)e^{ik(u+x)/2}\,du,$$

putting $u=2t-x$,

$$=\int_{-\pi}^{\pi} K(x,t)e^{ik(t+x)/2}\,dt.$$

Therefore

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} K(x,2t-x)e^{ik(x-t)}\,dx\,dt=0,$$

if k odd,

$$=\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} K(x,t)e^{ik(t-x)/2}\,dx\,dt,$$

if k even. Hence

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} K(x,2t-x)J_n(x-t)\,dx\,dt=\sum_k a_k\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} K(x,2t-x)e^{ik(x-t)}\,dx\,dt$$
$$=\sum_{k\text{ even}} a_k\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} K(x,t)e^{ik(x-t)/2}\,dx\,dt$$
$$=\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} K(x,t)H_n(x-t)\,dx\,dt.$$

LEMMA 4. *If R_n is the operator on $L^2[-\pi,\pi]$ with kernel*

$$\frac{4}{3}J_n(x-t)-\frac{1}{3}H_n(x-t)$$

then $R_n \leq I$, the identity operator.

Proof. It is sufficient to prove that the Fourier coefficients of

$$\frac{4}{3}J_n(t)-\frac{1}{3}H_n(t)$$

are all ≤ 1. If c_k is the kth Fourier coefficient, then

$$\begin{aligned}1-c_k&=1-\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{4}{3}J_n(t)-\frac{1}{3}H_n(t)\right)e^{-ikt}dt\\&=\frac{1}{2\pi}\int_{-\pi}^{\pi}J_n(t)\left(1-\frac{4}{3}e^{-ikt}+\frac{1}{3}e^{-2ikt}\right)dt\\&=\frac{1}{12\pi}\int_{-\pi}^{\pi}J_n(t)(e^{ikt/2}-e^{-ikt/2})^4dt\\&=\frac{4}{3\pi}\int_{-\pi}^{\pi}J_n(t)\sin^4\frac{kt}{2}dt\\&\geq 0.\end{aligned}$$

4. Proof of the result in case $p=2$. If S is the positive square root of T, then the operator SR_nS is symmetric and has a continuous kernel (see [1]). Also

$$SR_nS\leq T$$

and so, by Mercer's theorem (see [1]), $T-SR_nS$ has trace norm

$$\begin{aligned}\|T-SR_nS\|_{\mathrm{tr}}&=\frac{1}{2\pi}\int_{-\pi}^{\pi}K(x,x)\,dx-\frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}K(x,t)\left(\frac{4}{3}J_n(x-t)-\frac{1}{3}H_n(x-t)\right)dx\,dt\\&=\frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\left(K(x,x)-\frac{4}{3}K(x,t)+\frac{1}{3}K(x,2t-x)\right)J_n(x-t)\,dx\,dt\\&=\frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\left(\frac{1}{6}(K(x,2t-x)+K(x,3x-2t))\right.\\&\qquad\left.-\frac{2}{3}(K(x,t)+K(x,2x-t))+K(x,x)\right)J_n(x-t)\,dx\,dt\\&=\frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\left(\frac{1}{6}(K(x,x-2u)+K(x,x+2u))\right.\\&\qquad\left.-\frac{2}{3}(K(x,x-u)+K(x,x+u))+K(x,x)\right)J_n(u)\,dx\,du,\end{aligned}$$

putting $u=x-t$ and using the periodicity of $K(x,t)$,

$$=\frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}u^2\phi(x,u)J_n(u)\,dx\,du,$$

where $\phi(x,u)$ is continuous and vanishes when $u=0$, by Lemma 1,

$$=\frac{1}{2\pi^2}\int_{-\pi}^{\pi}\int_0^{\pi}u^2\phi(x,u)J_n(u)\,dx\,du,$$

since $\phi(x,u)$ is even in u. Given $\varepsilon>0$, choose $\delta>0$ such that $|\phi(x,u)|<\varepsilon$ whenever $|u|<\delta$. Then

$$\begin{aligned}\left|\frac{1}{2\pi^2}\int_{-\pi}^{\pi}\int_0^{\delta}u^2\phi(x,u)J_n(u)\,dx\,du\right|&<\frac{\varepsilon}{2\pi^2}\int_{-\pi}^{\pi}\int_0^{\delta}u^2J_n(u)\,dx\,du\\&=\frac{\varepsilon}{\pi}\int_0^{\delta}u^2J_n(u)\,du\\&<\frac{A\varepsilon}{n^2},\end{aligned}$$

where A is an absolute constant, by Lemma 2.

$$\left|\frac{1}{2\pi^2}\int_{-\pi}^{\pi}\int_{\delta}^{\pi}u^2\phi(x,u)J_n(u)\,dx\,du\right|\leq\frac{M}{\pi}\int_{\delta}^{\pi}u^2J_n(u)\,du,$$

where $M=\max|\phi(x,u)|$,

$$\begin{aligned}&\leq\frac{3M}{\pi n(2n^2+1)}\int_{\delta}^{\pi}\frac{u^2\,du}{\sin^4(u/2)}\\&<\frac{\varepsilon}{n^2}\end{aligned}$$

for all $n\geq$ some N. Therefore

$$\|T-SR_nS\|_{\mathrm{tr}}<\frac{(A+1)\varepsilon}{n^2}$$

for all $n\geq N$, and so

$$\|T-SR_nS\|_{\mathrm{tr}}=o\left(\frac{1}{n^2}\right)$$

as $n\to\infty$. However, SR_nS has rank $\leq 4n-3$, and so, by the Weyl–Courant minimax principle for trace norms (see [1]), we have

$$\sum_{4n-2}^{\infty}\lambda_k=o\left(\frac{1}{n^2}\right),$$

which gives

$$\lambda_n=o\left(\frac{1}{n^3}\right).$$

5. The case $p\geq 3$. The proof we have given for $p=2$ readily generalises to $p\geq 3$. One has to use higher order Jackson kernels,

$$J_{pn}(t)=A_{pn}\frac{\sin^{2p}(nt/2)}{\sin^{2p}(t/2)},$$

where A_{pn} is such that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} J_{pn}(t)\,dt=1.$$

The generalisation of Lemma 1 needed is

$$f(x)+\sum_{r=1}^{p}(-1)^r\binom{2p}{p+r}\Big/\binom{2p}{p}(f(x+rh)+f(x-rh))=o(h^p)$$

for 2π-periodic $f\in C^p$. For R_n one takes the operator with kernel

$$2\sum_{r=1}^{p}(-1)^{r-1}\binom{2p}{p+r}\Big/\binom{2p}{p}H_{pnr}(x-t),$$

where

$$H_{pnr}(t)=\sum_k a_{rk}e^{ikt}$$

if

$$J_{pn}(t)=\sum_k a_k e^{ikt}.$$

REFERENCES

[1] J. B. Reade, *Eigenvalues of positive definite kernels*, this Journal, 14 (1983), pp. 152–157.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1501-0011 $01.25/0

ANALYTIC PROPERTIES OF ARITHMETIC SUMS ARISING IN THE THEORY OF THE CLASSICAL THETA-FUNCTIONS*

BRUCE C. BERNDT† AND LARRY A. GOLDBERG‡

Abstract. In the transformation formulae for the logarithms of the classical theta-functions, there arise certain arithmetic sums that are analogous to Dedekind sums. In this paper, analytic properties of these arithmetic sums are established. In particular, reciprocity theorems are proved and representations as finite trigonometric sums are given. Moreover, certain infinite series and certain doubly infinite series are evaluated in closed form in terms of these arithmetic sums.

It is well known that the classical Dedekind sums $s(h,k)$ first arose in the transformation formulae of the logarithm of the Dedekind eta-function $\eta(z)$. (For an elaboration of this connection and for basic properties of Dedekind sums, consult the monograph of Rademacher and Grosswald [17].) In contrast to $\operatorname{Log}\eta(z)$, the logarithms of the classical theta-functions $\vartheta_2(0,q)$, $\vartheta_3(0,q)$ and $\vartheta_4(0,q)$ have scarcely been studied. (We use the notation of Whittaker and Watson [19, Chapt. 21] for the theta-functions.) In [5] and [8] we derived the transformation formulae for $\operatorname{Log}\vartheta_n(0,q)$, $n=2,3,4$. There are, in fact, 9 distinct transformation formulae depending upon parities of certain coefficients a,b,c and d in the modular transformation$(az+b)/(cz+d)$. Arising in the transformation formulae are 6 different arithmetic sums, which are thus analogues of $s(h,k)$. If h and k are integers with $k>0$, these 6 sums are defined by

$$(1)\qquad S(h,k)=\sum_{j=1}^{k-1}(-1)^{j+1+[hj/k]},$$

$$s_1(h,k)=\sum_{j=1}^{k}(-1)^{[hj/k]}\left(\left(\frac{j}{k}\right)\right),$$

$$s_2(h,k)=\sum_{j=1}^{k}(-1)^{j}\left(\left(\frac{hj}{k}\right)\right)\left(\left(\frac{j}{k}\right)\right),$$

$$s_3(h,k)=\sum_{j=1}^{k}(-1)^{j}\left(\left(\frac{hj}{k}\right)\right),$$

$$s_4(h,k)=\sum_{j=1}^{k-1}(-1)^{[hj/k]},$$

$$s_5(h,k)=\sum_{j=1}^{k}(-1)^{j+[hj/k]}\left(\left(\frac{j}{k}\right)\right).$$

Here, as usual, $[x]$ denotes the greatest integer not exceeding x, and $((x))=0$ or $x-[x]-\frac{1}{2}$, according as x is or is not an integer, respectively.

Rademacher [15], [16, pp. 578–584] briefly studied $\operatorname{Log}\vartheta_n(0,q)$, $n=2,3,4$. However, his approach was via the Dedekind eta-function, and so the sums defined above were not discerned by Rademacher. Some of these sums, or variants thereof, are

* Received by the editors February 15, 1982.

† Department of Mathematics, University of Illinois, Urbana, Illinois 61801.

‡ Department of Mathematics, Temple University, Philadelphia, Pennsylvania 19122.

mentioned in a paper of Hardy [11, pp. 121–123], [12, pp. 390–392], where reciprocity theorems are stated without proofs. However, Hardy did not observe the connections between his sums and theta-functions.

The sums $S(h,k)$ and $s_j(h,k)$ arise in the theory of $r_s(n)$, the number of representations of n as the sum of s squares. Hardy has established exact formulas for $r_s(n)$, $5 \leq n \leq 8$, and asymptotic formulas for $s > 8$, an account of which may be found in Knopp's book [13, Chapt. 5]. Employing the sums mentioned above, Goldberg [10] has shown that a substantial simplification in Hardy's proof can be effected. These sums also arise in the study of the Fourier coefficients of the reciprocals of $\vartheta_n(0,q)$, $n=2,3,4$ [9].

In this paper, however, we are primarily concerned with analytic properties of $S(h,k)$ and $s_n(h,k)$, $1 \leq n \leq 5$. First, we shall establish infinite trigonometric series representations for $S(h,k)$ and $s_n(h,k)$. Viewed in another way, we evaluate certain infinite series in closed form in terms of $S(h,k)$ and $s_n(h,k)$. Secondly, these infinite series representations are employed in deriving representations of $S(h,k)$ and $s_n(h,k)$ as finite trigonometric sums. Thirdly, it is shown that either type of representation can be utilized to establish reciprocity theorems for our sums. Fourthly, we sum certain nonabsolutely convergent double series in terms of $S(h,k)$ or $s_n(h,k)$. We then use reciprocity theorems to determine the "error" made in inverting the order of summation.

THEOREM 1. *Let h and k denote relatively prime integers with $k>0$. If $h+k$ is odd, then*

$$S(h,k)=\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{1}{2n-1}\tan\left(\frac{\pi h(2n-1)}{2k}\right); \tag{2}$$

if h is even and k is odd, then

$$s_1(h,k)=-\frac{2}{\pi}\sum_{\substack{n=1\\ 2n-1\not\equiv 0\ (\mathrm{mod}\ k)}}^{\infty}\frac{1}{2n-1}\cot\left(\frac{\pi h(2n-1)}{2k}\right); \tag{3}$$

if h is odd and k is even, then

$$s_2(h,k)=-\frac{1}{2\pi}\sum_{\substack{n=1\\ 2n\not\equiv 0\ (\mathrm{mod}\ k)}}^{\infty}\frac{1}{n}\tan\left(\frac{\pi hn}{k}\right); \tag{4}$$

if k is odd, then

$$s_3(h,k)=\frac{1}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\tan\left(\frac{\pi hn}{k}\right); \tag{5}$$

if h is odd, then

$$s_4(h,k)=\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{1}{2n-1}\cot\left(\frac{\pi h(2n-1)}{2k}\right); \tag{6}$$

and if h and k are odd, then

$$s_5(h,k)=\frac{2}{\pi}\sum_{\substack{n=1\\ 2n-1\not\equiv 0\ (\mathrm{mod}\ k)}}^{\infty}\frac{1}{2n-1}\tan\left(\frac{\pi h(2n-1)}{2k}\right). \tag{7}$$

Proof. We prove (2). If we employ in (1) the well-known Fourier expansion

$$(-1)^{[x]}=\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{\sin(2n-1)\pi x}{2n-1},$$

where x is not an integer, we find that

$$S(h,k)=-\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{1}{2n-1}\sum_{j=1}^{k-1}(-1)^j\sin\left(\frac{(2n-1)\pi hj}{k}\right). \tag{8}$$

If $m=(2n-1)h, 2n-1\not\equiv 0 \pmod k$, and h and k are of opposite parity, an elementary calculation gives

$$\sum_{j=1}^{k-1}(-1)^j\sin\left(\frac{\pi mj}{k}\right)=-\tan\left(\frac{\pi m}{2k}\right). \tag{9}$$

Substituting (9) into (8), we establish (2) immediately.

To prove (3), we first observe that

$$s_1(h,k)=\frac{1}{k}\sum_{j=1}^{k-1}j(-1)^{[hj/k]}$$

when h is even. The remainder of the proof is now quite similar to that of (2).

Likewise, the proof of (6) is similar to that of (2).

To prove (7), we first show that

$$s_5(h,k)=\frac{1}{k}\sum_{j=1}^{k-1}j(-1)^{j+[hj/k]}$$

when h and k are odd. Now proceed as in the proofs above.

We sketch the proof of (4). Since h is odd and k is even, we find that

$$s_2(h,k)=\frac{1}{k}\sum_{j=1}^{k-1}(-1)^j j\left(\left(\frac{hj}{k}\right)\right). \tag{10}$$

We next recall that

$$((x))=-\frac{1}{\pi}\sum_{n=1}^{\infty}\frac{\sin(2\pi nx)}{n}. \tag{11}$$

Using (11) in (10) and proceeding as in the proof of (2), we easily complete the proof of (4).

The proof of (5) is like that of (4) and utilizes (11).

A similar representation for $s(h,k)$ was established by Rademacher [14], [16, pp. 26–36] and rediscovered in [3].

We next establish analogues of a familiar representation of $s(h,k)$ as a finite trigonometric sum [17, p. 18].

THEOREM 2. *Let h and k be coprime integers with $k>0$. If $h+k$ is odd, then*

$$S(h,k)=\frac{1}{k}\sum_{j=1}^{k}\tan\left(\frac{\pi h(2j-1)}{2k}\right)\cot\left(\frac{\pi(2j-1)}{2k}\right); \tag{12}$$

if h is even and k is odd, then

$$s_1(h,k)=-\frac{1}{2k}\sum_{\substack{j=1\\ j\neq(k+1)/2}}^{k}\cot\left(\frac{\pi h(2j-1)}{2k}\right)\cot\left(\frac{\pi(2j-1)}{2k}\right); \tag{13}$$

if h is odd and k is even, then

$$s_2(h,k)=-\frac{1}{4k}\sum_{\substack{j=1\\ j\neq k/2}}^{k-1}\tan\left(\frac{\pi hj}{k}\right)\cot\left(\frac{\pi j}{k}\right); \tag{14}$$

if k is odd, then

$$s_3(h,k)=\frac{1}{2k}\sum_{j=1}^{k-1}\tan\left(\frac{\pi hj}{k}\right)\cot\left(\frac{\pi j}{k}\right); \tag{15}$$

if h is odd, then

$$s_4(h,k)=\frac{1}{k}\sum_{j=1}^{k}\cot\left(\frac{\pi h(2j-1)}{2k}\right)\cot\left(\frac{\pi(2j-1)}{2k}\right); \tag{16}$$

and if h and k are odd, then

$$s_5(h,k)=\frac{1}{2k}\sum_{\substack{j=1\\ j\neq(k+1)/2}}^{k}\tan\left(\frac{\pi h(2j-1)}{2k}\right)\cot\left(\frac{\pi(2j-1)}{2k}\right). \tag{17}$$

Proof. We establish (12). From (2),

$$S(h,k)=\frac{2}{\pi}\sum_{n=-\infty}^{\infty}\frac{1}{2n-1}\tan\left(\frac{\pi h(2n-1)}{2k}\right).$$

Now let $n=rk+j$, $-\infty<r<\infty$, $1\leq j\leq k$. After some elementary simplification, we find that

$$S(h,k)=\frac{1}{\pi k}\sum_{j=1}^{k}\tan\left(\frac{\pi h(2j-1)}{2k}\right)\sum_{r=-\infty}^{\infty}\frac{1}{r+(2j-1)/(2k)},$$

where the inner sum is to be interpreted symmetrically. If we now employ the familiar partial fraction decomposition for $\pi\cot(\pi x)$ on the right side above, we deduce (12) at once.

The proofs of (13)–(17) follow precisely along the same lines as the proof of (12), and so we omit them.

Either Theorem 1 or 2 may be employed with contour integration to establish reciprocity theorems for $S(h,k)$ and $s_n(h,k)$, $1\leq n\leq 5$. We shall use Theorem 1 to prove the reciprocity formulas of Theorem 3. The proofs utilizing Theorem 2 are very much akin to a corresponding proof of the reciprocity theorem for $s(h,k)$ found in Rademacher and Grosswald's book [17, pp. 21, 22].

THEOREM 3. *Let h and k be coprime, positive integers. Then if $h+k$ is odd,*

$$S(h,k)+S(k,h)=1; \tag{18}$$

if h and k are odd, then

$$s_5(h,k)+s_5(k,h)=\frac{1}{2}-\frac{1}{2hk}; \tag{19}$$

if h is even, then

$$s_1(h,k)-2s_2(k,h)=\frac{1}{2}-\frac{1}{2}\left(\frac{1}{hk}+\frac{k}{h}\right); \tag{20}$$

if k is odd, then

$$2s_3(h,k)-s_4(k,h)=1-\frac{h}{k}. \tag{21}$$

Proof. Let C_N denote a positively oriented circle of radius R_N, $1 \le N < \infty$, centered at the origin. We assume that the radii R_N increase to ∞ and are chosen so that the poles of $\tan(\pi hz)\tan(\pi kz)$ are at a distance from C_N greater than some fixed positive number for all N. Let

$$I_N=\frac{1}{2\pi i}\int_{C_N}\tan(\pi hz)\tan(\pi kz)\,\frac{dz}{z}.$$

Now on $0<\theta<\pi$, $\tan(Re^{i\theta})$ tends to i boundedly, and on $\pi<\theta<2\pi$, $\tan(Re^{i\theta})$ tends to $-i$ boundedly, as R tends to ∞. Hence, a short calculation shows that

$$\lim_{N\to\infty} I_n=-1. \tag{22}$$

The integrand of I_N has simple poles at $z=(2m-1)/(2h)$, $-\infty<m<\infty$, and at $z=(2n-1)/(2k)$, $-\infty<n<\infty$. The residues are easily found to be

$$-\frac{2}{\pi(2m-1)}\tan\left(\frac{\pi k(2m-1)}{2k}\right), \qquad -\infty<m<\infty,$$

and

$$-\frac{2}{\pi(2n-1)}\tan\left(\frac{\pi h(2n-1)}{2k}\right), \qquad -\infty<n<\infty,$$

respectively. Hence, by the residue theorem,

$$\begin{aligned} I_N= &-\frac{2}{\pi}\sum_{|(2m-1)/(2h)|<R_N}\frac{1}{2m-1}\tan\left(\frac{\pi k(2m-1)}{2h}\right)\\ &-\frac{2}{\pi}\sum_{|(2n-1)/(2k)|<R_N}\frac{1}{2n-1}\tan\left(\frac{\pi h(2n-1)}{2k}\right). \end{aligned} \tag{23}$$

Letting N tend to ∞ in (23) and combining the result with (22), we find that

$$-1=-\frac{4}{\pi}\sum_{m=1}^{\infty}\frac{1}{2m-1}\tan\left(\frac{\pi k(2m-1)}{2h}\right)-\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{1}{2n-1}\tan\left(\frac{\pi h(2n-1)}{2k}\right),$$

which is equivalent to (18) by Theorem 1.

Proofs of (19)–(21) can be given along the same lines as the proof of (18). The calculations in the proofs of (19) and (20) are slightly more difficult because the integrands have double poles as well as simple poles.

The reciprocity theorems (18), (20) and (21) were first discovered by Berndt [5], while (19) was initially observed by Goldberg [8]. Elementary proofs of (18), (20) and (21) have been given by Apostol and Vu [1]. All of these reciprocity formulas are, in fact, special cases of "three-term relations" that have been established by Goldberg [8]. We remark that either of the two methods of contour integration to which we referred above can be extended to produce three-term relations. Three-term relations for Dedekind sums have been proved via contour integration in [4]. Further generalizations of (18) can be found in [6] and [7].

Let h and k denote coprime, positive integers. Define

$$L_1(h,k)=\sum_{\substack{n=1\\(2m-1)}}^{\infty}\sum_{\substack{m=1\\k\neq(2n-1)h}}^{\infty}\frac{1}{((2m-1)k)^2-((2n-1)h)^2},$$

$$L_2(h,k)=\sum_{\substack{n=1\\2mk\neq}}^{\infty}\sum_{\substack{m=1\\(2n-1)h}}^{\infty}\frac{1}{(2mk)^2-((2n-1)h)^2},$$

$$L_3(h,k)=\sum_{\substack{n=1\\(2m-1)}}^{\infty}\sum_{\substack{m=1\\k\neq2nh}}^{\infty}\frac{1}{((2m-1)k)^2-(2nh)^2}.$$

In the next theorem, we shall evaluate these conditionally convergent double series in terms of the sums $S(h,k)$ and $s_n(h,k)$, $1\leq n\leq 5$. In Berndt's paper [2], similar series are evaluated in terms of Dedekind sums. We could use the same method here. However, we shall use a suggestion communicated to us by Sczech [18], instead.

THEOREM 4. *Let* $(h,k)=1$ *with* $h,k>0$. *Then*

$$(24)\qquad L_1(h,k)=\begin{cases}\dfrac{\pi^2}{16hk}S(h,k) & \text{if } h+k \text{ is odd},\\[2ex] \dfrac{\pi^2}{8hk}s_5(h,k)+\dfrac{\pi^2}{32h^2k^2} & \text{if } h \text{ and } k \text{ are odd},\end{cases}$$

$$(25)\qquad L_2(h,k)=\begin{cases}-\dfrac{\pi^2}{16hk}s_4(h,k)+\dfrac{\pi^2}{16h^2} & \text{if } h \text{ is odd},\\[2ex] \dfrac{\pi^2}{8hk}s_1(h,k)+\dfrac{\pi^2}{16h^2}\left(1+\dfrac{1}{2k^2}\right) & \text{if } h \text{ is even},\end{cases}$$

$$(26)\qquad L_3(h,k)=\begin{cases}\dfrac{\pi^2}{8hk}s_3(h,k) & \text{if } k \text{ is odd},\\[2ex] -\dfrac{\pi^2}{4hk}s_2(h,k)+\dfrac{\pi^2}{32h^2k^2} & \text{if } k \text{ is even}.\end{cases}$$

Proof. We shall prove only (25). The proofs of (24) and (26) follow along the same lines and, in some instances, are simpler.

Suppose first that h is odd. Then $2mk\neq(2n-1)h$ for each pair m,n of integers. Thus,

$$\begin{aligned}L_2(h,k)&=\frac{1}{8k^2}\sum_{n=1}^{\infty}\left\{\sum_{m=-\infty}^{\infty}\frac{1}{m^2-\{(2n-1)h/(2k)\}^2}+\frac{1}{\{(2n-1)h/(2k)\}^2}\right\}\\&=-\frac{\pi}{4hk}\sum_{n=1}^{\infty}\frac{1}{2n-1}\cot\left(\frac{\pi h(2n-1)}{2k}\right)+\frac{1}{2h^2}\sum_{n=1}^{\infty}\frac{1}{(2n-1)^2}\\&=-\frac{\pi^2}{16hk}s_4(h,k)+\frac{\pi^2}{16h^2},\end{aligned}$$

by (6).

Next, assume that h is even and write

$$(27)\quad L_2(h,k)=\sum_{\substack{n=1\\2n-1\not\equiv 0\,(\mathrm{mod}\,k)}}^{\infty}\sum_{m=1}^{\infty}\frac{1}{(2mk)^2-((2n-1)h)^2}$$
$$+\sum_{\substack{n=1\\2n-1\equiv 0\,(\mathrm{mod}\,k)}}^{\infty}\sum_{\substack{m=1\\2mk\neq(2n-1)h}}^{\infty}\frac{1}{(2mk)^2-((2n-1)h)^2}$$
$$=S_1+S_2,$$

say.

By the same argument as in the first case,

$$(28)\quad S_1=-\frac{\pi}{4hk}\sum_{\substack{n=1\\2n-1\not\equiv 0\,(\mathrm{mod}\,k)}}^{\infty}\frac{1}{2n-1}\cot\left(\frac{\pi h(2n-1)}{2k}\right)$$
$$+\frac{1}{2h^2}\sum_{\substack{n=1\\2n-1\not\equiv 0\,(\mathrm{mod}\,k)}}^{\infty}\frac{1}{(2n-1)^2}$$
$$=\frac{\pi^2}{8hk}s_1(h,k)+\frac{1}{2h^2}\left\{\sum_{n=1}^{\infty}\frac{1}{(2n-1)^2}-\frac{1}{k^2}\sum_{n=1}^{\infty}\frac{1}{(2n-1)^2}\right\}$$
$$=\frac{\pi^2}{8hk}s_1(h,k)+\frac{\pi^2}{16h^2}\left(1-\frac{1}{k^2}\right).$$

In S_2, set $2n-1=(2j-1)k$, $1\leq j<\infty$, to get

$$(29)\quad S_2=\frac{1}{4k^2}\sum_{j=1}^{\infty}\sum_{\substack{m=1\\2m\neq(2j-1)h}}^{\infty}\frac{1}{m^2-\{(2j-1)h/2\}^2}$$
$$=\frac{1}{8k^2}\sum_{\substack{j=1\\\pm 2m\neq(2j-1)h}}^{\infty}\left\{\sum_{m=-\infty}^{\infty}\frac{1}{m^2-\{(2j-1)h/2\}^2}+\frac{1}{\{(2j-1)h/2\}^2}\right\}$$
$$=\frac{1}{4h^2k^2}\sum_{j=1}^{\infty}\frac{1}{(2j-1)^2}+\frac{1}{2h^2k^2}\sum_{j=1}^{\infty}\frac{1}{(2j-1)^2}$$
$$=\frac{3\pi^2}{32h^2k^2}.$$

Putting (28) and (29) in (27), we complete the proof.

COROLLARY 5. *Let h and k denote coprime, positive integers. Then*

$$(30)\quad L_1(h,k)+L_1(k,h)=\frac{\pi^2}{16hk},$$

$$(31)\quad L_2(h,k)+L_3(k,h)=\frac{\pi^2}{16hk}.$$

Proof. To prove (30), combine the reciprocity theorems (18) and (19) with the evaluations (24). Similarly, to establish (31), combine the reciprocity formulas (20) and (21) with the evaluations (25) and (26).

Equalities (30) and (31) imply that the order of summation in $L_n(h,k)$, $1 \leq n \leq 3$, may not be inverted. Moreover, (30) and (31) indicate precisely the "error" made in such an inversion. Thus, interchanging m and n in $L_1(k,h)$ and $L_3(k,h)$, we find that, respectively,

$$\sum_{\substack{n=1 \\ (2m-1)}}^{\infty} \sum_{\substack{m=1 \\ k \neq (2n-1)h}}^{\infty} \frac{1}{((2m-1)k)^2-((2n-1)h)^2} - \sum_{\substack{m=1 \\ (2m-1)}}^{\infty} \sum_{\substack{n=1 \\ k \neq (2n-1)h}}^{\infty} \frac{1}{((2m-1)k)^2-((2n-1)h)^2} = \frac{\pi^2}{16hk}$$

and

$$\sum_{\substack{n=1 \\ 2mk \neq}}^{\infty} \sum_{\substack{m=1 \\ (2n-1)h}}^{\infty} \frac{1}{(2mk)^2-((2n-1)h)^2} - \sum_{\substack{m=1 \\ 2mk \neq}}^{\infty} \sum_{\substack{n=1 \\ (2n-1)h}}^{\infty} \frac{1}{(2mk)^2-((2n-1)h)^2} = \frac{\pi^2}{16hk}.$$

REFERENCES

[1] T. M. Apostol and T. H. Vu, *Elementary proofs of Berndt's reciprocity laws*, Pacific J. Math, 98 (1982), pp. 17–23.

[2] B. C. Berndt, *The evaluation of certain classes of nonabsolutely convergent double series*, this Journal, 6 (1975), pp. 966–977.

[3] ______, *Dedekind sums and a paper of G. H. Hardy*, J. London Math. Soc., (2) 13 (1976), pp. 129–136.

[4] ______, *Reciprocity theorems for Dedekind sums and generalizations*, Advances in Math., 23 (1977), pp. 285–316.

[5] ______, *Analytic Eisenstein series, theta-functions, and series relations in the spirit of Ramanujan*, J. Reine Angew. Math., 303/304 (1978), pp. 332–365.

[6] B. C. Berndt and U. Dieter, *Sums involving the greatest integer function and Riemann-Stieltjes integration*, J. Reine Angew. Math., 337 (1982), pp. 208–220.

[7] B. C. Berndt and R. J. Evans, *Problem E2758, solutions by D. M. Broline, F. S. Cater, L. Carlitz, L. L. Foster, F. D. Hammer, L. E. Mattics, and J. Silverman*, Amer. Math. Monthly, 87 (1980), pp. 404–405.

[8] L. A. Goldberg, *Analogues of the Petersson-Knopp identity for Dedekind sums, and other identities*, in preparation.

[9] ______, *On the Fourier coefficients of the reciprocals of theta-functions*, in preparation.

[10] ______, *Transformations of theta-functions and analogues of Dedekind sums*, Thesis, University of Illinois, Urbana 1981.

[11] G. H. Hardy, *On certain series of discontinuous functions connected with the modular functions*, Quart. J. Math., 36 (1905), pp. 93–123.

[12] ______, *Collected Papers*, vol. IV, Clarendon Press, Oxford, 1969.

[13] M. I. Knopp, *Modular Functions in Analytic Number Theory*, Markham, Chicago, 1970.

[14] H. Rademacher, *Egy Reciprocitásképletröl a Modulfüggevenyek Elméletéböl*, Mat. Fiz. Lapok, 40 (1933), pp. 24–34.

[15] ______, *Über die Transformation der Logarithmen der Thetafunktionen*, Math. Ann., 168 (1967), pp. 142–148.

[16] ______, *Collected Papers*, vol. II, MIT Press, Cambridge, MA, 1974.

[17] H. Rademacher and E. Grosswald, *Dedekind Sums*, Math. Assoc. of America, Washington, DC, 1972.

[18] R. Sczech, personal correspondence, November 25, 1978.

[19] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed., Cambridge Univ. Press, Cambridge, 1962.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics

0036-1410/84/1501-0012 $01.25/0

BANDWIDTH VERSUS TIME CONCENTRATION: THE HEISENBERG-PAULI-WEYL INEQUALITY*

MICHAEL G. COWLING[†] AND JOHN F. PRICE[‡]

Abstract. The main result is that for quite general weight functions v, w

$$\|f\|_2 \leq K\left(\|vf\|_p + \|w\hat{f}\|_q\right)$$

for all tempered distributions f for which, roughly speaking, the right side makes sense, where $1 \leq p, q \leq \infty$, K is a constant independent of f, and $\hat{f}$ is the Fourier transform of f. As a corollary, if $\theta, \phi \geq 0$ satisfy $\theta > 1/p^{\#}$ and $\phi > 1/q^{\#}$, where $t^{\#} = 2t/(t-2)$, there exists $K = K(p, q, \theta, \phi)$ such that

$$\|f\|_2 \leq K\left(\left\||x|^{\theta} f\right\|_p + \left\||y|^{\phi} \hat{f}\right\|_q\right) \tag{1}$$

for all f. In this case the inequality is equivalent to $\|f\|_2 \leq K\alpha^{-\alpha}(1-\alpha)^{\alpha-1}\||x|^{\theta}f\|_p^{\alpha}\||y|^{\phi}\hat{f}\|_q^{1-\alpha}$ where α satisfies $\alpha(\theta - 1/p^{\#}) = (1-\alpha)(\phi - 1/q^{\#})$. Hence it generalizes the classical uncertainty principle inequality (which is the case $p = q = 2$ and $\theta = \phi = 1$) and an inequality due to Hirschman (the case $p = q = 2$ and $\theta, \phi > 0$). Also (1) is trivially true when $\theta = 0$ and $p = 2$ or $\phi = 0$ and $q = 2$ and it is shown that it is not possible apart from these three cases.

One of the approaches to the main inequality is as follows: Suppose $s, t \in [1, 2]$ and E, F are subsets of $\mathbb{R}$ of finite measure. For all $f \in L^2$

$$\|f\|_2 \leq K\left\{\left(\int_{E'} |f(x)|^s dx\right)^{1/s} + \left(\int_{F'} |\hat{f}(y)|^t dy\right)^{1/t}\right\}$$

where $K = K(s, t, E, F)$ is independent of f and $'$ denotes complementation.

1. Introduction. Let L^p, $1 \leq p \leq \infty$, denote the usual Lebesgue spaces of complex-valued functions over the real line $\mathbb{R}$; denote their respective norms by $\|\cdot\|_p$. The Fourier transform $\hat{f}$ of f in L^1 is defined by $\hat{f}(y) = \int f(x)e^{-2\pi ixy}dx$. (Unless indicated otherwise, $\int \cdots dx$ will always denote Lebesgue integration over $\mathbb{R}$.) The Fourier transform of f in L^p, $1 < p \leq 2$, as a function in $L^{p'}$ will also be denoted by $\hat{f}$. (Throughout p' will be the usual conjugate exponent of p.) The starting point for this paper is the well-known inequality

$$\|xf\|_2 \cdot \|y\hat{f}\|_2 \geq (4\pi)^{-1}\|f\|_2^2, \qquad f \in L^2. \tag{1.1}$$

This inequality is of fundamental importance in quantum mechanics. In this case it is usual to normalize f so that $\|f\|_2 = 1$; as such it represents the state of a one-dimensional system. Proceeding with this interpretation, the first and second norms in (1.1) represent the standard deviations of the position and momentum observables (assuming they both have mean zero). In this way the inequality becomes the mathematical formulation of the quantum mechanical uncertainty principle first described by Heisenberg [9] in 1927. (The precise version (1.1) appears in Weyl [18, p. 77] where it is attributed to Pauli.)

Recently Fefferman and Phong [6] have given an application to the theory of partial differential equations.

The inequality is also of considerable importance in signal analysis [3], [13], [16] where it is sometimes referred to as the bandwidth theorem. Its role in this area is to

* Received by the editors December 22, 1982.

[†] Istituto di Matematica, Università di Genova, 16132 Genova, Italia. This author gratefully acknowledges the support of the University of New South Wales during the preparation of this paper.

[‡] School of Mathematics, University of New South Wales, Kensington, NSW 2033, Australia.

give precision to the statement that for signals of equal strengths (that is, equal L^2-norms), the more a signal is "concentrated" in time, the more its band is "dispersed", and vice versa. This interpretation seems to have been first pointed out by Gabor [7].

One drawback when attempting to apply the result in this area is that it only provides a reciprocity relation between time and frequency concentration for functions that decrease fairly rapidly to zero at infinity. With this in mind we investigate inequalities of the form (1.1) but with $\|vf\|_p$ and $\|w\hat{f}\|_q$ replacing the norms on the right, where v and w are nonnegative measurable functions. The general nature of our results is best seen by restricting the main theorems to the case where $v, w \colon x \mapsto |x|^\theta$. For example, from Lemma 2.1 and Theorem 5.1 we have

THEOREM 1.1. *Suppose* $1 \le p \le \infty$ *and* $0 \le \theta < \infty$. *Then there exists* K *such that*

$$\|f\|_2^2 \le K \big\| |x|^\theta f \big\|_p \big\| |y|^\theta \hat{f} \big\|_p \tag{1.2}$$

for all f *in* L^2 *provided* $\theta > 1/p^{\#} = (p-2)/2p$. *Otherwise no such inequality is possible* (*apart from* $\theta = 0$ *and* $p = 2$ *in which case both sides are equal with* $K = 1$).

In the sequel it turns out to be more fruitful to analyze a modification of the above form, namely

$$\|f\|_2 \le K\big(\|vf\|_p + \|w\hat{f}\|_q\big), \qquad f \in L^2. \tag{1.3}$$

In many cases this is equivalent to the corresponding "multiplicative" inequality as evidenced by Lemma 2.1. However, with the latter type we can go even further and show that no assumptions need be placed on the weights (and hence on f and its Fourier transform) in a neighbourhood of the origin (Corollaries 2.3 and 2.4). This further illustrates the direct relationship between the asymptotic behaviour of a function and the local smoothness of its transform.

When $\theta = 1$ and $p = 2$, inequality (1.2) is just the classical case (1.1). When $p = 2$ and $0 < \theta < \infty$, (1.2) is due to Hirschman [11]. To complete this introduction we outline a proof (of a mild extension) of the classical inequality.

THEOREM 1.2. *If* $1 \le p \le 2$ *and* $f \in L^2$ *is nonzero, then*

$$\|f\|_2^2 \le 4\pi \|xf\|_p \|y\hat{f}\|_p,$$

with equality if and only if $p = 2$ *and* f *is a constant multiple of* $\exp(kx^2)$ *with* $k < 0$.

Proof. It is enough to consider f in the Schwarz space $\mathcal{S}$. Since $2\pi i y \hat{f} = (f')^\wedge$, the right side becomes $2\|xf\|_p \|(f')^\wedge\|_p$. Hence

$$4\pi \|xf\|_p \|y\hat{f}\|_p \ge 2 \int |x\bar{f} f'|\, dx \ge \int x\big(|f^2|\big)'\, dx = \int |f|^2 dx,$$

as required, where the first step is the Hausdorff–Young inequality followed by Hölder's inequality.

If $p = 2$, it is easily seen that the constant is attained when and only when f as described in the statement. A similar analysis of the inequalities in the preceding paragraph when $1 \le p < 2$ shows that in this case the inequality is always strict. □

Better constants. The sharp form of the Hausdorff–Young inequality (due to Babenko and to Beckner [1]) asserts that when $1 < p < 2$ and $k_p = (p^{1/p}/p'^{1/p'})^{1/2}$, $\|\hat{f}\|_{p'} \le k_p \|f\|_p$ for all f in L^p with equality if and only if f is as in Theorem 1.2. If this is used in the proof of Theorem 1.2, the inequality becomes

$$\|f\|_2^2 \le 4\pi k_p \|xf\|_p \|y\hat{f}\|_p.$$

However it is still the case that this inequality is always strict when $1<p<2$ although it is likely that the constant given is not best possible.

In the sequel the constants we obtain are, in general, probably far from best possible. Hence any improvement due to using the sharp form of the Hausdorff–Young inequality is unlikely to be significant so we only use the classical form.

A useful reference for the classical inequality (1.1) and related results is Dym and McKean [4]. A family of inequalities related to those described in the abstract have been developed in [15] and used to estimate quantum mechanical Hamiltonians.

2. Some inequalities with weights. We shall be interested in inequalities relating $\|f\|_2$, $\||x|^\theta f\|_p$ and $\||y|^\phi \hat{f}\|_q$. Inequalities of the form

$$\|f\|_2 \le K\left\||x|^\theta f\right\|_p^\alpha \left\||y|^\phi \hat{f}\right\|_q^{1-\alpha} \tag{2.1}$$

for all $f \in L^2$ can only be true if the following relation between α, p, q, θ and ϕ holds:

$$\alpha(\theta - 1/p^\#) = (1-\alpha)(\phi - 1/q^\#). \tag{2.2}$$

For otherwise, by replacing f by its "normalized dilate" $D_\lambda f$, where $D_\lambda f(x) = \lambda^{-1/2} f(x/\lambda)$, and simplifying the expressions obtained, we may deduce that

$$\|f\|_2 \le K\lambda^{\alpha(\theta-1/p^\#)}\left\||x|^\theta f\right\|_p^\alpha \lambda^{-1(1-\alpha)(\phi-1/q^\#)}\left\||y|^\phi \hat{f}\right\|_q^{1-\alpha},$$

for all λ in $\mathbb{R}^+$. This is false unless the left-hand side is 0 or the right-hand side is $+\infty$. Nevertheless, it is possible to relate $\|f\|_2$ to $\||x|^\theta f\|_p$ and $\||y|^\phi \hat{f}\|_q$ in a different way, which seems more appropriate. We shall consider inequalities of the form

$$\|f\|_2 \le K\alpha^\alpha(1-\alpha)^{1-\alpha}\left\{\left\||x|^\theta f\right\|_p + \left\||y|^\phi \hat{f}\right\|_q\right\} \tag{2.3}$$

for all f in L^2. If (2.2) is verified, then (2.3) is equivalent to (2.1), while if (2.2) does not hold, we still obtain some information. The scope of Lemma 2.1 (below) is to show that (2.1) and (2.3) are equivalent if (2.2) holds. This ties our work to that of Hirschman [11], who treated the case where $p=2$, with the multiplicative inequality.

We shall actually work in the more general context of inequalities of the form

$$\|f\|_2 \le C\{\|vf\|_p + \|w\hat{f}\|_q\} \quad \text{for all } f \in L^2(\mathbb{R}), \tag{2.4}$$

and our first theorem about these follows Lemma 2.1.

LEMMA 2.1. *Suppose that* $1 \le p, q \le \infty$, *that* $0 \le \theta, \phi < \infty$, *that* $0 < \alpha < 1$, *and that* (2.2) *holds. Then the following inequalities are equivalent* (*where in each case f ranges over* L^2):

(i) $$\|f\|_2 \le K\left\||x|^\theta f\right\|_p^\alpha \left\||y|^\phi \hat{f}\right\|_q^{1-\alpha},$$

(ii) $$\|f\|_2 \le K\left\{\alpha\left\||x|^\theta f\right\|_p + (1-\alpha)\left\||y|^\phi \hat{f}\right\|_q\right\},$$

(iii) $$\|f\|_2 \le K\alpha^\alpha(1-\alpha)^{1-\alpha}\left\{\left\||x|^\theta f\right\|_p + \left\||y|^\phi \hat{f}\right\|_q\right\},$$

(iv) $$\|f\|_2 \le K\alpha^\alpha(1-\alpha)^{1-\alpha}\left\{\delta^{1-\alpha}\left\||x|^\theta f\right\|_p + \delta^{-\alpha}\left\||y|^\phi \hat{f}\right\|_q\right\}$$

for all δ *in* $\mathbb{R}^+$.

Proof. The general inequality, for α in $(0,1)$ and a, b in $\mathbb{R}^+$,

$$a^\alpha b^{1-\alpha} \le \alpha a + (1-\alpha)b$$

shows that (i) implies (ii). Next, if (ii) holds, then by replacing f by $D_\lambda f$ as in the discussion after (2.2), we find that, for all f in L^2,

$$\|f\|_2 \leq K\left\{\alpha\lambda^{\theta-1/p^{\#}}\left\||x|^{\theta}f\right\|_p+(1-\alpha)\lambda^{-\phi+1/q^{\#}}\left\||y|^{\phi}\hat{f}\right\|_q\right\}.$$

Choosing λ such that

$$\alpha\lambda^{\theta-1/p^{\#}}=\alpha^{\alpha}(1-\alpha)^{(1-\alpha)},$$

we obtain (iii). Replacing f by $D_\lambda f$ in (iii), and setting $\delta^{1-\alpha}$ equal to $\lambda^{\theta-1/p^{\#}}$ proves (iv). Finally, if (iv) holds, then we minimize the right-hand side by choosing

$$\delta=\left(\alpha\left\||y|^{\phi}\hat{f}\right\|_q\right)/\left((1-\alpha)\left\||x|^{\theta}f\right\|_p\right),$$

and obtain (i). □

Before we state Theorem 2.2, we recall some of the results of Slepian and Pollak [14] and of Landau and Pollak [12]. Let I_δ be the interval $[-\delta,\delta]$, and suppose that f is a nonzero L^2-function supported in I_δ. Then $\hat{f}$ extends to an entire function in the complex plane. Plancherel's formula tells us that

$$\left(\int dy|\hat{f}(y)|^2\right)^{1/2}=\left(\int dx|f(x)|^2\right)^{1/2},$$

but since f cannot vanish on any set of positive measure,

$$\left(\int_{I_\varepsilon} dy|\hat{f}(y)|^2\right)^{1/2}<\left(\int dx|f(x)|^2\right)^{1/2}.$$

The above mentioned authors quantify this inequality: they prove that there is a function $\gamma\colon \mathbb{R}^+\to(0,1)$ such that

$$\left(\int_{I_\varepsilon} dy|\hat{f}(y)|^2\right)^{1/2}\leq\gamma(\delta\varepsilon)\left(\int dx|f(x)|^2\right)^{1/2} \tag{2.5}$$

for all f in L^2 which vanish off I_δ. They show that $\gamma^2(\delta\varepsilon)$ is the largest eigenvalue of the integral equation

$$\lambda f(x)=\int_{I_\delta} f(w)\frac{\sin(2\pi\varepsilon[x-w])}{\pi[x-w]}dw.$$

In what follows $\mathcal{S}'$ is the space of tempered distributions; it is the dual of $\mathcal{S}$, the space of rapidly decreasing, infinitely differentiable functions. Also whenever $1\leq p\leq\infty$, p' denotes its usual conjugate.

THEOREM 2.2. *Let f be in $\mathcal{S}'$, let δ and ε be positive real numbers, and let s and t be in* $[1,2]$. *Suppose that outside the interval I_δ, f is given by an L^s-function, and that outside I_ε, $\hat{f}$ is given by an L^t-function. Then f is in L^2. Moreover, if*

$$A=\left(\int_{I'_\delta} dx|f(x)|^s\right)^{1/s}, \qquad B=\left(\int_{I'_\varepsilon} dy|\hat{f}(y)|^t\right)^{1/t},$$

then

$$\|f\|_2\leq\left[A^s+(2\delta)^{s\sigma}\alpha^{s}_{.}\right]^{\theta/s}\left[B^t+(2\varepsilon)^{t\tau}\beta^t\right]^{(1-\theta)/t},$$

where

$$\sigma=1/s-1/2,\quad \tau=1/t-1/2,\quad \theta=\tau/(\sigma+\tau),$$
$$\alpha=\left[1-\gamma(\delta\varepsilon)^2\right]^{-1}\left[(2\delta)^\tau B+\gamma(\delta\varepsilon)(2\varepsilon)^\sigma A\right]$$

and

$$\beta=\left[1-\gamma(\delta\varepsilon)^2\right]^{-1}\left[(2\varepsilon)^\sigma A+\gamma(\delta\varepsilon)(2\delta)^\tau B\right].$$

Proof. The proof splits into two stages. First, we show that f and $\hat{f}$ are given by locally square integrable functions, and deduce that $f\cdot\chi_{I_\delta}$ and $\hat{f}\cdot\chi_{I_\varepsilon}$ are square integrable. (χ_E is the characteristic function of E.) Then we estimate the L^2-norm of f.

It is possible to write f as the sum of a compactly-supported distribution and an L^s-function. Then $\hat{f}$ is the sum of a smooth function and an $L^{s'}$-function, so $\hat{f}$ is locally square integrable. Similarly, f is the sum of a smooth function and an $L^{t'}$-function and so is locally square integrable.

Let C and D be the numbers given by the rules

$$C=\left(\int_{I_\delta}dx|f(x)|^2\right)^{1/2},\qquad D=\left(\int_{I_\varepsilon}dy|\hat{f}(y)|^2\right)^{1/2}.$$

Now $f=f\cdot\chi_{I_\delta}+f\cdot\chi_{I'_\delta}$, so $f=(\hat{f}\cdot\chi_{I_\delta})\hat{}+(f\cdot\chi_{I'_\delta})\hat{}$ and $\hat{f}\cdot\chi_{I_\varepsilon}=(f\cdot\chi_{I_\delta})\hat{}\cdot\chi_{I_\varepsilon}+(f\cdot\chi_{I'_\delta})\hat{}\cdot\chi_{I_\varepsilon}$. Therefore

$$\begin{aligned}D&\le\left(\int_{I_\varepsilon}dy|(f\cdot\chi_{I_\delta})\hat{}(y)|^2\right)^{1/2}+\left(\int_{I_\varepsilon}dy|(f\cdot\chi_{I'_\delta})\hat{}(y)|^2\right)^{1/2}\\&\le\gamma(\delta\varepsilon)\|f\cdot\chi_{I_\delta}\|_2+(2\varepsilon)^\sigma\left(\int_{I_\varepsilon}dy|(f\cdot\chi_{I'_\delta})\hat{}(y)|^{s'}\right)^{1/s'}\\&\le\gamma(\delta\varepsilon)C+(2\varepsilon)^\sigma\|f\cdot\chi_{I'_\delta}\|_s,\end{aligned}$$

by the inequality (2.5), Hölder's inequality, and the Hausdorff-Young theorem. Similarly,

$$C\le\gamma(\delta\varepsilon)D+(2\delta)^\tau\|\hat{f}\cdot\chi_{I'_\varepsilon}\|_t.$$

Eliminating D, then C, we obtain the inequalities

$$C\le\left[1-\gamma(\delta\varepsilon)^2\right]\left[(2\varepsilon)^\sigma A\gamma(\delta\varepsilon)+(2\delta)^\tau B\right],\tag{2.6}$$
$$D\le\left[1-\gamma(\delta\varepsilon)^2\right]\left[(2\delta)^\tau B\gamma(\delta\varepsilon)+(2\varepsilon)^\sigma A\right].\tag{2.7}$$

Now we show that f is in L^2 and estimate its norm. First of all, f is in L^s, and

$$\begin{aligned}\|f\|_s&\le\left\{\|f\cdot\chi_{I_\delta}\|_s^s+\|f\cdot\chi_{I'_\delta}\|_s^s\right\}^{1/s}\\&\le\left\{(2\delta)^{\sigma s}\|f\cdot\chi_{I_\delta}\|_2^s+A^s\right\}^{1/s}=\left\{(2\delta)^{\sigma s}C^s+A^s\right\}^{1/s}.\end{aligned}\tag{2.8}$$

Similarly, $\hat{f}$ lies in L^t, and

$$\|\hat{f}\|_t \leq \left\{(2\varepsilon)^{\tau t} D^t + B^t\right\}^{1/t}. \tag{2.9}$$

The Hausdorff–Young theorem implies that f is in $L^{t'}(\mathbb{R})$, and

$$\|f\|_{t'} \leq \|\hat{f}\|_t.$$

Hölder's inequality now allows us to conclude that f is in $L^2(\mathbb{R})$, and

$$\|f\|_2 \leq \|f\|_s^{\theta} \|f\|_{t'}^{1-\theta} \leq \|f\|_s^{\theta} \|\hat{f}\|_t^{1-\theta},$$

with θ as in the enunciation of the theorem. The inequalities (2.6)–(2.9), together with this last inequality, yield the desired conclusion. □

We now present various corollaries of this result. The first of these, Corollary 2.3, is very general, while the second, Corollary 2.4, refers to weights of the form $|x|^{\alpha}$. It is very hard to work with the estimate of Theorem 2.2, and it is not a good estimate, so we shall not bother to keep track of constants in the rest of this section. In what follows, K is a number, independent of the function f involved, which may vary from line to line, and may depend on other parameters.

To state Corollary 2.3 we need a further definition. A (measurable) function v: $\mathbb{R} \to \mathbb{R}^+$ will be said to be *(E,p)-adapted* for some measurable set E in $\mathbb{R}$ if $C(v,E,p)$, given by the formulae

$$C(v,E,p) = \operatorname{ess\,sup}\left\{v(x)^{-1} : x \in E'\right\} \quad \text{if } 1 \leq p \leq 2, \tag{2.10}$$

$$C(v,E,p) = \left(\int_{E'} v(x)^{-p^{\#}} dx\right)^{1/p^{\#}} \quad \text{if } 2 < p \leq \infty, \tag{2.11}$$

is finite.

COROLLARY 2.3. *Suppose that* $1 \leq p,q \leq \infty$, *that* $0 \leq \delta, \varepsilon \leq \infty$, *and that* v: $\mathbb{R} \to \mathbb{R}^+$ *is* $(I_{\delta'} p)$*-adapted while* w: $\mathbb{R} \to \mathbb{R}^+$ *is* (I_ε, q)*-adapted. Let f be a tempered distribution which is given by a locally integrable function off* I_δ *and whose Fourier transform is given by a locally integrable function off* I_ε. *Then, for some constant* K,

$$\|f\|_2 \leq K\left(\|vf \cdot \chi_{I'_\delta}\|_p + \|w\hat{f} \cdot \chi_{I'_\varepsilon}\|_q\right).$$

In particular, $f \in L^2$ *whenever the right side is finite.*

Proof. If $1 \leq p \leq 2$, then

$$\left(\int_{|x| \geq \delta} dx |f(x)|^p\right)^{1/p} \leq \|v^{-1}\chi_{I'_\delta}\|_\infty \left(\int_{|x| \geq \delta} dx |v(x) f(x)|^p\right)^{1/p}$$

while if $p \geq 2$, then

$$\left(\int_{|x| \geq \delta} dx |f(x)|^2\right)^{1/2} \leq \left(\int_{|x| \geq \delta} dx\, v(x)^{-p^{\#}}\right)^{1/p\#} \left(\int_{|x| \geq \delta} dx |v(x) f(x)|^p\right)^{1/p}$$

by Hölder's inequality (with the obvious modification if $p = \infty$). The hypotheses of the corollary therefore imply those of Theorem 2.2 with s equal to $\min(p,2)$ and t equal to $\min(q,2)$. □

COROLLARY 2.4. *Suppose that* $1\leq p,q\leq\infty$, *and that* $0\leq\theta,\phi\leq\infty$. *If* $\theta>1/p^{\#}$ *and* $\phi>1/q^{\#}$, *then the following inequality holds*:

$$\|f\|_2\leq K\left(\left\||x|^{\theta}f\cdot\chi_{I'_\delta}\right\|_p+\left\||y|^{\phi}\hat{f}\cdot\chi_{I'_\varepsilon}\right\|_q\right).$$

This may be interpreted as follows: *if f is a distribution for which the right-hand side makes sense and is finite, then f is in* L^2 *and the inequality holds.*

Proof. The function $x\to|x|^{\theta}$ is (δ,p)-adapted for some positive δ if and only if $\theta>1/p^{\#}$. Thus if the right-hand side of the inequality is finite, we are in the situation dealt with in Corollary 2.3 and Theorem 2.2. The only problem is the evaluation of the constant. □

3. More inequalities. In this section we obtain the central inequality under conditions which are in part more general and in part more restrictive than those assumed for Corollary 2.3. The main difference is that the intervals I_δ and I_ε are replaced by arbitrary sets E_1,E_2 of finite measure. (This will also be done in §4 but here we shall be able to obtain estimates of the constant.) Corollary 2.3 required no a priori assumptions on f and $\hat{f}$ on I_δ and I_ε, respectively, which amounted to allowing the weights to vanish on those sets. In Theorem 3.4 below we do require the weights to satisfy certain mild conditions on E_1 and E_2. In view of the counterexample described in the last section, the hypotheses of Corollary 2.3 and Theorem 3.5 are quite reasonable.

The "bootstrap" methods of this section consist of repeated applications of the Hölder and Hausdorff–Young inequalities, separately applied on E_1,E_2 and their complements. They are well-suited to keeping track of the relevant constants.

Throughout E_1 and E_2 will be sets of finite measure, their measures being denoted by m_1 and m_2, respectively. Whenever the weights u_i: $\mathbb{R}\to\mathbb{R}^+$ $(i=1,2)$ are (E_i,q_i)-adapted, define b_1 and b_2 by

$$(3.1)\qquad b_1=\begin{cases} m_2^{-1/q_1^{\#}}\left\|u_1^{-1}\chi_{E_1'}\right\|_\infty & \text{if } 1\leq q_1\leq 2,\\ \left\|u_1^{-1}\chi_{E_1'}\right\|_{q_1^{\#}} & \text{if } 2<q_1\leq\infty,\end{cases}$$

and similarly for b_2.

LEMMA 3.1. *Let* u_i: $\mathbb{R}\to\mathbb{R}^+$ $(i=1,2)$ *be* (E_i,q_i)-*adapted. Suppose further that the numbers*

$$(3.2)\qquad a_1=m_2^{1/2}\left\|u_1^{-1}\chi_{E_1}\right\|_{q_1'},\qquad a_2=m_1^{1/2}\left\|u_2^{-1}\chi_{E_2}\right\|_{q_2'}$$

are finite. Then there is a constant K such that

$$(3.3)\qquad \|f\|_2\leq K\left(\|u_1f\|_{q_1}+\|u_2f\|_{q_2}\right)$$

for all tempered distributions f such that f and its Fourier transform are given by locally integrable functions. In particular, $f\in L^2$ *whenever the right side of* (3.3) *is finite.*

Constants. As we shall see, this inequality is a consequence of the following: Let $A_1=\|u_1f\|_{q_1}$ and $A_2=\|u_2\hat{f}\|_{q_2}$. If $1\leq q_1$, $q_2\leq 2$, then

$$(3.4)\quad \|f\|_2\leq\left[m_2^{1/q_1^{\#}}b_1A_1+m_1^{-1/q_1^{\#}}(a_2+b_2)A_2\right]^t\left[m_2^{-1/q_2^{\#}}(a_1+b_1)A_1+m_1^{1/q_2^{\#}}b_2A_2\right]^{1-t}$$

where $\frac{1}{2}=t/q_1+(1-t)/q_2'$, while if $1\leq q_1\leq\infty$ and $2<q_2\leq\infty$, then

$$(3.5)\qquad \|f\|_2\leq(a_1+b_1)A_1+b_2A_2.$$

When $2<q_1, q_2\leq\infty$, (3.5) can be replaced by

$$\|f\|_2\leq b_1A_1+(a_2+b_2)A_2. \tag{3.6}$$

Proof. With f as in the statement of the lemma, assume that A_1 and A_2 as defined in previous paragraph are finite. (Otherwise there is nothing to prove.) Our first step is to estimate $\hat{f}$ locally (that is, on E_2) by showing that

$$\|\hat{f}\chi_{E_2}\|_2\leq(a_1+b_1)A_1. \tag{3.7}$$

(Since f and $\hat{f}$ are measurable, all the integrals in the proof are defined and so, by reversing the order of the steps, we see that $f\in L^2$.)

Since $f=f\chi_{E_1}+f\chi_{E_1'}$, $\hat{f}=(f\chi_{E_1})\hat{}+(f\chi_{E_1'})\hat{}$ and so

$$\|\hat{f}\chi_{E_2}\|_2\leq\left(\int_{E_2}|(f\chi_{E_1})\hat{}|^2dy\right)^{1/2}+\left(\int_{E_2}|(f\chi_{E_1'})\hat{}|^2dy\right)^{1/2}. \tag{3.8}$$

Denote the latter two integrals by I_1 and I_2 respectively. Hölder's inequality applied twice and the Hausdorff–Young inequality once yield

$$I_1\leq m_2^{1/2}\|(f\chi_{E_1})\hat{}\|_\infty\leq m_2^{1/2}\|f\chi_{E_1}\|_1=m_2^{1/2}\int_{E_1}|f(x)|u_1(x)u_1(x)^{-1}dx\leq a_1A_1.$$

Assume now $1\leq q_1\leq 2$:

$$I_2\leq m_2^{-1/q_1^\#}\|(f\chi_{E_1'})\hat{}\|_{q_1'}\leq m_2^{-1/q_1^\#}\|f\chi_{E_1'}\|_{q_1}\leq b_1A_1.$$

On the other hand, if $2<q_1\leq\infty$,

$$I_2\leq\|(f\chi_{E_1'})\hat{}\|_2=\|f\chi_{E_1'}\|_2\leq\|fu_1\chi_{E_1'}\|_{2r}\|u_1^{-1}\chi_{E_1'}\|_{2r'}.$$

Choose $r=q_1/2\in(1,\infty]$. Hence $2r=q_1$ and $2r'=2q_1/(q_1-2)=q_1^\#$ so, once again, $I_2\leq b_1A_1$. When used in (3.8), these estimates for I_1 and I_2 establish (3.7), as required.

Our methods now differ for the three cases (3.4), (3.5) and (3.6). First assume $1\leq q_1, q_2\leq 2$; (3.7) is used to estimate $\|\hat{f}\|_{q_2}$ as follows:

$$\begin{aligned}\|\hat{f}\|_{q_2}\leq\|\hat{f}\chi_{E_2}\|_{q_2}+\|\hat{f}\chi_{E_2'}\|_{q_2}&\leq m_2^{-1/q_2^\#}\|\hat{f}\chi_{E_2}\|_2+m_1^{1/q_2^\#}b_2A_2\\&\leq m_2^{-1/q_2^\#}(a_1+b_1)A_1+m_1^{1/q_2^\#}b_2A_2.\end{aligned}$$

A similar inequality holds for $\|f\|_{q_1}$ with the subscripts 1 and 2 interchanged. This allows the final estimate for $\|f\|_2$ to be obtained. In fact, it is possible to estimate $\|f\|_p$ with $q_1\leq p\leq q_2'$:

$$\begin{aligned}\|f\|_p&\leq\|f\|_{q_1}^t\|f\|_{q_2'}^{1-t}\quad\text{where } 1/p=t/q_1+(1-t)/q_2'\\&\leq\|f\|_{q_1}^t\|\hat{f}\|_{q_2}^{1-t}\\&\leq\left[m_2^{1/q_1^\#}b_1A_1+m_1^{-1/q_1^\#}(a_2+b_2)A_2\right]^t\left[m_2^{-1/q_2^\#}(a_1+b_1)A_1+m_1^{1/q_2^\#}b_2A_2\right]^{1-t}.\end{aligned}$$

This yields (3.4) which in turn results in (3.3) with

$$K=\max\left\{m_2^{1/q_1^\#}b_1, m_1^{-1/q_1^\#}(a_2+b_2), m_1^{1/q_2^\#}b_2, m_2^{-1/q_2^\#}(a_1+b_1)\right\}.$$

Now suppose that $2\le q_2\le\infty$. Arguing as in the second estimate for I_2, $\|\hat f\chi_{E_2'}\|\le b_2A_2$. Combining this with (3.7) shows that

$$\|f\|_2=\|\hat f\|_2\le\big\|\hat f\chi_{E_2}\big\|_2+\big\|\hat f\chi_{E_2'}\big\|_2\le(a_1+b_1)A_1+b_2A_2,$$

which is (3.5). To obtain (3.3), take $K=\max\{a_1+b_1,b_2\}$. In the case where both $q_1,q_2\in(2,\infty]$,

$$\|f\|_2\le\min\{(a_1+b_1)A_1+b_2A_2,(a_2+b_2)A_2+b_1A_1\},$$

which completes the proof. □

The following lemma provides the transition from the previous lemma to the main result of the section, Theorem 3.3, by showing that, roughly speaking, the inequality remains valid when the weights are replaced by higher powers of themselves. Its proof is a simple application of Hölder's inequality.

LEMMA 3.2. *Let p,q,β satisfy $1\le p$, $q\le\infty$, $q\ne\infty$, $0<\beta<p/q$ and $(q-2)/\beta q=(p-2)/p$. Then*

$$\|w^\beta\phi\|_q\le\|w\phi\|_p^\beta\|\phi\|_2^{1-\beta}$$

for any measurable function ϕ.

Powers of weights. Suppose that

$$\|f\|_2\le K\big(\|w_1^\beta f\|_{q_1}+\|w_2^\beta\hat f\|_{q_2}\big).$$

By taking ϕ to be f, then $\hat f$, in the preceding lemma we have

$$\|f\|_2\le K^{1/\beta}\big(\|w_1f\|_{p_1}^\beta+\|w_2\hat f\|_{p_2}^\beta\big)^{1/\beta},$$

provided the exponents satisfy $0<\beta<\min\{p_1/q_1,p_2/q_2\}$, $q_i\ne\infty$ and $(q_i-2)/\beta q_i=(p_i-2)/p_i$ for $i=1,2$. On the other hand, if we have the multiplicative version of the inequality, namely

$$\|f\|_2^2\le K\|w_1^\beta f\|_{q_1}\|w_2^\beta\hat f\|_{q_2},$$

then under the conditions just described

$$\|f\|_2\le K^{1/\beta}\|w_1f\|_{p_1}\|w_2\hat f\|_{p_2}.$$

The first case forms the basis for the proof of Theorem 3.3 and the second for its corollary given in Remark 3.4. Before this, however, we illustrate the ideas with the simplest case, namely $p_1=p_2=2$.

The case $p_1=p_2=2$. When $p_1=p_2=2$ in 3.2, $q_1=q_2=2$ and $\beta\in(0,1)$. In combination with results Theorem 1.2 and Lemma 2.1, this yields

$$\|f\|_2^2\le(4\pi)^\alpha\big\||x|^\alpha f\big\|_2\big\||y|^\alpha\hat f\big\|_2$$

for all f in L^2 and $\alpha\ge1$. As a point of comparison, the argument used by Hirschman [11] (with the sharp form of the Hausdorff–Young inequality [1]) shows that

$$\|f\|_2^2\le H_\alpha\big\||x|^\alpha f\big\|_2\big\||y|^\alpha\hat f\big\|_2$$

where $H_\alpha=2\alpha e(8/e)^\alpha(\Gamma(1/2\alpha)/2\alpha)^{2\alpha}$. Hirschman's constant is better than ours and as $\alpha\to\infty$, $H_\alpha\sim\alpha(8/e)^\alpha\approx\alpha(2.94)^\alpha$.

THEOREM 3.3. *Let E_1, E_2 be measurable subsets of $\mathbb{R}$ with finite measures. Let $p_1, p_2 \in [1, \infty]$ and suppose that the weights $w_i \colon \mathbb{R} \to \mathbb{R}^+$ $(i = 1, 2)$ are (E_i, p_i)-adapted. If also there exists $\theta > 0$ such that*

$$\int_{E_i} w_i^{-\theta}\, dx < \infty, \tag{3.9}$$

then there are constants $K, \beta > 0$, such that

$$\|f\|_2 \leq K\left(\|w_1 f\|_{p_1}^{\beta} + \|w_2 \hat{f}\|_{p_2}^{\beta}\right)^{1/\beta}$$

for all tempered distributions f such that f and its transform are given by locally integrable functions.

Proof. The proof varies slightly among the three cases (i) $1 \leq p_1, p_2 \leq 2$, (ii) $2 \leq p_1, p_2 \leq \infty$ and (iii) $1 \leq p_1 < 2 < p_2 \leq \infty$. Since the third case contains all the features of the first two, it alone will be proved. Suppose we have β, q_1 and q_2 satisfying

$$0 < \beta \leq \theta / q_i', \tag{3.10}$$

$$(q_i - 2)/\beta q_i = (p_i - 2)/p_i, \tag{3.11}$$

$$\beta < p_i/q_i, \qquad q_i < \infty. \tag{3.12}$$

for $i = 1, 2$. From (3.9) and (3.10)

$$\left\|w_i^{-\beta} \chi_{E_i}\right\|_{q_i'} \leq m_i^{(\theta - \beta)/q_i'\theta} \left(\int_{E_i} w_i^{-\theta}\, dx\right)^{\beta/\theta} < \infty.$$

Since $\beta > 0$, (3.11) shows that $1 \leq q_1 < 2$ and $2 < q_2 \leq \infty$. Hence w_1^{β} is (E_1, q_1)-adapted since w_1 is (E_1, p_1)-adapted. Also (3.11) shows that w_2^{β} is (E_2, q_2)-adapted since w_2 is (E_2, p_2)-adapted. Thus Lemma 3.1 applies with the conclusion that

$$\|f\|_2 \leq K_1\left(\|w_1^{\beta} f\|_{q_1} + \|w_2^{\beta} \hat{f}\|_{q_2}\right)$$

for f in L^2. Application of the ideas following Lemma 3.2 based on (3.11) and (3.12) leads to the required inequality with $K = K_1^{1/\beta}$.

It remains to find β, q_1 and q_2 satisfying (3.10) to (3.12). Since $\varepsilon p_i/(1-\varepsilon)(p_i - 2)$, $\varepsilon p_i/(2-\varepsilon)(p_i - 2)$ and $\varepsilon p_i/(p_i - 2)$ all tend to 0 as $\varepsilon \to 0$, we may choose $\varepsilon_1, \varepsilon_2 \in (0, 1)$ so that

$$\frac{\varepsilon_1 p_1}{(2 - p_1)(1 - \varepsilon_1)}, \ \frac{\varepsilon_2 p_2}{(p_2 - 2)(1 - \varepsilon_2)} \leq \theta, \tag{3.13}$$

$$\frac{\varepsilon_1 p_1}{(2 - \varepsilon_1)(2 - p_1)} = \frac{\varepsilon_2 p_2}{(2 + \varepsilon_2)(p_2 - 2)}, \tag{3.14}$$

$$\frac{\varepsilon_1}{2 - p_1}, \ \frac{\varepsilon_2}{p_2 - 2} < 1. \tag{3.15}$$

Denote the number in (3.14) by β and define $q_1 = 2 - \varepsilon_1$ and $q_2 = 2 + \varepsilon_2$. Direct substitutions show that (3.13), (3.14) and (3.15) imply (3.10), (3.11) and (3.12) respectively, as required.

The only difference for the first two cases described at the beginning of the proof is that whenever p_1 (or p_2) equals 2, then q_1 (or q_2) is given the same value. (See the discussion below Lemma 3.2.) □

Remark 3.4. When working with the weights $w_1(x)=|x|^\theta$, $w_2(y)=|y|^\theta$ with $\theta,\phi\geq 0$, we can use Lemma 2.1 to transform the additive version of the inequality (3.3) to the multiplicative version. Hence the argument in the second part of the paragraph following Lemma 3.2 is now applicable. For these weights (3.9) is automatically satisfied when $E_1=E_2=[-1,1]$. Also in this case w_1 is (E_1,p)-adapted provided $\theta>1/p^{\#}$ and w_2 is (E_2,q)-adapted provided $\phi>1/q^{\#}$. Assuming these conditions, arguing as in the proof of Theorem 3.3, but based on the multiplicative version of Lemma 3.1, and applying Lemma 2.1 once more leads to

$$\|f\|_2\leq K\left(\left\||x|^\theta f\right\|_p+\left\||y|^\phi \hat{f}\right\|_q\right),$$

just as in Corollary 2.4.

4. A priori inequalities. In this section we prove some a priori inequalities for $L^2(\mathbb{R})$-functions which generalise the results of §§2 and 3. More precisely, we prove that if E and F are sets of finite measure, if $1\leq p$, $q\leq\infty$, and if v and w are weights which are respectively (E,p)- and (F,q)-adapted, then, for some constant K,

$$\|f\|_2\leq K\left(\|vf\chi_{E'}\|_p+\|w\hat{f}\chi_{F'}\|_q\right) \tag{4.1}$$

for all f in $L^2(\mathbb{R})$. We are unable to give any estimate whatever for the constant K. Examples of tempered distributions f for which the right side of (4.1) is finite, even zero, but for which $f\notin L^2$ are given in [19].

The proof of the a priori inequality (4.1) is based on the argument of §2, together with a more general version of the results of Pollak and Slepian [14] and Landau and Pollak [12]. We give this generalisation first, which is based on the following result due to Benedicks [2].

PROPOSITION 4.1. *If* $f\in L^2(\mathbb{R})$, $\operatorname{supp}(f)\subseteq E$, $\operatorname{supp}(f)\subseteq F$, *and* $m(E)+m(F)<\infty$, *then* $f=0$.

THEOREM 4.2. *Let E and F be subsets of* $\mathbb{R}$ *of finite measure. There exists a number* $\gamma(E,F)<1$ *such that*

$$\left(\int_F|\hat{f}(y)|^2dy\right)^{1/2}\leq\gamma(E,F)\left(\int|f(x)|^2dx\right)^{1/2}$$

for all f in $L^2(\mathbb{R})$ *whose supports are contained in E.*

Proof. Consider the operator T on L^2 given by the formula

$$Tf=\chi_E\cdot\mathfrak{F}^{-1}\left(\chi_F(f\chi_E)\hat{}\,\right),$$

where $\mathfrak{F}^{-1}$ denotes the inverse Fourier transform. This operator is compact and of positive type. (To see that it is compact, note that the kernel of the integral operator

$$f\to\mathfrak{F}^{-1}(\chi_F(f\chi_E))$$

belongs to $L^2(\mathbb{R}\times\mathbb{R})$.)

The operator norm of T is at most 1, so T admits a spectral decomposition with eigenvalues in [0, 1]. Let $\gamma(E,F)^2$ be the largest eigenvalue. If $\gamma(E,F)$ were equal to 1, then there would exist a nonzero function f such that

$$\langle Tf,f\rangle=\|f\|_2^2,$$

that is, such that

$$\int_F |(\chi_E f)\hat{\ }(y)|^2 dy = \|f\|_2^2.$$

Since for all g we have

$$\int_F |(\chi_E g)\hat{\ }(y)|^2 dy \leq \|\chi_E g\|_2^2$$

(with equality only if $\operatorname{supp}((\chi_E g)\hat{\ }) \subseteq F$) and

$$\|\chi_E g\|_2^2 \leq \|g\|_2^2,$$

(with equality only if $\operatorname{supp}(g) \subseteq E$), the existence of such an f would contradict Proposition 4.1. Thus $\gamma(E,F)<1$. Further, if $f=\chi_E f$, then

$$\int_F |\hat{f}(y)|^2 dy = \langle Tf,f\rangle \leq \gamma(E,F)^2\|f\|_2^2,$$

as required. □

THEOREM 4.3. *Let f be in $L^2(\mathbb{R})$, let s and t be in* $[1,2]$, *let E and F be sets of finite measure, and let*

$$A=\left[\int_{E'} dx|f(x)|^s\right]^{1/s}, \qquad B=\left[\int_{F'} dx|\hat{f}(y)|^2\right]^{1/t}.$$

Then

$$\|f\|_2 \leq \left[A^s+m(E)^{s\sigma}\alpha^s\right]^{\theta/s}\left[B^t+m(F)^{t\tau}\beta^t\right]^{(1-\theta)/t},$$

where

$$\sigma=1/s-1/2, \quad \tau=1/t-1/2, \quad \theta=\tau/(\sigma+\tau),$$
$$\alpha=\left[1-\gamma(E,F)^2\right]^{-1}\left[m(E)^\tau B+\gamma(E,F)m(F)^\sigma A\right],$$
$$\beta=\left[1-\gamma(E,F)^2\right]^{-1}\left[m(F)^\sigma A+\gamma(E,F)m(E)^\tau B\right].$$

Proof. The proof is but part of the proof of Theorem 2.2, and we omit it. □

COROLLARY 4.4. *Suppose that $1\leq p$, $q\leq\infty$, that E and F are subsets of $\mathbb{R}$ of finite measure, and that $v\colon \mathbb{R}\to\mathbb{R}^+$ is (E,p)-adapted while $w\colon\mathbb{R}\to\mathbb{R}^+$ is (F,q)-adapted. There exists a constant K such that*

$$\|f\|_2 \leq K\left(\|vf\chi_{E'}\|_p+\|w\hat{f}\chi_{F'}\|_q\right)$$

for all f in $L^2(\mathbb{R})$.

Proof. The corollary follows from Theorem 4.3 just as Corollary 2.3 follows from Theorem 2.2. □

It would be interesting to study further the constants $\gamma(E,F)$. Superficially it appears that $\gamma(E,F)\leq\gamma(m(E)m(F)/4)$, where the γ function on the right side is that considered by Pollak and Slepian [14] and Landau and Pollak [12].

5. Counterexamples. In this section we show that conditions similar to those assumed in Theorems 2.2 and 3.3 are necessary to establish the relevant inequalities. Throughout v and w are measurable functions from $\mathbb{R}$ to $\mathbb{R}^+$.

COUNTEREXAMPLE I. *Suppose continuous v, w satisfy $v(x)$, $w(x) \to 0$ as $x \to \infty$. If $0 < p$, $q \leq \infty$ there is no constant K such that*

$$\|f\|_2 \leq K\left(\|vf\|_p + \|w\hat{f}\|_q\right) \tag{5.1}$$

for all f in L^2.

Proof. Choose nonzero f in $\mathbb{S}$. For each $n \in \mathbb{Z}^+$ define $f_n: x \to \exp(2\pi inx)f(x-n)$. Then $\|f_n\|_2 = \|f\|_2$ while $\|vf_n\|_p$, $\|w\hat{f}_n\|_q \to 0$ as $n \to \infty$. □

COUNTEREXAMPLE II. *Given $1 \leq p$, $q \leq \infty$, suppose v, w satisfy*

$$\|\chi_{[0,\lambda]}v\|_p = o(\lambda^{1/2}) \quad \text{as } \lambda \to \infty, \tag{5.2}$$

$$w \text{ is of polynomial order as } x \to \infty, \tag{5.3}$$

or vice versa. Then there is no constant K such that

$$\|f\|_2^2 \leq K\|vf\|_p \cdot \|w\hat{f}\|_q \tag{5.4}$$

for all f in L^2.

Proof. The proof is based on the familiar Rudin–Shapiro construction. Choose nonzero f in $\mathbb{S}$ with support in $[0,1]$. Let $f_0 = g_0 = f$ and define sequences (f_n), (g_n) via the inductive step

$$f_{k+1} = f_k + \tau_{2^k} g_k, \qquad g_{k+1} = f_k - \tau_{2^k} g_k$$

for $k \in \mathbb{Z}^+$ where $\tau_a h: x \to h(x-a)$. Evidently

$$\|f_n\|_2 = 2^{n/2}\|f\|_2, \tag{5.5}$$

$$\|vf_n\|_p \leq \|f\|_\infty \|\chi_{[0,2^n]}v\|_p. \tag{5.6}$$

The critical property of the sequence (f_n) is $|\hat{f}_n| \leq 2^{(n+1)/2}|\hat{f}|$ which follows from the identity $|f_n|^2 + |g_n|^2 = 2^{n+1}|f|^2$ (see [10, (37.19)]). Hence

$$\|w\hat{f}_n\|_q \leq 2^{(n+1)/2}\|w\hat{f}\|_q. \tag{5.7}$$

Suppose (5.4) is valid for all f in L^2. From (5.3), (5.5), (5.6) and (5.7),

$$2^n \leq \text{const.} \|\chi_{[0,2^n]}v\|_p \cdot 2^{n/2},$$

which contradicts (5.2). □

Remark. If we drop condition (5.3) and assume that the weights v, w are "rapidly increasing", then for fairly general functions f, (5.1) and (5.4) are valid in the trivial sense that either $\|vf\|_p$ or $\|w\hat{f}\|_q = \infty$, or $f = 0$. Hardy's theorem ([8], see also [4, pp. 155–158]) gives us one example of this type of result: Suppose that

$$v(x) \geq A \cdot \exp(\alpha x^2), \qquad w(y) \geq B \cdot \exp(\beta y^2)$$

for $|x|$, $|y|$ sufficiently large where $A, B, \alpha, \beta > 0$. If f satisfies $\|vf\|_\infty$, $\|w\hat{f}\|_\infty < \infty$, then $f = 0$, or f is a constant multiple of $\exp(-\alpha x^2)$, or there are infinitely many such functions f in $\mathbb{S}$ according as $\alpha\beta > \pi^2$, $\alpha\beta = \pi^2$ or $\alpha\beta < \pi^2$.

In [19] we give the following extension: Suppose that v, w are as above and at least one of $p, q \in [1, \infty]$ is finite. If $\alpha\beta \geq \pi^2$, then the only f in $\mathbb{S}'$ satisfying $\|vf\|_p$, $\|w\hat{f}\|_q < \infty$ is $f = 0$, while if $\alpha\beta < \pi^2$ there are infinitely many such functions in $\mathbb{S}$.

Discussion. In the remainder of the section we are concerned with the inequality

$$\|f\|_2\le K\left(\left\||x|^\theta f\right\|_p+\left\||y|^\phi f\right\|_q\right) \tag{5.8}$$

for all f in L^2. Apart from the situation described in Corollary 2.4 or Remark 3.4, this inequality is also valid, but now in a trivial way, when $p=2$ and $\theta=0$ or $q=2$ and $\phi=0$. The following three counterexamples show that there are no other cases in which (5.8) is possible. As usual, we suppose $p,q\in[1,\infty]$ and $\theta,\phi\ge 0$.

COUNTEREXAMPLE II′. *If $\theta<1/p^{\#}$ and $\phi\ne 1/q^{\#}$, or vice versa, then* (5.8) *is not possible.*

Proof. Assume that (5.8) is valid for all f in L^2 with $\theta<1/p^{\#}$ and $\phi\ne 1/q^{\#}$. Substitute $D_\lambda f_n$ for f, where $D_\lambda f$ is the normalized dilate defined by $D_\lambda f(x)=\lambda^{-1/2}f(x/\lambda)$ (see §2) and f_n is as defined in the proof of Counterexample II. Then $2^{n/2}\le\text{const.}(\lambda^{\theta-1/p^{\#}}2^{n(\theta+1/p)}+\lambda^{-\phi+1/q^{\#}}2^{n/2})$ and hence

$$1\le\text{const.}\left((2^n\lambda)^{(\theta-1/p^{\#})}+\lambda^{-\phi+1/q^{\#}}\right).$$

If $-\phi+1/q^{\#}<0$, take $\lambda=2^n$ and if $-\phi+1/q^{\#}>0$, take $\lambda=(2/3)^n$. In either case we get a contradiction by letting $n\to\infty$.

COUNTEREXAMPLE III. *If $\theta=1/p^{\#}$ with $2<p\le\infty$ and $\phi\ne 1/q^{\#}$, or vice versa, no inequality of the form* (5.8) *is possible.*

Proof. When $\theta=1/p^{\#}$, replacement of f in (5.8) with $D_\lambda f$ gives

$$\|f\|_2\le K\left(\left\||x|^\theta f\right\|_p+\lambda^{-\phi+1/q^{\#}}\left\||y|^\phi f\right\|_q\right)$$

for all $\lambda>0$. But this requires

$$\|f\|_2\le K\left\||x|^\theta f\right\|_p \tag{5.9}$$

whenever $\phi\ne 1/q^{\#}$. Define f_n by $f_n(x)=x^{-1/2}$ on $[1,n]$ and 0 otherwise. Substitution in (5.9) results in $(\log n)^{1/2}\le K(\log n)^{1/p}$ for all n (with the obvious modification when $p=\infty$), an impossibility if $2<p\le\infty$. □

COUNTEREXAMPLE IV. *Suppose $\theta=1/p^{\#}$ with $2<p\le\infty$ and $\phi=1/q^{\#}$ with $2<q\le\infty$. Then* (5.8) *is not possible.*

Proof. Choose $\psi\in C^\infty(\mathbb{R})$ satisfying $\psi(x)=0$ for $x\le 0$, $\psi(x)=1$ for $x\ge 1$ and $0\le\psi\le 1$. Also choose $\alpha\in C^\infty(\mathbb{R})$ satisfying $\alpha(x)=2+x$ for $x\le 1$, $\alpha(x)=x$ for $x\ge 10$, and $x\le\alpha(x)\le 2+x$ for all x. For each $\varepsilon>0$ define

$$g_\varepsilon(y)=\psi(|y|)|y|^{-1/2}\left(\log\left(\alpha(|y|)\right)\right)^{-1/2}e^{-\varepsilon|y|}.$$

From Theorem 1 of Wainger [17] (with $k=1$, $l=0$, $\gamma=\frac{1}{2}$ and $b(y)=(\log(\alpha(|y|)))^{-1/2}$), $f(x)=\lim_{\varepsilon\to 0+}\mathfrak{F}^{-1}g_\varepsilon(x)$ is defined for all $x\ne 0$ and is infinitely differentiable apart from $x=0$. Furthermore, this theorem shows that as $x\to 0$

$$f(x)\sim|x|^{-1/2}\left(\log\left(\alpha(1/|x|)\right)\right)^{-1/2}\sim|x|^{-1/2}\log(|x|)^{-1/2}$$

and, since $f\in L^1$, that $\hat f(y)=\psi(|y|)|y|^{-1/2}(\log(\alpha(|y|)))^{-1/2}$. It is routine to check that $\||x|^\theta f\|_p$, $\||y|^\phi\hat f\|_q<\infty$ when $\theta=1/p^{\#}$, $2<p\le\infty$ and $\phi=1/q^{\#}$, $2<q\le\infty$ and yet $\|f\|_2=\infty$, which contradicts (5.8). (In fact, $\||x|^\theta f\|_p<\infty$ when $\theta>1/p^{\#}$ or $\theta\ge 1/p^{\#}$ and $2<p\le\infty$, and $\||y|^\phi\hat f\|_q<\infty$ when $\phi<1/q^{\#}$ or $\phi\le 1/q^{\#}$ and $2<q\le\infty$. Hence this function also provides a counterexample in these cases. It uses, however, more powerful

machinery than counterexamples II′ and III. Also note that the case $p=q=\infty$ and $\theta=\phi=\frac{1}{2}$ is easily disposed of by the function $f: x\to|x|^{-1/2}$ since its (distributional) Fourier transform is itself.) □

As indicated in the discussion below Counterexample II, when we put together the above bits and pieces we get the completion of the result started in Corollary 2.4 and 3.4.

THEOREM 5.1. *Suppose $p,q\in[1,\infty]$ and $\theta,\phi\geq 0$. There exists a constant K such that*

$$\|f\|_2\leq K\left(\left\||x|^\theta f\right\|_p+\left\||y|^\phi \hat{f}\right\|_q\right)$$

for all tempered distributions f with the property that f and $\hat{f}$ are locally integrable functions if and only if $\theta>1/p^\#$ and $\phi>1/q^\#$ or $(p,\theta)=(2,0)$ or $(q,\phi)=(2,0)$.

REFERENCES

[1] W. BECKNER, *Inequalities in Fourier analysis on* $\mathbb{R}^n$, Proc. Nat. Acad. Sci., 72 (1975), pp. 638–641.

[2] M. BENEDICKS, *On Fourier transforms of functions supported on sets of finite Lebesgue measure*, Res. Rep., TRITA-MAT-1974-5, Royal Institute of Technology, Stockholm.

[3] D. C. CHAMPENEY, *Fourier Transforms and Their Physical Applications*, Academic Press, New York, 1973.

[4] H. DYM AND H. P. MCKEAN, *Fourier Series and Integrals*, Academic Press, New York, 1972.

[5] W. G. FARIS, *Inequalities and uncertainty principles*, J. Math. Phys., 19 (1978), pp. 461–466.

[6] C. FEFFERMAN AND D. H. PHONG, *The uncertainty principle and sharp Gärding inequalities*, Comm. Pure Appl. Math., 34 (1981), pp. 285–331.

[7] D. GABOR, *Theory of communication*, J. Inst. Electr. Engrs. 93(3) (1946), pp. 429–457.

[8] G. H. HARDY, *A theorem concerning Fourier transforms*, J. London Math. Soc. 8 (1933), pp. 227–231.

[9] W. HEISENBERG, *Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik*, Z. Phys., 43 (1927), pp. 172–198.

[10] E. HEWITT AND K. A. ROSS, *Abstract Harmonic Analysis*, Vol. II, Springer-Verlag, New York, 1970.

[11] I. I. HIRSCHMAN JR., *A note on entropy*, Amer. J. Math., 79 (1957), pp. 152–156.

[12] H. J. LANDAU AND H. O. POLLAK, *Prolate spheroidal wave functions, Fourier analysis and uncertainty* (2), Bell System Tech. J., 40 (1961), pp. 65–84.

[13] A. PAPOULIS, *Signal Analysis*, McGraw-Hill, New York, 1977.

[14] H. O. POLLAK AND D. SLEPIAN, *Prolate spheroidal wave functions, Fourier analysis and uncertainty* (1), Bell System Tech. J., 40 (1961), pp. 43–64.

[15] J. F. PRICE, *Inequalities and local uncertainty principles*, J. Math. Physics, 24 (1983), pp. 1711–1714.

[16] D. E. VAKMAN, *Sophisticated Signals and the Uncertainty Principle in Radar*, Springer-Verlag, New York, 1968.

[17] S. WAINGER, *Special trigonometric series in k-dimensions*, Mem. Amer. Math. Soc., 59 (1965).

[18] H. WEYL, *The Theory of Groups and Quantum Mechanics*, Dover, New York, (transl. of 1930 German edition).

[19] M. G. COWLING AND J. F. PRICE, *Generalisations of Heisenberg's inequality*, Proc. Harmonic Anal. Conf. Cortona, 1982, to appear.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics

0036-1410/84/1501-0013 $01.25/0

CHEBYSHEV SYSTEMS OF MINIMAL DEGREE*

B. L. GRANOVSKY† AND ELI PASSOW‡

Abstract. Let $F=\{f_i\}_{i=0}^n$ be a set of continuous functions on $[a,b]$, and let $F^*=\{f_if_j\}_{i,j=0}^n$. We determine conditions on F which are necessary and sufficient for the set F^* to be a Chebyshev system on $[a,b]$ consisting of exactly $2n+1$ distinct functions. The results have applications in the field of experimental design.

1. Introduction.

In many problems of mathematical statistics and probability the matrix of moments, $M(\xi)$, occurs. Here $M(\xi)=\|m_{ij}\|_{i,j=0}^n$, $m_{ij}=\int_X f_i(x)f_j(x)\xi(dx)$, where $\{f_i\}$, $i=0,1,\cdots,n$, is a set of $n+1$ linearly independent continuous functions on a compact space, X, and ξ is a probability measure on X. In particular, in the theory of least-squares and experimental design such matrices are called information matrices or design matrices, and the measures ξ are called experimental designs.

One of the questions arising in this field is to find a solution ξ for the problem of moments $m_{ij}(\xi)=m_{ij}^*$ (m_{ij}^* are given), with the minimum number of points of support. (For the statistical significance of this problem and related results see [1, Ch. 10].) The solution clearly depends upon properties of the set of functions $\{f_if_j\}_{i,j=0}^n$, in particular whether this set forms a Chebyshev system. (A set $\{u_i\}_{i=0}^n$ of continuous functions on X is a *Chebyshev system* on X if every nontrivial "polynomial" $\sum_{i=0}^n C_iu_i(x)$ has at most n zeros on X. We call $n+1$ the *degree* of the Chebyshev system.) In this paper we give necessary and sufficient conditions on the set $\{f_i\}_{i=0}^n$ so that the set $\{f_if_j\}_{i,j=0}^n$ is a Chebyshev system of minimal degree.

2. Preliminary results.

Let $\nu(U)$ be the number of distinct elements of the set U. Let $U=U_n=\{u_i\}_{i=0}^n$ be a set of real numbers and denote by $U^*=U_n^*$ the set of all possible products of two elements of U_n; that is, $U^*=U_n^*=\{u_iu_j\}_{i,j=0}^n$. Our first result tells when $\nu(U_n^*)$ is minimal.

LEMMA 1. *Suppose $u_i\neq 0$, $i=0,1,\cdots,n$, and $|u_i|\neq|u_j|$, $i,j=0,1,\cdots,n$, $i\neq j$. Then $\nu(U_n^*)\geq 2n+1$, with equality if and only if the set U_n is of the form $U_n=\{\omega u^k\}_{k=0}^n$, where ω and u are some real numbers such that $u,\omega\neq 0$, $|u|\neq 1$.*

Proof. The sufficiency is obvious and we prove the necessity by induction. The assertion is trivial for $n=0$, so assume it is true for $n-1$. Without loss of generality, assume that $|u_0|<|u_1|<\cdots<|u_n|$. Then $U_n^*=U_{n-1}^*\cup\{u_ju_n\}_{j=0}^n$. Now $|u_{n-1}u_n|$ and u_n^2 are larger than the absolute values of all the terms of U_{n-1}^* and, by the induction hypothesis, $\nu(U_{n-1}^*)\geq 2n-1$. Hence, $\nu(U_n^*)\geq\nu(U_{n-1}^*)+2\geq 2n+1$, which proves the first part of the assertion. Now suppose that $\nu(U_n^*)=2n+1$. Then the above inequality implies that $\nu(U_{n-1}^*)$ must equal $2n-1$. Thus, by the second part of the induction hypothesis, $U_{n-1}=\{\omega u^k\}_{k=0}^{n-1}$ and, hence, $U_{n-1}^*=\{\omega^2u^k\}_{k=0}^{2n-2}$. It is only left to show that u_n too is of the desired form. Observe first that without loss of generality we can assume that $|u|>1$.

*Received by the editors March 5, 1982, and in final revised form June 27, 1983.

†Department of Mathematics, Technion-Israel Institute of Technology, Haifa, Israel. The work of this author was supported by Technion V. P. R. Fund—K. and M. Bank Research Fund.

‡Department of Mathematics, Temple University, Philadelphia, Pennsylvania 19122. The work of this author was supported in part by a Temple University Faculty Summer Research Fellowship and a Grant-in-Aid of Research. The research in this article was conducted partially while this author was on leave at the Technion in 1979–80 and was supported by the Lady Davis Foundation.

(For if $0<|u|<1$, then $\tilde{u}=u^{-1}$ and $\tilde{\omega}=\omega u^{n-1}$ will generate the same set $U_{n-1}=\{\omega u^k\}_{k=0}^{n-1}$.) So according to the assumed order of the elements of U_n, we have $u_k=\omega u^k$, $k=0,1,\cdots,n-1$. Now consider u_0u_n which is distinct from $u_{n-1}u_n$ and u_n^2 and, hence, must coincide with some element of U_{n-1}^*. Thus $u_0u_n=\omega^2u^j$ for some $0\leq j\leq 2n-2$, and from this it follows that $u_n=\omega u^k$ for some $n\leq k\leq 2n-2$ ($k\geq n$ because otherwise u_n would belong to U_{n-1}).

We now claim that $j=n$. For suppose that $j\geq n+1$. Then the products $u_{n-2}u_n$, $u_{n-1}u_n$ and u_n^2 would fail to be elements of U_{n-1}^*, so that U_n^* would have at least three more elements than U_{n-1}^* has. But we showed earlier that $\nu(U_n^*)=\nu(U_{n-1}^*)+2$. Therefore, $j=n$, $u_n=\omega u^n$, and the proof is complete.

We would now like to apply Lemma 1 to a set of functions. We wish to show that under appropriate hypotheses the set of all possible products of functions $\{f_0,f_1,\cdots,f_n\}$ consists of $2n+1$ distinct functions if and only if $f_{i_k}(x)=\omega(x)[u(x)]^k$, $k=0,1,\cdots,n$, for some permutation $\{i_k\}$ of $\{0,1,\cdots,n\}$. The obvious approach is to apply Lemma 1 point by point to define the functions $\omega(x)$ and $u(x)$. The potential difficulty is that Lemma 1 guarantees only that *some* rearrangement of $\{u_0,u_1,\cdots,u_n\}$ forms a geometric progression, and it is possible that different rearrangements hold at different points. This could be the case if the relationships which hold between the products (of the values of the functions) which reduce the number of these products to $2n+1$ differ at different points. If, however, we insist that the *same* relationships hold at each point—that is, if we demand that the relationships hold between the *functions*—then this difficulty will not arise, as is shown in the next lemma.

LEMMA 2. *Let $U_n=\{u_0,u_1,\cdots,u_n\}$, $V_n=\{v_0,v_1,\cdots,v_n\}$ be two sets of real numbers as in Lemma 1, and suppose $\nu(U_n^*)=2n+1$. Suppose that identical relationships hold between the elements of U_n^* and V_n^*; that is, $u_iu_j=u_ku_l$ (for some i,j,k,l) if and only if $v_iv_j=v_kv_l$. Then there exists a permutation $\{i_k\}_{k=0}^n$ of $\{0,1,\cdots,n\}$ such that simultaneously $u_{i_k}=\omega u^k$ and $v_{i_k}=tv^k$, $k=0,1,\cdots,n$.*

Proof. From Lemma 1 it follows that $U_n=\{\omega u^k\}_{k=0}^n$ and $V_n=\{tv^k\}_{k=0}^n$. Without loss of generality assume that $u_k=\omega u^k$, $k=0,1,\cdots,n$, so that, in particular, $u_0u_k=u_1u_{k-1}$, $k=1,2,\cdots,n$. Thus the same relationships must hold for the corresponding elements of V_n, that is, $v_0v_k=v_1v_{k-1}$, $k=1,2,\cdots,n$, and it follows from these relationships that $v_k=v_0(v_1/v_0)^k$, $k=1,2,\cdots,n$. Thus $v_k=tv^k$, $k=0,1,\cdots,n$, where $t=v_0$ and $u=v_1/v_0$.

Remark. Note in Lemma 1 that the order of the terms $\{|u_k|\}_{k=0}^n$ determines the order of the terms in the geometric progression $\{\omega u^k\}_{k=0}^n$. From Lemma 2 we see that if the products $\{u_iu_j\}$ and $\{v_iv_j\}$ satisfy identical relationships, then the order of the terms $|v_i|$ will either be identical to that of $|u_i|$ or exactly reversed; that is, if $|u_{i_0}|<|u_{i_1}|<\cdots<|u_{i_n}|$, then either $|v_{i_0}|<|v_{i_1}|<\cdots<|v_{i_n}|$ or $|v_{i_0}|>|v_{i_1}|>\cdots>|v_{i_n}|$.

3. The main results. Let $F=\{f_0,f_1,\cdots,f_n\}$ be a set of continuous functions on $[a,b]$, let $F^*=\{f_if_j\}_{i,j=0}^n$, and suppose that $f_k(x)=\omega(x)(u(x))^k$, $k=0,1,\cdots,n$. Then F^* will be of the form $F^*=\{\omega^2(x)(u(x))^k\}_{k=0}^{2n}$, so that $\nu(F^*)\leq 2n+1$. (In certain degenerate cases it is possible that $\nu(F^*)<2n+1$.) Our next theorem is a converse to this result.

THEOREM 1. *Let $F=\{f_0,f_1,\cdots,f_n\}$ be a set of continuous functions on $[a,b]$. Let $T=\{x\in[a,b]\colon f_i(x)\neq 0,\ |f_i(x)|\neq|f_j(x)|,\ i,j=0,1,\cdots,n,i\neq j\}$ and suppose that $\bar{T}$, the closure of T, is equal to $[a,b]$. If $\nu(F^*)=2n+1$, then $F=\{\omega(x)(u(x))^k\}_{k=0}^n$, where $\omega(x)\in C[a,b]$ and $\omega(x)\neq 0$, $x\in T$, while $u(x)\neq 0$, $|u(x)|\neq 1$, $x\in T$, and $u(x)$ is continuous on $[a,b]$, except possibly where $\omega(x)=0$.*

Proof. Since $\nu(F^*)=2n+1$, certain identities of the form $f_i f_j = f_k f_l$ exist. Let $x \in T$, $u_k = f_k(x)$, $k=0,1,\cdots,n$, and let $U_n=\{u_0,u_1,\cdots,u_n\}$. Then $\nu(U_n^*)=2n+1$, so that, by Lemma 1, there exist $\omega=\omega(x)\neq 0$, $u=u(x)\neq 0$, $|u(x)|\neq 1$, and a permutation $i_k=i_k(x)$, $k=0,1,\cdots,n$, of $\{0,1,\cdots,n\}$, such that $u_{i_k}=\omega u^k$, $k=0,1,\cdots,n$. If y is any other point of T and if we let $v_k=f_k(y)$, then $v_i v_j = v_k v_l$ if and only if $u_i u_j = u_k u_l$. Hence, by Lemma 2, the v's form a geometric progression in the same order as the u's. Since x and y are arbitrary in T, it follows that $f_{i_k}(x)=\omega(x)(u(x))^k$, $k=0,1,\cdots,n$, for all $x\in T$. In particular, $f_{i_0}(x)=\omega(x)$, $x\in T$, and because of the continuity of $f_{i_0}(x)$ on $[a,b]=\bar{T}$, $\omega(x)$ coincides with $f_{i_0}(x)$ on $[a,b]$. Thus ω is continuous on $[a,b]$. Now $u(x)=f_{i_1}(x)/\omega(x)$ holds for all $x\in T$, where both $f_{i_1}(x)$ and $\omega(x)$ are continuous on $[a,b]$. Therefore, $u(x)$ is also continuous on $[a,b]$, except possibly where $\omega(x)=0$.

THEOREM 2. *Let $F=\{f_0,f_1,\cdots,f_n\}$ be a set of continuous functions on $[a,b]$. Then all distinct functions of the set $F^*=\{f_i f_j\}_{i,j=0}^n$ form a Chebyshev system of minimal degree $2n+1$ on $[a,b]$ if and only if F is of the form $F=\{\omega(x)(u(x))^k\}_{k=0}^n$, where $\omega(x)$ and $u(x)$ are continuous functions on $[a,b]$, satisfying*

(i) *$\omega(x)\neq 0$, $x\in[a,b]$;*

(ii) *$u(x)$ is monotone on $[a,b]$.*

Proof. Assume that all distinct functions comprising F^* form a Chebyshev system of degree $2n+1$. We show first that the functions in F are linearly independent. Note that each f_k, $k=0,1,\cdots,n$, must have a finite number of distinct zeros, for otherwise F^* would not be a Chebyshev system. From this and the fact that $\{f_i\}_{i=0}^n$ are distinct continuous functions, it follows that for any fixed $k=0,1,\cdots,n$, the functions $f_i f_k$, $i=0,1,\cdots,n$, are distinct. Let $\Sigma_{i=0}^n a_i f_i(x)$ be a nontrivial polynomial. Then $f_k(x)\Sigma_{i=0}^n a_i f_i(x)=\Sigma_{i=0}^n a_i f_i(x) f_k(x)$ is a nontrivial polynomial formed from distinct functions $f_i f_k \in F^*$, $i=0,1,\cdots,n$. Hence $\Sigma_{i=0}^n a_i f_i(x) f_k(x)$ has at most $2n$ distinct zeros, so that $\Sigma_{i=0}^n a_i f_i(x)$ has a finite number of distinct zeros. Hence F is a linearly independent set of functions.

We now show that F is actually a Chebyshev system of degree n. Let $p(x)=\Sigma_{i=0}^n b_i f_i(x)$ be a nontrivial polynomial with r distinct zeros on $[a,b]$, at $x_1, x_2,\cdots,x_r$. We will show that $r\leq n$. Construct a nontrivial polynomial $q(x)=\Sigma_{i=0}^n c_i f_i(x)$ having n distinct zeros on $[a,b]$, all of them different from $x_1, x_2,\cdots,x_r$. Then $p(x)q(x)$ is a nontrivial polynomial formed from linear combinations of the functions $f_i f_j \in F^*$, so that, according to our assumption on F^*, the total number of distinct zeros of pq cannot exceed $2n$. Thus $r+n\leq 2n$, so that $r\leq n$. But p is an arbitrary polynomial, so that F is a Chebyshev system of degree n on $[a,b]$.

It follows from this that the complement of the set T in Theorem 1 is a finite set, since no function $f_i\in F$ can vanish at more than n points, and no two functions $|f_i|$, $|f_j|$, where $f_i, f_j\in F$, can agree at more than $2n$ points. Thus $T=[a,b]$, so that by Theorem 1, the system F is of the form $F=\{\omega u^k\}_{k=0}^n$. It remains only to show that the functions $\omega=\omega(x)$ and $u=u(x)$ satisfy conditions (i) and (ii).

Suppose that $\omega(x_0)=0$ for some $x_0\in[a,b]$. Then all of the functions $\omega(x)[u(x)]^k$, $k=0,1,\cdots,n$ vanish at x_0, so that any polynomial, $p(x)$, formed from these functions will also vanish at x_0. Now let $p(x)$ be a nontrivial polynomial having n distinct zeros in $[a,b]$, all different from x_0. Then p will have $n+1$ distinct zeros on $[a,b]$, so that $F=\{\omega u^k\}_{k=0}^n$ is not a Chebyshev system. This proves the necessity of (i), and from this, by Theorem 1, it follows that $u(x)$ is continuous on $[a,b]$. Suppose now that x_0, $x_1,\cdots,x_n$ are $n+1$ distinct points in $[a,b]$, and consider the system of linear equations

$$\omega(x_i)\sum_{k=0}^{n} a_k[u(x_i)]^k=0, \qquad i=0,1,\cdots,n.$$

If $u(x)$ is not monotone on $[a,b]$, then there exist x_j, x_l in $[a,b]$, $x_j \neq x_l$, for which $u(x_j)=u(x_l)$. But then the jth equation of this linear system will be a multiple of the lth equation, so that the determinant of this system will vanish. Hence, the system has a nontrivial solution, so that F is not a Chebyshev system, contradicting our earlier findings, and completing the proof of the necessity of the conditions. The sufficiency is evident from the above analysis.

COROLLARY. *Let* $0 \leq t_0 < t_1 < \cdots < t_n$ *be a set of integers and let* $F=\{x^{t_k}\}_{k=0}^{n}$. *Then* $F^* = \{x^{t_i}x^{t_j}\}_{i,j=0}^{n}$ *is a Chebyshev system of degree* $2n+1$ *on* $[-1,1]$ *if and only if* $t_k=kt$, $k=0,1,\cdots,n$, *where* t *is an odd integer.*

Proof. By Theorem 2 it is necessary and sufficient that $x^{t_k}=\omega(x)[u(x)]^k$, $k=0,1,\cdots,n$, where $\omega(x)$ and $u(x)$ satisfy conditions (i) and (ii) of that theorem. Thus, $\omega(x)=x^{t_0}$, $u(x)=x^{t_1-t_0}$, and it follows that the conditions (i) and (ii) are satisfied if and only if $t_0=0$ and $t_1=t$ is odd, since x^t is monotone on $[-1,1]$ if and only if t is odd.

Remark. From Theorem 2 it follows that for F^* to be a Chebyshev system of the minimal degree $2n+1$ it is necessary that F be a Chebyshev system of degree n. If F is as in the Corollary then, by [2], it is a Chebyshev system of degree n if and only if $t_0=0$ and t_k, $k=1,2,\cdots,n$, are alternately odd and even. The result of the corollary is that among the sequences $\{t_k\}$ of [2] only those which are of the form $t_0=0$, $t_k=kt$, t odd, provide the desired property of F^*.

REFERENCES

[1] S. KARLIN AND W. J. STUDDEN, *Tchebycheff Systems with Applications in Analysis and Statistics*, Interscience, New York, 1966.

[2] E. PASSOW, *Alternating parity of Tchebycheff Systems*, J. Approx. Theory, 9 (1973), 295–298.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1501-0014 $01.25/0

A NOTE ON HUDSON'S THEOREM ABOUT FUNCTIONS WITH NONNEGATIVE WIGNER DISTRIBUTIONS*

A. J. E. M. JANSSEN†

Abstract. We show that a (generalized) function f has a nonnegative Wigner distribution $W(f,f)$ if and only if f is a Gauss function (possibly degenerate). We prove, more generally, that the convolution of $W(f,f)$ with certain Gauss functions is nonnegative if and only if f is of the special type mentioned. As a consequence we have that the only (generalized) functions whose Wigner distributions are concentrated on a curve of a particular type are delta functions or exponentials $\exp(-\pi\alpha t^2+2\pi\beta t+\gamma)$ with α, β, γ complex, $\operatorname{Re}\alpha=0$. The main tool used is Moyal's formula for the Wigner distribution together with Bargmann's integral transform.

1. Introduction. For $f\in L^2(\mathbb{R})$, the Wigner distribution $W(f,f)$ of f is defined as

$$(1.1)\qquad W(x,y;f,f)=\int_{-\infty}^{\infty} e^{-2\pi iyt}f\left(x+\frac{1}{2}t\right)\overline{f\left(x-\frac{1}{2}t\right)}dt, \qquad (x\in\mathbb{R},y\in\mathbb{R}).$$

It is known that $W(f,f)$ is a continuous, bounded, real-valued function that may take negative values. The Wigner distribution was introduced by Wigner [15] as a device that allows one to express quantum mechanical expectation values in the same form as the averages of classical statistical mechanics. By means of the Wigner distribution one can describe Weyl's correspondence [7], [14] in the following elegant form (see for this e.g. [4]). If a: $\mathbb{R}^2\to\mathbb{R}$ is an observable, then the expectation value of a in the state f is given by

$$(1.2)\qquad \iint a(x,y)W(x,y;f,f)\,dx\,dy,$$

i.e., instead of substituting a particular point (x_0,y_0) of the phase plane in a (as one does in classical mechanics), one integrates a against the "density function" $W(f,f)$. More recently there has been considerable interest in the Wigner distribution as a tool for signal analysts to describe a signal in time and frequency simultaneously (cf. [3], [5]). In both quantum mechanics and signal analysis one likes to interpret $W(f,f)$ as a density function of two variables. Such an interpretation is awkward, since $W(f,f)$ may take negative values as already said. Nevertheless, there is a fairly extensive list of positivity properties of the Wigner distribution (cf. [3], [11]). These properties express that certain averages of the Wigner distribution are nonnegative. A typical example is: for any $f\in L^2(\mathbb{R})$ (cf. [2]),

$$(1.3)\qquad \iint \exp\left(-2\pi\delta(x-a)^2-2\pi\gamma(y-b)^2\right)W(x,y;f,f)\,dx\,dy\geq 0,$$

for all $\delta>0$, $\gamma>0$, $a\in\mathbb{R}$, $b\in\mathbb{R}$ where $\delta\gamma\leq 1$.

It is convenient to allow in this note certain generalized functions f which we shall describe in §2. We shall show that if $f\neq 0$ has a Wigner distribution that is nonnegative everywhere (in a generalized sense), then f is necessarily of the form

$$(1.4)\qquad f(t)=\exp(-\pi\alpha t^2+2\pi\beta t-\pi\gamma),$$

or

$$(1.5)\qquad f(t)=d\delta_a(t),$$

* Received by the editors November 9, 1981, and in revised form June 25, 1982.

† Philips Research Laboratories, 5600 MD Eindhoven, the Netherlands.

where α, β, γ, a, d are complex numbers with $\operatorname{Re}\alpha \geq 0$. If we restrict to $f \in L^2(\mathbb{R})$, this is known as Hudson's theorem [8]. The f's in (1.4) are what we call Gabor functions (although this name is usually reserved for the case that α is real and positive). We have for the f in (1.4), by calculation,

(1.6)

$$W(x,y;f,f) = \left(\frac{2}{\operatorname{Re}\alpha}\right)^{1/2} \exp\Big(-2\pi\operatorname{Re}\gamma + 2\pi(\operatorname{Re}\beta)^2/\operatorname{Re}\alpha - 2\pi(x - \operatorname{Re}\beta/\operatorname{Re}\alpha)^2\operatorname{Re}\alpha \\ - 2\pi(y + x\operatorname{Im}\alpha - \operatorname{Im}\beta)^2/\operatorname{Re}\alpha\Big),$$

if $\operatorname{Re}\alpha > 0$, and

$$(1.7) \qquad W(x,y;f,f) = \exp(-2\pi\operatorname{Re}\gamma + 4\pi x\operatorname{Re}\beta)\delta_0(y - \operatorname{Im}\beta + x\operatorname{Im}\alpha),$$

if $\operatorname{Re}\alpha = 0$. And for the f in (1.5), we have

$$(1.8) \qquad W(x,y;f,f) = |d|^2\exp(4\pi y\operatorname{Im}a)\delta_0(x - \operatorname{Re}a).$$

We shall show more generally that if $\delta\gamma > 1$, and (1.3) is nonnegative for all a and b, then f must be of the form (1.4) or (1.5). This result shows that Gabor functions and delta functions are fairly isolated objects in this kind of time-frequency analysis. As an application we show that if $W(f,f)$ is concentrated on a curve of a certain type, then f must be of the form (1.4) (with $\operatorname{Re}\alpha = 0$) or (1.5).

The key argument, due to Hudson (cf. [8]), is the observation that for $\gamma = \delta = 1$, the expression (1.3) can be written as $\exp(-\pi(a^2 + b^2))|G(a - ib)|^2$, where G is an entire function of order 2 (Bargmann transform of f). Now, $f \in L^2(\mathbb{R})$, $W(f,f) \geq 0$ everywhere implies that $G\,(a - ib) \neq 0$ for all a and b (unless $f \equiv 0$). And Hadamard's theorem can be used to show that G, and hence f, has a special form. Since we also want to discuss f's which are not necessarily square integrable, we consider in §2 the Bargmann transform in some detail for f's in a convenient set of generalized functions.

2. Preliminaries. A convenient theory of generalized functions for discussing the Wigner distribution was elaborated by De Bruijn (cf. [4]); we describe it here briefly. We don't want to use Schwartz' theory of tempered distributions since this theory has the disadvantage that functions like $f(t) = \exp(t)$ and $f(t) = \delta_i(t)$ cannot be considered. Also, the theory used in this note arises naturally in the context of the Bargmann transform which will be used later on. Our test function space S consists of all entire functions f for which there are $A > 0$, $B > 0$ such that $f(x + iy) = O(\exp(-\pi Ax^2 + \pi By^2))$. This space can be identified with the Gelfand–Shilov space $S^{1/2}_{1/2}$ (cf. [6], [9]). We may describe S as the set of all $f \in L^2(\mathbb{R})$ for which $(f, \psi_n) = O(\exp(-n\alpha))$ for some $\alpha > 0$. Here ψ_n are the Hermite functions, given by

(2.1)

$$\psi_n(x) = (-1)^n 2^{1/4}(4\pi)^{-n/2}(n!)^{-1/2} e^{\pi x^2}\left(\frac{d}{dx}\right)^n e^{-2\pi x^2} \qquad (x \in \mathbb{R},\ n = 0, 1, \cdots);$$

we have $H\psi_n = (n + \frac{1}{2})\psi_n$, where $H = (x^2 - 1/4\pi^2(d^2/dx^2))\pi$ is the Hermite operator. We denote the dual of S by S^* : an $F \in S^*$ is an antilinear continuous functional on S. We have $(F, \psi_n) = O(\exp(n\alpha))$ for all $\alpha > 0$, if $F \in S^*$. Yet another way to describe S and S^* is by means of the Bargmann transform (cf. [2], [12]): for $F \in S^*$ we let

$$(2.2) \qquad (BF)(z) = e^{\pi z^2/2}(F, g_{\bar{z}}) \qquad (z \in \mathbb{C}),$$

where, for $w\in\mathbb{C}$,

$$(2.3)\qquad g_w(t)=2^{1/4}\exp\left(-\pi(t-w)^2\right)\qquad (t\in\mathbb{C}).$$

We note that $(B\psi_n)(z)=(z\sqrt{\pi})^n/\sqrt{n!}$. Now B maps $S(S^*)$ one-to-one onto the set of all entire functions of order 2, type $<\pi/2$ (order 2, type $\leq\pi/2$). For details we refer to [12].

It is important to note that

$$(2.4)\qquad (F,G_1(a,b))=\exp\left(-\frac{\pi}{2}(a^2+b^2)\right)(BF)(a-ib)\qquad (a\in\mathbb{R},b\in\mathbb{R}),$$

where $G_\gamma(a,b)$ denotes for $\gamma>0$, $a\in\mathbb{R}$, $b\in\mathbb{R}$ the Gabor function,

$$(2.5)\qquad G_\gamma(a,b)(t)=\left(\frac{2}{\gamma}\right)^{1/4}\exp\left(-\pi\gamma^{-1}(t-a)^2+2\pi ibt-\pi iab\right)\qquad (t\in\mathbb{R}),$$

whose Wigner distribution is given by

$$(2.6)\qquad W(x,y;\,G_\gamma(a,b),\,G_\gamma(a,b))=2\exp\left(-2\pi\gamma^{-1}(x-a)^2-2\pi\gamma(y-b)^2\right)$$
$$(x\in\mathbb{R},y\in\mathbb{R}).$$

We further have

(2.7)

$$(BF)(z)=2^{1/4}(1+\alpha)^{-1/2}\exp\left(\frac{1}{2}\frac{1-\alpha}{1+\alpha}\pi z^2+\frac{2\pi\beta z}{1+\alpha}-\pi\left(\gamma-\beta^2(1+\alpha)^{-1}\right)\right)\qquad (z\in\mathbb{C}),$$

and

$$(2.8)\qquad (BF)(z)=2^{1/4}d\exp\left(-\frac{1}{2}\pi z^2+2\pi az-\pi a^2\right)\qquad (z\in\mathbb{C}),$$

where F is the f of (1.4) and (1.5) respectively. We conclude that if $P(z)=az^2+bz+c$ with $|a|\leq\pi/2$, $b\in\mathbb{C}$, $c\in\mathbb{C}$, then there is exactly one F of the form (1.4) or (1.5) such that $(BF)(z)=\exp(P(z))$.

We shall also need the operator $e^{-\alpha H}$, which can be defined on S and S^* for $\operatorname{Re}\alpha\geq 0$. We have

$$(2.9)\qquad B(e^{-\alpha H}F)(z)=e^{-\alpha/2}(BF)(ze^{-\alpha})\qquad (z\in\mathbb{C}),$$

for $F\in S^*$ (cf. [2], [12]). For $\alpha>0$, $e^{-\alpha H}$ is De Bruijn's smoothing operator N_α (cf. [4]); the kernel K_α of N_α is given by

$$(2.10)\quad K_\alpha(z,t)=(\sinh\alpha)^{-1/2}\exp\left(\frac{-\pi}{\sinh\alpha}\left((z^2+t^2)\cosh\alpha-2zt\right)\right)\qquad (z\in\mathbb{R},t\in\mathbb{R}).$$

The Wigner distribution can also be defined for $F\in S^*$; it thus becomes a generalized function of two variables. An important formula is due to Moyal (cf. [4]): if $F\in S^*, f\in S$, then

$$(2.11)\qquad (W(F,F),\,W(f,f))=|(F,f)|^2.$$

Note now that (1.3) follows from (2.6) and (2.11) in case $\delta=\gamma^{-1}$.

We shall also use the formula

(2.12)

$$W(x,y;\,N_\alpha f,N_\alpha f)=(2\sinh\alpha)^{-1}\exp(-2\pi(x^2+y^2)\tanh\alpha)$$

$$\cdot\iint\exp\left(-2\pi\coth\alpha\left((z-x/\cosh\alpha)^2+(w-y/\cosh\alpha)^2\right)\right)$$

$$\cdot W(z,w;f,f)\,dz\,dw$$

for $x\in\mathbb{R}$, $y\in\mathbb{R}$; this is just another way to write [4, Thm. 16.1]. Here $f\in S$, but it is easy to extend (2.12) so that it holds for $F\in S^*$ (cf. [10], where things like these are treated in detail).

3. The main result. In [9], a generalized function Φ of 2 variables is called nonnegative ($\Phi\geq 0$), if $(\Phi,\varphi)\geq 0$ for every nonnegative test function φ of two variables. It can be shown from the Riesz representation theorem (also cf. [9, App. 4]) that for such a Φ there is a unique Borel measure μ_Φ on $\mathbb{R}^2$, such that

$$\iint\exp(-\pi\varepsilon(x^2+y^2))d\mu_\Phi(x,y)<\infty\quad\text{for all }\varepsilon>0,$$

and such that $(\Phi,\varphi)=\iint\overline{\varphi(x,y)}\,d\mu_\Phi(x,y)$ for all test functions φ. This notion of nonnegativity agrees with the familiar notion of nonnegativity, a.e., if Φ is an ordinary function.

THEOREM 1. *Let $F\in S^*$, and assume that $W(F,F)\geq 0$. Then F is of the form* (1.4) *or* (1.5).

Proof. Let $\Phi:=W(F,F)$, and assume that $F\neq 0$. This implies by (2.11) that $\Phi\neq 0$, whence $\mu_\Phi\neq 0$. We conclude from (2.11) and (2.6) that

$$(3.1)\qquad |(F,G_1(a,b))|^2=(W(F,F),\,W(G_1(a,b),\,G_1(a,b)))$$

$$=\iint\exp\left(-2\pi(x-a)^2-2\pi(y-b)^2\right)d\mu_\Phi(x,y)>0,$$

for all $a\in\mathbb{R}$, $b\in\mathbb{R}$. That is, $(F,G_1(a,b))\neq 0$ for all $a\in\mathbb{R}$, $b\in\mathbb{R}$. We see from (2.4) that $(BF)\,(z)\neq 0$ for all $z\in\mathbb{C}$. Since BF is an entire function of order 2, type $\leq\pi/2$, we conclude that BF is of the form $(BF)(z)=\exp(P(z))$, where $P(z)=az^2+bz+c$, with $|a|\leq\pi/2$. Hence, by (2.7) and (2.8), and injectivity of B, F is of the form (1.4) or (1.5). This completes the proof. □

As an incidental note we remark that with a similar method one can show the following . Assume that $F\in S^*$ has a radially symmetric Wigner distribution. Then F is a multiple of a Hermite function ψ_n. Here we call a generalized function Φ of two variables radially symmetric if $(\Phi,\varphi\circ U_\theta)=(\Phi,\varphi)$ for all test functions φ and all $\theta\in\mathbb{R}$, where $(\varphi\circ U_\theta)(x,y)=\varphi(x\cos\theta+y\sin\theta,\,-x\sin\theta+y\cos\theta)$ for $(x,y)\in\mathbb{R}^2$. For the proof one observes that, by radial symmetry of $W(F,F)$ and $W(G_1(0,0),\,G_1(0,0))$ and (3.1), the expression $|(F,G_1(a,b))|^2$ only depends on a^2+b^2. This implies that $|(BF)(z)|$ only depends on $|z|$, whence, by the maximum modules principle, $(BF)(z)=cz^n$ for some $c\in\mathbb{C}$, $n=0,1,\cdots$. Hence $F=d\psi_n$ for some $d\in\mathbb{C}$. Also see [11], [13], where it is proved that

$$W(x,y;\,\psi_n,\psi_n)=2(-1)^n\exp(-2\pi(x^2+y^2))L_n(4\pi(x^2+y^2)),$$

with L_n the nth Laguerre polynomial.

It is fairly easy to generalize the previous theorem as follows.

THEOREM 2. *Let $F \in S^*$, $\delta > 0$, $\gamma > 0$, $\delta\gamma > 1$, and assume that F is not of the form* (1.4) *or* (1.5). *Then the convolution of $W(F,F)$ with* $\exp(-2\pi\delta x^2 - 2\pi\gamma y^2)$ *takes negative values.*

Proof. We see from (2.9) that $N_\alpha F$ is not of the form (1.4) or (1.5) if $\alpha > 0$. Hence, by the previous theorem, $W(N_\alpha F, N_\alpha F)$ takes negative values. Then (2.12) shows that the convolution of $W(F,F)$ and $\exp(-2\pi \coth\alpha(x^2+y^2))$ takes negative values. This proves the theorem in case $\gamma = \delta = \coth\alpha$.

In general we can express, by a transformation of variables and (3.2) below, the convolution of $\exp(-2\pi\delta x^2 - 2\pi\gamma y^2)$ and $W(F,F)$ at the point (a,b), as the inner product of $\exp(-2\pi\rho((x-a\varepsilon)^2+(y-b\varepsilon^{-1})^2))$ and $W(Z_\varepsilon F, Z_\varepsilon F)$. Here $\rho = (\delta\gamma)^{1/2}$, $\varepsilon = (\delta/\gamma)^{1/4}$ and Z_ε is the operator defined by $(Z_\varepsilon f)(t) = \varepsilon^{-1/2} f(\varepsilon^{-1} t)$ for $f \in S$, and extended in the obvious way (cf. [10, 1.15]) to S^*. We use here that for $f \in S$, $x \in \mathbb{R}$, $y \in \mathbb{R}$,

$$W(\varepsilon^{-1}x, \varepsilon y; f, f) = W(x, y; Z_\varepsilon f, Z_\varepsilon f), \tag{3.2}$$

a formula that can be generalized straightforwardly so as to hold for $f \in S^*$ as well. It is clear that if F is not of the form (1.4) or (1.5), then neither is $Z_\varepsilon F$. Since we can find an $\alpha > 0$ such that $\rho = \coth\alpha$, we conclude from the special case already treated that the proof is complete. □

4. An application. It is believed that the only curve a Wigner distribution can be concentrated on is a straight line[1]; this is true only if certain restrictions on the curve are imposed (cf. the examples at the end of this section). We shall give a proof for the following simple case. Let C be a continuously differentiable curve in the plane with parametrization $\gamma\colon \mathbb{R} \to \mathbb{R}^2$, where we assume that $|\gamma'(t)| > 0$ for all t. Assume that for all $t_0 \in \mathbb{R}$ there is a straight line l passing through $\gamma(t_0)$, but not tangent to C, such that there is $\varepsilon > 0$, $\delta > 0$, with the property that the distance between $\gamma(s)$ and $l \geq \varepsilon$, if $|\gamma(s) - \gamma(t_0)| \geq \delta$. This condition is satisfied, e.g., if C is the graph of a continuously differentiable function defined on $\mathbb{R}$. Now let $F \in S^*$ be a function whose Wigner distribution is concentrated on C in the following sense: there is a continuous function $h\colon C \to \mathbb{R}$, such that $h(\gamma(t)) = O(\exp(\varepsilon|\gamma(t)|^2))$ for all $\varepsilon > 0$, and

$$(W(F,F), \varphi) = \int_{-\infty}^{\infty} h(\gamma(t))\varphi(\gamma(t))|\gamma'(t)|\,dt \tag{4.1}$$

for all test functions φ of two variables. We shall show that this implies that F is of the form (1.4) (with $\operatorname{Re}\alpha = 0$) or (1.5), so that, in particular, C is a straight line. To this end let $\gamma(t_0) = (a,b)$ be a point on C and consider for $\operatorname{Re}\alpha > 0$ the function $g_{\alpha,a,b}$, given by

$$g_{\alpha,a,b}(t) = \exp(-\pi\alpha(t-a)^2 + 2\pi i b t - \pi i a b) \qquad (t \in \mathbb{R}), \tag{4.2}$$

whose Wigner distribution $W_{\alpha,a,b}$ is given by

(4.3)

$$W_{\alpha,a,b}(x,y) = \left(\frac{2}{\operatorname{Re}\alpha}\right)^{1/2} \exp(-2\pi(x-a)^2 \operatorname{Re}\alpha - 2\pi(y-b+(x-a)\operatorname{Im}\alpha)^2/\operatorname{Re}\alpha).$$

We have by (2.11) and (4.1)

$$0 \leq (W(F,F), W_{\alpha,a,b}) = \int_{-\infty}^{\infty} h(\gamma(t)) W_{\alpha,a,b}(\gamma(t))|\gamma'(t)|\,dt. \tag{4.4}$$

[1] Cf. [1]. I thank Alan Weinstein for calling my attention to this paper.

Now let l be the line through $\gamma(t_0)$ whose existence is assured by our assumptions, and take α such that $\{(x,y)|y=b-(x-a)\operatorname{Im}\alpha\}$ is the graph of l. (If l is parallel to the y-axis we can use a similar argument with

$$\left(\frac{2}{\gamma}\right)^{1/2}\exp\left(-\pi\gamma^{-1}(t-a)^2+2\pi ibt-\pi iab\right)$$

instead of $g_{\alpha,a,b}$, where we take $\gamma\to 0$). If we let $\operatorname{Re}\alpha\to 0$, the right-hand side of (4.4) tends to $C_0h(\gamma(t_0))|\gamma'(t_0)|$, where $C_0>0$ is a number that depends only on the angle between l and the tangent line at C through $\gamma(t_0)$. Hence $h(\gamma(t_0))\geq 0$. We easily see from our theorems and (1.6)–(1.8) that F is of the form (1.4) (with $\operatorname{Re}\alpha=0$) or (1.5).

Notes. The condition "h continuous" can be relaxed to "h measurable" at the expense of elegance of the proof. It is furthermore likely that the conditions on the curve C can be relaxed somewhat as well. On the other hand, consider the function $f=\Sigma_n\delta_n$, whose Wigner distribution is given by $\frac{1}{2}\Sigma_{k,l}(-1)^{kl}\delta_{k/2}\otimes\delta_{l/2}$, where the summations are over all integers (this follows from a straightforward calculation and the Poisson summation formula, written in the form $\Sigma_n\delta_n(x)=\Sigma_n e^{-2\pi inx}$). The points of the lattice $(\frac{k}{2},\frac{l}{2})$ can be joined by a smooth curve C; such a C does not satisfy our assumptions, of course. Another objection is that the function h cannot be continuous in this case. This is not a serious point, however, as can be shown as follows. Let k_0: $\mathbb{R}\to\mathbb{R}$ be continuous, and assume that k_0 vanishes outside $[-\frac{1}{8},\frac{1}{8}]$. The Wigner distribution of k_0*f (where f is as above and $*$ denotes convolution) is obtained by convolving $W(f,f)$ and $W(k_0,k_0)$ with respect to the first variable (cf. [5, 4.1]). We get

$$W(x,y;\,k_0*f,k_0*f)=\frac{1}{2}\sum_{k,l}(-1)^{kl}W\left(x-\frac{k}{2},y;\,k_0,k_0\right)\delta_{l/2}(y) \tag{4.5}$$

(this formula can also be derived by directly using the Poisson summation formula). Since $W(k_0,k_0)$ is concentrated in the strip $[-\frac{1}{8},\frac{1}{8}]\times\mathbb{R}$, we see that $W(k_0*f,k_0*f)$ is concentrated in the set $\{(x+\frac{k}{2},\frac{l}{2})||x|\leq\frac{1}{8},\,k\in\mathbb{Z},\,l\in\mathbb{Z}\}$. The components of this set can be embedded in a smooth curve, and the function h now becomes continuous, since $W(k_0,k_0)$ is continuous.

A second example showing that one has to be careful with the statement, "$W(f,f)$ cannot be concentrated on a curve unless this curve is a straight line," is the function $f(t)=\cos 2\pi t$, whose Wigner distribution equals $\frac{1}{4}(\delta_{2\pi}(y)+\delta_{-2\pi}(y)+2\delta_0(y)\cos 4\pi x)$. Now $W(f,f)$ is concentrated on the three lines $y=0$, $y=\pm 2\pi$, and these lines can be embedded in a smooth curve.

REFERENCES

[1] N. L. BALAZS, *Weyl's association, Wigner's function and affine geometry*, Physica, 102A (1980), pp. 236–254.

[2] V. BARGMANN, *On a Hilbert space of analytic functions and an associated integral transform, part* I, Comm. Pure Appl. Math., 14 (1961), pp. 187–214.

[3] N. G. DE BRUIJN, *Uncertainty principles in Fourier analysis*, in Inequalities, O. Shisha, ed., Academic Press, New York, 1967, pp. 57–71.

[4] ______, *A theory of generalized functions, with applications to Wigner distribution and Weyl correspondence*, Nieuw Archief voor Wiskunde, 21 (1973), pp. 205–280.

[5] T. A. C. M. CLAASEN AND W. F. G. MECKLENBRÄUKER, *The Wigner distribution—A tool for time-frequency signal analysis, Parts* I, II *and* III, Philips J. Res., 35 (1980), pp. 217–250, 276–300, 372–389.

[6] I. M. GELFAND AND G. E. SHILOV, *Generalized Functions*, Vol. 2, Academic Press, New York, 1969.

[7] A. GROSSMAN, G. LOUPIAS AND E. M. STEIN, *An algebra of pseudo-differential operators and quantum mechanics in phase space*, Ann. Inst. Fourier, 18 (1969), pp. 343–368.

[8] R. L. HUDSON, *When is the Wigner quasi-probability density non-negative*, Rep. Math. Phys., 6 (1974), pp. 249–252.

[9] A. J. E. M. JANSSEN, *Application of the Wigner distribution to harmonic analysis of generalized stochastic processes*, MC-tract 114, Mathematisch Centrum, Amsterdam, 1979.

[10] ______, *Convolution theory in a space of generalized functions*, Proc. K.N.A.W., Ser. A, 82 (1979), pp. 283–305.

[11] ______, *Positivity of weighted Wigner distributions*, this Journal, 12 (1981), pp. 752–758.

[12] ______, *Bargmann transform, Zak transform and coherent states*, J. Math. Phys., 23 (1982), pp. 720–731.

[13] J. PEETRE, *The Weyl transform and Laguerre polynomials*, Matematiche (Catania), 27 (1972), pp. 301–323.

[14] H. WEYL, *Theory of Groups and Quantum Mechanics*, Dover, New York, 1950.

[15] E. WIGNER, *On the quantum correction for thermodynamic equilibrium*, Phys. Rev., 40 (1932), pp. 749–759.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics

0036-1410/84/1501-0015 $01.25/0

GENERALIZED FOCK SPACES AND ASSOCIATED OPERATORS*

FRANK M. CHOLEWINSKI†

Abstract. A class of generalized Fock spaces associated with Bessel functions is studied. The generalized Fock space is a Hilbert space of even entire functions weighted by a modified Bessel function of the third kind, whereas ordinary Fock space is a Hilbert space of entire functions of several complex variables weighted by a Gaussian kernel. The generalized Fock space has a reproducing kernel which is a modified Bessel function of the first kind.

Commutator relations between the Schrödinger radial kinetic energy operator and multiplication by z^2 lead to a generalized class of Weyl relations for the Bessel functions.

1. Introduction. In a series of papers, V. Bargmann [2]–[4] studied a family of Hilbert spaces, whose elements are entire functions of n complex variables. These Hilbert spaces are associated with Fock's [10] realization of the creation and annihilation operators of Bose particles in quantum field theory.

If q and p are selfadjoint operators on a Hilbert space $\mathcal{H}$ satisfying the canonical commutation rule

$$[p,q]=-iI, \quad \text{with Planck's constant } h=2\pi, \tag{1.1}$$

and if

$$P=2^{-1/2}(q+ip) \quad \text{and} \quad Q=2^{-1/2}(q-ip),$$

then $P^*=Q$, $Q^*=P$ and

$$[P,Q]=I. \tag{1.2}$$

Fock [10] introduced the operator solution $P=\frac{\partial}{\partial Q}$ of the commutation rule (1.2) and applied it to quantum field theory. Bargmann obtained a realization of Fock space $\mathcal{F}$ as a space of entire functions weighted by a Gaussian function.

In this paper a Hilbert space of entire functions on which the square of the position operator and the generalized radial kinetic energy operators are adjoints is obtained. The Hilbert space is weighted by a modified Bessel function of the third kind. As a matter of convenience, we deal primarily with the one variable case throughout the paper.

Fock space (also known as Fischer space) $\mathcal{F}$ is the Hilbert space of entire functions with inner product given by

$$(f|g)=\frac{1}{\pi}\int_C f(z)\overline{g(z)}e^{-|z|^2}\,dx\,dy, \qquad z=x+iy, \tag{1.3}$$

where C denotes the complex numbers. Thus the growth of functions in $\mathcal{F}$ is dominated by $\exp(|z|^2/2)$. Let $f,g\in\mathcal{F}$, with Taylor series expansions

$$f(x)=\sum_{n=0}^{\infty} a_n z^n \quad \text{and} \quad g(z)=\sum_{n=0}^{\infty} c_n z^n.$$

*Received by the editors January 5, 1982, and in revised form September 23, 1982.

†Department of Mathematical Sciences, Clemson University, Clemson, South Carolina 29631.

Then

$$(f|g)=\sum_{n=0}^{\infty} a_n\bar{c}_n n! \tag{1.4}$$

and

$$\|f\|^2=\sum_{n=0}^{\infty} |a_n|^2 n! \tag{1.5}$$

The existence of a reproducing kernel in Fock space is of fundamental importance. Let

$$\mathcal{K}(z,w)=e^{z\bar{w}}, \qquad z,w\in C. \tag{1.6}$$

Then

$$f(w)=\left(f(z)|e^{z\bar{w}}\right) \quad \text{for all } w\in C. \tag{1.7}$$

Thus, the Dirac delta function in Fock space is the exponential function. If we define the multiplication and differentiation operators on $\mathcal{F}$ by

$$(Zf)(z)=zf(z) \quad \text{and} \quad Df(z)=\frac{df(z)}{dz},$$

then

$$(Df|g)=(f|zg),$$

that is, D and Z are adjoints. Furthermore, D and Z satisfy the commutation rule

$$[D,Z]=I$$

of the annihilation and creation operators for "bosons" in quantum theory.

Let ν be a fixed positive number. The generalized Schrödinger radial kinetic energy operator is given by

$$\Delta_x=\frac{d^2}{dx^2}+\frac{2\nu}{x}\frac{d}{dx}. \tag{1.8}$$

Δ_x is the familiar radial part of the Laplace operator on n-dimensional Euclidean space E^n, with $2\nu=n-1$. The operator Δ_x is also known as the Euler or Bessel operator; see Weinstein [22]. In this paper we construct a Hilbert space of entire functions on which Δ_x and x^2 are adjoints.

Let

$$\mathbf{K}_\nu(x)=\mathbf{K}(x)=x^{1/2-\nu}K_{\nu-1/2}(x) \tag{1.9}$$

where $K_{\nu-1/2}(x)$ is a modified Bessel function of the third kind. The generalized Fock space $\mathcal{F}_\nu$, introduced in this paper, is the Hilbert space of even entire functions with inner product given by

$$(f|g)=\int_C f(z)\overline{g(z)}\,d\mathfrak{m}_\nu(z) \tag{1.10}$$

where

$$(1.11)\qquad d\mathfrak{m}_\nu(z)=\frac{\mathbf{K}_\nu(|z|^2)r^{4\nu+1}}{\pi 2^{\nu-1/2}\Gamma(\nu+1/2)}dr\,d\theta,\qquad z=re^{i\theta}.$$

If $f,g\in\mathcal{F}_\nu$ have Taylor series expansions given by

$$f(z)=\sum_{n=0}^{\infty}a_nz^{2n}\quad\text{and}\quad g(z)=\sum_{n=0}^{\infty}c_nz^{2n},$$

then

$$(1.12)\qquad (f|g)_\nu=\sum_{n=0}^{\infty}a_n\bar{c}_nb_{2n}(\nu)$$

where

$$(1.13)\qquad b_{2n}(\nu)=2^{2n}\frac{n!\Gamma(n+\nu+1/2)}{\Gamma(\nu+1/2)}.$$

The generalized Fock space $\mathcal{F}_\nu$ also has a reproducing kernel. Let

$$(1.14)\qquad \mathcal{K}_\nu(z,w)=\mathbf{I}_\nu(z\bar{w}),\qquad z,w\in C,$$

where

$$\mathbf{I}_\nu(z)=2^{\nu-1/2}\Gamma(\nu+1/2)z^{1/2-\nu}I_{\nu-1/2}(z)$$

and $I_{\nu-1/2}(z)$ is a modified Bessel function of the first kind of order $\nu-\frac{1}{2}$. If $f\in\mathcal{F}_\nu$, then we have

$$(1.15)\qquad f(w)=(f(z)|\mathbf{I}_\nu(z\bar{w}))_\nu\quad\text{for all } w\in C.$$

Thus the modified Bessel function serves as the generalized Dirac delta function in $\mathcal{F}_\nu$.

The Schrödinger radial kinetic energy operator and multiplication by z^2 are defined on $\mathcal{F}_\nu$ by

$$\Delta_zf(z)=\frac{d^2f(z)}{dz^2}+\frac{2\nu}{z}\frac{d}{dz}f(z)$$

and

$$(Q^2)f(z)=z^2f(z).$$

In $\mathcal{F}_\nu$ we have

$$(\Delta_zf|g)=(f|z^2f).$$

Thus Δ_z and z^2 are adjoints in $\mathcal{F}_\nu$.

Moreover, Δ_z and z^2 satisfy a commutation rule,

$$(1.16)\qquad [\Delta_z,z^2]=4\left(\nu+\frac{1}{2}\right)I+4z\frac{d}{dz}.$$

This commutator rule leads to a generalized class of Weyl relations for the Bessel functions.

2. Definitions and preliminary results. Let ν be a fixed positive number. We set

$$d\mu_\nu(x)=\frac{x^{2\nu}dx}{2^{\nu-1/2}\Gamma(\nu+1/2)}. \tag{2.1}$$

Let

$$\mathbf{I}_\nu(x)=\mathbf{I}(x)=2^{\nu-1/2}\Gamma\left(\nu+\frac{1}{2}\right)X^{1/2-\nu}I_{\nu-1/2}(x)$$

where $I_{\nu-1/2}(x)$ is a modified Bessel function of the first kind. Then we have

$$\mathbf{I}_\nu(x)=\sum_{n=0}^{\infty}\frac{x^{2n}}{b_{2n}(\nu)}, \tag{2.2}$$

with $b_{2n}(\nu)$ given by (1.12). Further we let

$$\mathbf{J}_\nu(x)=2^{\nu-1/2}\Gamma\left(\nu+\frac{1}{2}\right)x^{1/2-\nu}J_{\nu-1/2}(x) \tag{2.3}$$

and

$$\mathbf{K}_\nu(x)=x^{1/2-\nu}K_{\nu-1/2}(x) \tag{2.4}$$

where $J_{\nu-1/2}(x)$ is the ordinary Bessel function of order $\nu-1/2$, and $K_{\nu-1/2}(x)$ is a modified Bessel function of the third kind [9, p. 5].

It readily follows that

$$\Delta_x\mathbf{I}(wx)=w^2\mathbf{I}(wx),\qquad \Delta_x\mathbf{K}(wx)=w^2\mathbf{K}(wx)\quad \Delta_x\mathbf{J}(wx)=-w^2\mathbf{J}(wx). \tag{2.5}$$

Further, we define

$$D(x,y,z)=\frac{2^{(3\nu-5/2)}\Gamma(\nu+1/2)^2}{\Gamma(\nu)\pi^{1/2}}(xyz)^{1-2\nu}\Delta(x,y,z)^{2\nu-2} \tag{2.6}$$

where $\Delta(x,y,z)$ is the area of a triangle whose sides are x, y, z if there is such a triangle and otherwise $D(x,y,z)$ is zero. Then we have that

$$\int_0^\infty \mathbf{J}(zt)D(x,y,z)\,d\mu(z)=\mathbf{J}(xt)\mathbf{J}(yt) \tag{2.7}$$

and

$$\int_0^\infty \mathbf{J}(xt)\mathbf{J}(yt)\mathbf{J}(zt)\,d\mu(t)=D(x,y,z), \tag{2.8}$$

valid for $0<x,y<\infty$, $0\leq t<\infty$; see [20, p. 411].

Next we define $L_\nu^p(0,\infty)$, $1\leq p<\infty$, as the Banach space of measurable functions on $(0,\infty)$ for which

$$\|f\|_p=\left[\int_0^\infty|f(x)|^p d\mu(x)\right]^{1/p}<\infty.$$

We let $\mathcal{H}_\nu$ denote the Hilbert space $L_\nu^2(0,\infty)$.

DEFINITION 2.1. Let ϕ be a locally integrable function on $[0,\infty)$. We define the generalized translation function $\phi(x \circledS y)$ by

$$\phi(x \circledS y)=\int_0^\infty\phi(z)D(x,y,z)\,d\mu(z). \tag{2.9}$$

Equation (2.5) yields

$$\mathbf{I}_\nu(y\Delta_x^{1/2})\mathbf{J}_\nu(x)=J(x \textcircled{Δ} y)=\mathbf{J}(x)\mathbf{J}(y). \tag{2.10}$$

Thus it follows that operationally

$$\mathbf{I}_\nu(y\Delta_x^{1/2})\phi(x)=\phi(x \textcircled{Δ} y); \tag{2.11}$$

see [7]; that is, $\mathbf{I}_\nu(y\Delta_x^{1/2})$ is the generalized translation operator. Moreover,

$$\|\phi(x \textcircled{Δ} y)\|_2=\|\phi(x)\|_2 \quad \text{on } \mathcal{H}_\nu. \tag{2.12}$$

The generalized translation operator is extended to even entire functions as follows: If $f(x)$ be an even entire function with Taylor expansion $f(x)=\Sigma a_n x^{2n}$, then

$$f(z \textcircled{Δ} w)=\sum_{n=0}^{\infty} a_n(z \textcircled{Δ} w)^{2n} \tag{2.13}$$

where we define

$$\begin{aligned}(w \textcircled{Δ} z)^{2n}&=\mathbf{I}_\nu(w\Delta_z^{1/2})z^{2n}\\ &=z^{2n}{}_2F_1\left(-n,-n-\nu+\frac{1}{2};\nu+\frac{1}{2};\frac{w^2}{z^2}\right)\\ &=\sum_{k=0}^{n}\binom{n}{k}\frac{\Gamma(n+\nu+1/2)\Gamma(\nu+1/2)}{\Gamma(n-k+\nu+1/2)\Gamma(k+\nu+1/2)}z^{2(n-k)}w^{2k};\end{aligned} \tag{2.14}$$

see [6, p. 7] for the details. The generalized translation given by (2.13) is the natural notion of translation on $\mathcal{F}_\nu$, since $\mathcal{F}_\nu$ is not closed under ordinary translation of variables. If $f(x)\in\mathcal{F}_\nu$ is an even entire function, then $f(z \textcircled{$\Delta$} w)$ is also an even entire function. However, $f(z \textcircled{$\Delta$} w)$ is not necessarily an element of $\mathcal{F}_\nu$.

In the estimation of certain integrals, we will need the following asymptotic expansion for the modified Bessel functions of the first and third kind:

$$\begin{aligned}I_\gamma(z)=(2\pi z)^{-1/2}\Bigg\{e^z&\left[\sum_{m=0}^{M-1}(-1)^m(\gamma,m)(2z)^{-m}+O\left(|z|^{-M}\right)\right]\\ &+ie^{-z+i\gamma}\left[\sum_{m=0}^{M-1}(\gamma,m)(2z)^{-m}+O\left(|z|^{-M}\right)\right]\Bigg\},\\ &\qquad -\tfrac{1}{2}\pi<\arg z<\tfrac{1}{2}\pi,\end{aligned} \tag{2.15}$$

$$K_\gamma(z)=\left[\frac{\pi}{2z}\right]^{1/2}e^{-z}\left\{\sum_{m=0}^{M-1}(\gamma,m)(2z)^{-m}+O\left(|z|^{-M}\right)\right\},\quad -\tfrac{3}{2}\pi<\arg z<\tfrac{3}{2}\pi; \tag{2.16}$$

see Erdélyi [9, p. 36]. In these formulas (γ,m) is the Hankel symbol

$$(\gamma,m)=\frac{\Gamma(1/2+\gamma+m)}{m!\Gamma(1/2+\gamma-m)}.$$

3. Generalized Fock space.

DEFINITION 3.1. Let $\nu>0$. The space $\mathcal{F}_\nu$ is the Hilbert space of even entire functions on C, the complex numbers, with inner product defined by

$$(f|g)_\nu=\int_C f(z)\bar{g}(z)\,d\mathfrak{m}_\nu(z) \tag{3.1}$$

where $d\mathfrak{m}_\nu(z)$ is given by (1.11). An even entire function f belongs to $\mathcal{F}_\nu$ if and only if $\|f\|_\nu^2=(f|f)_\nu<\infty$. If $f(z)=\Sigma a_n z^{2n}\in\mathcal{F}_\nu$, we define

$$f_\#(z)=\sum_{n=0}^{\infty}\bar{a}_n z^{2n}. \tag{3.2}$$

THEOREM 3.2. *If $f,g\in\mathcal{F}_\nu$ with $g(z)=\Sigma_{n=0}^{\infty}c_n z^{2n}$, then*

$$(f(z)|g(z))_\nu=\sum_{n=0}^{\infty}a_n\bar{c}_n b_{2n}(\nu) \tag{3.3}$$

and

$$(f(z)|g(z))_\nu=f(\Delta_\nu^{1/2})g_\#(z)|_{z=0}. \tag{3.4}$$

Proof. We have

$$I=\int_C f(z)g(z)\,d\mathfrak{m}_\nu(z)=\sum_{n,m=0}^{\infty}a_n\bar{c}_n\int_C z^{2n}z^{-2m}\,d\mathfrak{m}_\nu(z).$$

The term-by-term integration is justified by the absolute convergence of the integral and by the Tonelli–Hobson theorem. We also note that $\mathbf{K}_\nu(|z|^2)$ is positive. Let $z=re^{i\theta}$. Then

$$\begin{aligned}\int_C z^{2n}z^{-2m}\,d\mathfrak{m}_\nu(z)&=\int_0^\infty r^{2(n+m+2)}\frac{\mathbf{K}_\nu(r^2)2r\,dr}{2^{\nu-1/2}\Gamma(\nu+1/2)}\frac{1}{2\pi}\int_0^{2\pi}e^{i(n-m)\theta}\,d\theta\\&=\delta_{n,m}\int_0^\infty r^{4(n+\nu)}\frac{\mathbf{K}_\nu(r^2)2r\,dr}{2^{\nu-1/2}\Gamma(\nu+1/2)}\\&=\frac{2^{2n}n!\Gamma(n+\nu+1/2)}{\Gamma(\nu+1/2)}\delta_{n,m},\end{aligned} \tag{3.5}$$

where $\delta_{n,m}$ is the Kronecker delta function. The evaluation of the integral is given by Erdélyi et al. [9, p. 51]. It follows that

$$I=\sum_{n=0}^{\infty}a_n\bar{c}_n b_{2n}(\nu).$$

In order to prove the second part of the theorem, we note that $g(z)$ can be rewritten as

$$g(z)=\sum_{n=0}^{\infty}\frac{\Delta^n g\cdot(0)}{b_{2n}(\nu)}z^{2n}=\sum_{n=0}^{\infty}c_n z^{2n};$$

see J. Delsarte [8] or Cholewinski [5, p. 57]. Using the first part of the theorem and (3.5), we find that

$$(f(z)|g(z))_\nu=\sum_{n,m=0}^{\infty}\frac{\overline{\Delta^m g\cdot(0)}}{b_{2m}}\left(z^{2n}|z^{2m}\right)_\nu$$

$$=\sum_{n=0}^{\infty}a_n\overline{\Delta^n g}\cdot(0)=\sum_{n=0}^{\infty}a_n\Delta^n g_{\#}(z)|_{z=0}=f(\Delta^{1/2})g_{\#}(z)|_{z=0}.$$

COROLLARY 3.3. *Let $e_n(z)=z^{2n}/b_{2n}^{1/2}$. Then the family $A_\nu=\{e_n(z)\}_{n=0}^{\infty}$ forms an orthonormal basis for $\mathfrak{F}_\nu$.*

Proof. From the proof of Theorem 3.2, we have

$$(3.6)\qquad \left(z^{2n}|z^{2m}\right)_\nu=b_{2n}(\nu)\delta_{n,m},$$

or equivalently,

$$(3.7)\qquad (e_n|e_m)_\nu=\delta_{n,m}.$$

Thus (3.3) is an expression of Parseval's identity, and therefore A_ν is complete in $\mathfrak{F}_\nu$. The Fourier coefficient of $f\in\mathfrak{F}_\nu$ is given by

$$(3.8)\qquad (f|e_n)_\nu=a_nb_{2n}(\nu)^{1/2}.$$

DEFINITION 3.4. An even entire function $f(z)=\Sigma a_nz^{2n}$ is said to be of growth $\{\rho,\tau\}$ if and only if

$$(3.9)\qquad \limsup_{n\to\infty}\frac{2n}{e\rho}|a_n|^{\rho/2n}\le\tau.$$

Thus $f(z)$ is entire and of order <2, or f is of order 2 and of type $\le\tau$. It follows that $f\in\{\rho,\tau\}$ if and only if, for every $\varepsilon>0$,

$$f(z)=O\left(e^{(\tau+\varepsilon)|z|^2}\right),\qquad |z|\to\infty.$$

THEOREM 3.5. *If $f\in\mathfrak{F}_\nu$, then f is of growth $\{\rho,\tau\}\le\{2,\tfrac{1}{2}\}$.*

Proof. Since $f\in\mathfrak{F}_\nu$, we have by (3.3)

$$(3.10)\qquad |a_n|^2 2^{2n}n!\Gamma\left(\nu+\frac{1}{2}+n\right)=o(1).$$

Using Stirling's formula a simple calculation shows that

$$\limsup_{n\to\infty}\frac{2n\log 2n}{\log 1/|a_n|}\le 2.$$

Thus f is of order ≤ 2. Let $T(n)=\frac{n}{e}|a_n|^{1/n}$. Another application of Stirling's formula yields

$$\lim_{n\to\infty}\frac{n^2}{e^2\left[2^{2n}n!\Gamma(n+\nu+1/2)\right]^{1/n}}=\frac{1}{4}.$$

By (3.10), we have $|a_n|^2 2^{2n}n!\Gamma(\nu+\frac{1}{2}+n)\le 1$ for n sufficiently large. It follows that

$$\limsup_{n\to\infty}T(n)^2\le\tfrac{1}{4}.$$

Hence f is of growth $\le\{2,\tfrac{1}{2}\}$.

Corollary 3.6. *If f is an even entire function of growth $\{\rho,\tau\}<\{2,\frac{1}{2}\}$, then $f\in\mathcal{F}_\nu$.*
Proof. Since

$$|f(z)|^2\mathbf{K}_\nu\left(|z|^2\right)=O\left(e^{2(\tau+\varepsilon)|z|^2}e^{-|z|^2}\right)$$

where $\tau+\varepsilon<\frac{1}{2}$, it follows that the integral form of the norm of f is finite.

Thus it follows that any even entire function of order $\rho<2$ is in $\mathcal{F}_\nu$. Furthermore, the even entire function of order $\rho=2$ for which $\tau<\frac{1}{2}$ also belong to $\mathcal{F}_\nu$. The function $f(z)=e^{z^2/2}$ is of growth $\{2,\frac{1}{2}\}$. However, $e^{z^2/2}\notin\mathcal{F}_\nu$.

Let $f(z)=\sum_{n=0}^{\infty}a_nz^{2n}$ be an element of $\mathcal{F}_\nu$. Using the Cauchy–Schwarz inequality, we find that

$$|f(z)|^2\leq\left[\sum_{n=0}^{\infty}|a_n|^2b_{2n}\right]\left[\sum_{n=0}^{\infty}\frac{|z|^{4n}}{b_{2n}}\right]=\|f\|_\nu^2\mathbf{I}_\nu\left(|z|^2\right).$$

Therefore

$$|f(z)|\leq\|f\|_\nu\left\{\mathbf{I}_\nu\left(|z|^2\right)\right\}^{1/2}. \tag{3.11}$$

It follows that strong convergence in $\mathcal{F}_\nu$ implies local uniform convergence on C. Using the standard asymptotic expansion

$$I_{\nu-1/2}(|z|)\sim\frac{e^{|z|}}{(2\pi|z|)^{1/2}}\quad\text{as } z\to\infty,$$

we obtain

$$|f(z)|\leq Me^{|z|^2/2}\|f\|_\nu. \tag{3.12}$$

Let $\mathcal{F}$ denote the ordinary Fock space of entire functions on C given by (1.3) or (1.4) with norm $\|f\|$.

Theorem 3.7. *Under the natural injection $\mathcal{F}_\nu$ is a subspace of $\mathcal{F}$ and $\|f\|\leq\|f\|_\nu$.*

Proof. Now $\|z^{2n}\|_\nu^2=b_{2n}(\nu)$ and $\|z^{2n}\|^2=(2n)!$. By Legendre's duplication formula, we have

$$(2n)!=2^{2n}n!\frac{\Gamma(n+1/2)}{\Gamma(1/2)}=b_{2n}(0).$$

It easily follows that $(2n)!\leq b_{2n}(\nu)$. Hence

$$\|f\|=\sum_{n=0}^{\infty}|a_n|^2(2n)!<\sum_{n=0}^{\infty}|a_n|^2b_{2n}(\nu)=\|f\|_\nu^2. \tag{3.13}$$

4. The reproducing kernel for $\mathcal{F}_\nu$. Let F be a class of functions defined on a set E, forming a Hilbert space. A function $\mathcal{K}(x,y)$ of x and y in E is called a reproducing kernel of F if the following two properties hold:

(a) For every $y\in E$, $\mathcal{K}(x,y)$ as a function of x belongs to F.

(b) *The reproducing property.* For every $y\in E$ and $f\in F$, we have

$$f(y)=(f(x)|\mathcal{K}(x,y))_x.$$

It is well known that if a reproducing kernel exists it is unique. Furthermore, for the existence of a reproducing kernel $\mathcal{K}(x,y)$, it is necessary and sufficient that for every y in E, $f(y)$ is a continuous linear functional on F; see Aronszajn [1] for the basic theory.

Inequality (3.12) shows that the map $f \to f(z)$, $z \in C$, is a continuous linear functional on $\mathcal{F}_\nu$. Thus $\mathcal{F}_\nu$ has a reproducing kernel.

THEOREM 4.1. *Let*

$$\mathcal{K}_\nu(z,w) = \mathbf{I}_\nu(z\bar{w}), \qquad z, w \in C. \tag{4.1}$$

Then $\mathcal{K}_\nu$ is a reproducing kernel for $\mathcal{F}_\nu$.

Proof. Using the standard asymptotic expansions for $I_{\nu-1/2}$ and $K_{\nu-1/2}$, it follows that $\mathcal{K}_\nu(z,w) \in \mathcal{F}_\nu$ as a function of z for $w \in C$. Let $z, w \in C$ then

$$\mathbf{I}_\nu(z\bar{w}) = \sum_{n=0}^{\infty} \frac{\bar{w}^{2n}}{b_{2n}(\nu)} z^{2n}. \tag{4.2}$$

If $f(z) = \Sigma a_n z^{2n} \in \mathcal{F}_\nu$, it follows from Theorem 3.2 that

$$(f(z) | \mathbf{I}_\nu(z\bar{w}))_\nu = \sum_{n=0}^{\infty} a_n \frac{w^{2n}}{b_{2n}} b_{2n} = f(w).$$

COROLLARY 4.2. *The set $\xi_\nu = \{\mathbf{I}_\nu(z\bar{w}) | w \in C\}$ is complete in $\mathcal{F}_\nu$.*

This is a well-known property of reproducing kernels in general. The linear combinations of elements in ξ_ν are the generalized exponential polynomials associated with the Δ_x operator. We call a finite linear combination of elements of ξ_ν an I-function. Let $\underline{w} = \{w_1, w_2, \cdots, w_m\} \in C^m$ with $w_1 \neq w_2 \neq \cdots \neq w_m$ and let $\underline{a} = \{a_1, \cdots, a_m\} \in C^m$, $\underline{a} \neq \underline{0}$. Let

$$g_{\underline{a}}(z) = \sum_{k=1}^{m} a_k \mathbf{I}_\nu(z\bar{w}_k). \tag{4.3}$$

Then

$$\|g_{\underline{a}}\|_\nu^2 = \sum_{k,j=1}^{m} \mathbf{I}_\nu(\bar{w}_k w_j) a_k \bar{a}_j > 0.$$

Thus

$$\mathcal{G}(\underline{w}) = \left[\mathbf{I}_\nu(w_i \bar{w}_j)\right]_{i,j=1}^{m}$$

is a positive hermitian matrix. Hence, we obtain the following result.

COROLLARY 4.3. *Let $\underline{w} = (w_1, w_2, \cdots, w_m) \in C^m$, with distinct elements w_k. Then*

$$\det\left[\mathbf{I}(w_i \bar{w}_j)\right]_{i,j=1}^{m} > 0.$$

Taking the product $\mathcal{F}_{\nu_1} \otimes \mathcal{F}_{\nu_2} \otimes \cdots \otimes \mathcal{F}_{\nu_n}$, it follows that

$$\mathcal{K}_{\underline{\nu}}(\underline{z}, \underline{w}) = \mathbf{I}_{\nu_1}(z_1\bar{w}_1)\mathbf{I}_{\nu_2}(z_2\bar{w}_2) \cdots \mathbf{I}_{\nu_n}(z_n\bar{w}_n) = \mathbf{I}_{\underline{\nu}}(\underline{z} \cdot \underline{\bar{w}}),$$

where $\underline{\nu} = (\nu_1, \nu_2, \cdots, \nu_n)$ with positive ν_k, $\underline{z}$ and $\underline{w} \in C^n$, is also a positive matrix. Thus Corollary 4.3 can be extended to read

$$\det\left[\mathbf{I}_{\underline{\nu}}(\underline{z}_i \cdot \underline{\bar{z}}_j)\right]_{i,j=1}^{m} > 0, \tag{4.4}$$

with distinct elements $\{\underline{z}_i\}_{i=1}^{m}$ belonging to C^m, and m an arbitrary positive integer.

This result is essentially known for $\underline{w}$ a real vector and follows from the fact that $f(x)=e^{-x^2}$ is a variation diminishing kernel for the Hankel convolution. It appears to be a new property with $\underline{w}\in C^m$.

PROPOSITION 4.4. *Let $\mathcal{F}$ denote the ordinary Fock space in one variable. Then*

$$\mathcal{F}=\mathcal{F}_0\dot{+}\{z\mathcal{F}_1\}, \tag{4.5}$$

where $\dot{+}$ denotes the orthogonal sum. Therefore

$$e^{z\bar{w}}=\mathbf{I}_0(z\bar{w})+z\bar{w}\mathbf{I}_1(z\bar{w}). \tag{4.6}$$

Proof. Certainly the even and odd entire functions form complementing subspaces of $\mathcal{F}$. Since $(2n)!=b_{2n}(0)$ and $(2n+1)!=b_{2n}(1)$, it follows that

$$e^{zw}\big|_{\text{even functions}}=\mathbf{I}_0(z\bar{w})=\cosh z\bar{w},$$

and

$$e^{z\bar{w}}\big|_{\text{odd functions}}=\mathbf{I}_1(z\bar{w})=\frac{\sinh z\bar{w}}{z\bar{w}}.$$

Equation (4.6) is a general addition property of reproducing kernels on complementary subspaces. In ordinary Fock space, it is the elementary identity

$$e^{z\bar{w}}=\cosh z\bar{w}+\sinh z\bar{w}.$$

5. Unitary equivalence of $L_n^2(0,\infty)$ and $\mathcal{F}_\nu$. The generalized heat polynomials and their Appell transforms play a key role in establishing the unitary equivalence of $L_n^2(0,\infty)=\mathcal{H}_\nu$ and $\mathcal{F}_\nu$.

The generalized heat polynomial is given by

$$P_{n,\nu}(x,t)=e^{t\Delta_x}x^{2n}=\sum_{k=0}^{n}2^{2k}\binom{n}{k}\frac{\Gamma(\nu+1/2+n)}{\Gamma(n-k+\nu+1/2)}x^{2(n-k)}t^k. \tag{5.1}$$

The Appell transform of $P_{n,\nu}(x,t)$ is given by

$$W_{n,\nu}(x,t)=\frac{t^{-2n}}{(2t)^{\nu+1/2}}e^{-x^2/4t}P_{n,\nu}(x,-t). \tag{5.2}$$

As shown in [7], $P_{n,\nu}(x,t)$ and $W_{n,\nu}(x,t)$ are biorthogonal in the sense that

$$\frac{1}{2^{2n}}\int_0^\infty W_{n,\nu}(x,t)P_{m,\nu}(x,-t)\,d\mu_\nu(x)=b_{2n}(\nu)\delta_{n,m}. \tag{5.3}$$

In this paper we are mainly interested in the particular case of $t=\frac{1}{2}$. A sequence of generalized Laguerre functions is defined by

$$h_{n,\nu}(x)=(-1)^n e^{-x^2/4}P_{n,\nu}(x,-\tfrac{1}{2})=(-1)^n e^{x^2/4}W_{n,\nu}(x,\tfrac{1}{2}). \tag{5.4}$$

Since $P_{n,\nu}(x,-\frac{1}{2})=(-1)^n2^nn!L_n^{\nu-1/2}(x^2/2)$, where $L_n^{\nu-1/2}$ denotes the usual Laguerre polynomial of degree n, it follows that

$$h_{n,\nu}(x)=2^{2n}n!e^{-x^2/4}L_n^{\nu-1/2}\left(\frac{x^2}{2}\right), \tag{5.5}$$

a familiar generalized Laguerre function.

THEOREM 5.1. *If $\psi \in \mathcal{H}_\nu$, then*

$$\psi(x) = \sum_{n=0}^{\infty} a_n h_{n,\nu}(x), \quad \textit{with} \tag{5.6}$$

$$a_n b_{2n}(\nu) = (\psi | h_{n,\nu}) \quad \textit{and} \tag{5.7}$$

$$\|\psi\|^2 = \sum_{n=0}^{\infty} |a_n|^2 b_{2n}(\nu). \tag{5.8}$$

Proof. It follows from (5.3), that

$$\int_0^\infty h_{n,\nu}(x) h_{m,\nu}(x)\, d\mu_\nu(x) = b_{2n}(\nu)\delta_{n,m}. \tag{5.9}$$

Let $\psi \in \mathcal{H}_\nu$. By the completeness of the generalized Laguerre functions (5.5) in $\mathcal{L}^2\{(0,\infty); x^{2\nu}\,dx\}$, we have

$$\psi(x) = \sum_{n=0}^{\infty} a_n h_{n,\nu}(x) \quad \text{in } \mathcal{H}_\nu.$$

Equations (5.7) and (5.8) are simple consequences of (5.9). In fact, (5.7) gives the Fourier coefficient, whereas (5.8) is Parseval's identity.

D. T. Haimo has used similar expansions with variable t; see [12] and [13].

We define a kernel U_ν by

$$U_\nu(z,x) = e^{-(x^2-2z^2)/4} \mathbf{J}_\nu(zx), \tag{5.10}$$

where $z \in C$ and $x \geq 0$. We have that $U_\nu(z,x)$ is a generating function for the generalized Laguerre functions.

THEOREM 5.2. *For $0 \leq x < \infty$, and all complex z,*

$$U_\nu(z,x) = \sum_{n=0}^{\infty} \frac{h_{n,\nu}(x)}{b_{2n}(\nu)} z^{2n}. \tag{5.11}$$

Proof. We have

$$U_\nu(z,x) = e^{x^2/4} G_\nu(iz, x; \tfrac{1}{2}) \tag{5.12}$$

where G_ν is the source solution of the generalized heat equation; see [5]. It follows that

$$\begin{aligned} U_\nu(z,x) &= e^{x^2/4} \sum_{n=0}^{\infty} \frac{W_{n,\nu}(x; 1/2)}{2^{2n} b_{2n}(\nu)} (-1)^n z^{2n} \\ &= \sum_{n=0}^{\infty} \frac{h_{n,\nu}(x)}{b_{2n}(\nu)} z^{2n}; \end{aligned}$$

see [7] or [13, p. 739].

COROLLARY 5.3.

$$\Delta_z^n U_\nu(z,x)|_{z=0} = h_{n,\nu}(x). \tag{5.13}$$

THEOREM 5.4. *For all complex z and w,*

$$\mathbf{I}_\nu(zw)=\int_0^\infty U_\nu(z,x)U_\nu(w,x)\,d\mu_\nu(x). \tag{5.14}$$

Proof. The integral may be written as

$$e^{(z^2+w^2)/2}\int_0^\infty e^{-x^2/2}\mathbf{J}_\nu(zx)\mathbf{J}_\nu(wx)\,d\mu_\nu(x)=e^{(z^2+w^2)/2}e^{-(z^2+w^2)/2}\mathbf{I}_\nu(zw)$$
$$=\mathbf{I}_\nu(zw);$$

see Erdélyi et al. [9, p. 50] for the evaluation of the integral.

From (5.14) it readily follows that

$$\begin{aligned}\mathbf{I}_\nu(z\bar{w})&=\int_0^\infty U_\nu(z,x)U_\nu(\bar{w},x)\,d\mu_\nu(x)\\&=\int_0^\infty U_\nu(z,x)\overline{U_\nu(w,x)}\,d\mu_\nu(x).\end{aligned} \tag{5.15}$$

In particular,

$$\mathbf{I}_\nu(|z|^2)=\int_0^\infty |U_\nu(z,x)|^2 d\mu_\nu(x). \tag{5.16}$$

Thus we have the result.

COROLLARY 5.5. *For all complex z, $U_\nu(z,x)\in\mathcal{H}_\nu$ and*

$$\|U_\nu(z,\cdot)\|^2=\mathbf{I}_\nu(|z|^2). \tag{5.17}$$

The transformation $f=U_\nu\psi$ is defined by

$$f(z)=\int_0^\infty U_\nu(z,x)\psi(x)\,d\mu_\nu(x)=U_\nu\psi\cdot(z) \tag{5.18}$$

for $\psi\in\mathcal{H}_\nu$. Using Corollary 5.5 and the Cauchy–Schwarz inequality, we obtain

$$|f(x)|\le\mathbf{I}_\nu(|z|^2)^{1/2}\|\psi\|. \tag{5.19}$$

By differentiating under the integral (which is easily justified: see [2, p. 191]) we see that f is an entire function of growth $\le\{2,\frac{1}{2}\}$.

THEOREM 5.6. *The transformation $U_\nu\psi=f$ is a unitary mapping of $\mathcal{H}_\nu$ onto $\mathcal{F}_\nu$. Moreover, the basis elements are related by*

$$U_\nu h_{n,\nu}=z^{2n}. \tag{5.20}$$

Proof. Equation (5.20) can be obtained as a standard result from tables of integrals [9]. However, it follows directly from Theorems 5.1 and 5.2 and Corollary 5.5, for

$$U_\nu h_{n,\nu}(z)=(U_\nu(z,\cdot)|h_{n,\nu})=\sum_{m=0}^\infty\frac{z^{2m}}{b_{2m}(\nu)}(h_{m,\nu}|h_{n,\nu})=z^{2n}\quad\text{for all } n.$$

Consequently U_ν maps the linear manifold determined by the family $\{h_{n,\nu}\}_{n=0}^\infty$ onto the even polynomials in $\mathcal{F}_\nu$. Therefore U_ν maps a dense set in $\mathcal{H}_\nu$ onto a dense set in $\mathcal{F}_\nu$.

Further, if $\psi\in\mathcal{H}_\nu$, then $\psi(x)=\sum_{n=0}^{\infty}a_nh_{n,\nu}(x)$, and

$$\begin{aligned} f(z)=U_\nu\psi\cdot(z)&=\int_0^\infty U_\nu(z,x)\psi(x)\,d\mu_\nu(x)\\ &=\sum_{n=0}^{\infty}a_n\int_0^\infty U_\nu(z,x)h_{n,\nu}(x)\,d\mu_\nu(x)\\ &=\sum_{n=0}a_nz^{2n}, \end{aligned}$$

the interchange of summation and integration being easily justified by (5.19) and the Tonelli–Hobson theorem. Now

$$\|f\|^2=\|U_\nu\psi\|_\nu^2=\sum_{n=0}^{\infty}|a_n|^2b_{2n}(\nu)=\|\psi\|^2,$$

by (5.8). It follows that U_ν is a unitary transformation of $\mathcal{H}_\nu$ onto $\mathcal{F}_\nu$. Clearly

$$U_\nu^{-1}f\cdot(x)=\sum_{n=0}^{\infty}a_nh_{n,\nu}(x),\quad \text{where}\, f(z)=\sum a_nz^{2n}\in\mathcal{F}_\nu.$$

Letting $B_w(x)=\overline{U_\nu(w,x)}$ for all $w\in C$, we can rewrite the integral (5.15) as

$$\mathbf{I}_\nu(z\bar{w})=U_\nu B_w\cdot(z). \tag{5.21}$$

Thus the linear manifold determined by the family $\{B_w\}_{w\in C}$ is mapped onto the family of I-functions in $\mathcal{F}_\nu$.

Definition 5.7. Let $\mathcal{B}$ denote the family of even entire functions $f(z)$ such that

$$|f(z)|=O\left(e^{1/2|z|^2-\alpha|z|}\right) \tag{5.22}$$

for every $\alpha>0$.

The family $\mathcal{B}$ is dense in $\mathcal{F}_\nu$ for the even polynomials are in $\mathcal{B}$.

Proposition 5.8. *For all complex w,*

$$\mathcal{B}\cdot\mathbf{I}_\nu(z\bar{w})\subset\mathcal{F}_\nu. \tag{5.23}$$

Proof. Using the standard estimates of the Bessel functions, we have

$$|f(z)|^2|\mathbf{I}(z\bar{w})|^2\mathbf{K}_\nu\left(|z|^2\right)=O(e^{2|z|\{|w|-\alpha\}})$$

for every $\alpha>0$. Thus the integral (1.10) which gives the norm of $f(z)\mathbf{I}(z\bar{w})$ converges.

The family $\mathcal{B}$ simplifies the presentation of the inverse operator U_ν^{-1}: $\mathcal{F}_\nu\to\mathcal{H}_\nu$. Theorem 5.2 suggest that U_ν^{-1} should be given by the integral equation

$$U_\nu^{-1}f\cdot(x)=\int_C\overline{U_\nu(z,x)}f(z)\,d\mathfrak{m}_\nu(z). \tag{5.24}$$

Since $U_\nu(z,x)\notin\mathcal{F}_\nu$, the integral does not necessarily converge. However, if $f\in\mathcal{B}$, we have the following result.

Theorem 5.9. *If $f=\sum a_nz^{2n}\in\mathcal{B}$, then*

$$U_\nu^{-1}f\cdot(x)=\int_C\overline{U_\nu(z,x)}f(z)\,d\mathfrak{m}_\nu(z)=\sum_{n=0}^{\infty}a_nh_{n,\nu}(x)=\psi(x). \tag{5.25}$$

Proof. Using the standard estimates for the Bessel functions we have, for $f \in \mathfrak{B}$,

$$|U_\nu(z,x)||f(x)|\mathbf{K}_\nu(|z|^2) \leq Me^{-(\alpha - x)|z|}$$

for every $\alpha > 0$. Thus the integral (5.25) converges locally uniformly and absolutely, and therefore the following interchange of integration and summation is valid. We have

$$U_\nu^{-1}f(x) = \sum_{n=0}^{\infty} \frac{h_{n,\nu}(x)}{b_{2n}(\nu)} \int_C \bar{z}^{2n} f(z)\, d\mathfrak{m}_\nu(z) = \sum_{n=0}^{\infty} \frac{h_{n,\nu}(x)}{b_{2n}(\nu)} \left(f(z)|z^{2n} \right)_\nu$$

$$= \sum_{n=0}^{\infty} a_n h_{n,\nu}(x),$$

by Theorem 5.2 and (3.8).

COROLLARY 5.10. *If $f \in \mathfrak{F}_\nu$, then*

$$U_\nu^{-1} f \cdot (x) = \text{L.i.m.} \int_C U_\nu(z,x) f_n(z)\, d\mathfrak{m}_\nu(z), \tag{5.26}$$

where the sequence $\{f_n\} \subset \mathfrak{B}$ and converges to f in $\mathfrak{F}_\nu$.

This is clear since $\mathfrak{B}$ is dense in $\mathfrak{F}_\nu$.

6. Operators on $\mathfrak{F}_\nu$. Since $\mathfrak{F}_\nu$ has a reproducing kernel, each operator on $\mathfrak{F}_\nu$ has in general an associated kernel, whence an operator on $\mathfrak{F}_\nu$ is given by an integral equation with a suitable kernel. The generalized Schrödinger radial kinetic energy operator $\Delta_z(\nu)$, the operator of multiplication-by-z^2, and functions of these operators are studied in this section.

The mapping U_ν induces an isomorphism between the linear operators on $\mathfrak{F}_\nu$ and those on $\mathcal{H}_\nu$. The induced mapping is also unitary on the bounded operators. The correspondence is given by

$$U_\nu T U_\nu^{-1} = \tau \tag{6.1}$$

and

$$U_\nu^{-1} \tau U_\nu = T, \tag{6.2}$$

where T and τ are operators on $\mathcal{H}_\nu$ and $\mathfrak{F}_\nu$, respectively.

In our study of Δ_z and z^2, we need the following identity.

THEOREM 6.1. *If $f(z) = \Sigma a_n z^{2n} \in \mathfrak{F}_\nu$, then*

$$\|z^2 f(z)\|_\nu^2 = \|\Delta f(\underline{z})\|_\nu^2 + b_2 \|f\|_\nu^2 + 8 \sum_{n=0}^{\infty} n|a_n|^2 b_{2n}(\nu) \tag{6.3}$$

where both sides either have the same finite value or are infinite.

Proof. Let $f \in \mathfrak{F}_\nu$ with $\|f\|_\nu^2 = \Sigma_{n=0}^{\infty} a_n^2 b_{2n}(\nu)$. Then $z^2 f(z) = \Sigma_{n=0}^{\infty} a_n z^{2(n+1)}$, and

$$\|z^2 f(z)\|_\nu^2 = \sum_{n=0}^{\infty} |a_n|^2 b_{2(n+1)} = \sum_{n=0}^{\infty} 4(n+1)\left(n+\nu+\frac{1}{2}\right)|a_n|^2 b_{2n}. \tag{6.4}$$

Moreover,

$$\Delta f(z) = \sum_{n=0}^{\infty} 4n\left(n+\nu-\frac{1}{2}\right) a_n z^{2(n-1)}, \tag{6.5}$$

and therefore

$$\text{(6.6)} \qquad \|\Delta f(z)\|_\nu^2 = \sum_{n=0}^{\infty} \left|4n\left(n+\nu-\frac{1}{2}\right)\right|^2 |a_n|^2 b_{2(n-1)}$$
$$= \sum_{n=0}^{\infty} 4n\left(n+\nu-\frac{1}{2}\right)|a_n|^2 b_{2n}.$$

Since $4(n+1)(n+\nu+\frac{1}{2})=4n(n+\nu-\frac{1}{2})+4(\nu+\frac{1}{2})+8n$, it follows that

$$\|z^2 f(z)\|_\nu^2 = \|\Delta f(z)\|_\nu^2 + b_2(\nu)\|f\|_\nu^2 + 8\sum_{n=0}^{\infty} n|a_n|^2 b_{2n}(\nu).$$

Certainly, $z^2 f(z) \in \mathfrak{F}_\nu$ implies that the right-hand side of (6.3) is finite. Suppose $\Delta f(z) \in \mathfrak{F}_\nu$. Then by (6.6), $\Sigma 4n(n+\nu+\frac{1}{2})|a_n|^2 b_{2n}(\nu) < \infty$, which implies that $8\Sigma n|a_n|^2 b_{2n}(\nu)$ is finite. It follows that $z^2 f(z) \in \mathfrak{F}_\nu$.

Using induction, we obtain the following extension of the theorem.

COROLLARY 6.2. *For every positive integer k,*

$$\text{(6.7)} \qquad \|z^{2k} f(z)\|_\nu^2 = \sum_{l=1}^{k} \|\Delta^{2(k-l)} f(z)\|_\nu^2 b_2^{l-1}(\nu)$$
$$+ b_2^k\|f\|_\nu^2 + 8\sum_{n=0}^{\infty} n\left\{\sum_{l=0}^{k-1} b_2^{k-l-1}|a_{k-l}|^2 b_{2n}\right\},$$

where a_j with a negative subscript is equal to zero.

We need the following elementary lemma, which is the analogue of a result of Bargmann [2] for ordinary Fock space, and which holds in any reproducing kernel space (apart from the assertion of local uniform convergence).

THEOREM 6.3. *Let $\{f_n\}_1^\infty \subset \mathfrak{F}_\nu$ be such that:*

(1) *$\|f_n\|_\nu \leq M$ for some positive M and every n;*

(2) *$\{f_n(z)\}_1^\infty$ is convergent for all $z \in C$.*

Then $\{f_n\}_1^\infty$ has a weak limit $f \in \mathfrak{F}_\nu$, and $\{f_n(z)\}_1^\infty$ converges locally uniformly to $f(z)$. If, in addition,

$$\text{(3)} \qquad \lim_{n\to\infty} \|f_n\|_\nu = \|f\|_\nu,$$

then the sequence converges to f in $\mathfrak{F}_\nu$.

DEFINITION 6.4. Operators Q^2 and $\mathfrak{D}_\nu$ are defined for $f \in \mathfrak{F}_\nu$ by

$$\text{(6.8)} \qquad (Q^2 f)(z) = z^2 f(z) \quad \text{if } z^2 f(z) \in \mathfrak{F}_\nu,$$

and

$$\text{(6.9)} \qquad (\mathfrak{D}_\nu f)(z) = \Delta_z f(z) \quad \text{if } \Delta_z f(z) \in \mathfrak{F}_\nu,$$

where $\Delta_z = d^2/dz^2 + 2\nu/z \; d/dz$.

Let $D(Q^2)$ and $D(\mathfrak{D}_\nu)$ denote the domains of Q^2 and $\mathfrak{D}_\nu$, respectively.

THEOREM 6.5. *The operators Q^2 and $\mathfrak{D}_\nu$ are closed densely defined operators on $\mathfrak{F}_\nu$ such that*

$$\text{(6.10)} \qquad D(Q^2) = D(\mathfrak{D}_\nu),$$

and

$$\mathfrak{D}_\nu^* = Q^2 \quad \textit{and} \quad (Q^2)^* = \mathfrak{D}_\nu. \tag{6.11}$$

Proof. Clearly, Q^2 and $\mathfrak{D}_\nu$ are densely defined, for the set of even polynomials is contained in each of their domains. Let $\{(f_n, z^2 f_n)\}_1^\infty$ be a sequence in the graph of Q^2 and let $(f_n, z^2 f_n) \Rightarrow (g, h) \in \mathfrak{F}_\nu \times \mathfrak{F}_\nu$. Now $\lim \|f_n\|_\nu = \|g\|_\nu$ and $\lim \|z^2 f_n\|_\nu = \|h\|_\nu$. Since strong convergence implies pointwise convergence, we have, for every $z \in C$, $h(z) = \lim z^2 f_n(z) = z^2 \lim f_n(z) = z^2 g(z)$. It follows from Theorem 6.3 that Q^2 is closed.

By Theorem 6.1, we see that $D(Q^2) = D(\mathfrak{D}_\nu)$.

Let $g \in D((Q^2)^*)$. Then there exists a $h \in \mathfrak{F}_\nu$ such that

$$\big(z^2 f(z) | g(z)\big)_\nu = \big(f(z) | h(z)\big)_\nu \tag{6.12}$$

for every $f \in D(Q^2)$. Let $g(z) = \Sigma g_n z^{2n}$ and $h(z) = \Sigma h_n z^{2n}$. Then by (6.12) we have

$$\sum_{n=0}^{\infty} a_n \bar{g}_{n+1} b_{2(n+1)} = \sum_{n=0} a_n \bar{h}_n b_{2n}.$$

In particular, taking $f(z) = z^{2n}$, we get $g_{n+1} b_{2(n+1)} = h_n b_{2n}$, and therefore $h_n = 4(n+1)(n+\nu+\frac{1}{2}) g_{n+1}$. Hence $h(z) = \Delta_z g(z)$ and $g \in D(\mathfrak{D}_\nu)$. Hence

$$D\big((Q^2)^*\big) \subset D(\mathfrak{D}_\nu) \quad \text{and} \quad (Q^2)^* \subset \mathfrak{D}_\nu.$$

Since $(Q^2)^*$ is densely defined (even polynomials $\subset D((Q^2)^*)$), it is also closed. Next we let $f \in D(Q^2)$ and $g \in D(\mathfrak{D}_\nu)$. A simple calculation shows that $(z^2 f | g)_\nu = (f | \Delta g)_\nu$ and therefore $g \in D((Q^2)^*)$ and $(Q^2)^* g = \Delta g$. Thus $\mathfrak{D}_\nu \subset (Q^2)^*$, which implies that $(Q^2)^* = \mathfrak{D}_\nu$, and therefore $\mathfrak{D}_\nu$ is also closed. Finally $\mathfrak{D}_\nu^* = (Q^2)^{**} = Q^2$ since Q^2 is closed.

Proposition 6.6. *The operator Q^2 has an inverse on $\mathfrak{F}_\nu$.*

Proof. By Theorem 6.1, we have

$$\|Q^2 f(z)\|_\nu^2 \geq b_2(\nu) \|f\|_\nu^2.$$

Therefore $Q^2 f = 0$ implies $f = 0$. We have domain $(Q^2)^{-1} = \operatorname{Range} Q^2$ and $(Q^2)^{-1} Q^2 f = f$ for every $f \in D(Q^2)$.

Let T denote an arbitrary densely defined operator on $\mathfrak{F}_\nu$ with domain $D(T)$, and let $M(z,w)$ be a function defined on $C \times C$ such that $M(z,w)$ belongs to $\mathfrak{F}_\nu$ for every $w \in C$.

Definition 6.7. The operator T is said to correspond to the kernel $M(z,w)$, written $T \sim M(z,w)$, if for every $f \in D(T)$

$$Tf \cdot (w) = \big(f(z) | M(z,w)\big)_\nu. \tag{6.13}$$

T is said to correspond to $M(z,w)$ in the maximal sense, written $T \cong M(z,w)$, if $D(T)$ consists of all f in $\mathfrak{F}_\nu$ such that $(f(z)|M(z,w))_\nu$ is again an element of $\mathfrak{F}_\nu$, when considered as a function of w, and if for every f in $D(T)$, (6.13) holds.

Bounded linear operators always correspond to kernels in the maximal sense. To a given kernel there is a unique operator which corresponds to it in the maximal sense.

The following well-known properties hold for an arbitrary reproducing kernel space. We state them for $\mathfrak{F}_\nu$.

Properties 6.8. Let $T \cong M(z,w)$.

(a) If T^* is the adjoint of T, then $T^* \cong M^*(z,w) = \overline{M(w,z)}$.

(b) If $T_1 \cong M_1(z,w)$ and $T_2 \cong M_2(z,w)$, then

$$T_1T_2 \cong M(z,w) = \left(M_1(s,w)|\overline{M_2(z,s)}\right)_\nu.$$

(c) If $T \cong M(z,w)$, then $T = T^*$ is equivalent to $M(w,z) = M(z,w)$, that is, the Hermitian symmetry of the kernel $M(z,w)$.

(d) If $T \cong M(z,w)$, then T is positive if and only if $M(z,w)$ is a positive matrix.

(e) $T \cong M(z,w)$ if and only if the reproducing kernel $\mathcal{K}_\nu(z,w) = \mathbf{I}_\nu(z\bar{w})$ belongs to $D(T^*)$ for every w in C and $M(z,w) = T^*\mathcal{K}_\nu(z,w)$.

See Aronszajn [1, p. 371] or Meschkowski [17] for the general theory.

Let L be a unitary operator on $\mathcal{F}_\nu$, that is, $LL^* = L^*L = I$. By properties (a) and (b) we have

$$\mathbf{I}_\nu(z\bar{w}) = \int_C L(s,w)L(z,s)\,d\mathfrak{m}_\nu(s) \tag{6.14}$$

where

$$L \cong L(z,w) = L^*\mathbf{I}(s\bar{w}) = \left(L^*\mathbf{I}(s\bar{w})|\mathbf{I}(s\bar{z})\right)_\nu = \left(\mathbf{I}(s\bar{w})|L\mathbf{I}(s\bar{z})\right)_\nu.$$

This is an analogue of Bargmann's result on ordinary Fock space; see [2, p. 195].

Definition 6.9. Let $\phi \in \mathcal{B}$ and let ϕ denote the operator of multiplication by $\phi(z)$ on $\mathcal{F}_\nu$, that is,

$$\phi(Q)f(z) = \phi(z)f(z) \quad \text{if } \phi f \in \mathcal{F}_\nu.$$

By Proposition 5.8, $\phi(Q)$ is densely defined, for the I-functions are in its domain.

Definition 6.10. Let $\phi \in \mathcal{B}$. We define $\phi(\mathcal{D}_\nu^{1/2})$ on $\mathcal{F}_\nu$ by

$$\phi(\mathcal{D}_\nu^{1/2})f\cdot(w) = \int_C \overline{\phi_\#(z)}\mathbf{I}_\nu(w\bar{z})\,d\mathfrak{m}_\nu(z) \tag{6.15}$$

where the domain of $\phi(\mathcal{D}_\nu^{1/2})$ is the set of all f in $\mathcal{F}_\nu$ such that the function given by (6.14) is again in $\mathcal{F}_\nu$. Thus $\phi(\mathcal{D}_\nu^{1/2}) \cong \phi_\#(z)\mathbf{I}(z\bar{w}) = M_{\phi(\mathcal{D}_\nu^{1/2})}(z,w)$.

If $\phi(z)$ is an even polynomial, then our definition agrees with the usual definition of $\phi(\Delta_z)$. Since $\phi \in \mathcal{B}$, the integral (6.14) converges for all $f \in \mathcal{F}_\nu$.

Taking $f(z) = \mathbf{I}(\bar{a}z)$ in (6.14), a calculation shows that

$$\int_C \left| \int_C \mathbf{I}(\bar{a}z)\bar{\phi}_\#(z)\mathbf{I}(a\bar{z})\,d\mathfrak{m}_\nu(z) \right|^2 d\mathfrak{m}_\nu(w) = |\phi_\#(a)|^2\mathbf{I}_\nu\left(|a|^2\right). \tag{6.16}$$

Consequently, the I-functions are in the domain of $\phi(\mathcal{D}_\nu^{1/2})$, and therefore $\phi(\mathcal{D}_\nu^{1/2})$ is densely defined.

Theorem 6.11. *Let* $\phi \in \mathcal{B}$. *Then*

(a) $\phi(Q)$ *and* $\phi(\mathcal{D}_\nu^{1/2})$ *are closed operators,*

(b) $\phi(Q)^* = \phi_\#(\mathcal{D}_\nu^{1/2})$ *and* $\phi(\mathcal{D}_\nu^{1/2})^* = \phi_\#(Q)$,

and

(c) $\phi(Q) \cong M_{\phi(Q)}(z,w) = \phi(\bar{w})\mathbf{I}(z\bar{w})$.

Proof. First of all we will show that $\phi(Q)$ is closed. Let $\{(f_n,\phi f_n)\}_1^\infty$ be a sequence in the graph of $\phi(Q)$ such that $(f_n,\phi f_n) \cong (f,h)$ in $\mathcal{F}_\nu \times \mathcal{F}_\nu$. By inequality (3.11), $f_n(z)$ and $\phi(z)f_n(z)$ converge pointwise to $f(z)$ and $h(z)$, respectively. It also follows that $\phi(z)f_n(z) \to \phi(z)f(z)$ pointwise. Using Theorem 6.3, we get $h = \phi f$, and therefore the graph of $\phi(Q)$ is closed.

Next we let $\{(f_n,\phi(\mathcal{D}_\nu^{1/2})f_n)\}_1^\infty$ be a sequence in the graph of $\phi(\mathcal{D}_\nu^{1/2})$ which converges to (f,h) in $\mathcal{F}_\nu\times\mathcal{F}_\nu$. Then

$$\phi(\mathcal{D}_\nu^{1/2})f_n(w)=\big(f_n(z)|\phi_\#(z)\mathbf{I}(z\bar{w})\big)_\nu\to\big(f(z)|\phi_\#(Uz)\mathbf{I}(z\bar{w})\big),$$

since strong convergence implies weak convergence. By inequality (3.12), we also have $\phi(\mathcal{D}_\nu^{1/2})f_n\cdot(w)\to h(w)$. Hence

$$h(w)=\big(f(z)|\phi_\#(z)\mathbf{I}_\nu(z\bar{w})\big)_\nu\in\mathcal{F}_\nu.$$

It follows that $f\in D(\phi(\mathcal{D}_\nu^{1/2}))$ and $h=\phi(\mathcal{D}_\nu^{1/2})f$. Thus $\phi(\mathcal{D}_\nu^{1/2})$ is closed.

Further,

$$M_{\phi(\mathcal{D}_\nu^{1/2})^*}(z,w)=\overline{M_{\phi(\mathcal{D}_\nu^{1/2})}(w,z)}=\bar{\phi}_\#(w)\mathbf{I}(z\bar{w}).$$

Thus we find that for all $f\in D(\phi(Q))$,

$$\begin{aligned}\big(f(z)|\phi(\bar{w})\mathbf{I}(z\bar{w})\big)_\nu&=\phi_\#(w)\big(f(z)|\mathbf{I}(z\bar{w})\big)\\&=\phi_\#(w)f(w)\in\mathcal{F}_\nu.\end{aligned}$$

By Property 6.8 (a), we get $\phi(\mathcal{D}_\nu^{1/2})^*=\phi_\#(Q)$. Likewise $\phi(Q)^*=\phi_\#(\mathcal{D}_\nu^{1/2})$. Part (c) of the theorem is clear.

PROPOSITION 6.12. *If $\phi_a(z)=\mathbf{I}(az)$, $a\in C$, then*

$$\phi_a(\mathcal{D}_\nu^{1/2})f\cdot(w)=\big(f(z\,\text{\textcircled{Δ}}\,a)|\mathbf{I}(z\bar{w})\big)_\nu,$$

for all $f\in D(\phi_a(\mathcal{D}_\nu^{1/2}))$.

Proof. We have $\phi_a\in\mathcal{B}$. Therefore for $f\in D(\phi_a(\mathcal{D}_\nu^{1/2}))$, with $f(z)=\Sigma a_nz^{2n}$, it follows that

$$\begin{aligned}\phi_a(\mathcal{D}_\nu^{1/2})f\cdot(w)&=\big(f(z)|\mathbf{I}(\bar{a}z)\mathbf{I}(z\bar{w})\big)_\nu\\&=\big(f(z)|\mathbf{I}(z(\bar{a}\,\text{\textcircled{Δ}}\,\bar{w}))\big)_\nu\\&=\sum_{n=0}^{\infty}\frac{(a\,\text{\textcircled{Δ}}\,w)^{2n}}{b_{2n}}\big(f(z)|z^{2n}\big)_\nu\\&=\sum_{n=0}^{\infty}a_n(a\,\text{\textcircled{Δ}}\,w)^{2n}\\&=f(a\,\text{\textcircled{Δ}}\,w)\in\mathcal{F}_\nu.\end{aligned}$$

Hence $f(a\,\text{\textcircled{$\Delta$}}\,w)=(f(a\,\text{\textcircled{Δ}}\,z)|\mathbf{I}(z\bar{w}))$ and the proof of the proposition is complete.

7. Commutation relations for D_x and x^2. A simple calculation shows that the commutator of Δ_x and x^2 is given by

$$[\Delta_x,x^2]=\Delta_xx^2-x^2\Delta_x=4(\nu+\tfrac{1}{2})I+4x\frac{d}{dx}. \tag{7.1}$$

In this section we work with general analogues of Δ_x and x^2 on a general Hilbert space. Due to the difficulties with domains of definition of unbounded operators, some of the following results are of a heuristic nature. We shall formally treat the operators like bounded operators and use a formal algebraic procedure to derive the results. In most cases, the stated equality of operators is equality on a linear manifold in the intersection of the domains of the given operators.

Let P and Q be operators on a Hilbert space $\mathcal{H}$ which satisfy the Fock commutation rule,

$$[P,Q]=1 \quad \text{on } D(PQ)\cap D(QP), \tag{7.2}$$

and suppose Q has an inverse. Let $\nu>0$. We define

$$D_\nu = P^2 + 2\nu Q^{-1}P, \tag{7.3}$$

and

$$B = QP. \tag{7.4}$$

Proposition 7.1. *Let (P,Q) satisfy the Fock commutation rule. Then*

(a) $[P,Q^2]=2Q,$

(b) $[P^2,Q]=2P,$

(c) $[Q^{-1}P,Q]=Q^{-1},$

(d) $[P^2,Q^2]=2I+4QP,$

(e) $[D_\nu,Q^2]=b_2(\nu)+4QP,$

(f) $[D_\nu^n,Q^2]=nb_2(\nu)D_\nu^{n-1}+4(BD_\nu^{n-1}+D_\nu BD_\nu^{n-2}+\cdots+D_\nu^{n-1}B).$

Proof. (a) through (d) follow by elementary calculations. We will consider (e). Now

$$\begin{aligned}
[D_\nu,Q^2] &= (P^2+2\nu Q^{-1})Q^2 - Q^2(P^2+2\nu Q^{-1}P)\\
&= P^2Q^2+2\nu Q^{-1}PQ^2-Q^2P^2-2\nu QP\\
&= 2I+4QP+2\nu(Q^{-1}PQ^2-QP) \quad \text{by (d)}\\
&= 2I+4QP+2\nu(Q^{-1}(I+QP)Q-QP) \quad \text{by (7.2)}\\
&= 2I+4QP+4\nu I\\
&= b_2(\nu)I+4QP.
\end{aligned}$$

Using (e), we find that

$$[D_\nu^2,Q^2]=[D_\nu,Q^2]D_\nu+D_\nu[D_\nu,Q^2].$$

Let $A=[D_\nu,Q^2]$. Then by induction we obtain

$$[D_\nu^n,Q^2]=AD_\nu^{n-1}+D_\nu AD_\nu^{n-2}+\cdots+D_\nu^{n-1}A. \tag{7.5}$$

Now $D_\nu^kAD_\nu^{n-k-1}=b_2(\nu)D_\nu^{n-1}+4D_\nu^kBD_\nu^{n-k-1}$. Substituting in (7.5), we obtain

$$[D_\nu^n,Q^2]=nb_2(\nu)D^{n-1}+4(BD_\nu^{n-1}+D_\nu BD^{n-2}+\cdots+D^{n-1}B),$$

which is (f).

THEOREM 7.2. *We have*

(a) $[D_\nu, B] = 2D_\nu$,

(b) $D_\nu^k B D_\nu^{n-k-1} = BD_\nu^{n-1} + 2kD^{n-1}$,

(c) $[D_\nu^n, Q^2] = \{4n(n+\nu-1/2) + 4nB\} D_\nu^{n-1}$,

for every positive integer n.

Proof.

$$\begin{aligned}[D, QP] &= D_\nu QP - QPD_\nu \\ &= [P^2, Q]P + 2\nu Q^{-1}PQP - 2\nu QPQ^{-1}P \\ &= 2P^2 + 2\nu(Q^{-1} + P)P - 2\nu QPQ^{-1}P \quad \text{by 7.1 (c)}, \\ &= 2P^2 + 2\nu Q^{-1}P + 2\nu P^2 - 2\nu(PQ - I)Q^{-1}P \\ &= 2P^2 + 4\nu Q^{-1}P = 2D_\nu.\end{aligned}$$

Part (b) easily follows from (a).

Using (b), we obtain

$$\begin{aligned}\sum_{k=0}^{n-1} D_\nu^k B D_\nu^{n-k-1} &= \sum_{k=0}^{n-1} \left(BD_\nu^{n-1} + 2kD_\nu^{n-1}\right) \\ &= nBD_\nu^{n-1} + n(n-1)D_\nu^{n-1}.\end{aligned}$$

Consequently,

$$\begin{aligned}[D_\nu^n, Q^2] &= nb_2(\nu)D_\nu^{n-1} + \{4nB + 4n(n-1)\}D_\nu^{n-1} \\ &= [4n(n+\nu-1/2) + 4nB]\, D_\nu^{n-1},\end{aligned}$$

which is part (c). We note that $\Delta_x x^{2n} = 4n(n+\nu-1/2)x^{2(n-1)}$.

Let

$$\mathbf{I}_\nu(az) = \sum_{n=0}^{\infty} \frac{a^{2n}}{b_{2n}(\nu)} z^{2n}.$$

Since $\mathbf{I}_\nu(0) = 1$, $\mathbf{I}$ is a unit in the integral domain of formal power series over C. We define

$$\mathbf{I}_\nu(az)^{(-1)} = \sum_{n=0}^{\infty} \frac{\mathfrak{R}_{2n}^\nu(a)}{b_{2n}(\nu)} z^{2n}. \tag{7.6}$$

Then $\mathfrak{R}_0^\nu(a) = 1$ and

$$\begin{aligned}1 = \mathbf{I}(az)\mathbf{I}(az)^{(-1)} &= \sum_{n=0}^{\infty} \frac{a^{2n}}{b_{2n}} z^{2n} \sum_{n=0}^{\infty} \frac{\mathfrak{R}_{2n}^\nu(a)}{b_{2n}} z^{2n} \\ &= \sum_{n=0}^{\infty} \left\{ \sum_{k=0}^{n} \frac{b_{2n}}{b_{2(n-k)}b_{2k}} \mathfrak{R}_{2k}^\nu(a) a^{2(n-k)} \right\} \frac{z^{2n}}{b_{2n}}.\end{aligned}$$

It follows that

$$(7.7)\qquad \sum_{k=0}^{n}\binom{n}{k}\frac{\Gamma(n+\nu+1/2)\Gamma(\nu+1/2)}{\Gamma(n-k+\nu+1/2)\Gamma(k+\nu+1/2)}\mathfrak{R}^{\nu}_{2k}(a)a^{2(n-k)}=0$$

for $n\geq 1$. Thus $\{\mathfrak{R}^{\nu}_{2n}(a)\}_{n=0}^{\infty}$ is a sequence of even polynomials in a, determined by

$$(7.8)\quad \mathfrak{R}^{\nu}_{2n}(a)=(-1)\sum_{k=0}^{n-1}\binom{n}{k}\frac{\Gamma(n+\nu+1/2)\Gamma(\nu+1/2)}{\Gamma(n-k+\nu+1/2)\Gamma(k+\nu+1/2)}\mathfrak{R}^{\nu}_{2k}(a)a^{2(n-k)},$$

with $\mathfrak{R}^{\nu}_{0}(a)=1$.

The function $\mathbf{I}_\nu(az)^{(-1)}$ occurs in the generalized Weyl relations for Bessel functions.

LEMMA 7.3.

$$[\mathbf{I}_\nu(aD_\nu^{1/2}),Q^2]=a^2\mathbf{I}_\nu(aD_\nu^{1/2})+\frac{2a^2B}{2\nu+1}\mathbf{I}_{\nu+1}(aD_\nu^{1/2}).$$

Proof. Using part (c) of Theorem 7.2, we obtain

$$\left[\frac{D_\nu^n}{b_{2n}},Q^2\right]=\frac{4n(n+\nu-1/2)}{b_{2n}}D_\nu^{n-1}+\frac{4nB}{b_{2n}}D_\nu^{n-1}.$$

Summing, we get

$$\begin{aligned}[\mathbf{I}_\nu(aD_\nu^{1/2}),Q^2]&=a^2\mathbf{I}_\nu(aD_\nu^{1/2})+B\sum_{n=0}^{\infty}\frac{4nD_\nu^{n-1}}{b_{2n}}a^{2n}\\&=a^2\mathbf{I}_\nu(aD_\nu^{1/2})+\frac{2a^2B}{2\nu+1}\mathbf{I}_{u+1}(aD_\nu^{1/2}).\end{aligned}$$

THEOREM 7.4 (generalized Weyl relation). *Let $a,b\in C$. Then*

$$\begin{aligned}&\mathbf{I}_\nu(bQ)\mathbf{I}_\nu(aD_\nu^{1/2})\\&=\mathbf{I}_\nu(aD_\nu^{1/2})\mathbf{I}_\nu\left(\left\{(bQ)^2-(ab)^2-\frac{2(ab)^2}{2\nu+1}\mathbf{I}_\nu(aD_\nu^{1/2})^{(-1)}B\mathbf{I}_{\nu+1}(aD_\nu^{1/2})\right\}^{1/2}\right).\end{aligned}$$

Proof. Using Lemma 7.3, we obtain with the aid of (7.5)

$$(7.9)\quad \mathbf{I}_\nu(aD_\nu^{1/2})^{(-1)}Q^2\mathbf{I}_\nu(aD_\nu^{1/2})=Q^2-a^2-\frac{2a^2}{2\nu+1}\mathbf{I}_\nu(aD_\nu^{1/2})^{(-1)}B\mathbf{I}_{\nu+1}(aD_\nu^{1/2}),$$

which implies

$$(7.10)\quad \mathbf{I}_\nu(aD_\nu^{1/2})^{(-1)}Q^{2n}\mathbf{I}_\nu(aD_\nu^{1/2})=\left\{Q^2-a^2-\frac{2a^2}{2\nu+1}\mathbf{I}_\nu(aD_\nu^{1/2})B\mathbf{I}_{\nu+1}(aD_\nu^{1/2})\right\}^n.$$

Multiplying by $b/b_{2n}(\nu)$ and summing yields

$$\begin{aligned}(7.11)\qquad &\mathbf{I}_\nu(aD_\nu^{1/2})^{(-1)}\mathbf{I}_\nu(bQ)\mathbf{I}_\nu(aD_\nu^{1/2})\\&=\mathbf{I}_\nu\left(\left\{(bQ)^2-(ab)^2-\frac{2(ab)^2}{2\nu+1}\mathbf{I}_\nu(aD_\nu^{1/2})^{(-1)}B\mathbf{I}_{\nu+1}(aD_\nu^{1/2})\right\}^{1/2}\right),\end{aligned}$$

and the theorem follows upon multiplication by $\mathbf{I}_\nu(aD_\nu^{1/2})$.

COROLLARY 7.5.
(a)

$$\exp(b^2Q^2)\mathbf{I}_\nu(aD_\nu^{1/2})$$
$$=I_\nu(aD_\nu^{1/2})\exp\left\{(bQ)^2-(ab)^2-\frac{2(ab)^2}{2\nu+1}\mathbf{I}_\nu(aD_\nu^{1/2})^{(-1)}B\mathbf{I}_{\nu+1}(aD_\nu^{1/2})\right\}.$$

(b) *If* $f(z)=\sum_{n=0}^{\infty}a_nz^{2n}$ *is an entire function, then*

$$f(bQ)\mathbf{I}_\nu(aD_\nu^{1/2})$$
$$=\mathbf{I}_\nu(aD_\nu^{1/2})f\left(\left\{(bQ)^2-(ab)^2-\frac{2(ab)^2}{2\nu+1}\mathbf{I}_\nu(aD_\nu^{1/2})^{(-1)}B\mathbf{I}_{\nu+1}(aD_\nu^{1/2})\right\}^{1/2}\right).$$

Proof. These relations follow from (7.10).

The final results of this section are based on the work of H. Tillmann; see [19]. Let P be a closed densely defined operator on a separable Hilbert space $\mathcal{H}$ with adjoint $P^*=Q$, and suppose that $[P,Q]=I$ on $\mathcal{M}=D(PQ)\cap D(QP)$. Then the numbers operator $B=QP$ is selfadjoint and has a pure discrete spectrum sp $QP=\{0,1,2,\cdots\}$, and each eigenvalue has the same multiplicity.

Let

$$\mathcal{M}_n=\{f\in\mathcal{H}\,|\,QPf=nf\}. \tag{7.12}$$

Then $\mathcal{M}_n=\{f\in\mathcal{H}\,|\,PQf=(n+1)f\}$ and

$$P:\mathcal{M}_{n+1}\to\mathcal{M}_n \quad\text{and}\quad Q:\mathcal{M}_n\to\mathcal{M}_{n+1} \tag{7.13}$$

are one-to-one, onto, for each $n=0,1,2,\cdots$. If we let $E(\mathcal{M}_n)$ denote the projection of $\mathcal{H}$ on $\mathcal{M}_n$, then the spectral resolution of the numbers operator is given by

$$QP=\int\lambda\,dE(\lambda)=\sum_{n=0}^{\infty}nE(\mathcal{M}_n). \tag{7.14}$$

Let $\{e_{\alpha,0}\}$ be an orthonormal basis for $\mathcal{M}_0=\{f\in\mathcal{H}\,|\,QPf=0\}$. Then the vectors

$$e_{\alpha,n}=\frac{1}{(n!)^{1/2}}Q^ne_{\alpha,0}, \qquad n=1,2,\cdots, \tag{7.15}$$

exist, $\{e_{\alpha,n}\}_\alpha$ is an orthonormal basis for $\mathcal{M}_n$, and $\{e_{\alpha,n}\}_{\alpha,n}$ is an orthonormal basis for $\mathcal{H}$. Furthermore, the operators P and Q are determined by

$$Pe_{\alpha,n}=n^{1/2}e_{\alpha,n-1}, \qquad Qe_{\alpha n}=(n+1)^{1/2}e_{\alpha,n+1} \tag{7.16}$$

with $e_{\alpha,-1}=0$, for $n=0,1,2,\cdots$, and $D(P)=D(Q)$ is given by

$$D(Q)=\left\{f\in\mathcal{H}\,\middle|\,\sum_{\alpha,k}(k+1)^2|(f|e_{\alpha,k})|^2<\infty\right\}, \tag{7.17}$$

THEOREM 7.6. *Let* D_ν *be given by* (7.3). *Then* Q^2D_ν *has an extension to a positive selfadjoint operator, which we also denote by* Q^2D_ν, *with*

$$Q^2D_\nu=\int\lambda(\lambda+2\nu-1)\,dE(\lambda)=\sum_{k=0}k(k+2\nu-1)E(\mathcal{M}_k), \tag{7.18}$$

and

$$D(Q^2D_\nu)=\left\{f\in\mathcal{H}\,\middle|\,\sum_{\alpha,k}k^2(k+2\nu-1)^2|(f|e_{\alpha,k})|^2<\infty\right\}. \tag{7.19}$$

Proof. Since $D(D_\nu)\subset\mathcal{H}\ominus\mathcal{M}_1$, we have $D(Q^2D_\nu)\subset D(D_\nu)\subset\mathcal{H}\ominus\mathcal{M}_1$. However,

$$Q^2D_\nu\subset Q^2P^2+2\nu QP=Q(PQ-I)P+2\nu QP=(QP)^2+(2\nu-1)QP.$$

Further, $f(\lambda)=\lambda^2+(2\nu-1)\lambda$ is a Borel function on $(-\infty,\infty)$. Thus by the operator calculus (see Nagy [18]), we have

$$\begin{aligned} f(Q)=\int f(\lambda)\,dE(\lambda)&=\sum_{k=0}^{\infty}k(k+2\nu-1)E(\mathcal{M}_k)\\ &=(QP)^2+(2\nu-1)QP=Q^2D_\nu. \end{aligned} \tag{7.20}$$

By the operator calculus, it also follows that the domain of Q^2D_ν is given by (7.19).

Proceeding in the same manner, we obtain the following result.

COROLLARY 7.7. *We have*

$$D_\nu Q^2=\sum_{n=0}^{\infty}(n+2)(n+2\nu+1)E(\mathcal{M}_n) \tag{7.21}$$

with

$$D(D_\nu Q^2)=\left\{f\,\middle|\,\sum_{n,\alpha}\{(n+2)(n+2\nu+1)\}^2|(f|e_{\alpha,k})|^2<\infty\right\}.$$

Writing $f(\lambda)=\lambda^2+(2\nu-1)\lambda=(\lambda+(\nu-1/2))^2-(\nu-1/2)^2$, we obtain from (7.18) that

$$\left(\nu-\frac{1}{2}\right)^2+Q^2D_\nu=\sum_{k=0}^{\infty}\left\{k+\left(\nu-\frac{1}{2}\right)\right\}^2E(\mathcal{M}_k). \tag{7.22}$$

Thus Q^2D_ν is a generalized correlation of the numbers operator $B=QP$.

Finally, the elementary equality

$$x^2\Delta_x(\nu)x^k=k(k+2\nu-1)x^k$$

relates the eigenvalues of Q^2D_ν with the Bessel coefficients.

8. Generalized Fock spaces of several variables. The generalized Fock space theory presented in this paper occurs naturally in n-dimensional systems invariant under the orthogonal group. In systems presenting a multiple orthogonal invariance, the following product Fock spaces occur.

Let n be a positive integer and let $\underline{\nu}=(\nu_1,\nu_2,\cdots,\nu_n)$ with $\nu_k>0$ for all k. We let $\underline{m}=(m_1,m_2,\cdots,m_n)\in\mathcal{N}^n$, where $\mathcal{N}$ denotes the nonnegative integers. We define

$$b_{2\underline{m}}(\bar{\nu})=\prod_{k=1}^{m}b_{2m_k}(\nu_k), \tag{8.1}$$

where

$$b_{2m_k}(\nu_k)=2^{2m_k}m_k!\frac{\Gamma(m_k+\nu_k+1/2)}{\Gamma(\nu_k+1/2)}$$

is the Bessel coefficient,

$$\mathbf{K}_{\underline{\nu}}(\underline{z})=\mathbf{K}_{\nu_1}(z_1)\mathbf{K}_{\nu_2}(z_2)\cdots\mathbf{K}_{\nu_n}(z_n),\qquad \underline{z}\in C^n, \tag{8.2}$$

and

$$\mathbf{I}_{\underline{\nu}}(\underline{z})=\mathbf{I}_{\nu}(z_1)\mathbf{I}_{\nu_2}(z_2)\cdots\mathbf{I}_{\nu_n}(z_n). \tag{8.3}$$

We also introduce the product measure

$$d\mathfrak{m}_{\underline{\nu}}(\underline{z})=d\mathfrak{m}_{\nu_1}(z_1)d\mathfrak{m}_{\nu_2}(z_2)\cdots d\mathfrak{m}_{\nu_n}(z_n). \tag{8.4}$$

The Fock space $\mathfrak{F}_{\underline{\nu}}$ is the Hilbert space of "even" entire functions on C^n with inner product given by

$$\big(f(\underline{z})|g(\underline{x})\big)_{\underline{\nu}}=\int_{C^n}f(\underline{z})\bar{g}(\underline{z})\,d\mathfrak{m}_{\underline{\nu}}(\underline{z}). \tag{8.5}$$

If $f,g\in\mathfrak{F}_{\underline{\nu}}$ have Taylor series expansions

$$f(\underline{z})=\sum_{\underline{m}\in\mathcal{N}^n}a_{\underline{m}}\underline{z}^{2\underline{m}}\quad\text{and}\quad g(\underline{z})=\sum_{\underline{m}\in\mathcal{N}^n}c_{\underline{m}}\underline{z}^{2\underline{m}}, \tag{8.6}$$

where $\underline{z}^{2\underline{m}}=z_1^{2m_1}z_2^{2m_2}\cdots z_n^{2m_n}$, then

$$(f|g)_{\underline{\nu}}=\sum_{\underline{m}\in\mathcal{N}^n}a_{\underline{m}}\bar{c}_{\underline{m}}b_{2\underline{m}}(\underline{\nu}). \tag{8.7}$$

The reproducing kernel in $\mathfrak{F}_{\underline{\nu}}$ is given by

$$\mathcal{K}_{\underline{\nu}}(\underline{z},\underline{w})=\mathbf{I}_{\nu_1}(z_1\bar{w}_1)\mathbf{I}_{\nu_2}(z_2\bar{w}_2)\cdots\mathbf{I}_{\nu_n}(z_n\bar{w}_n)=:\mathbf{I}_{\underline{\nu}}(\underline{z}\cdot\underline{\bar{w}}). \tag{8.8}$$

We have

$$f(\underline{w})=\big(f(\underline{z})|K_{\underline{\nu}}(\underline{z},\underline{w})\big)_{\underline{\nu}}\quad\text{for all } \underline{w}\in C^n, \tag{8.9}$$

and therefore the product of the modified Bessel functions is the generalized Dirac delta function in $\mathfrak{F}_{\underline{\nu}}$. Moreover, $(f(\underline{z})|g(\underline{z}))_{\underline{\nu}}=f(\Delta_{\underline{z}}^{1/2})g_{\#}(\underline{z})|_{\underline{z}=0}$.

Further, the basic inequality

$$|f(\underline{z})|\le\|f\|_{\underline{\nu}}\prod_{k=1}^{n}\mathbf{I}_{\nu_k}\big(|z_k|^2\big)^{1/2} \tag{8.10}$$

is also valid. If $f\in\mathfrak{F}_{\underline{\nu}}$, then the growth of f is $\le(\underline{2},\underline{\tfrac{1}{2}})$ where $\underline{2}=(2,\cdots,2)$ is the associated order and $\underline{\tfrac{1}{2}}=(\tfrac{1}{2},\cdots,\tfrac{1}{2})$ is the associated type.

We define

$$\Delta_{\underline{z}}(\underline{\nu})=\big(\Delta_{z_1}(\nu_1),\Delta_{z_2}(\nu_2),\cdots,\Delta_{z_n}(\nu_n)\big), \tag{8.11}$$

$$\Delta_{\underline{z}}^{\underline{m}}f(\underline{z})=\Delta_{z_1}^{m_1}\Delta_{z_2}^{m_2}\cdots\Delta_{z_n}^{m_n}f(\underline{z}),\qquad \underline{m}\in\mathcal{N}^n, \tag{8.12}$$

$$\underline{z}^{2\underline{m}}f(\underline{z})=z_1^{2m_1}z_2^{2m_2}\cdots z_n^{2m_n}f(\underline{z}). \tag{8.13}$$

Letting $\underline{1}=(1,1,\cdots,1)\in\mathcal{N}^n$, we have that

$$\big(\Delta_{\underline{z}}^{\underline{1}}f(\underline{z})|g(\underline{z})\big)_{\underline{\nu}}=\big(f(\underline{z})|\underline{z}^{\underline{2}}g(\underline{z})\big)_{\nu}; \tag{8.14}$$

thus $\Delta_{\underline{z}}^{\underline{1}}$ and $\underline{z}^{\underline{2}}$ are adjoints in $\mathfrak{F}_{\underline{\nu}}$.

Let $d\mu_{\underline{\nu}}(\underline{x})=d\mu_{\nu_1}(x_1)\,d\mu_{\nu_2}(x_2)\cdots d\mu_{\nu_n}(x_n)$, and let $\mathcal{H}_{\underline{\nu}}=L^2_{\underline{\nu}}[R^n_+,d\mu_{\underline{\nu}}(\underline{x})]$. Further, define generalized Laguerre functions by

$$h_{\underline{m},\underline{\nu}}(\underline{x})=\prod_{k=1}^{m} h_{n_k,\nu_k}(x_k),\qquad \underline{x}=(x_1,\cdots,x_n). \tag{8.15}$$

Then the $\{h_{\underline{n},\underline{\nu}}|\underline{m}\in\mathfrak{N}^n\}$ form a complete orthogonal family in $\mathcal{H}_{\underline{\nu}}$ and

$$\int_{R^n_+} h_{\underline{m},\underline{\nu}}(\underline{x})h_{\underline{n},\underline{\nu}}(\underline{x})\,d\mu_{\underline{\nu}}(\underline{x})=b_{2\underline{n}}(\underline{\nu})\delta_{\underline{n},\underline{m}}. \tag{8.16}$$

Let

$$U_{\underline{\nu}}(\underline{z},\underline{x})=e^{-(\underline{x}\cdot\underline{x}-2\underline{z}\cdot\underline{z})/4}\mathbf{J}_{\underline{\nu}}(\underline{z}\cdot\underline{\bar{x}}) \tag{8.17}$$

where $\underline{z}\cdot\underline{z}=z_1^2+z_2^2+\cdots+z_n^2$, in the exponential. Then for all $z\in C^n$ and $\underline{x}\in R^n_+$, we have

$$U_{\underline{\nu}}(\underline{z},\underline{x})=\sum_{m\in\mathfrak{N}^n}\frac{h_{\underline{\underline{m}},\underline{\nu}}(\underline{x})}{b_{2\underline{n}}(\underline{\nu})}\underline{z}^{2\underline{m}} \tag{8.18}$$

and

$$\mathbf{I}_{\underline{\nu}}(\underline{z}\cdot\underline{w})=\int_{R^n_+} U_{\underline{\nu}}(\underline{z}\cdot\underline{x})U_{\underline{\nu}}(\underline{w},\underline{x})\,d\mu_{\underline{\nu}}(\underline{x}). \tag{8.19}$$

Finally we let

$$f(\underline{z})=\int_{R^n_+} U_{\underline{\nu}}(\underline{z},\underline{x})\psi(\underline{x})\,d\mu_{\underline{\nu}}(\underline{x})=U_{\underline{\nu}}\psi\cdot(\underline{z})\quad\text{for all }\psi\in\mathcal{H}_{\underline{\nu}}. \tag{8.20}$$

Then $U_\nu\psi=f$ is a unitary mapping of $\mathcal{H}_\nu$ onto $\mathfrak{F}_\nu$ and $U_\nu h_{\underline{m},\underline{\nu}}=\underline{z}^{2\underline{m}}$.

With these basic definitions and results, the main results of this paper can be extended to $\mathfrak{F}_{\underline{\nu}}$.

REFERENCES

[1] N. ARONSZAJN, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1948), pp. 337–404.

[2] V. BARGMANN, *On a Hilbert space of analytic functions and an associated integral transform, Part* I, Comm. Pure Appl. Math., 14 (1961), pp. 187–214.

[3] ______, *Remarks on a Hilbert space of analytic functions*, Proc. Nat. Acad. Sci. USA, 48 (1962), pp. 199–204.

[4] ______, *Group representations on Hilbert spaces of analytic functions*, in Analytic Methods in Mathematical Physics, R. P. Gilbert and R. G. Newton, eds., Gordon and Breach, New York, 1968, pp. 27–63.

[5] F. M. CHOLEWINSKI, *A Hankel Convolution Complex Inversion Theory*, Memoir 58, American Mathematical Society, Providence, RI, 1965.

[6] F. CHOLEWINSKI AND D. T. HAIMO, *The Weierstrass–Hankel convolution transform*, J. Analyse Math., 17 (1966), pp. 1–58.

[7] ______, *Classical analysis and the generalized heat equation*, SIAM Rev., 10 (1968), pp. 67–80.

[8] J. DELSARTE, *Sur une extension de la formule de Taylor*, J. Math. Pures Appl., 17 (1936), pp. 213–231.

[9] A. ERDÉLYI et al., *Higher Transcendental Functions*, Vol. 2, McGraw-Hill, New York, 1953.

[10] V. FOCK, *Verallgemeinerung und Lösung der Diracschen statistischen Gleichung*, Z. Phys., 49 (1928), pp. 339–357.

[11] D. L. GUY, *Hankel multiplier transformations and weighted p-norms*, Trans. Amer. Math. Soc., 95 (1960), pp. 137–189.

[12] D. T. HAIMO, L^2 *expansions in terms of generalized heat polynomials and of their Appell transforms*, Pacific J. Math., 15 (1965), pp. 865–875.

[13] ______, *Expansions in terms of generalized heat polynomials and of their Appell transforms*, J. Math. and Mech., 15 (1966), pp. 735–758.

[14] I. I. HIRSCHMAN, JR., *Variation diminishing Hankel transforms*, J. Analyse Math., 8 (1960-61), pp. 307–336.

[15] C. ITZYKSON, *Remarks on boson commutation relations*, Comm. Math. Phys., 4 (1967), pp. 92–122.

[16] J. L. LIONS, *Opérateurs de Delsarte et problèmes mixtes*, Bull. Soc. Math. France, 84 (1956), pp. 9–95.

[17] H. MESCHKOWSKI, *Hilbertsche Räume Mit Kernfunction*, Springer-Verlag, Berlin, 1962.

[18] B. SZ.-NAGY, *Spektraldarslellung linearer Transformationers des Hilbertschen Raumes*, Springer-Verlag, Berlin, 1942.

[19] H. G. TILLMANN, *Zur Eindeutigkeit der Losungen der quantenmechanischen Vertauschungsrelationen*, Acta. Sci. Math., 24 (1963), pp. 258–270.

[20] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge Univ. Press, Cambridge, 1959.

[21] A. WEINSTEIN, *Generalized axially symmetric potential theory*, Bull. Amer. Math. Soc., 59 (1953), pp. 20–38.

[22] ______, *The generalized radiation problem and the Euler–Poisson–Darboux equation*, Summa Brasiliensis, 3 (1955), pp. 125–145.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1501-0016 $01.25/0

AN INEQUALITY FOR THE BESSEL FUNCTION $J_\nu(\nu x)$*

R. B. PARIS[†]

Abstract. An inequality for the Bessel function $J_\nu(\nu x)$, $\nu>0$, $0<x\le 1$ involving both upper and lower bounds is derived. Inequalities for the modified Bessel functions are also obtained.

Most of the known inequalities for the Bessel function $J_\nu(\nu x)$, when the argument is less than the order, are upper bounds. The purpose of this note is to draw attention to the simple inequality involving both upper and lower bounds

$$(1) \qquad 1\le\frac{J_\nu(\nu x)}{x^\nu J_\nu(\nu)}\le e^{\nu(1-x)}, \qquad \nu>0, \quad 0<x\le 1.$$

The inequality (1) is quite sharp in the limit $x\to 1$, although it does not actually provide any useful information when $x=1$. This results, of course, from the fact that the expressions for the bounds themselves contain the term $J_\nu(\nu)$. Kapteyn's inequality [5, p. 268] (extended to $\nu>0$ by Siegel [3])

$$J_\nu(\nu x)<\frac{x^\nu \exp\left\{\nu(1-x^2)^{1/2}\right\}}{\left[1+(1-x^2)^{1/2}\right]^\nu}, \qquad \nu>0, \quad 0<x\le 1,$$

for example, is not particularly precise as $x\to 1$, though it provides a better upper bound than (1) in the limit $x\to 0$ whenever $2^\nu J_\nu(\nu)>1$, or roughly when $\nu>\frac{3}{2}$. The inequality [5, p. 49]

$$J_\nu(\nu x)<\frac{(\nu x/2)^\nu}{\Gamma(\nu+1)}, \qquad \nu>0, \quad x>0$$

is similarly not very sharp for $x\to 1$ (especially for large ν), but is in general more restrictive in the limit $x\to 0$ than the right-hand side of (1). However, as the chief interest in such inequalities is for functions of argument almost equal to their order, (1) would seem to be of possible interest.

To derive (1), we first observe that

$$(2) \qquad 0<\frac{x\nu}{2\nu+2}<\frac{J_{\nu+1}(\nu x)}{J_\nu(\nu x)}<\frac{x\nu}{\nu+2}<1, \qquad \nu>0, \quad 0<x\le 1.$$

The lower limit follows from the recurrence relation $J_\nu(z)+J_{\nu+2}(z)=2(\nu+1)J_{\nu+1}(z)/z$ and the fact that both $J_\nu(z)$ and $J_{\nu+2}(z)$ are positive for $\nu>0$ and $0<z\le 1$, since the

*Received by the editors June 15, 1982, and in revised form September 13, 1982.

[†]Association Euratom-CEA, Dèpartement de Recherches sur la Fusion Contrôlée, Centre d'Etudes Nuclèaires, 92260 Fontenay-aux-Roses, France.

smallest positive zeros $j_{\nu,1}$ and $j_{\nu+2,1}$ of $J_\nu(z)$ and $J_{\nu+2}(z)$ respectively satisfy $j_{\nu+2,1}>j_{\nu,1}>\nu$. The upper limit can be deduced from Sonine's integral [5, p. 373]

$$\begin{aligned}J_{\nu+1}(\nu x)&=\nu x\int_0^{\pi/2}J_\nu(\nu x\sin\theta)\sin^{\nu+1}\theta\cos\theta\,d\theta\\&<\nu xJ_\nu(\nu x)\int_0^{\pi/2}\sin^{\nu+1}\theta\cos\theta\,d\theta\\&=\frac{\nu x}{\nu+2}J_\nu(\nu x),\qquad \nu>0,\quad 0<x\le 1,\end{aligned}$$

since $J_\nu(z)$ is an increasing function of its argument for $0<z<\nu$.

Then using the recurrence relation $J_{\nu+1}(z)/J_\nu(z)=\nu/z-J_\nu'(z)/J_\nu(z)$, we obtain from (2) the result

$$0<\frac{1}{x}-\frac{J_\nu'(\nu x)}{J_\nu(\nu x)}<1,\qquad \nu>0,\quad 0<x<1,$$

where the prime denotes differentiation with respect to the argument. Integrating over the interval $[x,1]$ then yields

$$0<\int_x^1\left\{\frac{1}{x}-\frac{J_\nu'(\nu x)}{J_\nu(\nu x)}\right\}dx<1-x,$$

whence the inequality (1) immediately follows. This result can be sharpened by employing the more restrictive bounds in (2) to find

$$\exp\left\{\frac{\nu^2(1-x^2)}{4\nu+4}\right\}\le\frac{J_\nu(\nu x)}{x^\nu J_\nu(\nu)}\le\exp\left\{\frac{\nu^2(1-x^2)}{2\nu+4}\right\},\qquad \nu>0,\quad 0<x\le 1.$$

It is possible to establish in a similar manner the following inequalities for the modified Bessel functions, using the results $0<I_{\nu+1}(x)/I_\nu(x)<1$ for $x>0$, $\nu>-\frac{1}{2}$ [4] and $K_{\nu+1}(x)/K_\nu(x)>1$ for $x>0$, $\nu>-\frac{1}{2}$. This latter inequality follows immediately from Schläfi's integral representation for $K_\nu(x)$ [5, p. 181]

$$K_\nu(x)=\int_0^\infty\exp(-x\cosh t)\cosh\nu t\,dt.$$

Then integrating over the interval $[x,y]$, where $y>x>0$, we find

$$\frac{K_\nu(x)}{K_\nu(y)}>e^{y-x}\left(\frac{x}{y}\right)^\nu,\qquad y>x>0,\quad \nu>-\frac{1}{2}\tag{3}$$

and

$$\left(\frac{x}{y}\right)^\nu e^{x-y}<\frac{I_\nu(x)}{I_\nu(y)}<\left(\frac{x}{y}\right)^\nu,\qquad y>x>0,\quad \nu>-\frac{1}{2}.\tag{4}$$

The inequalities (3) and (4) have been derived previously by Ross [2] and Bordelon [1] by different methods. It should be remarked, however, that the upper bound of (4) is sharper than that given by Ross, since it does not contain the additional factor $e^{y-x} > 1$.

REFERENCES

[1] D. J. Bordelon, *Solution to Problem* 72-15, *Inequalities for special functions, D. K. Ross*, SIAM Rev., 15 (1973), pp. 666–668.

[2] D. K. Ross, *Solution to Problem* 72-15, *Inequalities for special functions, D. K. Ross*, SIAM Rev., 15 (1973), pp. 668–670.

[3] K. M. Siegel, *An inequality involving Bessel functions of argument nearly equal to their order*, Proc. Amer. Math. Soc., 4 (1953), pp. 858–859.

[4] R. P. Soni, *On an inequality for modified Bessel functions*, J. Math. and Phys., 44 (1965), pp. 406–407.

[5] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, Cambridge, 1966.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics

0036-1410/84/1501-0017 $01.25/0

ON THE SQUARE OF THE ZEROS OF BESSEL FUNCTIONS*

ÁRPÁD ELBERT† AND ANDREA LAFORGIA‡

Abstract. Let $j_{\nu k}$ denote the kth positive zero of the Bessel function $J_\nu(x)$ of the first kind. We define the function $j_{\nu\kappa}$ for all $\kappa>0$ in such a way that $j_{\nu\kappa}$ is the kth positive zero of the cylinder function $C_\nu(x)=\cos\alpha J_\nu(x)-\sin\alpha Y_\nu(x)$ by some α and k, for $0\leq\alpha<\pi$ and $k=1,2,\cdots$. Let $\kappa_0=\inf\{\kappa;\kappa>0,j^1_{\nu\kappa}>1\}$ where the (prime) indicates the derivative with respect to ν, then we find $0<\kappa_0<1$ (for $\nu\geq0$).

Our main result is that the function $j^2_{\nu\kappa}$ is a convex function of ν for $\nu\geq0$ and $\kappa\geq\kappa_0$. This result proves also the conjecture of J. T. Lewis and M. E. Muldoon [SIAM J. Math. Anal. 8 (1977), pp. 171–178], that $j^2_{\nu k}$ is convex for $\nu\geq0$ and $k=1,2,\cdots$. Finally we give some applications of this result and we show that the validity of this convexity cannot be extended to the whole interval $-\kappa\leq\nu<\infty$.

1. Introduction. Many authors have studied the monotonic properties of the kth positive zero $j_{\nu k}$ of the Bessel function $J_\nu(x)$ of the first kind where $k=1,2,\cdots$. More precisely, R. McCann in [6] and J. T. Lewis and M. E. Muldoon in [4] showed independently that the function $j_{\nu k}/\nu$ decreases as ν increases and $\nu>0$. Later, E. Makai [5] proved the same property with an ingenious application of the Sturm comparison theorem. In [4] the authors conjectured that $j_{\nu k}$ is a concave function and $j^2_{\nu k}$ is a convex function of ν, when $\nu>0$, at least in the case $k=1$, and they showed the convexity of $j^2_{\nu 1}$ only for $3\leq\nu<\infty$. One of the present authors proved in [1] that $j_{\nu k}$ is concave on the extended domain of the definition of $j_{\nu k}$, $-k\leq\nu<\infty$.

In this work we generalize the notation of $j_{\nu k}$ to $j_{\nu\kappa}$ with $\kappa>0$ and real. Our main result is that function $j^2_{\nu\kappa}$ is convex with respect to ν for $\nu\geq0$ and $\kappa\geq\kappa_0$ with some κ_0, $0<\kappa_0<1$. This result cannot be extended to the whole domain of the definition of $j_{\nu\kappa}$, at least not for $k=2,3,\cdots$. Finally we give some applications of the results obtained here.

2. On the inequality $j'=dj_{\nu\kappa}/d\nu>1$. From [8, p. 508] we know that

$$\frac{d}{d\nu}j_{\nu k}=2j_{\nu k}\int_0^\infty K_0(2j_{\nu k}\sinh t)e^{-2\nu t}\,dt,\qquad k=1,2,\cdots,\tag{2.1}$$

where $K_0(u)$ is the modified Bessel function of order zero, and it has the following integral representation [8, p. 446]:

$$K_0(u)=\int_0^\infty e^{-u\cosh z}\,dz.\tag{2.2}$$

For $\nu\geq0$ we use $c_{\nu k}$ to denote the kth positive zero of the general cylinder function

$$C_\nu(x)=\cos\alpha J_\nu(x)-\sin\alpha Y_\nu(x),$$

where α is fixed, $0\leq\alpha<\pi$ and $Y_\nu(x)$ is the Bessel function of the second kind. The definition may be extended to negative values of ν in such a way that $c_{\nu k}$ varies

*Received by the editors September 18, 1981, and in revised form July 16, 1982. This research was supported by Consiglio Nazionale delle Ricerche, Italy.

†Mathematical Institute of the Hungarian Academy of Sciences, 1053 Budapest, Reáltanoda u.13–15, Hungary.

‡Dipartimento di Matematica dell'Università, Via Carlo Alberto 10, 10123 Torino, Italy.

continuously with ν, $c_{\nu k}\to 0$ when $\nu\to\alpha/\pi-k$, and on the interval

$$\frac{\alpha}{\pi}-k<\nu<\frac{\alpha}{\pi}-k+1,$$

$c_{\nu k}$ is the first positive zero of $C_\nu(x)$; see [8, p. 508; 2].

The function $c_{\nu k}$ satisfies the differential equation (2.1) if we change $j_{\nu k}$ to $c_{\nu k}$ there. This fact suggests the following generalization: let $j_{\nu\kappa}$ be the solution of the differential equation

$$\frac{d}{d\nu}j=2j\int_0^\infty K_0(2j\sinh t)e^{-2\nu t}dt \tag{2.3}$$

for $\kappa>0$ with the boundary condition

$$\lim_{\nu\to-\kappa+0} j(\nu)=0.$$

Then for $\kappa=1,2,\cdots$ we obtain the known functions $j_{\nu k}$. If $k-1<\kappa<k$, we have $j_{\nu\kappa}=c_{\nu k}$ with $\alpha=(k-\kappa)\pi$. For example, $j_{\nu,k-1/2}=y_{\nu k}$, $k=1,2,\cdots$, where $y_{\nu k}$ denotes the kth zero of $Y_\nu(x)$. It is not difficult to show that the right-hand side of (2.3) is Lipschitzian with respect to j for $j>0$. Concerning the case $j=0$, we have $\lim_{\nu\to-\kappa+0}j_{\nu\kappa}=0$ for every $\kappa>0$, and hence the relation $\lim_{\nu\to-\kappa+0}j_{\nu\kappa'}=0$ implies $\kappa'=\kappa$. Therefore we have the uniqueness of the solutions for any initial value problem. Moreover this uniqueness implies that if $0<\kappa'<\kappa''$, then

$$j_{\nu\kappa'}<j_{\nu\kappa''}, \qquad \nu>-\kappa'. \tag{2.4}$$

In what follows we shall need the next result.

LEMMA. *If* $0\le\nu<\infty$ *and* $j_{\nu\kappa}\ge\nu+\frac{1}{4}$, *then*

$$j'=\frac{d}{d\nu}j_{\nu\kappa}>1.$$

Proof. Let us consider the domain $D=\{(\nu,j);\ 0\le\nu<\infty,\ j\ge\nu+\frac{1}{4}\}$. By (2.3) we must show that

$$I=I(\nu,j)=2j\int_0^\infty K_0(2j\sinh t)e^{-2\nu t}dt>1, \quad \text{if } (\nu,j)\in D. \tag{2.5}$$

Making the change of variable $u=2j\sinh t$, we have

$$I=\int_0^\infty K_0(u)\frac{e^{-2\nu\,\mathrm{arc\,sinh}(u/2j)}}{\sqrt{1+u^2/4j^2}}du. \tag{2.6}$$

On the other hand we know [8, p. 388]

$$\int_0^\infty K_0(u)e^{-u}du=1,$$

and hence it is sufficient to show that

$$\frac{e^{-2\nu\,\mathrm{arc\,sinh}(u/2j)}}{\sqrt{1+u^2/4j^2}}>e^{-u} \quad \text{for } u>0, \quad (\nu,j)\in D.$$

This is equivalent to

$$\nu\frac{\log\left(x+\sqrt{1+x^2}\right)}{x}+\frac{1}{4}\frac{\log(1+x^2)}{x}<j,$$

where $x=u/2j$. Since for $x>0$

$$\frac{\log\left(x+\sqrt{1+x^2}\right)}{x}<1;\ \frac{\log(1+x^2)}{x}<1,$$

the lemma follows.

By computation we could provide a larger domain for the validity of the inequality $j'>1$. Let us choose an initial point $(0,j_0)\in D$. Then the solution $j=j(\nu;0,j_0)$ remains in D for all $\nu>0$ and this solution can be continued to the left until $\nu=-\kappa(j_0)$, where $\lim_{\nu\to-\kappa(j_0)+0}j(\nu;0,j_0)=0$. Hence we have $j(\nu;0,j_0)=j_{\nu,\kappa(j_0)}$.

On account of the uniqueness of the solutions of the differential equation (2.3) we have that the function $\kappa=\kappa(j_0)$ is an increasing function of j_0. By the lemma we conclude that $(\nu,j_{\nu,\kappa(1/4)})\in D$ for all $\nu\geq 0$ and therefore $j'_{\nu\kappa}>1$ for all $\kappa\geq\kappa(\frac{1}{4})$. Since $j_{01}=2.40\cdots>1/4$, we have $0<\kappa(\frac{1}{4})<1$. If κ_0 is defined by

$$\kappa_0=\inf\{\kappa;\kappa>0, j'_{\nu\kappa}>1,\text{ for all }\nu\geq 0\}\tag{2.7}$$

then we get $0\leq\kappa_0<\kappa(\frac{1}{4})$. By (2.6) we have $\lim_{j\to+0}I(0,j)=0$; hence by (2.3) $j'(0,j_0)<1$ if j_0 is sufficiently small and therefore $\kappa_0>0$.

Remark 2.1. A consequence of the definition of κ_0 by (2.7) is

$$j_{\nu\kappa}>j_{0\kappa}+\nu,\qquad \nu>0,\quad \kappa\geq\kappa_0.\tag{2.8}$$

This inequality generalizes the similar result in [2] obtained only for $k=1,2,\cdots$.

Remark 2.2. Concerning the role played by κ_0 we have

$$\lim_{\nu\to\infty}\frac{j_{\nu\kappa}}{\nu}=1,\qquad \kappa\geq\kappa_0.\tag{2.9}$$

In fact, by (2.8) and (2.4) the function $j_{\nu\kappa}$ satisfies the inequalities

$$j_{0\kappa}+\nu<j_{\nu\kappa}<j_{\nu,[\kappa]+1},$$

where $[x]$ denotes the greatest integer less than or equal to x.

Tricomi's asymptotic formula [7] states for $k=1,2,\cdots$

$$j_{\nu k}=\nu+a_k\nu^{1/3}+O(\nu^{-1/3}),\qquad \nu\to\infty,$$

where a_k is independent of ν and then (2.9) follows. On the other hand by (2.5) one has

$$I(\nu,\nu)=2\nu\int_0^\infty K_0(2\nu\sinh t)e^{-2\nu t}dt$$

and recalling that the function $K_0(u)$ is decreasing as u increases we get

$$I(\nu,\nu)<2\nu\int_0^\infty K_0(2\nu t)e^{-2\nu t}dt=\int_0^\infty K_0(u)e^{-u}du=1.$$

Hence the solution $j(\nu;\nu_0,\nu_0)$ cannot cross the line $j=\nu$ any more; thus $j(\nu,\nu_0,\nu_0)<\nu$ for $\nu>\nu_0$ and $j(\nu;\nu_0,\nu_0)=j_{\nu,\tilde{\kappa}}$ with some $\tilde{\kappa}=\tilde{\kappa}(\nu_0)<\kappa_0$.

3. The main result. J. T. Lewis and M. E. Muldoon [4] proved the convexity of the function $j_{\nu 1}^2$ for $\nu \geq 3$. Now we prove the following more general result.

THEOREM. *Let the function $j_{\nu\kappa}$ be defined as above. Then $j_{\nu\kappa}^2$ is a convex function of ν for $\nu \geq 0$ and for every $\kappa \geq \kappa_0$.*

Proof. It is sufficient to show that

$$\left(\frac{j^2}{2}\right)'' = j'^2 + jj'' > 0. \tag{3.1}$$

Using the differential equation (2.3) we have by differentiation

$$j'' = 2j' \int_0^\infty K_0(2j \sinh t) e^{-2\nu t} dt + 2j \int_0^\infty K_0'(2j \sinh t) 2j' \sinh t e^{-2\nu t} dt \tag{3.2}$$
$$-4j \int_0^\infty K_0(2j \sinh t) t e^{-2\nu t} dt.$$

In view of (2.3) we can write the three terms on the right-hand side of (3.2) in the following form

$$j'' = \frac{j'^2}{j} + I_1 - I_2. \tag{3.3}$$

By the substitution $u = 2j \sinh t$ the integral I_1 becomes

$$I_1 = 2j' \int_0^\infty K_0'(u) \phi\left(\frac{u}{2j}\right) du,$$

where

$$\phi(x) = \frac{x e^{-2\nu \operatorname{arc\,sinh} x}}{\sqrt{1+x^2}}.$$

An integration by part gives

$$I_1 = 2j' \left[K_0(u) \phi\left(\frac{u}{2j}\right)\right]_0^\infty - \frac{j'}{j} \int_0^\infty K_0(u) \phi'\left(\frac{u}{2j}\right) du, \tag{3.4}$$

with

$$\phi'(x) = \frac{1 - 2\nu x \sqrt{1+x^2}}{(1+x^2)^{3/2}} e^{-2\nu \operatorname{arc\,sinh} x}, \tag{3.5}$$

and then

$$\phi'(x) < \frac{e^{-2\nu \operatorname{arc\,sinh} x}}{(1+x^2)^{3/2}} < 1, \qquad x > 0, \quad \nu \geq 0.$$

Recalling that

$$K_0(x) = \begin{cases} O(\log(1/x)), & x > 0, \quad x \sim 0, \\ o(e^{-x}), & x \gg 1, \end{cases}$$

we have that the first term in the right-hand side of (3.4) is zero. Then we get

$$I_1 = \frac{-j'}{j}\int_0^\infty K_0(u)\phi'\left(\frac{u}{2j}\right)du. \tag{3.6}$$

Similarly, for I_2 we have

$$I_2 = 2\int_0^\infty K_0(u)\frac{\operatorname{arc\,sinh}(u/2j)}{\sqrt{1+u^2/4j^2}}e^{-2\nu \operatorname{arc\,sinh}(u/2j)}du = \frac{1}{j}\int_0^\infty K_0(u)u\psi\left(\frac{u}{2j}\right)du \tag{3.7}$$

where

$$\psi(x) = \frac{1}{\sqrt{1+x^2}}\,\frac{\operatorname{arc\,sinh} x}{x}e^{-2\nu \operatorname{arc\,sinh} x}.$$

It is easy to see that

$$\psi(x) < \psi(0) = 1, \qquad x > 0,$$

hence

$$I_2 < \frac{1}{j}\int_0^\infty K_0(u)u\,du = \frac{1}{j}, \tag{3.8}$$

where the value of the integral may be found from [8, p. 388]. By (3.3), taking into account (3.6), (3.7), (3.8) and $j' > 1$ we have in (3.1)

$$\left(\frac{j^2}{2}\right)'' > j'^2 - j'\int_0^\infty K_0(u)\phi'\left(\frac{u}{2j}\right)du,$$

and therefore it is sufficient to show that

$$I_3 = j' - \int_0^\infty K_0(u)\phi'\left(\frac{u}{2j}\right)du > 0.$$

By (2.6) and (3.5) we have

$$I_3 = \int_0^\infty K_0(u)\frac{e^{-2\nu \operatorname{arc\,sinh}(u/2j)}}{\sqrt{1+u^2/4j^2}}\left[1 + \frac{1-2\nu(u/2j)\sqrt{1+u^2/4j^2}}{1+u^2/4j^2}\right]du,$$

where the quantity between the brackets is clearly positive for $u > 0$. This completes the proof of the convexity of $j_{\nu\kappa}^2$ for $\kappa \geq \kappa_0$ and $\nu \geq 0$. □

COROLLARY. *In the particular case* $\kappa \equiv k = 1, 2, \cdots$, *the function* $j_{\nu\kappa}^2$ *is convex for* $\nu \geq 0$.

4. Concluding remarks. Since $j_{\nu\kappa}^2$ is convex, the graph of $j_{\nu\kappa}^2$ lies below the chord joining the points $(0, j_{0\kappa}^2)$ and $(\nu^*, j_{\nu^*\kappa}^2)$. This gives the inequality

$$\frac{j_{\nu\kappa}^2 - j_{0\kappa}^2}{\nu} < \frac{j_{\nu^*\kappa}^2 - j_{0\kappa}^2}{\nu^*}, \qquad 0 < \nu < \nu^*,$$

i.e., the function $(j_{\nu\kappa}^2 - j_{0\kappa}^2)/\nu$ increases as ν increases, for $\nu > 0$. Next we consider the chord joining the points $(0, j_{0\kappa}^2)$ and $(\frac{1}{2}, j_{1/2,\kappa}^2)$ on the graph of $j_{\nu\kappa}^2$ as a function of ν. The convexity of the graph gives

$$j^2_{\nu\kappa} < j^2_{0\kappa} + 2\nu\left[\kappa^2\pi^2 - j^2_{0\kappa}\right], \qquad 0<\nu<\tfrac{1}{2},$$

where the inequality becomes equality only for $\nu=0$ and $\nu=1/2$. Similarly, by the convexity of $j^2_{\nu\kappa}$ it follows that

$$j^2_{\nu\kappa} > j^2_{0\kappa} + 2 j_{0\kappa}\nu\left[\frac{dj_{\nu\kappa}}{d\nu}\right]_{\nu=0}, \qquad \nu>0, \quad \kappa\geq\kappa_0,$$

and since $j'_{\nu\kappa}>1$ for $\kappa\geq\kappa_0$,

$$j^2_{\nu\kappa} > j^2_{0\kappa} + 2 j_{0\kappa}\nu, \qquad \nu>0.$$

The convexity of $j^2_{\nu\kappa}$ can be used to find many other inequalities too.

Finally one might ask the natural question, whether the validity of the convexity could be extended to the whole domain of definition of $j_{\nu\kappa}$, i.e., to the interval $(-\kappa,\infty)$. Let us consider only the cases when κ is a natural number, i.e, $\kappa\equiv k=1,2,\cdots$. Then for the zeros $j_{\nu k}$ of the Bessel function $J_\nu(x)$ we have [8, p. 15]

$$\begin{aligned}0&=(\nu+k)\Gamma(\nu+1)\left(\frac{j_{\nu k}}{2}\right)^{-\nu}J_\nu(j_{\nu k})\\ &=(\nu+k)\left[1-\frac{(j_{\nu k}/2)^2}{1!(\nu+1)}+\cdots+(-1)^{k-1}\frac{(j_{\nu k}/2)^{2(k-1)}}{(k-1)!(\nu+1)\cdots(\nu+k-1)}\right]\\ &\quad+(-1)^k\frac{(j_{\nu k}/2)^{2k}}{k!(\nu+1)\cdots(\nu+k-1)}\left[1-\frac{(j_{\nu k}/2)^2}{(k+1)(\nu+k+1)}+\cdots\right].\end{aligned}$$

Hence in the right neighborhood of $\nu=-k$ we have

$$\tag{4.1} \begin{aligned}\frac{(j_{\nu k}/2)^{2k}}{\nu+k}&=k!(k-1)!(1-\varepsilon)\left(1-\frac{\varepsilon}{2}\right)\cdots\left(1-\frac{\varepsilon}{k-1}\right)\\ &\quad\times\frac{1+\dfrac{(j_{\nu k}/2)^2}{k-1-\varepsilon}+\cdots+\dfrac{(j_{\nu k}/2)^{2(k-1)}}{(k-1)!(k-1-\varepsilon)\cdots(1-\varepsilon)}}{1-\dfrac{(j_{\nu k}/2)^2}{(k+1)(1+\varepsilon)}+\cdots},\end{aligned}$$

where $\varepsilon=\nu+k$.

Letting $\varepsilon\to0$ and $j_{\nu k}\to0$, we obtain

$$\lim_{\nu\to-k+0}\frac{(j_{\nu k}/2)^{2k}}{\nu+k}=k!(k-1)!.$$

We can write this relation in the form

$$\left(\frac{j_{\nu k}}{2}\right)^2=[k!(k-1)!(\nu+k)]^{1/k}[1+o(1)], \qquad \nu\to-k.$$

In the case $k=2,3,\cdots$ by (4.1), we have the more precise approximation

$$\frac{(j_{\nu k}/2)^{2k}}{\nu+k}=k!(k-1)!\left\{1+\frac{2k}{k^2-1}[k!(k-1)!\varepsilon]^{1/k}[1+o(1)]\right\}.$$

Hence

$$\left(\frac{j_{\nu k}}{2}\right)^2=[k!(k-1)!(\nu+k)]^{1/k}+\frac{2}{k^2-1}[k!(k-1)!(\nu+k)]^{2/k}[1+o(1)],$$
$$\nu\to -k, \quad k=2,3,\cdots.$$

For $k=1$ we get

$$\left(\frac{j_{\nu 1}}{2}\right)^2=\nu+1+\frac{1}{2}(\nu+1)^2[1+o(1)], \qquad \nu\to -1.$$

These approximations indicate that the function $j_{\nu k}^2$ cannot be convex on the whole interval $(-k,\infty)$ for $k=2,3,\cdots$. Whether the function $j_{\nu 1}^2$ is convex on $(-1,0)$, too, is not known, but we expect that it is.

Acknowledgments. We are grateful to a referee for helpful comments. The second author is grateful for the hospitality extended by the Mathematical Institute of the Hungarian Academy of Sciences, where he was a visitor when this paper was written.

REFERENCES

[1] Á. Elbert, *Concavity of the zeros of Bessel functions*, Studia Sci. Math. Hungar., 12 (1977), pp. 81–88.

[2] A. Laforgia and M. E. Muldoon, *Inequalities and approximations for zeros for Bessel functions of small order*, this Journal, 14 (1983), pp. 383–388.

[3] ———, *Monotonicity and concavity properties of zeros of Bessel functions*, J. Math. Anal. Appl., to appear.

[4] J. T. Lewis and M. E. Muldoon, *Monotonicity and convexity properties of zeros of Bessel functions*, this Journal, 8 (1977), pp. 171–178.

[5] E. Makai, *On zeros of Bessel functions*, Univ. Beograd publ. Elektrotechn, Fak. Ser. Mat. fiz. N. 602–633, (1978), pp. 109–110.

[6] R. McCann, *Inequalities for the zeros of Bessel functions*, this Journal, 8 (1977), pp. 166–170.

[7] F. G. Tricomi, *Sulle funzioni di Bessel di ordine e argomento pressoché uguali*, Atti Acc. Sci. Torino Cl. Sci. Fis. Mat. Nat., 83 (1949), pp. 3–20.

[8] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge Univ. Press, Cambridge, 1944.

SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1501-0018 $01.25/0

ERRATA: AN INTEGRAL EQUATION CONNECTED WITH THE JACOBI POLYNOMIALS*

B. F. LOGAN†

In the abstract, (ii) should read

$$k(t)=a|t|^{-\nu}+b|t|^{-\nu}\operatorname{sgn} t, \qquad |t|>0$$

and after (ii) delete "nonzero" before "real numbers".

In (1.13), $P_n^{\alpha,\beta}(t)$ should read $P_n^{(\alpha,\beta)}(t)$.

Equation (2.23) should read

$$\tilde{k}_{\alpha,\beta}(x)=-\tilde{k}_{\alpha,\beta}(-x).$$

Equation (2.31) should read

$$\left(\frac{d}{dx}\right)^n x^m\tilde{k}_{\alpha,\beta}(x)=(-1)^n(\nu-m)_n x^{m-n}\tilde{k}_{\alpha,\beta}(x).$$

For clarity, a "cut" in the integral sign ($\fint$), should be inserted to indicate a Cauchy principal value (at $t=x$) in equations (3.19), the first integral of (3.26), (4.8), (7.5), (7.7), (7.14), (7.15), (8.61), (8.64), (11.22), (11.23), and (11.24).

In the heading of §8.5, n should be replaced by ν. In the second line of (8.5.10), the coefficient of $-xf(x)$ should be $(2-\lambda-\mu)$.

In (10.10) the second factor of the integrand should be $(1+t)^{\mu/2-1/4}$.

On the right in (10.18), the exponent n should be replaced by m.

*This Journal, 14 (1983), pp. 269–322.

†Bell Laboratories, Murray Hill, New Jersey 07974.

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1502-0001 $01.25/0

FREQUENCY PLATEAUS IN A CHAIN OF WEAKLY COUPLED OSCILLATORS, I.*

GEORGE BARD ERMENTROUT† AND NANCY KOPELL‡

Abstract. A chain of $n+1$ weakly coupled oscillators with a linear gradient in natural frequencies is shown to exhibit "frequency plateaus," or sequences of oscillators having the same frequency, with a jump in frequency from one plateau to another. We first show that the equations for the coupled oscillators admit an invariant $(n+1)$-torus on which the equations have a special form, one in which an n-dimensional subsystem is approximately invariant. We then show that when the linear gradient becomes too steep to allow phaselocking, there emerges a large-scale invariant circle in this n-dimensional system which corresponds to the existence of a pair of plateaus, and whose homotopy class within the n-torus corresponds to the position of the frequency jump. Also discussed are the effects of anisotropic and nonuniform coupling.

1. Introduction.

We shall study a chain of $n+1$ weakly coupled oscillators which are uniformly close. For much of the paper, we shall assume that the coupling is nearest neighbor, isotropic (symmetric), homogeneous in k and linear. Thus, the kth oscillator satisfies an equation of the form

$$(1.1)_k \qquad X_k' = F(X_k) + \varepsilon R_k(X_k, \varepsilon) \equiv F_k(X_k, \varepsilon)$$

where $X_k \in R^m$, F: $R^m \to R^m$ and (1.1), with $\varepsilon = 0$, has a stable limit cycle solution of period $2\pi/\omega_0$. The full equations are

$$(1.2) \qquad X_k' = F_k(X_k) + \varepsilon D(X_{k+1} - 2\gamma X_k + X_{k-1}), \qquad X_0 = 0 = X_{n+2},$$

where D is an $m \times m$-matrix, $\varepsilon \ll 1$ and $\gamma = 0$ or 1. If $\gamma = 1$, the coupling is of the kind associated with diffusion; if $\gamma = 0$, the coupling is of "direct" type used to describe some electrical interactions.

Let ω_k be the frequency of the limit cycle of $(1.1)_k$. By hypothesis, $\omega_k = \omega_0 + O(\varepsilon)$. We first show, in §2, that there is an $(n+1)$-dimensional submanifold of $R^{m(n+1)}$ which is attracting and invariant under (1.2). This manifold is an $(n+1)$-dimensional torus T^{n+1}; we prove that variables $\theta_1, \theta_2, \cdots, \theta_{n+1}$ may be chosen on the torus so that, if $\phi_k \equiv \theta_{k+1} - \theta_k$, then the equations for θ_1 and the $\{\phi_k\}$ take the form

$$(1.3) \qquad \theta_1' = \omega_1 + \varepsilon H(\phi_1) + O(\varepsilon^2),$$

$$(1.4) \qquad \begin{aligned} &\phi_k' = \varepsilon[\Delta_k + H(\phi_{k+1}) + H(-\phi_k) - H(\phi_k) - H(-\phi_{k-1})] + O(\varepsilon^2), \\ &H(-\phi_0) = 0 = H(\phi_{n+1}). \end{aligned}$$

Here H is 2π-periodic and $\varepsilon\Delta_k = \omega_{k+1} - \omega_k$. The $O(\varepsilon^2)$ terms may depend on all the variables $\theta_1, \phi_1, \cdots, \phi_n$. H depends on D, on the form of the coupling and on the dynamics of (1.1) in the neighborhood of the limit cycles. Note that the equations for the $\{\phi_k\}$ are, to lowest order, independent of θ_1. Thus, through $O(\varepsilon)$, we may treat the phase space as T^n, with variables $\phi_1, \cdots, \phi_n$.

*Received by the editors September 14, 1982.

†National Institutes of Health, NIADDK, Mathematical Research Branch, Bethesda, Maryland 20205. Present address: Department of Mathematics (Applied), University of Pittsburgh, Pittsburgh, Pennsylvania 15261.

‡Department of Mathematics, Northeastern University, Boston, Massachusetts 02115. The research of this author was supported in part by the National Science Foundation under grant MCS-8102078.

The results of §2 are rigorous generalizations of calculations made by Neu [1], [2], Holmes [3], and Holmes and Rand [4]. Neu's calculations [1] were for a general pair of oscillators with diffusive coupling; Holmes and Rand [4] computed ϕ' for a pair of Van der Pol oscillators, also with diffusive coupling. Holmes [3] worked out examples in which $H(\phi)=\sin\phi$. Now $\sin\phi$ is an odd function of its argument. Also, for two coupled oscillators, H may just as well be odd, since from (1.5) we have that $\phi'=\varepsilon[\Delta-2H_0(\phi)]+O(\varepsilon^2)$, where H_0 is the odd part of H. However, H need not in general be odd. In §2 we give examples to illustrate which features of the dynamics or coupling lead to a function H which is odd. We compute H for $\lambda-\omega$ oscillations and Van der Pol oscillations (in the nearly sinusoidal regime) with various kinds of coupling.

The symmetry, or lack thereof, of H turns out to play an important role in the behavior of (1.4). In this paper, we shall study only the case H odd; later papers will take up the effects of lack of symmetry. If H is assumed to be odd, the governing equations immediately become simpler: letting $\tau=\varepsilon t$ and $\dot{\phi}\equiv d\phi/d\tau=(1/\varepsilon)(d\phi/dt)\equiv(1/\varepsilon)\phi'$, to lowest order, (1.4b) becomes

$$(1.5) \qquad \dot{\boldsymbol{\phi}}=\beta\boldsymbol{\Delta}+K\mathbf{H}(\boldsymbol{\phi})$$

where $\boldsymbol{\phi}=(\phi_1,\cdots,\phi_n)^t$, $\beta\boldsymbol{\Delta}=(\Delta_1,\cdots,\Delta_n)^t$, $\mathbf{H}(\boldsymbol{\phi})=(H(\phi_1),\cdots,H(\phi_n))^t$ and K is a tridiagonal matrix with $K_{ii}=-2$, $K_{i+i,i}=K_{i,i+1}=1$. The parameter $\beta\in R^1$ has been introduced, so we may consider (1.5) as a one-parameter family of equations with $\boldsymbol{\Delta}$ fixed and β measuring the strength of the "detuning."

We prove in §3 that for β sufficiently small, there is a unique stable equilibrium point for the n-dimensional system (1.5) which corresponds to "phase-locked" behavior, i.e., all the oscillators move at the same frequency, with fixed (in time) phase differences between any pair. (For the full $(n+1)$-dimensional equations (1.3), (1.4), the critical point of (1.4) or (1.5) corresponds to a stable limit cycle whose period is the shared period of the coupled oscillators.) The main result, proved in §§3 and 4, concerns "frequency plateaus" which emerge for (1.3), (1.4) when the stable critical point of (1.5) disappears. By a frequency plateau we mean a sequence of oscillators whose frequency is the same; this does not mean that the phase differences within the plateau are constant in time. It is shown that when the stable critical point coalesces with another critical point and disappears (as β is increased), a *large* amplitude stable limit cycle for (1.5) emerges (*not* by a Hopf bifurcation); this can be interpreted to correspond to the existence of a pair of frequency plateaus with different frequencies. The homotopy class of this cycle (as a point set within T^n) indicates the position of the discontinuity in frequency. For this we need more assumptions on H (it must be qualitatively similar to $\sin\phi$) and $\boldsymbol{\Delta}$ which we detail in §3. The methods used involve the construction of a large invariant region for (1.5) on which a set of inequalities hold. These inequalities are reminiscent of those used by Hirsch [5] in his study of cooperative systems. The proof also requires algebraic results about matrices of the form KA where A is diagonal; these are given in the Appendix.

The existence of the large amplitude limit cycle for (1.5) is done in §3; the relation of this to frequency plateaus is discussed in §4. Also done in §4 are a calculation of the size of the frequency jump as a function of the amount of detuning, and numerical computations showing the existence of further plateaus as the spread of natural frequencies increases. Section 5 contains calculations concerning related models: we consider the effect of anisotropy in the coupling, and a gradient in the strength of coupling. For these cases, we consider only phase-locked solutions.

Other papers treating phase-locking in coupled nonlinear oscillators are [6]–[11]. References [2], [3], [8], [9] deal with more than two oscillators. Of these, the approach of Holmes et al. [3], [4], [8] and Hoppensteadt and Keener [9] are closest to ours, using equations governing phase differences. Hoppensteadt and Keener derive their equations under the assumptions that each oscillator is a perturbation of a harmonic oscillator. Their analysis then requires them to make further assumptions about the algebraic relationship of the frequencies; these assumptions are unnecessary in our formulation. References [1], [2], [3] noted that if the natural frequencies of a pair of coupled oscillators are too far apart, the oscillators may lose synchrony. To the best of our knowledge, there has not yet been a mathematical analysis of the fact that, when there are many oscillators, the loss of synchrony can be local, i.e., the frequency may be constant over many oscillators.

This paper was partially motivated by certain phenomena observed in mammalian small intestine, which consists of layers of smooth muscle fiber. It is known that the muscle fibers support travelling waves of electrical activity which run from the oral to the aboral end [12]–[15]. These, in turn, trigger waves of muscular contractions [12], [13] via high frequency electrical spikes. The spikes, which have much higher frequency, are considered to be consequences of the slow waves, so we are concerned only with the slow electrical waves.

The connection with the above mathematics is as follows: If a section of the intestine is sliced into pieces of length 1–3 cms., each piece is capable of supporting spontaneous oscillations at a constant frequency, with a wave form that is close to sinusoidal [15]. (The origin of these oscillations is controversial [13].) Furthermore, over a substantial section of the intestine there is a linear gradient in the frequency of these oscillations, higher in the oral end than in the aboral. In vivo, the measured electrical activity along the (intact) intestine displays the frequency plateaus discussed in this paper. (There are usually more than two plateaus.)

In [16]–[20], this system was modelled by a chain of loosely coupled Van der Pol or related oscillators in the sinusoidal (nonrelaxation) regime, and simulated either digitally or electronically. These papers showed that, with a variety of different couplings (usually anisotropic), and with gradients in frequencies and couplings, frequency plateaus can be produced. Such plateaus share with the physiological data the property that the plateaus lie above the curve of natural (uncoupled) frequencies. (See Fig. 1.1.)

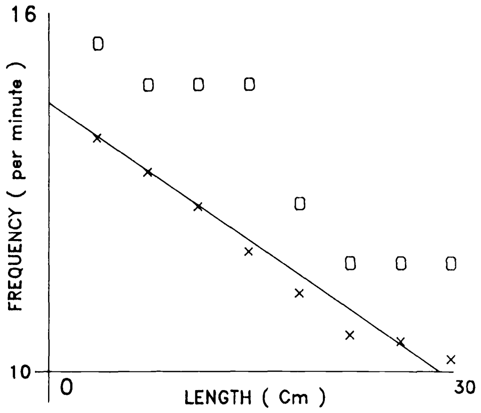


FIG. 1.1. *A schematic representation of frequency measurements in an intact mammalian intestine (top, piecewise constant), and after cutting a* 30-*cm. segment into* 8 *slices. Diagram after Diamont and Bortoff* [15]. *The positions of the plateaus do not remain constant in time* [15].

This paper is the beginning of an attempt to understand in a more general context the underlying reasons for the existence and properties of frequency plateaus. For example, we wish to show that the observations of [16]–[20] can be accounted for by phase models, with all the relevant information about the oscillators encoded in a set of 2π-periodic functions H (which may depend on k). This first paper is aimed primarily at the existence of plateaus. There are other aspects of the physiological data and simulations that cannot be accounted for if H is assumed to be odd and the coupling is isotropic. In particular, if H is odd, the coupling is uniform and isotropic, the natural frequency gradient is linear, and β is small enough that phase-locking occurs, then the phase-locked frequency is the average of the natural frequencies; if β is large enough so there are plateaus, these plateaus must be arranged symmetrically with respect to the average frequency (not above the curve of natural frequencies). Even if nonisotropic or nonuniform coupling is allowed, it is shown in §5 that the phase-locked frequency lies strictly between the highest and the lowest of the natural frequencies. We show in a later paper [21] that plateaus lying above the curve of natural frequencies can be derived from a phase model, provided that H is allowed to have a nonodd component, and n is large. Ultimately, this physiological system should be understood in terms of a continuum model.

2. Equations on an invariant torus. In this section we show that, for ε sufficiently small, there is an $(n+1)$-dimensional invariant submanifold $T^{n+1}(\varepsilon)$ of $R^{m(n+1)}$ which is an $(n+1)$-dimensional torus. On $T^{n+1}(\varepsilon)$, the motion is parametrized by phases θ_k associated to each oscillator. We also show that, to lowest order in ε, the equations have a special form which will enable us to analyze their behavior as the amount of detuning is increased.

It is easy to show that there is an invariant torus $T^{n+1}(\varepsilon)$ if ε is sufficiently small. For if $\varepsilon=0$, the cross products of the limit cycles of (1.1) for each X_k forms such a torus T^{n+1}. Furthermore, since each limit cycle is exponentially stable, this invariant manifold is "normally hyperbolic," i.e., in a neighborhood of T^{n+1}, trajectories approach the invariant manifold at an exponential rate. (See [22], [23] for more precise and general definitions.) It follows that there is an ε_0 such that, for $\varepsilon\leq\varepsilon_0$, the invariant manifold persists [22], [23], i.e., there is an invariant $T^{n+1}(\varepsilon)$ close to T^{n+1}.

We now show that coordinates $\theta_1, \phi_1,\cdots,\phi_n$ may be chosen on $T^{n+1}(\varepsilon)$ so that the equations for $\{\phi_k\}$ have the form (1.4). We first make a preliminary change of variables:

LEMMA 2.1. *Suppose that*

$$X'=F(X) \tag{2.1}$$

has a stable limit cycle with period $2\pi/\omega_0$, where $X\in R^m$ and F: $R^m\to R^m$ is C^∞. Then there exist smooth coordinates $\theta\in S^1$, $Y\in R^{m-1}$ in a neighborhood of the limit cycle of (2.1) *such that* (2.1) *becomes*

$$\theta'=\omega_0, \qquad Y'=L(\theta)Y+O(|Y|^2) \tag{2.2}$$

where the $O(|Y|^2)$ term may depend on θ.

Proof. The basic idea is to use coordinates in a neighborhood H of the limit cycle that are adapted to certain codimension-1 submanifolds which are known in the context of oscillations as "isochrons" [24], [25] and more generally as "leaves" of a "foliation" [23]. These leaves are transverse to the limit cycle and have the properties that each leaf gets sent onto another leaf under the action of the differential equation, and that any two points on the same leaf approach each other exponentially as $t\to\infty$. It can be shown that there are such manifolds, and that they vary smoothly with points on the limit cycle [23]. $\theta(X)$ is defined by requiring that the motion of (2.1) be uniform on the limit cycle, and θ be constant on each leaf of the foliation. ($\theta=0$ is chosen arbitrarily.) Since the flow takes each leaf into another leaf at each fixed time, (2.2) holds not only on the limit cycle, but in the entire neighborhood. Also, since the foliation is smooth, $\theta(X)$ is smooth. The Y coordinate may be defined more arbitrarily on each leaf, provided only that $Y=0$ on the limit cycle and $Y(X)$ is smooth. □

Lemma 2.1 shows that there is a smooth transformation $X=G(\theta,Y)$ which takes (1.1), with $\varepsilon=0$, into (2.2). Denote the Jacobian matrix by $J(\theta,Y)$. In a neighborhood of the limit cycles, J is invertible, so (1.2), $k\neq 1$, $n+1$ may be written as

$$\begin{pmatrix}\theta_k'\\ Y_k'\end{pmatrix}=J^{-1}(\theta_k,Y_k)\{F_k(G(\theta_k,Y_k))$$
$$+\varepsilon D[G(\theta_{k+1},Y_{k+1})-2\gamma G(\theta_k,Y_k)+G(\theta_{k-1},Y_{k-1})]\}.$$

There are similar equations for $k=1$, $n+1$. By hypothesis,

$$\begin{matrix}(2.3\text{a})\\(2.3\text{b})\end{matrix}\qquad J^{-1}(\theta_k,Y_k)F_k(G(\theta_k,Y_k\ ))=\begin{pmatrix}\omega_0+O(\varepsilon)\\ L(\theta_k)Y_k+O(|Y_k|^2,\varepsilon)\end{pmatrix}.$$

The right-hand side of (2.3a) may be written as

$$\omega_k+\varepsilon\bar{R}_k(\theta_k,Y_k,\varepsilon)$$

where

$$\int_0^{2\pi}\bar{R}_k(\theta_k,0,\varepsilon)\,d\theta_k=O(\varepsilon)$$

and ω_k, as stated before, is the frequency of the limit cycle of $(1.1)_k$. Let $h(\theta_i,\theta_k)$ denote the θ component of $J^{-1}(\theta_k,Y_k)DG(\theta_i,Y_i)$ at $Y_i=0=Y_k$. h is 2π-periodic in each of its arguments. Also let $\phi_k\equiv\theta_{k+1}-\theta_k$ and $S_k=Y_k/\varepsilon$. (The latter change of variables "blows up" an ε-neighborhood of $T^{n+1}(\varepsilon)$.) Then (2.3) becomes

$$\theta_1'=\omega_1+\varepsilon h(\theta_2,\theta_1)+O(\varepsilon^2), \tag{2.4a}$$

$$\begin{aligned}\phi_k'=\varepsilon[\Delta_k+h(\theta_{k+2},\theta_{k+1})-2\gamma h(\theta_{k+1},\theta_{k+1})+h(\theta_k,\theta_{k+1})&\\ -h(\theta_{k+1},\theta_k)+2\gamma h(\theta_k,\theta_k)-h(\theta_{k-1},\theta_k)]+O(\varepsilon^2)&,\end{aligned} \tag{2.4b}$$

$$S_k'=O(1)$$

where $\varepsilon\Delta_k\equiv\omega_{k+1}-\omega_k$ and the $O(\varepsilon^2)$ terms may depend on all the variables θ_1, $\{\phi_k\}$ and $\{S_k\}$. (Equation (2.4b) is true for $k=2,\cdots,n-1$. To get the equations for $k=1$ and $k=n$, set $h(\theta_0,\theta_1)=0=h(\theta_{n+1},\theta_n)$.) Note that, to lowest order, the right-hand side of (2.4a, b) is independent of the $\{S_k\}$. Thus (2.4) may be thought of as the dynamical system on $T^{n+1}(\varepsilon)$. (There is a dependence on $\{S_k\}$ in the $O(\varepsilon^2)$ term. However, on the invariant manifold, $S_k=\bar{S}_k(\theta_1,\cdots,\theta_{n+1})$, and so the $\{S_k\}$ may be eliminated.)

Note also that there are two time scales in (2.4a, b): $\theta_1'=O(1)$ in ε and $\phi_k'=O(\varepsilon)$ for all k. Thus, the $\{\phi_k\}$ form an n-dimensional "slow system" within $T^{n+1}(\varepsilon)$. However, there is not necessarily an n-dimensional submanifold of $T^{n+1}(\varepsilon)$ invariant under

(2.4). Nevertheless, using averaging theory, the difference in time scales can be exploited to write equations for the $\{\phi_k\}$ which, to lowest order, are independent of θ_1. Denote by (2.5) equation (2.4) with the expressions θ_{k+1}, θ_{k+2} and $\theta_{\ k-1}$ replaced by $\theta_k+\phi_k$, $\theta_k+\phi_k+\phi_{k+1}$ and $\theta_k-\phi_{k-1}$ respectively. Using the fact that $\theta_k=\omega_0 t+O(\varepsilon)$ and $\phi'_k=O(\varepsilon)$ for all k, we may now use the averaging theorem [26]. This theorem asserts that there is a near-identity change of coordinates such that, in the new coordinates the right-hand side of (2.5) may be replaced, to lowest order in ε, by its average with respect to t over one period. But, by the periodicity of h and the fact that $\phi'_k=O(\varepsilon)$,

$$(2.6)\quad \frac{\omega_0}{2\pi}\int_0^{2\pi/\omega_0} h(\theta_k+\phi_k+\phi_{k+1},\theta_k+\phi_k)\,dt=\frac{\omega_0}{2\pi}\int_0^{2\pi/\omega_0} h(\theta_k+\phi_{k+1},\theta_k)\,dt+O(\varepsilon)$$

$$=\frac{1}{2\pi}\int_0^{2\pi} h(\theta_k+\phi_{k+1},\theta_k)\,d\theta_k+O(\varepsilon).$$

A similar computation holds for the other terms of (2.5). Define

$$(2.7)\qquad H(\phi)\equiv\frac{1}{2\pi}\int_0^{2\pi}[h(\theta+\phi,\theta)-\gamma h(\theta,\theta)]\,d\theta.$$

We have shown the following:

THEOREM 2.1. *There is an* $(n+1)$*-dimensional submanifold* $T^{n+1}(\varepsilon)$ *invariant under* (1.2). *Variables* θ_1, $\phi_1,\cdots,\phi_n$ *may be chosen on* $T^{n+1}(\varepsilon)$ *so that, on the invariant manifold,* (1.2) *has the form* (1.3), (1.4), *with* H 2π*-periodic.*

We now explicitly calculate the function H for several classes of examples. The first has a natural polar coordinate system representation. However, as we shall see, the natural representation is not the one used in the proof of Theorem 2.1. Consider $m=2$ and

$$(2.8)\qquad F_k\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}\lambda & -\omega\\ \omega & \lambda\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix},\qquad D=\begin{pmatrix}d_1 & d_2\\ d_3 & d_4\end{pmatrix}$$

where $\lambda=1-(x^2+y^2)$, $\omega=\omega_k+\hat{\omega}(x^2+y^2)$, $\hat{\omega}$: $R^1\to R^1$, $\hat{\omega}(1)=0$, and $\omega_k=\omega_0+O(\varepsilon)$ for all k. In the usual polar coordinates $(x=r\cos\theta, y=r\sin\theta)$, $X'=F(X)$ is

$$r'=r\lambda(r^2),\qquad \theta'=\omega(r^2).$$

Thus $\omega(r)$ is an amplitude dependent angular frequency. The representation used in Theorem 2.1 has the form $\theta'=\omega_k$, where ω_k is a constant (independent of amplitude). To achieve this, we make the coordinate change $(\theta,r)\to(\bar{\theta},r)$, where $\bar{\theta}=\theta+\mu(r)$ and

$$\mu(r)=\int_1^r \frac{\hat{\omega}(\bar{r})}{\bar{r}(1-\bar{r}^2)}\,d\bar{r}.$$

Again we let S_k be a "blown up" normal coordinate, i.e., $r_k=1+\varepsilon S_k$. Using trigonometric identities, it can be checked that, in S_k, θ_k coordinates, (1.2) is

$$(2.9)\quad \theta'_k=\omega_k+\varepsilon[\hat{\omega}'(1)S_k+H_1(\theta_{k-1},\theta_k)+H_1(\theta_{k+1},\theta_k)]+O(\varepsilon^2),\qquad k=2,\cdots,n,$$

$$(2.10)\quad S'_k=-2S_k+H_2(\theta_{k-1},\theta_k)+H_2(\theta_{k+1},\theta_k),\qquad k=2,\cdots,n,$$

where

$$
\begin{aligned}
H_1(\theta_{k\pm1},\theta_k)={}&d_1\sin(\theta_{k\pm1}-\theta_k)+(d_4-d_1)\sin\theta_{k\pm1}\cos\theta_k\\
&+(d_1-d_4)\sin\theta_k\cos\theta_k+d_2-(d_2+d_3)\cos^2\theta_k\\
&-d_2\cos(\theta_{k\pm1}-\theta_k)+(d_3+d_2)\cos\theta_{k\pm1}\cos\theta_k,\\
H_2(\theta_{k\pm1},\theta_k)={}&d_1\cos(\theta_{k\pm1}-\theta_k)+(d_4-d_1)\sin\theta_k\sin\theta_{k\pm1}\\
&-d_1-(d_4-d_1)\sin^2\theta_k\\
&+d_2\sin(\theta_{k\pm1}+\theta_k)+(d_3-d_2)\sin\theta_k\cos\theta_{k\pm1}\\
&-(d_2+d_3)\cos\theta_k\sin\theta_k.
\end{aligned}
$$

Now $\bar{\theta}_k'=\theta_k'+\mu'(r_k)r_k'=\theta_k'+\varepsilon\mu'(1)S_k'+O(\varepsilon^2)=\theta_k'-\frac{1}{2}\varepsilon\hat{\omega}'(1)S_k'+O(\varepsilon^2)$ and $\bar{\phi}_k\equiv\bar{\theta}_{k+1}-\bar{\theta}_k$. Using (2.9), (2.10) and averaging as before the equations for ϕ_k', we get

Proposition 2.1. *For example* (2.8),

$$
\begin{aligned}
H(\bar{\phi}_k)={}&\left[(d_1+d_4)\frac{\hat{\omega}'(1)}{4}+\frac{(d_3-d_2)}{2}\right]\left[\cos(\bar{\phi}_k)-\gamma\right]\\
&+\left[(d_2-d_3)\frac{\hat{\omega}'(1)}{4}+\frac{(d_1+d_4)}{2}\right]\sin\bar{\phi}_k.
\end{aligned}
\tag{2.11}
$$

Remark. $H(\phi)$ is an odd function only when the coefficient of $\cos(\phi_k)-\gamma$ vanishes. This can happen, for example, if $\hat{\omega}'(1)=0$ and $d_2=d_3$. $\hat{\omega}'(1)=0$ implies that (infinitesimally) there is no frequency dependence on amplitude, while $d_2=d_3$ if the "diffusion" matrix D is symmetric. It is interesting to note that the frequency dependence on amplitude and the nonsymmetry of D may cancel each other to produce a function H which is odd.

We now consider a chain of coupled Van der Pol oscillators in the almost-sinusoidal regime, i.e.,

$$\ddot{X}+\delta(X^2-1)\dot{X}+X=0 \tag{2.12}$$

with $\delta\ll 1$. Using polar coordinates $X=r\cos\theta$, $\dot{X}=-r\sin\theta$, (2.12) is

$$
\begin{aligned}
r_t&=\delta(1-r^2\cos^2\theta)(r\sin^2\theta),\\
\theta_t&=1+\frac{\delta}{r}(1-r^2\cos^2\theta)(r\sin\theta\cos\theta).
\end{aligned}
\tag{2.13}
$$

By averaging techniques [26], it can be seen that (2.13) is equivalent to

$$r_t=\delta r\left[\frac{1}{2}-\frac{r^2}{8}\right]+O(\delta^2),\qquad \theta_t=1+O(\delta^2).$$

Thus, for fixed δ small, (2.12) is equivalent (up to $O(\delta^2)$) to a system of the form (2.8) with the special property that $\hat{\omega}=0$. Allowing detuning and coupling, the full equations have the form:

(2.14)

$$
\begin{aligned}
\ddot{X}_k+\delta(X_k^2-1)\dot{X}_k+(1+\delta\omega_k)X_k={}&\varepsilon\big[b(\dot{X}_{k+1}-2\dot{X}_k+\dot{X}_{k-1})\big]\\
&+c(X_{k+1}-2X_k+X_{k-1})+d(\ddot{X}_{k+1}-2\ddot{X}_k+\ddot{X}_{k-1}).
\end{aligned}
$$

The terms involving the b, c, d represent, respectively, resistive, inductive and capacitive coupling. As before, $\omega_{k+1}-\omega_k=O(\varepsilon)$. Then $H(\phi)$ can be computed as above, and we get

$$H(\phi)=b\sin\phi+(c-d)[\cos\phi-1]. \tag{2.15}$$

Note that, since $\hat{\omega}=0$, all the terms of $H(\phi)$ come from the coupling, and not from the frequency dependence on amplitude.

Remark. In [4], Rand and Holmes compute H for a pair of coupled Van der Pol oscillators, for δ fixed and small. Their formulation is somewhat different, but in terms of our notation, they allow εb and $\varepsilon(c-d)$ to go to zero at different rates as $\varepsilon\to 0$. If the coupling involves substantial resistance, i.e., if $\varepsilon(c-d)\to 0$ at least as fast as εb (as $\varepsilon\to 0$), then for small ε, their result agrees with ours; i.e., to lowest order, H is a multiple of $\sin\phi$. (When $n=2$, the even part of $H(\phi)$ disappears from (1.4), so (2.15) is effectively $b\sin\phi$.) However, if the resistive coupling is significantly smaller than the combined effect of inductive and capacitive coupling, then a more complicated expression may be obtained which is equivalent to the result of carrying out the computation of $H(\phi)$ to order ε^2, with $(c-d)=O(1)$ and $b=O(\varepsilon)$. We note that the same expression is obtained if one works with oscillators of the form (2.8), since (2.12) has been approximated by such an oscillator.

3. Existence of a large amplitude invariant circle. We now restrict ourselves to functions $H(\phi)$ which are odd, i.e. $H(-\phi)=-H(\phi)$, and consider (1.5). Since H is 2π-periodic as well as odd, we have $H(0)=H(\pi)=0$. We shall assume about H that it is qualitatively like $H=\sin\phi$, i.e., that $H>0$ for $0<\phi<\pi$, $H<0$ for $-\pi<\phi<0$, that H has a single maximum M and a single minimum m at ϕ_M and ϕ_m respectively, that H' is monotone increasing from ϕ_m to 0, and that H' is convex on $(\phi_m, 0)$, i.e. that $H'''(\phi)\neq 0$ for $\phi\in(\phi_m, 0)$.

LEMMA 3.1. *For fixed Δ, there exists β_0 such that, for $\beta<\beta_0$, (1.5) has 2^n critical points. Of these critical points, one is a sink and n are saddle points having one positive eigenvalue and $n-1$ eigenvalues with negative real part.*

Proof. The critical points of (1.5) are solutions to

$$H(\phi)=K^{-1}(-\beta\Delta). \tag{3.1}$$

Equation (3.1) has a solution if every component of $K^{-1}(-\beta\Delta)$ lies between m and M. Let

$$\beta_0=\max\{\beta | m\leq K^{-1}(-\beta\Delta)_i\leq M\ \forall i\}.$$

If $\beta<\beta_0$, then for each i there are two distinct solutions $\phi_i^{\pm}(\beta)$ to $H(\phi)=K^{-1}(-\beta\Delta)_i$ with $|\phi_i^{\pm}|<\pi$; ϕ_i^- denotes the solution with smaller absolute value. (Note that $H'(\phi_i^-)>0$ and $H'(\phi_i^+)<0$. See Fig. 3.1.) Thus there are 2^n critical points. Let $\xi_i=\xi_i(\beta)$, $i=1,\cdots,n$, denote the critical point whose kth component $\xi_{ik}(\beta)$ is ϕ_k^-, $k\neq i$ and $\xi_{ii}=\phi_i^+$; $\xi_0(\beta)$ is the critical point with kth component $\xi_{0k}(\beta)=\phi_k^-$ for all k. We will show that ξ_0 is a sink, and ξ_i is a saddle having exactly one eigenvalue with positive real part.

The linearization of (1.5) around one of the critical points ξ has matrix $K\mathbf{H}'(\xi)$ where $\mathbf{H}'(\xi_i)$ denotes the $n\times n$ diagonal matrix whose kth entry is $H'(\xi_{ik})$. Now if $\xi=\xi_0$, then the kth entry is $H'(\xi_{0k})=H'(\phi_k^-)>0$ for all k. By Proposition A.1 (see Appendix), the eigenvalues of $K\mathbf{H}'(\xi_0)$ all have negative real parts, so ξ_0 is a sink. If $\xi=\xi_i$ for some i, then $H'(\xi_{ii})<0$, but $H'(\xi_{ik})>0$ for $k\neq i$. Thus, by Proposition A.3 ξ_i

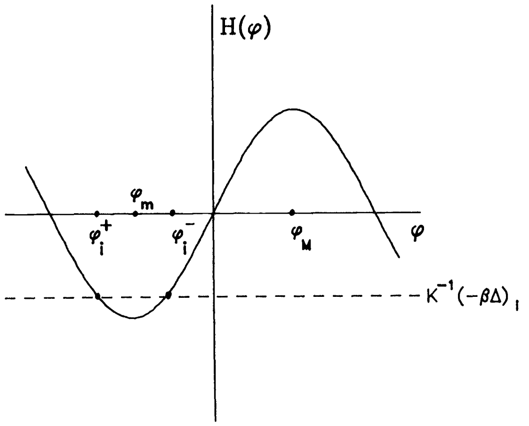


FIG. 3.1. *The two possible choices* $\phi_i^{\pm}$ *for the ith component of a critical point of* (1.5).

is a saddle having exactly one eigenvalue with positive real part. (Note that these stability properties of the critical points cannot change as β increases unless $H'(\xi_{ik}(\beta))$ changes sign for some i; this does not happen for $\beta<\beta_0$.) □

We now further restrict our attention to a linear gradient in frequency; such a gradient is equivalent to a constant vector $\mathbf{\Delta}$ for (1.5). The vector $-\beta(1,1,\cdots,1)^t$ corresponds to a linear *decrease* in frequency for increasing k, as in the measurements on mammalian intestine. For simplicity, we assume n is odd, so there is a unique "middle" phase difference ϕ_j. The main result is as follows. We shall later show that the theorem implies the existence of a pair of frequency plateaus, with a jump in frequency between the j and $(j+1)$st oscillators.

THEOREM 3.1. *Suppose that* $\mathbf{\Delta}=-(1,1,\cdots,1)^t$ *and that* $n=2j-1$ *in* (1.5). *Then for* $\beta\leq\beta_0$, $\beta_0-\beta$ *sufficiently small, the closure of the two branches of the unstable manifold of* ξ_j *forms a smooth attracting invariant circle which is homotopic to the circle* $\phi_k=0$, $k\neq j$, $0\leq\phi_j\leq 2\pi$. *This invariant manifold persists for* $\beta>\beta_0$, $\beta-\beta_0$ *sufficiently small.* (*See Fig.* 3.2.)

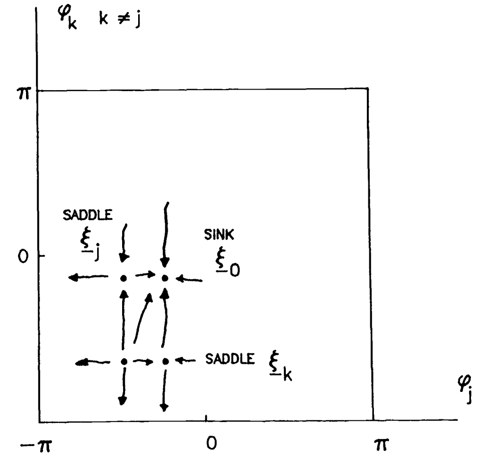


FIG. 3.2. *Schematic representation of the dynamics of* (1.5), *with a unique sink* ξ_0 *and a saddle* ξ_j *which coalesces with* ξ_0 *as* $\beta\to\beta_0$. *The two* (*one-dimensional*) *branches of the unstable manifold of* ξ_j *form a smooth invariant circle.*

Proof. We require several lemmas:

LEMMA 3.2. *Assume the hypotheses of Theorem* 3.1. *Then*

(i) $\phi_k^{\pm}(\beta)<0\ \forall k$ *and all* $\beta\le\beta_0$.

(ii) $m<K^{-1}(-\beta\Delta)_k<0\ \forall k\neq j,\ \beta\le\beta_0$. *For* $k=j$, $m=K^{-1}(-\beta_0\Delta)_j$.

(iii) *The eigenvector* $\mathbf{v}_j$ *of the unique positive eigenvalue of* ξ_j *satisfies* $\operatorname{sgn} v_{jk}=-\operatorname{sgn} v_{jj}\ \forall k\neq j$ *and all* $\beta\le\beta_0$.

Proof. (i) The critical points are solutions to (3.1) and $K^{-1}(\Delta)$ has kth component $k(n+1-k)/2$. Thus all the components of $K^{-1}(-\beta\Delta)$ are negative. Since $H(\phi)>0$ for $0<\phi<\pi$ and $H(\phi)<0$ for $-\pi<\phi<0$, the solutions to $H(\phi)=-\beta k(n+1-k)/2$, with $|\phi|<\pi$, are negative.

(ii) If $n=2j-1$, then $k(n+1-k)/2$ takes its largest value for $k=j$. (β_0 is then defined by $m=-\beta_0 j(j+1)/2$.)

(iii) This follows from Proposition A.5 (see Appendix), as soon as we establish that $\mathbf{H}'(\xi_j)$ has the form $\operatorname{diag}(a_1,a_2,\cdots,a_{j-1},a_j,a_{j-1},\cdots,a_1)$, where $a_k>0$ for $k\neq j$, $a_j<0$, $a_1>a_2>\cdots>a_{j-1}$, and $a_{k-1}+a_{k+1}<2a_k$ for $k=2,\cdots,j-1$. Now $\mathbf{H}'(\xi_j)$ is a diagonal matrix whose kth entry is $H'(\xi_{jk})$, where ξ_{jk}, the kth component of ξ_j, is $\phi_k^-(\beta)$ for $k\neq j$ and $\phi_k^+(\beta)$ for $k=j$. ($\phi_k^{\pm}(\beta)$ are defined by $H(\phi_k^{\pm})=K^{-1}(-\beta\Delta)_k=-\beta k(n+1-k)/2$.) Thus $a_k=a_{n+1-k}$. The signs of the a_k follow from the definition of ξ_j. Furthermore, $k(n+1-k)/2$ is an increasing function of k for $k<j$, so $|\phi_k^-(\beta)|$ increases with k (i.e, $\phi_k^-=-|\phi_k^-|$ decreases with k). Since H' is monotone increasing on $[\phi_m,0]$, this implies that $a_1>a_2>\cdots>a_{j-1}$. Also, the convexity condition for H' (i.e., $H'''\neq0$ on $(\phi_m,0)$) implies that $(a_{k-1}+a_{k+1})/2<a_k$. □

From Lemma 3.2(ii), we see that $|\phi_j^+-\phi_1^-|\to0$ as $\beta\to\beta_0$. Thus, as $\beta\to\beta_0$, all critical points coalesce in pairs, and for $\beta>\beta_0$ there are no solutions to (3.1). (Recall that each of the 2^n critical points has as its kth component either ϕ_k^+ or ϕ_k^-; thus each point is matched with another point with which it agrees except at the jth component.) The critical point ξ_j has the distinction of being the one that coalesces with the sink ξ_0; its components agree with those of ξ_0 except for the jth, with $\xi_{0j}=\phi_j^-$ and $\xi_{jj}=\phi_j^+$.

We shall focus separately on the two branches of the unstable manifold of ξ_j, which we shall refer to as the left or right branch, depending on whether the jth component v_{jj} of the tangent vector $\mathbf{v}_j$ is negative or positive. We shall show, for $\beta_0-\beta$ sufficiently small, that both of these have the sink in their closure, and hence form an invariant circle. The next lemma deals with the right branch. This is the easier part, since for $\beta_0-\beta$ small, ξ_j and ξ_0 are close, with ξ_0 to the right of ξ_j.

LEMMA 3.3. *For* $\beta_0-\beta$ *sufficiently small, the right branch of the unstable manifold of* ξ_j *contains* ξ_0 *in its closure. Furthermore, at* ξ_0 *this manifold is tangent to the eigenspace of the least negative eigenvalue of* $K\mathbf{H}'(\xi_0)$.

Proof. ξ_0 and ξ_j coalesce as $\beta\to\beta_0$. The techniques of [27] show that, under certain hypotheses, this implies that for $\beta_0-\beta$ sufficiently small, there is a trajectory joining ξ_0 and ξ_j. The unstable manifold of ξ_j is one-dimensional, so that trajectory must be the unstable manifold of ξ_j. It follows from the construction of this trajectory that its tangent at $\xi_0(\beta)$ is the eigenvector of the unique eigenvalue of $K\mathbf{H}'(\xi_0)$ which tends to 0 as $\beta\to\beta_0$.

The hypotheses on (1.6) needed to apply the technique of [27] are those of [27, Thm. 2.2]: we write (1.5) as

$$\dot\phi=K\mathbf{H}'(\xi_{\beta_0})(\phi-\xi_{\beta_0})+(\beta-\beta_0)\Delta+Q(\phi-\xi_{\beta_0},\phi-\xi_{\beta_0})+\rho \tag{3.2}$$

where $\xi_{\beta_0}=\xi_0(\beta_0)=\xi_j(\beta_0)$ is the saddle-sink at the critical value of β, Q is a vector-valued quadratic form containing the terms quadratic in $\phi-\xi_{\beta_0}$ and independent of β, and $\rho=o(\beta-\beta_0,|\phi-\xi_{\beta_0}|^2)$. Then we must have

(i) $K\mathbf{H}'(\xi_{\beta_0})$ has rank $n-1$.

(ii) $[K\mathbf{H}'(\xi_{\beta_0}), \Delta]$ has rank n, where $[P, Z]$ denotes the $n\times(n+1)$-matrix formed by adjoining the n-vector Z to the $n\times n$-matrix P as the last column.

(iii) $[K\mathbf{H}'(\xi_{\beta_0}), Q(V,V)]$ has rank n, where V is an eigenvector of the zero eigenvalue of $K\mathbf{H}'(\xi_{\beta_0})$ and $Q(V,V)$ is the n-vector obtained by evaluating the quadratic form Q on the vector V.

Now (i) and (ii) follow from (i) and (ii) of Proposition A.2. To establish (iii), we note that $V=(v_1,\cdots,v_n)^t$ with $v_k=0$, $k\neq j$, and $v_j=1$. Hence $Q(V,V)$ contains exactly those terms depending only on ϕ_j (and not $\phi_k, k\neq j$). In particular, there are no such terms in the kth equation of (3.2) with $k\neq j, j\pm 1$. For $k=j\pm 1$, the kth coordinate $Q(V,V)_k$ of $Q(V,V)$ is $\frac{1}{2}H''(\phi_j^-(\beta_0))(\phi_j-\phi_j^-(\beta_0))^2$;

$$Q(V,V)_j=-H''\big(\phi_j^-(\beta_0)\big)\big(\phi_j-\phi_j^-(\beta_0)\big)^2.$$

Thus $Q(V,V)$ is a multiple of $Z=(z_1,\cdots,z_n)^t$ with $z_j=-2$, $z_{j-1}=z_{j+1}=1$, $z_k=0$, $k\neq j, j\pm 1$. Then (iii) also follows from (ii) of Proposition A.2. □

We now turn to the left branch of the unstable manifold of ξ_j. For β near β_0, the ϕ_j component must change by nearly 2π before entering the sink ξ_0. Thus we shall need estimates on this branch that are not local. These estimates are contained in the following: Let $\sigma_k=H(\phi_k^{\pm})-H(\phi_{k+1}^{\pm})$.

LEMMA 3.4. *Let* $R=\{(\phi_1,\cdots,\phi_n)|\phi_{n+1-k}=\phi_k\ \forall k;\ \phi_k^-\leq\phi_k<\phi_M,\ k\neq j,\ H(\phi_j)>H(\phi_j^-),\ \dot{\phi}_j<0,\ H(\phi_k)\leq H(\phi_{k+1})+\sigma_k,\ k\leq j-1\}$. *Then the left branch is contained in* R. *All trajectories which start in* R *tend to the critical point* ξ_0 *as* $t\to\infty$.

Proof. We shall show that (i) the above statement is true for a neighborhood of ξ_j (i.e.,$\xi_j\in\bar{R}$, the closure of R, and the left branch points into R), and (ii) R is invariant under (1.5) for $t>0$, with all trajectories tending towared ξ_0. We note that if $\Delta=-(1,1,\cdots,1)^t$ the invariance of (1.5) under $\phi_k\leftrightarrow\phi_{n+1-k}$ implies that the points on the one-dimensional unstable manifold of ξ_j satisfy $\phi_{n+1-k}=\phi_k$ for all k. Furthermore, on the (initial piece of the) left branch, $\dot{\phi}_j<0$ by hypothesis, and $\dot{\phi}_k>0$ $k\neq j$ by Lemma 3.2. Since $\phi_k=\phi_k^-(k\neq j)$ at the critical point, we have $\phi_k^-<\phi_k<\phi_M$ $(k\neq j)$; also $\dot{\phi}_j<0$ implies $H(\phi_j)>H(\phi_j^-)=H(\phi_j^+)$. (See Fig. 3.3.)

To finish (i), we have left to show that

$$H(\phi_k)\leq H(\phi_{k+1})+\sigma_k,\qquad k=1,\cdots,j-1,\tag{3.3}$$

along the left branch of the unstable manifold, and in a neighborhood of ξ_j. By definition, $H(\phi_k)=H(\phi_{k+1})+\sigma_k$ at the critical point. Thus, it suffices to show that

$$H'(\phi_k)\dot{\phi}_k\leq H'(\phi_{k+1})\dot{\phi}_{k+1},\qquad k=1,\cdots,j-1,$$

along this part of the unstable manifold. Equivalently, we may show that

$$a_k v_{jk}\leq a_{k+1}v_{j,k+1},\qquad k=1,\cdots,j-1,\tag{3.4}$$

where $a_k=H'(\phi_k^-)$, $k\neq j$, $a_j=H'(\phi_j^+)$, and $(v_{j,1},\cdots,v_{j,n})=\mathrm{v}_j$ is the eigenvector of the eigenvalue $\lambda>0$ of $K\mathbf{H}'(\xi_j)$. But the $\{a_k\}$ and $\{v_{jk}\}$ then satisfy the hypotheses of Proposition A.5, so (3.4) holds.

We now go to (ii). The relationship $\phi_{n+1-k}=\phi_k$ for all k is invariant under (1.5) if $\Delta=-(1,1,\cdots,1)^t$, so we shall assume it. We first show that a trajectory cannot leave R through the boundaries $\phi_k=\phi_k^-$ or $\phi_k=\phi_M$, $k\neq j$. The vector field (1.5) does not cross $\phi_k=\phi_M$ for any k. For at $\phi_k=\phi_M$,

$$\dot{\phi}_k=-\beta+H(\phi_{k-1})-2M+H(\phi_{k+1});$$

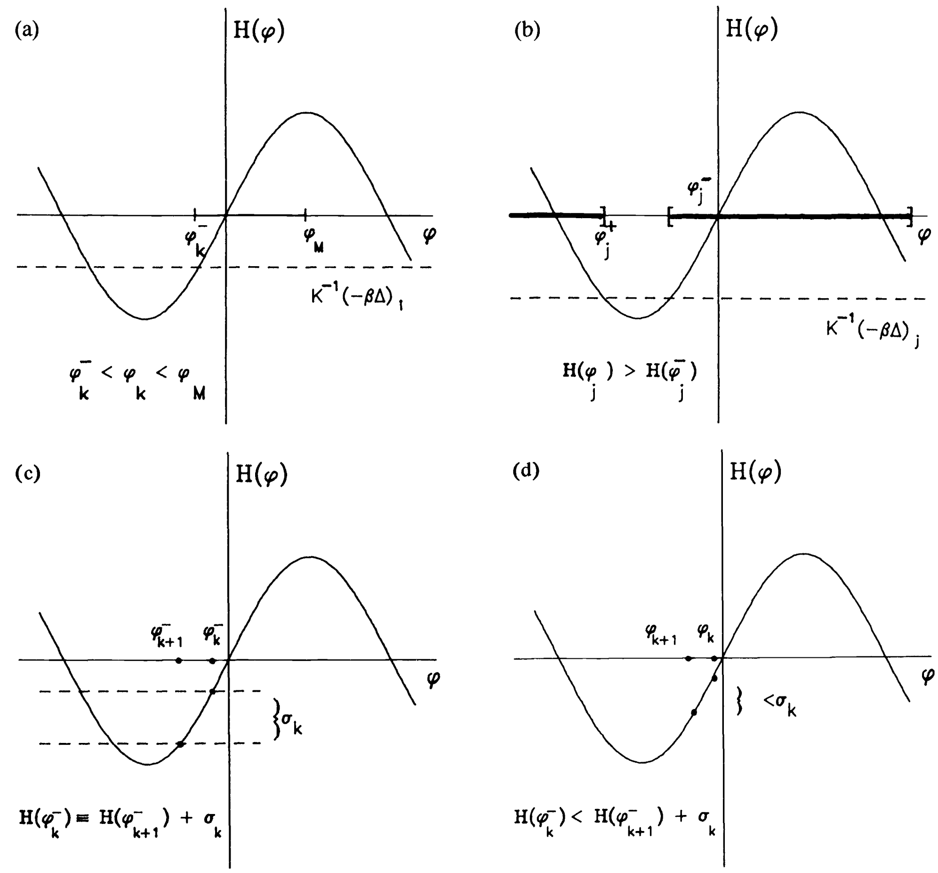


FIG. 3.3. *Some of the constraints on the* $\{\phi_k\}$ *in order that* $\phi\in R$: (a) $\phi_k^-<\phi_k<\phi_M$; (b) $H(\phi_j)>H(\phi_j^-)$; (d) $H(\phi_k)<H(\phi_{k+1})+\sigma_k$, *where* σ_k *is defined as in* (c).

since $H(\phi_{k\pm1})\le M$, $\dot\phi_k<0$. The surface $\phi_k=\phi_k^-$ *can* be crossed by the vector field, but not inside R. For

$$\dot\phi_k=-\beta+H(\phi_{k-1})-2H(\phi_k)+H(\phi_{k+1}). \tag{3.5}$$

The right-hand side of (3.5) vanishes at all critical points. If $\phi_{k\pm1}^-<\phi_{k\pm1}<\phi_M$, we have $H(\phi_{k\pm1})>H(\phi_{k\pm1}^-)$; so if we also have $\phi_k=\phi_k^-$, then $\dot\phi_k>0$ and the vector field points into R. Note that this argument works even if $k=j\pm1$, because all that is needed is that $H(\phi_{k\pm1})\ge H(\phi_{k\pm1}^{\pm})$. Now on the surface $\phi_k=\phi_k^-$, we may have $\dot\phi_k=0$ at some time t_0, i.e., if $\phi_{k\pm1}=\phi_{k\pm1}^-$ (ϕ_j^+ if $k\pm1=j$). But

$$\dot\phi_{k+1}=-\beta+H(\phi_k)-2H(\phi_{k+1})+H(\phi_{k+2}) \tag{3.6}$$

and $H(\phi_{k+2})\ge H(\phi_{k+2}^-)$. Hence, if $\phi_k=\phi_k^-$ and $\phi_{k+1}=\phi_{k+1}^-$ (ϕ_j^+ if $k+1=j$), we have $\dot\phi_{k+1}>0$ (so $H(\phi_{k+1})>H(\phi_{k+1}^-)$ for $t>t_0, t-t_0$ sufficiently small) unless $H(\phi_{k+2})=H(\phi_{k+2}^-)$. Following this argument, we conclude that unless $\phi_k=\phi_k^-$ for all $k\ne j$ and $\phi_j=\phi_j^+$, even if $\dot\phi_k=0$ for some time t, we will have $\phi_k^-\le\phi_k$ for succeeding times.

We next show that trajectories may not exit through surfaces of the form

$$H(\phi_k)=H(\phi_{k+1})+\sigma_k, \qquad k=1,\cdots,j-1. \tag{3.7}$$

For suppose that (3.7) holds for some k at some t_0. Then (3.5), (3.6) become

(3.8a) $$\dot{\phi}_k = -\beta + H(\phi_{k-1}) - H(\phi_k) - \sigma_k,$$

(3.8b) $$\dot{\phi}_{k+1} = -\beta + \sigma_k - H(\phi_{k+1}) + H(\phi_{k+2}).$$

We now use the inequalities (3.3) for $k \pm 1$. These imply that

(3.9a) $$\dot{\phi}_k \leq -\beta - \sigma_k + \sigma_{k-1}, \qquad k = 1, \cdots, j-1,$$

(3.9b) $$\dot{\phi}_{k+1} \geq -\beta + \sigma_k - \sigma_{k+1}, \qquad k = 1, \cdots, j-2.$$

But (3.7) and hence (3.8) hold at the critical point ξ_j, where $\dot{\phi}_k = \dot{\phi}_{k+1} = 0$. Inserting the components of ξ_j into (3.8), we get that

$$-\beta - \sigma_k + \sigma_{k-1} = 0 = -\beta + \sigma_k - \sigma_{k+1}.$$

Thus (3.9)

$$\dot{\phi}_k \leq 0, \; \dot{\phi}_{k+1} \geq 0, \qquad k = 1, \cdots, j-2.$$

This implies that, even if (3.7) holds for some t, the trajectory does not exit R through the surface (3.7) with $k = 1, \cdots, j-2$. For $k = j-1$, the deductions from (3.8a), (3.9a) are still valid. We replace (3.8b), (3.9b) by

(3.10) $$\begin{aligned} \dot{\phi}_j &= -\beta + H(\phi_{j-1}) - 2H(\phi_j) + H(\phi_{j-1}) \\ &= -\beta + 2\sigma_{k-1}. \end{aligned}$$

As before, by inserting $\phi = \xi_j$ into (3.10) we see that the right-hand side of (3.10) is zero, i.e., $\dot{\phi}_j = 0$. Since $\dot{\phi}_{j-1} \leq 0$, we have that the trajectory does not exit R through the surface (3.7) with $k = j-1$.

Trajectories may also not exit through the surface

(3.11) $$0 = -\beta + H(\phi_{j-1}) - 2H(\phi_j) + H(\phi_{j+1})$$

along which $\dot{\phi}_j = 0$. For we have just seen that (3.11) is equivalent to (3.7) for $k = j-1$, and that a trajectory may not exit through this surface.

Finally, trajectories may not exit through the surface $\phi_j = \phi_j^+$ or $\phi_j = \phi_j^-$, the boundaries of $H(\phi_j) > H(\phi_j^\pm)$. At ξ_j, $\phi_j = \phi_j^+$; since $\dot{\phi}_j < 0$, ϕ_j must decrease monotonely, and so cannot pass through $\phi = \phi_j^+$. Also, ϕ_j cannot decrease past $\phi_j = \phi_j^- - 2\pi$. For at $\phi_j = \phi_j^-$ (mod 2π),

(3.12) $$\dot{\phi}_j = -\beta + 2H(\phi_{j-1}) - 2H(\phi_j^-).$$

Since the right-hand side of (3.12) vanishes at the critical point, and $H(\phi_{j-1}) \geq H(\phi_{j-1}^-)$, the right-hand side of (3.12) is ≥ 0 at $\phi_j = \phi_j^-$ (mod 2π). But $\dot{\phi}_j \leq 0$, so trajectories cannot reach $\phi_j = \phi_j^-$ (mod 2π) unless $\phi_{j-1} = \phi_{j-1}^-$. Furthermore, by (3.5) with $k = j-1$, $\phi_j = \phi_j^-$ (mod 2π), and $\phi_{j-1} = \phi_{j-1}^-$, we have $\dot{\phi}_{j-1} > 0$ unless $\phi_{j-2} = \phi_{j-2}^-$; for later times, this implies that $\dot{\phi}_j > 0$, and so contradicts $\dot{\phi}_j \leq 0$. Hence $\phi_{j-2} = \phi_{j-2}^-$. A similar argument shows that if $\phi_j = \phi_j^-$, then $\phi_k = \phi_k^-$ for all k. Thus trajectories of R do not pass through $\phi_j = \phi_j^- - 2\pi$, but rather tend to ξ_0 as $t \to \infty$. □

Lemmas 3.3 and 3.4 together show that the two branches of the unstable manifold of ξ_j form an invariant circle. We next show that the circle is smooth. Since (1.5) is C^∞, so is the unstable manifold [23]; thus, smoothness need only be proved at ξ_0 where the branches join. We know from Lemma 3.3. that the right branch approaches ξ_0 tangent to the (left branch of the) eigenspace of the eigenvalue λ_0 which is closest to zero. In such a circumstance, the degree of contact of the trajectory with the eigenspace is bounded below by the ratio λ_1/λ_0, where λ_1 is the next smallest (in absolute value) eigenvalue of (1.5) at ξ_0; this ratio goes to ∞ as $\beta \to \beta_0$. Thus, to prove that the invariant circle is smooth at ξ_0 (with arbitrary smoothness for $\beta_0 - \beta$ sufficiently small), it suffices to prove

LEMMA 3.5. *The left branch of the unstable manifold of ξ_j enters ξ_0 tangent to the* (*right branch of the*) *eigenspace of the eigenvalue* λ_0.

Proof. Generically, trajectories approaching a sink do approach tangent to the eigenvector of the least negative eigenvalue. The exceptional trajectories approach tangent to the span of the remaining eigenspaces. We shall show that trajectories of R are not exceptional.

By Lemma 3.4, trajectories in R satisfy $\phi_k > \phi_k^-$ $(k \neq j)$, where ϕ_k^- is the kth coordinate of ξ_0; hence, as a trajectory in R approaches ξ_0, we have $\dot{\phi}_k < 0$ for all k. Now ξ_0 is a hyperbolic critical point, so trajectories near it behave like those of the linearization of (1.5) around ξ_0. Since trajectories in the eigenspace of a pair of complex eigenvalues oscillate around the critical point, and we have $\dot{\phi}_k < 0$ for all k, the trajectories in question must in fact approach ξ_0 tangent to the span of the eigenspaces of the remaining (real) eigenvalues. Because the real eigenvalues are ordered, trajectories of the linear system approach tangent to exactly one eigenspace, and, furthermore, to an eigenvector within that eigenspace. Thus, to rule out that trajectories of R are exceptional, it suffices to show, for any real eigenvalue $\lambda \neq \lambda_0$ and associated eigenvector $Z = (z_1, \cdots, z_n)$, that the z_i's cannot all have the same sign. ($\mathrm{Sgn}\, z_k \equiv \mathrm{sgn}\, z_1$ for all k is necessary if we are to have $\dot{\phi}_k < 0$ for all k.) But the linearization of (1.5) around ξ_0 has the form KA where $A = \mathrm{diag}(a_1, a_2, \cdots, a_n)$ with $a_k > 0$ $(k \neq j)$, $a_{n+1-k} = a_k$. For $\beta = \beta_0$, $a_j \equiv H'(\phi_j(\beta_0)) = 0$, so the result follows from Proposition A.6. For $\beta - \beta_0$ sufficiently small, it follows by continuity. □

To finish Theorem 3.1, it remains to show that the smooth attracting invariant manifold persists for $\beta > \beta_0$, $\beta - \beta_0$ sufficiently small, and that the circle is homotopic to $\phi_k = 0$ for all $k \neq j$. To prove the first assertion, we perturb (1.5) around $\beta = \beta_0$. For the invariant manifold to persist and be C^r, a certain "Lyapunov-type number" must be $< 1/r$ [22]; this number measures the ratio of the asymptotic (exponential) rate of contraction on the manifold to that of the asymptotic rate of approach to the manifold. This number is determined only by the ω-limit set on the invariant manifold, which, for (1.5) $\beta = \beta_0$, is the unique sink-saddle. For this case, the tangential contraction rate tends to zero as $\beta \to \beta_0$ from below, but the normal contraction rate stays bounded away from zero. (Equivalently,only one eigenvalue of the linearization at ξ_0 tends to zero as the sink and saddle coalesce.) Thus, the invariant manifold persists for $\beta > \beta_0$ and can be made arbitrarily smooth by taking $\beta - \beta_0$ small.

To see that the invariant circle is homotopic to the circle $\phi_k = 0$ for all $k \neq j$, we recall that, along the left branch of the unstable manifold of ξ_j, we have $\phi_k^- < \phi_k < \phi_M$, $k \neq j$. Also, the right branch is arbitrarily small for $\beta_0 - \beta$ small. Thus, as ϕ_j changes by 2π along the closure of the two branches, ϕ_k stays in a neighborhood of $\phi_k = 0$ having length less than 2π. It follows that the closure of the trajectories can be deformed into a circle for which $\phi_k = 0$, for all $k \neq j$. □

Remark. The attracting invariant circle of (1.5) (or equivalently (1.4b)) corresponds to an attracting 2-dimensional torus for (1.4), with variables θ_1 and ϕ_j. The dynamics on this torus is an $O(\varepsilon^2)$ perturbation of an uncoupled flow, with $\theta_1(t)$ satisfying $\theta_1' = \omega_1 + \varepsilon H(\phi_1)$ and $\phi_1(t)$, $\phi_j(t)$ the values along the (slow) limit cycle of (1.5), written in the original time variable t.

4. Frequency plateaus. In §3, we proved the existence of an attracting invariant circle for (1.5) on T^n. We now show why this circle corresponds to a pair of frequency plateaus with a break between the j th and $(j+1)$st oscillators. (Recall that $n+1=2j$.)

The "frequency" of an oscillator coupled to others requires a definition; one reasonable definition is

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \theta_k'\,dt \tag{4.1}$$

over some trajectory of (1.3), (1.4), provided that (4.1) converges. Note that this definition yields θ' if θ' is constant, and is, a priori, dependent on the trajectory. To compute (4.1) requires going to the full equations (1.3), (1.4). However, to lowest order, the frequency difference

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \phi_k'\,dt \tag{4.2}$$

can be computed from trajectories of (1.5). For any trajectory in the basin of attraction of the limit cycle of (1.5), (4.2) reduces to

$$\frac{\varepsilon}{T_0}\int_0^{T_0}\dot{\phi}_k\,d\tau$$

where $T_0=T_0(\beta)$ is the period of the limit cycle, and the integration of $\dot{\phi}_k$ is done along the limit cycle. By the fundamental theorem of calculus, (4.2) may be written as

$$\frac{\varepsilon}{T_0}\left[\hat{\phi}_k(T_0)-\hat{\phi}_k(0)\right]$$

where $\hat{\phi}_k(\tau)$ is the covering map of $\phi_k(\tau)$ (i.e., values of $\hat{\phi}_k(\tau)$ are not identified mod 2π). It was shown in §3 that the invariant circle is homotopic to the circle $\phi_k=0$, $k\neq j$, $0\leq\phi_j\leq 2\pi$. Thus $\hat{\phi}_k(T_0)=\phi_k(T_0)$, $k\neq j$. (We may assume that $\hat{\phi}_k(0)=\phi_k(0)$ by choice of covering map.) But $\phi_k(T_0)=\phi_k(0)$ by the periodicity of ϕ along this solution. Thus (4.1) vanishes for $k\neq j$, i.e., for $1\leq k\leq j$, the frequency of the kth oscillator is independent of k; similarly, this is true for $j+1\leq k\leq n$. However, for $k=j$, we have $\hat{\phi}_j(T_0)=\hat{\phi}_j(0)+2\pi$. This implies that the jump in frequency between the $(j+1)$st$+j$th oscillators is $2\pi\varepsilon/T_0$ (in ordinary time).

On each of the two plateaus, the phase differences ϕ_k are periodic in time rather than constant in time. That is, the oscillators remain phase-locked "on the average" rather than at every instant; some authors refer to this phenomenon as "phase-trapping" [28]. Furthermore, the frequency on each of the "plateaus" is not exactly constant, for (1.5) are valid only up to $O(\varepsilon)$ (in the scaled time, or $O(\varepsilon^2)$ in the original time scale). For $H=\sin\phi$, plateaus emerge when $\beta\geq\beta_0=2/j^2=8/(n+1)^2$. This implies that the total change in frequency from oscillators 1 to $n+1$, for $\beta\approx\beta_0$, is $n\varepsilon\beta_0=8\varepsilon n/(n+1)^2=O(\varepsilon,\frac{1}{n})$. Thus, for a fixed total change in frequency, the larger the n, the harder it is to phase-lock. This contrasts with a nonodd function H, e.g. $H=\sin\phi+\delta[\cos\phi-1]$ for which the total change in frequency just prior to loss of phase-locking is $O(\varepsilon)$, but does not go to zero as $n\to\infty$ [21].

To understand how the size of the frequency jump varies as β increases, we first note that $T_0\to\infty$ as $\beta\to\beta_0$. Furthermore, we claim that T_0 varies like $1/\sqrt{\beta-\beta_0}$. For consider the phase-locked solution $\phi(\tau)$ to (1.5), $\beta=\beta_0$, and choose a small interval I around the unique critical point. For $\beta-\beta_0$ sufficiently small, the large interval S^1-I is traversed in a finite amount of time (bounded above independent of β). Within I, a

β-dependent coordinate ψ may be chosen so that the equation takes the form $\dot{\psi}=\psi^2+\nu(\beta)$ where $\nu(\beta_0)=0$, $\nu'(\beta_0)>0$. If $a,b>0$, the time it takes ψ to go from $-a$ to $+b$ is

$$\frac{1}{\nu}\left[\tan^{-1}\left(\frac{\psi}{\nu}\right)\right]_{\psi=-a}^{\psi=b}=O\left(\frac{1}{\sqrt{\beta-\beta_0}}\right).$$

Since the time it takes to traverse I dominates the finite time to cross S^1-I, we see that $T_0=O(1/\sqrt{\beta-\beta_0})$ as $\beta\to\beta_0^+$.

The above computation shows that, as $\beta\to\beta_0^+$, the period T_0 passes continuously to $+\infty$ from a finite number. Thus the jump $2\pi/T_0$ in frequency between the two plateaus changes continuously as β is varied and tends to zero as $\beta\to\beta_0^+$. In particular, there need be no rational relationship between the frequencies of the two plateaus. The calculation also suggests at first glance that the frequency jump is never piecewise constant as β is changed (for $\beta-\beta_0$ small). However, this last conclusion is suspect: as mentioned above, the calculations are accurate only up to $O(\varepsilon^2)$ (in the original time scale). For fixed ε small and $\beta\to\beta_0^+$, the effects of the nonzero ε could lead to piecewise constancy of the frequencies over some (small) intervals in β.

In Fig. 4.1 we show numerical calculations of equations (1.5) for β near β_0 and a larger value of β, i.e., a steeper gradient in natural frequency. Note that more plateaus emerge. We conjecture that when there are $k+1$ plateaus, there is a k-dimensional subtorus T^k of T^n corresponding to k degrees of freedom at the jumps. It is less clear how to analytically define the frequencies on these plateaus.

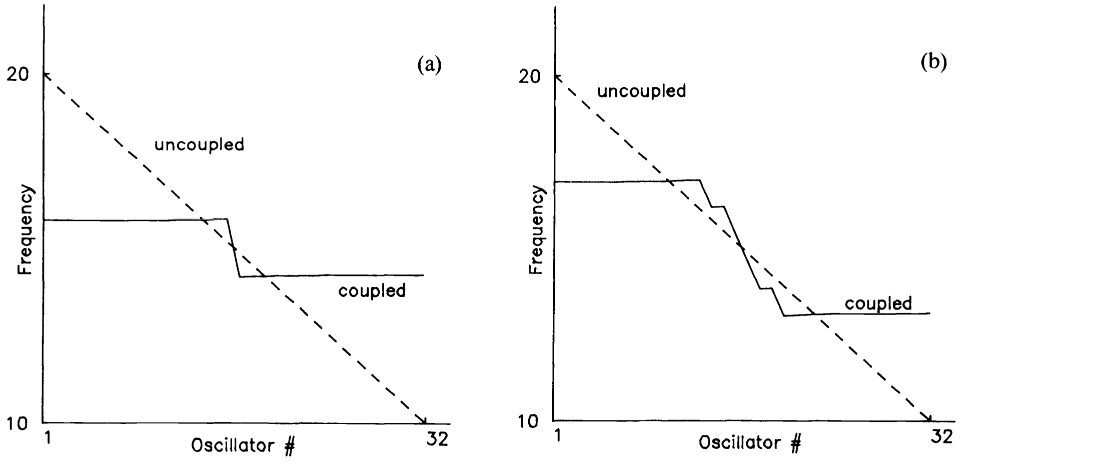


FIG. 4.1. *Frequency vs. k for* $\dot{\phi}_k=-10/31+\delta[\sin\phi_{k+1}-2\sin\phi_k+\sin\phi_{k-1}]$ *for* (a) $\delta=32$, (b) $\delta=18$. *Note that decreasing* δ *and leaving the frequency difference* 10/31 *the same is equivalent* (*under a change of time scale*) *to increasing the frequency difference.*

5. Nonuniform or nonisotropic coupling. In this section we consider some of the effects of relaxing the hypotheses that the coupling be isotropic and uniform; we still assume that only nearest neighbors are coupled, and that the coupling is weak.

5.1. Nonisotropic coupling. In the previous sections, we assumed that adjacent oscillators have symmetric influences on one another. Suppose instead that the forward coupling has a constant (independent of k) ratio α to the backward coupling (see Fig. 5.1.a). The equations for the ϕ_k (with H odd as before) then have the form

$$
\begin{aligned}
(5.1)\qquad & \dot{\phi}_1 = -\beta + H(\phi_2) - (\alpha+1)H(\phi_1), \\
& \dot{\phi}_k = -\beta + H(\phi_{k+1}) - (\alpha+1)H(\phi_k) + \alpha H(\phi_{k-1}), \qquad k=2,\cdots,n-1, \\
& \dot{\phi}_n = -\beta - (\alpha+1)H(\phi_n) + \alpha H(\phi_{n-1}).
\end{aligned}
$$

These equations reduce to (1.5) when $\alpha=1$. Note that $\alpha>1$ implies that forward coupling is stronger and $\alpha<1$ means backward coupling is stronger.

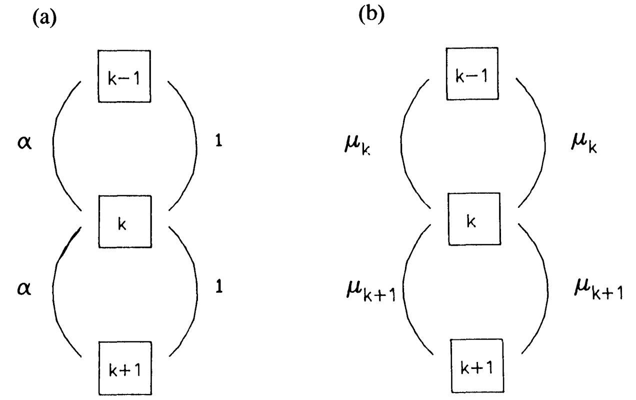


FIG. 5.1. (a) *Nonisotropic coupling. The forward coupling has a constant ratio* α *to the backward coupling.* (b) *Nonuniform coupling. There is a gradient in coupling strength, e.g. the diffusion coefficient associated with pair of cells varies with* k.

PROPOSITION 5.1. *Let* $y_k = H(\phi_k)$. *Then the critical point of* (5.1) *satisfies*

$$
(5.2)\qquad y_k = -\frac{\{n+1-k+k\alpha^{n+1}-(n+1)\alpha^k\}}{(1-\alpha)(1-\alpha^{n+1})}\beta.
$$

Proof. Insert (5.2) in (5.1) and check. □

Once we know the phase-locked solution of (5.1), we may compute the frequency of entrainment from the first equation of (1.4): when there is phase-locking, the frequency is $\dot{\theta}_k$ for any k, and

$$
\begin{aligned}
(5.3)\qquad \dot{\theta}_1 &= \omega_1 + \varepsilon H(\phi_1) + O(\varepsilon^2) \\
&= \omega_1 + \varepsilon\frac{[n(\alpha-1)+\alpha(1-\alpha^n)]}{(1-\alpha)(1-\alpha^{n+1})}\beta + O(\varepsilon^2).
\end{aligned}
$$

(For $\alpha=1$, (5.3) reduces to $\dot{\theta}_1 = \omega_1 - \varepsilon n\beta/2$, the average of the frequencies. (5.2) reduces to $y_k = -\beta k(n+1-k)/2$.)

Figures 5.2 and 5.3 graph y_k vs. k for several α, and $\dot{\theta}_1$ vs. α when $n=9$ and $n=29$. We see from the formulas and the pictures that one effect of e.g. increasing α is to skew the peak of the graph of y_k vs. k toward the higher k (lower frequency) end. thus, when β is large enough that phase-locking is no longer possible, we would expect a break in frequency to occur at the lower frequency end. Changing α also changes the frequency of the phase-locked solution. For example, if $\alpha>1$, then for n large the frequency is close to, but less than, ω_1. Figure 5.4 shows a pair of graphs of frequency vs. k for frequency gradients sufficient to form plateaus. Note that the plateaus are not symmetric with respect to average frequency.

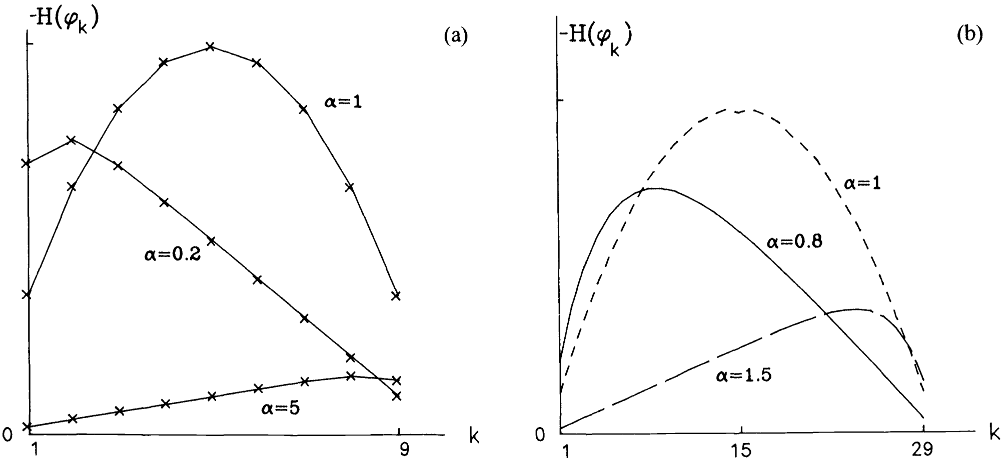


FIG. 5.2. *The graphs of* $-H(\phi_k)$ *vs.* k *for various* α. (a) $n=9$, (b) $n=29$.

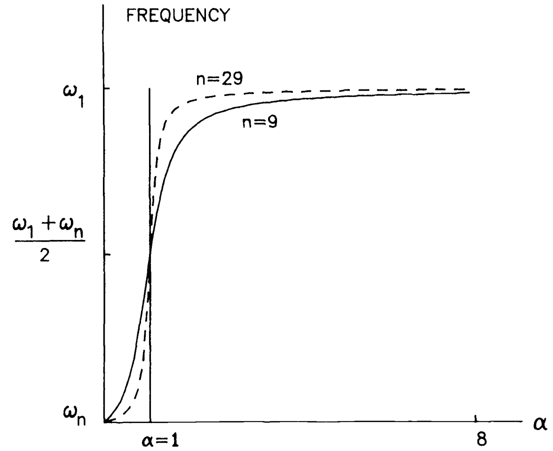


FIG. 5.3. *The frequency of the phase-locked solution as a function of the amount of anisotropy,* $n=9$ *and* $n=29$.

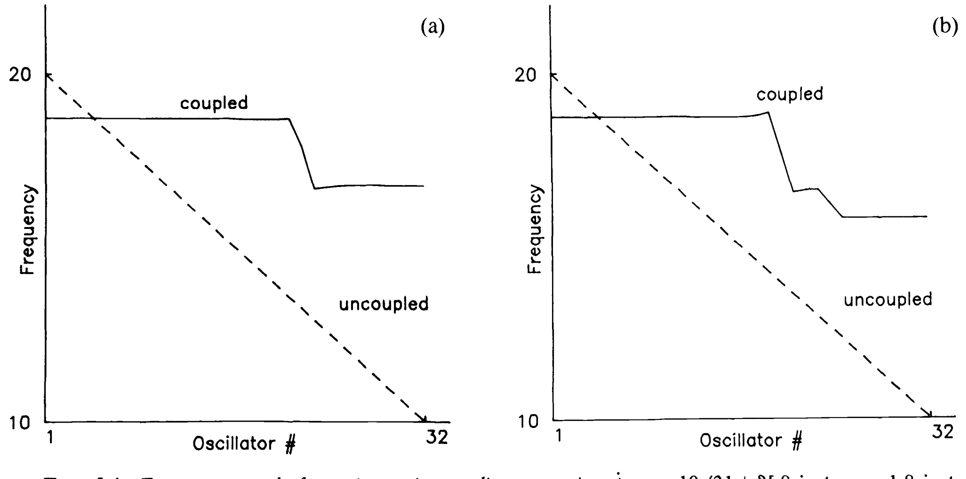


FIG. 5.4. *Frequency vs.* k *for anisotropic coupling, equation* $\dot{\phi}_k = -10/31 + \delta[.8\sin\phi_{k+1} - 1.8\sin\phi_k + \sin\phi_{k-1}]$, *with* a) $\delta=24$ *and* b) $\delta=10$. *The forward coupling is stronger, and the plateaus are shifted upward.*

5.2. Nonuniform coupling. We now assume that the coupling is isotropic, but varies with k. That is, we suppose that the coupling is diffusive, but that a different diffusion coefficient is associated with each pair of oscillators. (See Fig. 5.1.b.) The phase difference equations now take the form

$$\begin{aligned}(5.4)\qquad &\dot{\phi}_k = -\beta + \mu_{k+1}H(\phi_{k+1}) - 2\mu_k H(\phi_k) + \mu_{k-1}H(\phi_{k-1}),\\ &H(-\phi_0) = 0 = H(\phi_{n+1}).\end{aligned}$$

The critical point of (5.4) may easily be found: If we let $w_k = \mu_k H(\phi_k)$, then, at phaselocking, the w_k satisfy

$$O = \beta\Delta + K\mathbf{W}$$

where $\Delta = -(1,1,\cdots,1)^t$, $\mathbf{W} = (w_1,\cdots,w_n)^t$ and K is as in §3. Hence, as before, $w_k = -\beta k(n+1-k)/2$, so the critical point is given by

$$(5.5)\qquad H(\phi_k) = \frac{-\beta k(n+1-k)}{2\mu_k}.$$

It can be seen from (5.5) that a gradient in coupling changes the value k_0 of k at which $\max_k H(\phi_k)$ occurs. (If $\mu(x)$ is monotone increasing (resp. decreasing), k_0 decreases (resp. increases).) This suggests that if β is increased sufficiently to prevent phaselocking, and a pair of plateaus results, then the break will be in the high frequency range for $\mu(x)$ increasing, and low frequency range for $\mu(x)$ decreasing.

The frequency of the phase-locked solution is computed from

$$\dot{\theta}_1 = \omega_1 + \varepsilon\mu_1 H(\phi_1).$$

From (5.5), we see that $\mu_1 H(\phi_1)$ is independent of the coupling coeffients $\{\mu_k\}$, so that the frequency is the same as for uniform isotropic coupling, i.e. $\dot{\theta}_1 = \omega_1 - \varepsilon n\beta/2$.

5.3. Nonsymmetric coupling function *H*. Finally, we give a few simulations (Fig. 5.5) to show an effect of allowing $H(\phi)$ to be nonodd. Note that if $H(\phi) = \sin(\phi + \Phi_0)$ for $\Phi_0 > 0$, the plateaus may lie entirely above the line of natural frequencies.

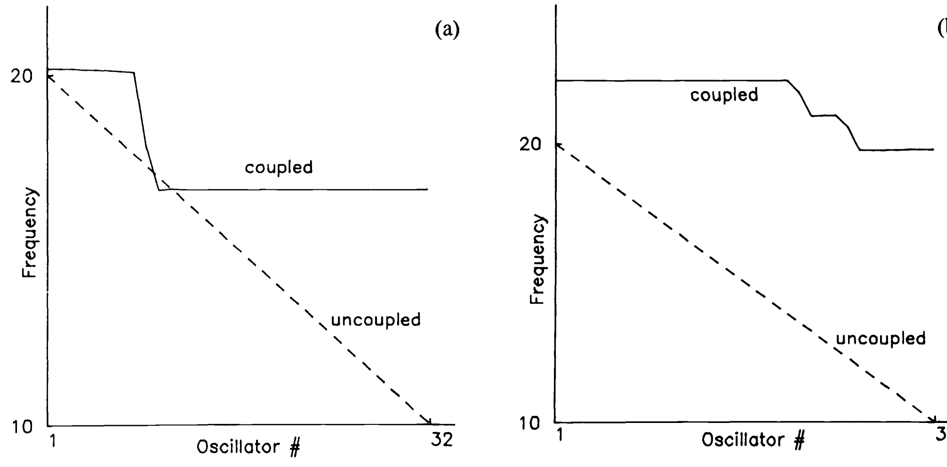


FIG. 5.5. *Frequency vs. k for* $H(\phi) = \sin(\phi + \Phi_0)$, $\Phi_0 > 0$. (a) $\dot{\phi}_k = -10/31 + 14[\sin(\phi_{k+1} + .2) - 2\sin(\phi_k + .2) + \sin(\phi_{k-1} + .2)]$. (b) $\dot{\phi}_k = -10/31 + 11[.2\sin(\phi_{k+1} + .4) - 1.2\sin(\phi_k + .4) + \sin(\phi_{k-1} + .4)]$.

Appendix. We wish to prove some results about eigenvalues and eigenvectors of matrices of the form KA, where K is the $n\times n$ tridiagonal matrix with $K_{ii}=-2$, $K_{i,i+1}=K_{i+1,i}=1$ and A is a diagonal matrix $A=\operatorname{diag}(a_1,a_2,\cdots,a_n)$.

PROPOSITION A.1. *No eigenvalue of KA is pure imaginary. If $a_k>0$ for all k, then all eigenvalues of KA have negative real parts.*

Proof. KA is a tridiagonal matrix with $(KA)_{ii}=-2a_i$, $(KA)_{i,i+1}=a_{i+1}$, $(KA_{i+1,i})=a_i$. By the Gershgorin theorem [A1], any eigenvalue λ of KA must satisfy

$$|\lambda+2a_k|\le 2|a_k| \tag{A.1}$$

for some k. Since the a_k are real, (A.1) rules out pure imaginary eigenvalues. If $a_k>0$ for all k, then (A.1) implies that $\operatorname{Re}\lambda\le 0$. Now $\operatorname{Re}\lambda=0$ can happen only if λ is pure imaginary, or if $\lambda=0$. But $\det(KA)=\det A\cdot\det K\ne 0$, so $\lambda\ne 0$ and $\operatorname{Re}\lambda<0$. □

PROPOSITION A.2. (i) *Suppose that $a_j=0$ for some j, $a_k\ne 0$ for $k\ne j$. Then there exists a unique zero eigenvalue of KA.*

(ii) *Let $[KA,Z]$ denote the $n\times(n+1)$-matrix obtained from KA by adding the n-vector Z as the last column. If $\{a_k\}$ is as above, then $[KA,Z]$ has rank n for $Z=(1,1,\cdots,1)^t$ and $Z=(z_1,\cdots,z_n)^t$ with $z_j=-2$, $z_{j+1}=z_{j-1}=1$ and $z_k=0$, $k\ne j, j\pm 1$.*

Proof. (i) $\operatorname{Det}(KA)=\det K\cdot\det A=0$, so KA has a zero eigenvalue. It can be checked by direct computation that KA has a unique null-vector $V=(v_1,\cdots,v_n)$, with $v_j=1$, $v_k=0$, $k\ne j$. (The equations for the $\{v_k\}$ split into two systems for $v_1,\cdots,v_{j-1}$ and v_{j-1}, v_{j+1}, $v_{j+2},\cdots,v_n$ respectively. The first has $v_1=\cdots=v_{j-1}=0$ as its only solution; using $v_{j-1}=0$, the other system has $v_{j+1}=\cdots=v_n=0$ as its only solution.) Furthermore, if W is the unique null-vector of $(KA)^t$, it is easy to show that $W_j\ne 0$, and hence $V\cdot W\ne 0$. This implies that KA has a simple zero eigenvalue.

(ii) The rank of $[KA,Z]$ is the dimension of the span of its columns. Since the kth column of KA is a_k times the kth column of K, the rank of $[KA,Z]$ is the same as that of $[KI_j,Z]$, where I_j is the identity matrix except for the jth column, which is zero. To show that $[KI_j,Z]$ has rank n, it suffices to show that $W\cdot Z\ne 0$, where $W=(w_1,\cdots,w_n)$ now denotes the null-vector of $(KI_j)^t$. Thus $w_1,\cdots,w_n$ satisfy the equations

$$\begin{aligned} &-2w_1+w_2=0,\\ &w_{k-1}-2w_k+w_{k+1}=0, \qquad k\ne 1,n,\\ &w_{n-1}-2w_n=0 \end{aligned} \tag{A.2}$$

with the jth equation omitted. From (A.2), we see that w_1 determines $w_2,\cdots,w_j$; indeed, (A.2) implies that $w_k=kw_1$ for $k\le j$. Similarly, w_n determines $w_j,\cdots,w_{n-1}$. It follows that the $\{w_k\}$ all have the same sign, and so $W\cdot(1,1,\cdots,1)\ne 0$.

To see that $W\cdot Z\ne 0$ for $z_j=-2$, $z_{j-1}=z_{j+1}=1$, $z_k=0$, $k\ne j, j\pm 1$, we note that $W\cdot Z=0$ implies that (A.2) is supplemented by the jth equation $w_{j-1}-2w_j+w_{j+1}=0$. But the full set of equations $k=1,\cdots,n$ of (A.2) is the system $KW=0$. Since K is nonsingular, and W is nontrivial, this is impossible. □

PROPOSITION A.3. *Suppose that $a_j<0$ for some j and $a_k>0$ for all $k\ne j$. Then there exists a unique eigenvalue of KA with a positive real part.*

Proof. We define a path K_ζ between KA and $K\bar{A}$, where $\bar{A}=\operatorname{diag}(a_1,a_2,\cdots,|a_j|,\cdots,a_n)$ as follows: $K_\zeta=KA$ except for the jth column, and $(K_\zeta)_{ij}=\zeta(KA)_{ij}$. Thus $K_{-1}=K\bar{A}$ and $K_1=KA$. The only value of ζ for which $\det K_\zeta=0$ is $\zeta=0$. By Proposition A.2, K_ζ has a simple zero eigenvalue at $\zeta=0$. Since, by Proposition A.1, all eigenvalues of K_ζ, $\zeta<0$ have negative real part, then for $\zeta>0$, K_ζ must have a unique positive eigenvalue. Since K_ζ has no pure imaginary eigenvalues for any ζ, this is the only eigenvalue with positive real part. □

The comments of Charles Johnson were helpful in proving the following:

PROPOSITION A.4. *Assume that* $A=\operatorname{diag}(a_1,\cdots,a_{j-1},-a_j,a_{j+1},\cdots,a_n)$, *with* $n+1=2j$, $a_k>0$ *for all* k. *Then the unique positive eigenvalue* λ *of* KA *satisfies* $\lambda\leq 2a_j$.

Proof. Instead of KA, we shall consider $B=D^{-1}KA\bar{D}$, where $\bar{D}=\operatorname{diag}(d_1,d_2,\cdots,d_{j-1},1,d_{j+1},\cdots,d_n)$, with $d_i=\sqrt{a_j/a_i}$. B is a tridiagonal matrix with $B_{ii}=-2a_i$, $B_{i,i+1}=B_{i+1,i}=\sqrt{a_ia_{i+1}}$, $i=1,\cdots,j-2$ and $i=j+2,\cdots,n$. The 3×3-matrix $B_{ik}, j-1<i, k<j+1$, is

$$\begin{pmatrix} -2a_{j-1} & -\sqrt{a_{j-1}a_j} & 0 \\ \sqrt{a_{j-1}a_j} & 2a_j & \sqrt{a_ja_{j+1}} \\ 0 & -\sqrt{a_ja_{j+1}} & -2a_{j+1} \end{pmatrix}. \tag{A.3}$$

To get the estimate $\lambda\leq 2a_j$, we shall estimate the spectrum of $C=\frac{1}{2}(B+B^t)$, and then relate this to the spectrum of B. Since B is symmetric except in the 3×3-block (A.3), $B=C$ outside of that block. $C_{ik}, j-1<i, k<j+1$, is given by the 3×3 diagonal matrix $\operatorname{diag}(-2a_{j-1},2a_j,-2a_{j+1})$. Thus C splits into the direct sum of two $(j-1)\times(j-1)$-matrices C_1, C_2 and the 1×1-matrix with entry $2a_j$. Thus the spectrum of C consists of $2a_j$ plus the spectrum of the C_i. We now show that the spectrum $\sigma(C_1)$ of C_1 is entirely negative. Let $U=(u_1,\cdots,u_{j-1})$. Then

$$\begin{aligned}\langle C_1U,U\rangle &= -2\sum_{i=1}^{j-1} a_iu_i^2+2\sum_{i=1}^{j-2}\sqrt{a_ia_{i+1}}\,u_iu_{i+1} \\ &= -\sum_{i=1}^{j-2}\left(\sqrt{a_i}\,u_i-\sqrt{a_{i+1}}\,u_{i+1}\right)^2-a_1u_1^2-a_{j-1}u_{j-1}^2<0,\end{aligned}$$

so C_1 is negative definite. Similarly, so is C_2. Thus $\max\sigma(C)=2a_j$.

Let $(\cdot,\cdot)$ denote the usual complex inner product and $\mathcal{F}(C)=\{(Cz,z)|z\in\mathbb{C},\|z\|=1\}$. Since C is a real symmetric (hence Hermitian) matrix, $\mathcal{F}(C)$ is the convex hull of $\sigma(C)$ [A2]. In particular, $\max\mathcal{F}(C)=\max\sigma(C)$. But C is the symmetrization of B, so $(Cz,z)=\operatorname{Re}(Bz,z)$. Thus $\max\mathcal{F}(C)=\max\operatorname{Re}\mathcal{F}(B)$. But for any matrix B, $\sigma(B)=\mathcal{F}(B)$. Hence $\max\operatorname{Re}\sigma(B)\leq\max\operatorname{Re}\mathcal{F}(B)$. Since B is known to have a unique real positive eigenvalue λ, it follows that $\lambda\leq\max\sigma(C)=2a_j$. □

PROPOSITION A.5. *Let* $A=\operatorname{diag}(a_1,\cdots,a_{j-1},a_j,a_{j-1},\cdots,a_1)$, *with* $a_k>0$, $k\neq j$ *and* $a_j<0$. *Assume that* $a_1\geq a_2\geq\cdots\geq a_{j-1}$ *and that* $a_{k-1}+a_{k+1}<2a_k$ *for all* $k=2,\cdots,j-2$. *Let* $V=(v_1,\cdots,v_n)$ *be the eigenvector of the unique positive eigenvalue* λ *of* KA. *Then* $\operatorname{sgn}v_j=-\operatorname{sgn}v_k$ *for all* $k\neq j$. *Also,* $a_kv_k\leq a_{k+1}v_{k+1}$ *for* $k\leq j-1$.

Proof. The eigenvector V is a nontrivial solution to

$$\begin{aligned} &-(2a_1+\lambda)v_1+a_2v_2=0, \\ &a_{k-1}v_{k-1}-(2a_k+\lambda)v_k+a_{k+1}v_{k+1}=0, \\ &a_{n-1}v_{n-1}-(2a_n+\lambda)v_n=0. \end{aligned} \tag{A.4}$$

If $v_1>0$, then $v_2>0$ by the first equation of (A.4). Also

$$v_2=\frac{2a_1+\lambda}{a_2}v_1>\frac{2a_2+\lambda}{a_2}v_1>2v_1. \tag{A.5}$$

Furthermore,

$$\text{(A.6)}\qquad v_{k+1}-v_k=\frac{(2a_k+\lambda)v_k-a_{k-1}v_{k-1}}{a_{k+1}}-v_k$$

$$=\frac{2a_k+\lambda-a_{k+1}}{a_{k+1}}\left[v_k-\left(\frac{a_{k-1}}{2a_k+\lambda-a_{k+1}}\right)v_{k-1}\right].$$

Now $\lambda>0$, $a_k>a_{k+1}$ implies that $(2a_k+\lambda-a_{k+1})/a_{k+1}>1$, $k=2,\cdots,j-2$. Also, $a_{k-1}/(2a_k+\lambda-a_{k+1})<a_{k-1}/(2a_k-a_{k+1})<1$. (The last inequality is equivalent to $a_{k-1}+a_{k+1}<2a_k$ which holds by hypothesis.) Hence, from (A.6) we have

$$v_{k+1}-v_k>v_k-v_{k+1}$$

which implies that $\operatorname{sgn} v_k=\operatorname{sgn} v_1$, $k=1,\cdots,j-1$. By symmetry, $v_{n+1-k}=v_k$. Finally, for $k=j$ we see that

$$\text{(A.7)}\qquad a_{j-1}v_{j-1}+a_{j+1}v_{j+1}=(2a_j+\lambda)v_j.$$

The left-hand side of (A.6) is positive, since $v_{j\pm1}$, $a_{j\pm1}>0$. By Proposition A.4, $\lambda\le|2a_j|$. Since $a_j<0$, this implies that $v_j<0$.

To show that

$$\text{(A.8)}\qquad a_kv_k\le a_{k+1}v_{k+1},\qquad k=1,\cdots,j-1,$$

we first consider $k=j-1$. By (A.7) and symmetry, $2a_{j-1}v_{j-1}=(2a_j+\lambda)v_j\le 2a_jv_j$ (since $\lambda>0, v_j<0$). For $k=1$, (A.8) follows from (A.5). Also

$$a_{k+1}v_{k+1}=(2a_k+\lambda)v_k-a_{k-1}v_{k-1}\ge a_kv_k+(a_kv_k-a_{k-1}v_{k-1}).$$

Thus, by induction, we have (A.8) for $k=2,\cdots,j-2$. □

PROPOSITION A.6. *Let* $A=\operatorname{diag}(a_1,\cdots,a_j,a_{j-1},\cdots,a_1)$ *with* $a_j=0$, $a_k>0$, $k\ne j$. *Let* $Z=(z_1,\cdots,z_n)$ *be an eigenvector of any real eigenvalue* $\lambda<0$ *of* KA. *Then for some* k_1, k_2, $\operatorname{sgn} z_{k_1}\ne\operatorname{sgn} z_{k_2}$.

Proof. If for some $k\ne j$ we have $\operatorname{sgn} z_k\ne\operatorname{sgn} z_1$, we are done. Otherwise consider (A.4) with $k=j$:

$$a_{j-1}z_{j-1}+a_{j+1}z_{j+1}=\lambda z_j.$$

Since $a_{j\pm1}>0$, $\lambda<0$, we have $\operatorname{sgn} z_j\ne\operatorname{sgn} z_{j\pm1}$. □

Acknowledgments. We wish to thank Dan Fischer, Morris Hirsch, David Rand, Reuben Rosales and Art Winfree for their comments and thought-provoking questions. A talk by James Keener (Salt Lake City, 1976) helped inspire this work.

REFERENCES

[1] J. NEU, *Coupled chemical oscillators*, SIAM J. Applied Math., 37 (1979), pp. 307–315.

[2] ______, *Large populations of coupled chemical oscillators*, SIAM J. Appl. Math., 38 (1980), pp. 305–316.

[3] P. J. HOLMES, *Phase locking and chaos in coupled limit cycle oscillators*, preprint.

[4] R. H. RAND AND P. J. HOLMES, *Bifurcation of periodic motions in two weakly coupled Van der Pol oscillators*, Internat. J. Nonlinear Mech. 15 (1980), pp. 387–399.

[5] M. HIRSCH, *Systems of differential equations which are competitive or cooperative*, I. *Limit sets*, this Journal, 13 (1982), pp. 167–179.

[6] M. KAWATO, M. SOKABE AND R. SUZUKI, *Synergism and antagonism of neurons caused by an electrical synapse*, Biol. Cybernet., 34 (1979), pp. 81–89.

[7] M. KAWATO AND R. SUZUKI, *Two coupled neural oscillators as a model of the circadian pacemaker*, J. Theoret. Biol., 86 (1980), pp. 547–575.

[8] A. H. COHEN, P. J. HOLMES AND R. H. RAND, *The nature of the coupling between segmental oscillators of the lamprey spinal generator*, J. Math. Biol., 13 (1982), pp. 345–369.

[9] F. C. HOPPENSTEADT AND J. P. KEENER, *Phase locking of biological clocks*, J. Math. Biol., 15 (1982), pp. 339–349.

[10] M. ASHKENAJI AND H. G. OTHMER, *Spatial patterns in coupled biochemical oscillators*, J. Math. Biol., 5 (1978), pp. 305–350.

[11] O. E. ROSSLER, *Chemical turbulence: Chaos in a simple reaction-diffusion system*, Z. Naturforsch., 31 (1976), p. 1168.

[12] J. CONNOR, *On exploring the basis for slow potential oscillations in the mammalian stomach and intestine*, J. Experiment. Biol., 81 (1979), pp. 153–173.

[13] E. DANIEL AND S. SARNA, *The generation and conduction of activity in smooth muscle*, Ann. Rev. Pharmacol. Toxicol., 18 (1978), pp. 145–66.

[14] J. CONNOR, A. MANGEL AND B. NELSON, *Propagation and entrainment of slow waves in cat small intestine*, Amer. J. Physiol., 237 (1979), pp. C237–C246.

[15] N. DIAMANT AND A. BORTOFF, Amer. J. Physiol., 216 (1969), pp. 301–307.

[16] N. E. DIAMANT, P. K. ROSE AND E. J. DAVIDSON, *Computer simulation of intestinal slow-wave frequency gradient*, Amer. J. Physiol., 219 (1970), pp. 1684–1690.

[17] S. K. SARNA, E. E. DANIEL AND Y. J. KINGMA, *Simulation of slow-wave electrical activity of small intestine*, Amer. J. Physiol., 221 (1971), pp. 166–175.

[18] B. ROBERTSON-DUNN AND D. A. LINKENS, *A mathematical model of the slow-wave electrical activity of the human small intestine*, Med. Biol. Engrg., 12 (1974), pp. 750–757.

[19] B. H. BROWN, H. L. DUTHIE, A. R. HORN, AND R. H. SMALLWOOD, *A linked oscillator model of electrical activity of human small intestine*, Amer. J. Physiol., 229 (1975), pp. 384–388.

[20] R. J. PATTON AND D. A. LINKENS, *Hodgkin–Huxley type electronic modelling of gastrointestinal electrical activity*, Med. Biol. Engrg. Computing, 16 (1978), pp. 195–202.

[21] G. B. ERMENTROUT AND N. KOPELL, *Frequency plateaus in a chain of coupled oscillators*, II, to appear.

[22] N. FENICHEL, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193–226.

[23] M. W. HIRSCH, C. C. PUGH AND M. SHUB, *Invariant Manifolds*, Lecture Notes in Mathematics, 583, Springer-Verlag, New York, 1977.

[24] J. GUCKENHEIMER, *Isochrons and phaseless sets*, J. Math. Biol., 1 (1975), pp. 259–273.

[25] A. WINFREE, *The Geometry of Biological Time*, Springer-Verlag, New York, 1980.

[26] J. HALE, *Ordinary Differential Equations*, John Wiley, New York, 1969.

[27] N. KOPELL AND L. N. HOWARD, *Bifurcations and trajectories joining critical points*, Adv. in Math., 18 (1975), pp. 306–358.

[28] R. KRONAUER, C. CZEISLER, S. PILATO, M. MOORE-EDE AND E. WEITZMAN, *Mathematical model of the human circadian system with 2 interacting oscillators*, Am. J. Phys. 242 (1982), pp. R3–R17.

[A1] M. MARCUS, *Basic Theorems in Matrix Theory*, National Bureau of Standards Applied Math Series, 57. U.S. Government Printing Office, Washington, DC, 1964.

[A2] P. R. HALMOS, *Finite Dimensional Vector Spaces*, Van Nostrand, New York, 1958.

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics

0036-1410/84/1502-0002 $01.25/0

NEWTON'S ALGORITHM AND CHAOTIC DYNAMICAL SYSTEMS*

M. HURLEY† AND C. MARTIN†‡

Abstract. We show how Newton's method for finding the roots of a real function f leads to chaotic dynamics (infinitely many periodic points and positive topological entropy) for a large class of functions f.

Introduction. This paper contains a description of how "chaotic" dynamics arise in a class of discrete time dynamical systems on the real line. Much is known about these types of dynamics; expositions of the general theory can be found in [CE], [BGMY] and [G]. Our specific concern is the study of the dynamics of the maps defined by the Newton algorithm for finding the real zeros of a function. In this paper we give a simple geometrical description of how complicated dynamics must arise when Newton's method is applied to any map that has at least three simple real roots. In particular we show that if any root-finding algorithm which shares certain features of the Newton algorithm is applied to a map that has three or more simple real roots, then the function defined by the algorithm will have periodic points of all periods and topological entropy greater than zero. (Definitions of the technical terms are given below.) In fact, we calculate lower bounds for the topological entropy which are larger than those that are obtained by applying the estimates for the general case that are found in [BGMY] and [B]. We also indicate one way of constructing algorithms that avoid this complicated behavior.

Given a continuously differentiable function f: $R \to R$, the Newton function for f, Nf, is given by

$$Nf = x - \left[\frac{f(x)}{f'(x)}\right].$$

As first year calculus students are shown, if x_0 is close enough to a root p of f and x_k is defined to be $Nf(x_{k-1})$ for all $k \geq 1$, then the sequence $\{x_k\}$ will converge to p, but for certain "bad" choices of x_0 the algorithm may "blow up" or it may generate a periodic sequence and thus fail to converge to any root of f. The study of this nonconvergence of Newton's method has a long history. B. Barna, in a series of papers [Ba1]-[Ba4], analyzed the situation when f is a polynomial all of whose roots are real. More recently there have been S. Smale's paper [S], and a paper written by D. Saari and J. Urenko [SU]. The latter paper contains an elementary description of how to use symbolic dynamics to describe the orbit structure of Nf for a large collection of maps f. Some of our results overlap those of Saari and Urenko; consequently our proofs (especially in §2) are somewhat abbreviated. We strongly recommend their paper to the reader who is interested in a more detailed explanation.

A key concept in the study of discrete dynamical systems is that of *chaos* or *sensitive dependence on initial conditions*. There have been several definitions of chaos,

*Received by the editors May 18, 1981, and in revised form July 22, 1982.

†Department of Mathematics, Case Institute of Technology, Case Western Reserve University, Cleveland, Ohio 44106.

‡The research of this author was supported in part by NASA grants 2384 and NAG 2-82, Office of Naval Research contract N000 14-80C-0199, and Department of Energy contract DE-A 01-80RA5256.

for example, in [BBW]:

DEFINITION 0.1. A discrete time dynamical system g is said to display *chaos* if it has a periodic solution of each sufficiently high period and an uncountable family of aperiodic solutions with the property that if x and y are distinct members of the family then there exist $\delta>0$ and a cofinal sequence of k's such that $d(g^k(x),g^k(y))>\delta$. (Here $g^k(x)$ is the kth iterate of x by $g, g^2(x)=g(g(x)), g^3(x)=g(g^2(x))$ and so on; x is periodic of period $m>1$ if $x=g^m(x)$ but $x\neq g^j(x)$ for $1\leq j<m$.)

Other definitions are found in [LY] and [G]. One problem with requiring the existence of periodic points in definitions of chaos is that they do not generalize well to dimensions bigger than 1. A map can be made more complicated and at the same time eliminate all periodic points. (Let g: $S^1\to S^1$ be an irrational rotation; then $f\times g$ has no periodic orbits for any f.) One is left in the awkward situation of having "chaotic" maps that are at least as complicated as "nonchaotic" ones. One way of avoiding such problems is to replace periodicity conditions with the requirement that the *topological entropy* be strictly positive. There are several equivalent definitions of topological entropy; the following one is due to R. Bowen [Bo2]. Suppose (X,d) is a compact metric space and that g maps X into itself. Let a positive constant δ and a positive integer n be given. A subset E of X is said to be *(n,δ)-separated* if for any pair of distinct points x,y in E, $d(g^j(x),g^j(y))>\delta$ for some j in $[0,n)$. Note that since X is compact, any (n,δ)-separated set must be finite.

DEFINITION 0.2. Let X and g be as above. The topological entropy, $h(g)$, of g is given by

$$h(g)=\lim_{\delta\to 0}\ \limsup_{n\to\infty}\frac{1}{n}\log(s(n,\delta))$$

where $s(n,\delta)$ is the maximal cardinality of any (n,δ)-separated subset of X.

Remark 0.3. The definition of topological entropy can be interpreted as follows: one observes the dynamical system (X,g) and would like to know "how many" different types of behavior there are. One might attempt to obtain an answer by viewing the system for a finite amount of time, $[0,n)$, through instruments that have imperfect resolution, so that any two initial points whose first n iterates stay within δ of each other are indistinguishable. In this setting, the largest number of "distinct" orbits that can be observed is $s(n,\delta)$. One then computes the asymptotic exponential growth rate of this number as length of observation and resolution approach the ideal case. One can see intuitively that the number $h(g)$ should be larger in a case where there are points x with the property that the approximate location of $g^k(x)$ is not well predicted by knowledge of the approximate location of x; such a map is said to display sensitive dependence on initial conditions.

Remark 0.4. Definition 0.2 may not make sense if X is not compact since the numbers $s(n,\delta)$ might be infinite. In our main applications we will use Definition 0.2 where g is some Newton function Nf and X is a compact, Nf-invariant subset of R. Occasionally we will make enough assumptions about f to ensure that Nf is well defined on the one-point compactification of R, and in this case we can let X be the circle $R\cup\{\infty\}$. (Definitions of topological entropy for noncompact spaces have been given but we do not need to use them in this paper.) For further details on this, see [DGS, Chapt. 14] or [Bo1].

Remark 0.5. There are connections between topological entropy and the collection of periodic orbits of a dynamical system. For one class, the so-called Axiom A diffeomorphisms, it has been shown that $h(g)$ is equal to the asymptotic exponential growth rate in the number of fixed points of g^n as n goes to infinity. In fact, the lower bounds for entropy of Newton functions that we give in §3 are lower bounds on the growth of the number of fixed points of $(Nf)^n$. On the other hand, whenever f and g are maps on compact spaces, $h(f\times g)\geq h(g)$ [DGS], so one cannot reduce topological entropy by adding factors to a given map. In particular, if f is an irrational rotation of a circle, then it is possible for $h(f\times g)$ to be large even though $f\times g$ has no periodic orbits of any period.

Remark 0.6. For us, then, a chaotic dynamical system is one whose topological entropy is positive. There is a weakness to this definition, namely that it allows the chaotic behavior that causes topological entropy to be positive to occur on small sets (say of Lebesgue measure zero). In [S] Smale remarks that this is the situation for the map Nf whenever f is a polynomial with all of its roots real. The physical significance of this type of chaos is uncertain; see [CE, p. 21], for a fuller discussion.

On the other hand there are a great many Newton functions Nf for which there are entire line segments composed of points x such that $(Nf)^k(x)$ does not converge to a root of f. The easiest way for this to happen is for Nf to have an attracting periodic point. In §4 we show how such attractors can arise and we give some examples.

The outline of the paper is as follows. In §1 we list some general properties of Newton functions and indicate certain feature of their graphs. Section 2 contains our description of how periodic orbits must occur for Nf provided that f has at least three real roots. In §3 we use these results to obtain an estimate of the topological entropy of Nf. Section 4 is a discussion of the phenomenon of attracting periodic points, and in §5 we indicate a way of constructing a root-finding algorithm for polynomials that is nonchaotic (there are practical problems with the implementation of this algorithm; they are discussed in §5).

As we noted above, the chaotic dynamics of Newton's method have been described by other authors, Barna and Saari–Urenko in particular. What is new in our approach is, first, the emphasis on topological entropy, and second, the description of how chaos can be caused not only by f having several real roots, but also by f having a critical point larger (smaller) than the largest (smallest) root of f (plus certain other technical conditions; see §3 for a full description).

1. Basic properties of Newton functions. Throughout this paper f will denote a map from R to R. For convenience we shall assume that f has two continuous derivatives. We also make the following nondegeneracy assumptions.

Assumption 1.1. (a) If $f(x)=0$, then $f'(x)\neq 0$.

(b) If $f'(x)=0$, then $f''(x)\neq 0$.

As before, $Nf=x-[f(x)/f'(x)]$ denotes the Newton function associated with f. The fundamental property of Nf is that it transforms the problem of finding roots of $f(x)$ into the problem of finding attracting fixed points of Nf. Note $(Nf)'=ff''/(f')^2$.

Remark 1.2. $f(x)=0$ if and only if $Nf(x)=x$. Moreover, if $f(p)=0$ then $(Nf)'(p)=0$, so $(Nf)^k(x)\to p$ for all x near p. We now list some properties of the function Nf that will be used in later sections. The proofs are elementary calculations which we leave to the reader.

Remark 1.3. Nf has a vertical asymptote at each real solution $x=c$ of $f'(x)=0$. At each such point $\lim_{x\to c^+}Nf(x)=-\lim_{x\to c^-}Nf(x)=\pm\infty$.

If $c_1<c_2$ are consecutive roots of $f'(x)$, then the interval (c_1,c_2) is called a *band* for Nf. If $f'(x)$ has a largest (respectively, smallest) root c (resp. b), then the interval (c,∞) (resp. $(-\infty,b)$) is called an *extreme band* for Nf.

Remark 1.4. If (c_1,c_2) is a band for Nf that contains a root of $f(x)$ then $\lim_{x\to c_1^+} Nf(x) = +\infty$, $\lim_{x\to c_2^-} Nf(x) = -\infty$.

Remark 1.5. If (c_1,c_2) is a band for Nf that does not contain a root of $f(x)$, then $\lim_{x\to c_1^+} Nf(x) = \lim_{x\to c_2^-} Nf(x) = \pm\infty$.

Remark 1.6. If a band for Nf does not contain a root of $f(x)$, then one of the adjoining bands or extreme bands also contains no roots of $f(x)$. This last property often holds for extreme bands as well (for example, when f is a polynomial), but it may fail for extreme bands, as the example $f(x) = xe^{-x}$ shows. Later, in §2, we will give sufficient conditions that Remark 1.6 holds for extreme bands as well as for bands, namely that $f''(x)$ is bounded away from 0 as $|x| \to \infty$.

Remark 1.7. Nf is undefined when $f'(x) = 0$; the local extrema of Nf are the points where $f(x)f''(x) = 0$.

Combining these facts, we see that a typical graph of $Nf(x)$ looks like the graph in Fig. 1. In that figure the bands are labelled B_1 through B_5 and the extreme bands are B_0 and B_6.

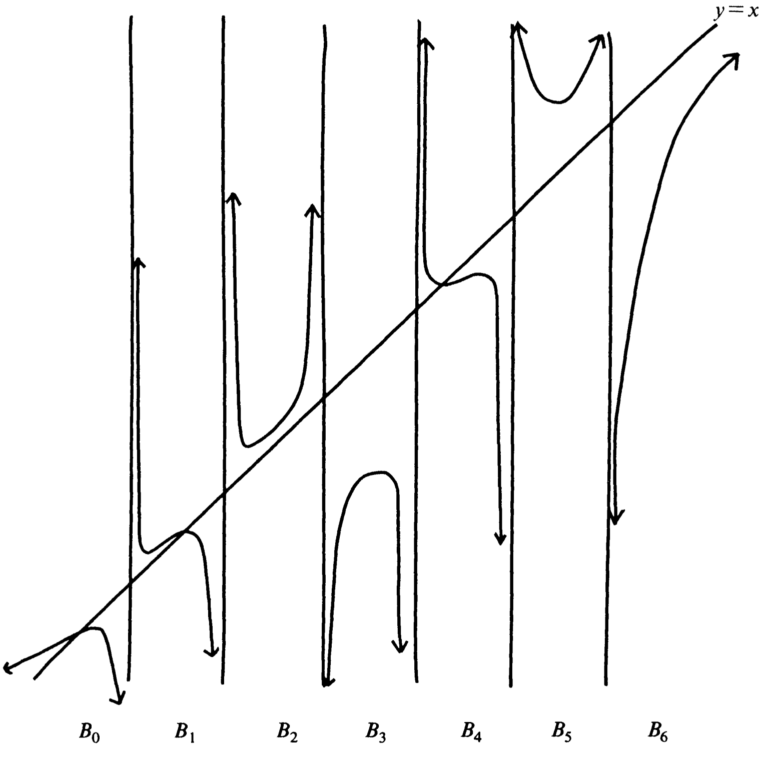


FIG. 1

2. Nonconvergence of $(Nf)^k(x)$—periodic orbits. The simplest way that the sequence $(Nf)^k(x)$ may fail to converge to a root of f is when this sequence becomes undefined at some finite stage, i.e., when some iterate of Nf maps x onto a vertical asymptote of Nf. By the mean value theorem this type of nonconvergence must occur whenever f has more than one real root.

The next simplest way that nonconvergence of $(Nf)^k(x)$ can occur is when x is periodic or eventually periodic under Nf: x is *periodic* if $Nf(x) \neq x$ but $(Nf)^k(x) = x$ for some k bigger than one, its *period* is the least value of k for which this holds; x is *eventually periodic* if $(Nf)^m(x)$ is periodic for some m. A point p of period k is called *attracting* (respectively, *repelling*) if the distance from x to p is greater than (respectively, less than) the distance from $(Nf)^k(x)$ to $p=(Nf)^k(p)$ for all x in some open interval containing p. Note that p is attracting if $|(d/dx)(Nf)^k(p)|<1$ and is repelling if $|(d/dx)(Nf)^k(p)|>1$ (it may be attracting, repelling, or neither if this derivative is equal to 1). In §4 we will discuss the existence of attracting periodic orbits for Nf.

For the remainder of this section we concern ourselves with estimating the number of periodic orbits of Nf in terms of the number of roots of $f(x)$. We begin with a simple lemma.

LEMMA 2.1. *If (c_1,c_2) is a band containing a root p of $f(x)$, then (c_1,c_2) contains a period two orbit $\{z,Nf(z)\}$.*

Proof. From our knowledge of the shape of the graph of Nf it is not hard to see that the graph of $(Nf)^2$ must look something like the graph in Fig. 2; that graph shows that the points labelled z and $Nf(z)$ form a period two orbit. Note that the graph shows that $(d/dx)(Nf)^2(z)$ is at least 1.

The following lemma is our basic tool for counting periodic points; a more sophisticated version (Theorem 3.1) will be the key to our estimates of topological entropy.

LEMMA 2.2. *Let g be a map $R\to R$ and suppose that $I_1, I_2,\cdots,I_k$ $(k\geq 2)$ are pairwise disjoint compact intervals with the restriction of g to I_j continuous for each j and the union of all the I_j's contained in $g(I_m)$ for each m. Then g has periodic points of all periods. In fact, g has at least k^n points satisfying $g^n(x)=x$ for each n and the closure of the set composed of all these periodic points is uncountable and is mapped onto itself by g.*

Proof. An elementary argument shows that if a compact interval I is mapped continuously over itself by some function h, then h must have a fixed point in I. Our assumptions suffice to ensure that for any j between 1 and k and any positive integer n, there is a compact subinterval of I_j that is mapped continuously over itself by g^n. Indeed, we can require more. Let $T(n;k)$ denote the set of all finite sequences $(x_i)_{i=0}^n$ with each x_i chosen from the set $\{1,\cdots,k\}$. Given such a sequence, we can find a compact interval J such that

(i) g^i is continuous on J for $1\leq i\leq n$,

(ii) $g^i(J)$ is contained in I_{x_i} for $0\leq i\leq n$, and in fact $g^n(J)=I_{x_n}$ (g^0 is the identity map).

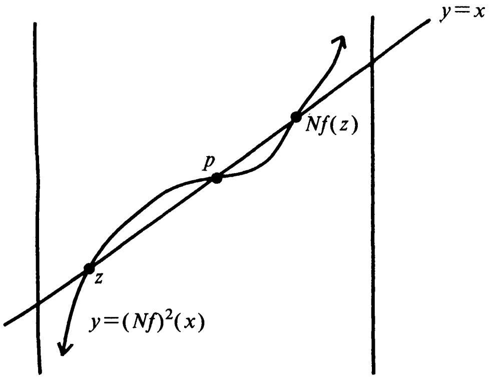


FIG. 2

By (i) and (ii), if $x_0=x_n$, then g^n maps J continuously over itself, so g^n has a fixed point $P((x_i))$ in J. Moreover, if (y_i) is a second sequence with $y_0=y_n$, but with $y_j \neq x_j$ for some j, then the disjointness of the intervals I_i combined with (ii) ensures that $P((x_i)) \neq P((y_i))$. Thus the number of fixed points of g^n is at least as great as the number of sequences (x_i) in $T(n;k)$ with $x_0=x_n$, and this number is clearly k^n. See [SU] or [BGMY] for more details. The final assertion of Lemma 2.2 follows from the fact that the set $T(\infty;k)=\{(x_i)_{i=0}^{\infty} | \text{each } x_i \in 1,2,\cdots,k\}$ is uncountable (the finite intersection property for compact sets ensures that given any (x_i) in $T(\infty;k)$ there will be at least one point $Q((x_i))$ with $g^i(Q((x_i)))$ contained in I_{x_i} for each i).

PROPOSITION 2.3. *Suppose $f(x)$ has at least* 4 (*distinct*) *roots. Then Nf has periodic points of all periods*; *the set of points not converging to a fixed point of Nf is uncountable.*

Remark. Barna [Ba2]-[3] gives a proof of this fact in the case that $f(x)$ is a polynomial with all real roots.

Proof. If a band B contains a fixed point, then Nf maps B continuously onto R (Remark 1.4). If $f(x)$ has at least 4 roots, then there are at least two such bands. With a little care one can choose compact subintervals I_1 and I_2 of these bands satisfying the assumptions of Lemma 2.2, and so there are periodic points of all periods. To see that there is an uncountable set not converging to a fixed point, note that the subset $S=\{(x_i)|x_i \text{ is not eventually constant}\}$ of $T(\infty;2)$ is also uncountable (for each m the set $\{(x_i)|x_m=x_{m+1}=\cdots\}$ is countable), and if (x_i) is in S, then $Q((x_i))$ (defined in the proof of Lemma 2.2) does not approach a fixed point under iteration by Nf.

In fact, we can sharpen this last result.

THEOREM 2.4. *Define integers α, β as follows*:

(1) *α is the number of extreme bands for Nf that*

(i) *contain no fixed points of Nf, and*

(ii) *are mapped onto R by Nf*

(*so* $\alpha=0, 1,$ *or* 2).

(2) *β is the number of bands for Nf that contain fixed points of Nf.*

If $\alpha+\beta$ is at least 2, *then there is an uncountable set of points x such that $(Nf)^k(x)$ does not converge to a fixed point of Nf as k goes to infinity. If in addition β is at least* 1, *then Nf has periodic points of all periods.*

Proof. Suppose $\alpha+\beta \geq 2$. If β is 0, then α must be 2, so Nf has no fixed points at all and the assertion is trivially true. If $\beta \geq 2$, the result follows from Proposition 2.3. The remaining case is $\beta=1$, $\alpha=1$ or 2. Here one uses the assumption that there is an extreme band B with $Nf(B)=R$ to select two compact intervals, I_1 in B and I_2 in the band containing the fixed point, such that

$$I_1 \cup I_2 \subset Nf(I_2) \quad \text{and} \quad I_2 \subset Nf(I_1).$$

(This is like the selection of the intervals in the proof of Lemma 2.2.) The remainder of the proof in this case is like that of Lemma 2.2, although the details (which we omit) are a bit more cumbersome.

Remark 2.5. It is not hard to find fairly general conditions on f that ensure that condition (1)(ii) in Theorem 2.4 is met whenever (1)(i) is met. For instance, one can require that $f''(x)$ be bounded away from 0 for $|x|$ large and x in the extreme band under consideration. (In particular, this is the case whenever f is a polynomial of degree at least 2.) To see that this condition on $f''(x)$ suffices, suppose for definiteness that $B=(c, \infty)$ (the other case is similar). The assumption on f'' then shows that either f and f' both approach ∞ or else they both approach $-\infty$ as x goes to ∞. Thus for x large $f(x)/f'(x)$ is positive, and so $Nf(x)$ is less than x for all $x>c$. Now by 1.3 we see that $Nf(x)$ goes to $-\infty$ as x approaches c from the right. On the other hand, since $f'(x)\to\pm\infty$ as $x\to\infty$, we can use l'Hôpital's rule to show that $Nf(x)=[xf'(x)-f(x)]/f'(x)$ goes to infinity as x does.

Remark 2.6. In the case $\alpha=2$, $\beta=0$ of 2.4, Nf will have periodic points of all periods greater than one. The proof of this result requires slightly different techniques than those used in the proof of 2.4, and so we omit the proof.

Remark 2.7. It is possible for Nf to have periodic points of all periods >1 even when $\alpha+\beta=0$. We shall give an example of this at the end of the next section (Example 3.5).

3. Topological entropy. In what follows we indicate a way of getting a lower bound on the number of fixed points of $(Nf)^k$ in terms of the numbers α and β defined above. We then use these bounds and the fact that under certain conditions the topological entropy of Nf if bounded below by $\lim_{k\to\infty}\frac{1}{k}$ log(the number of fixed points of $(Nf)^k$) to obtain lower bounds for topological entropy. This will show that the topological entropy of Nf is strictly positive under fairly mild conditions on f. We refer the reader to the introduction for the definition of topological entropy and other related definitions.

Let f, Nf, α, and β be as in Theorem 2.4, and define $M=\beta+2$. We define an $M\times M$ matrix A, called the *structure matrix*,

$$A=\left(\begin{array}{c|c} \multicolumn{2}{c}{V} \\ \hline W_\alpha & W_\beta \end{array}\right)$$

where V is $2\times M$,

$$V=\begin{pmatrix} 0 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & 1 & \cdots & 1 \end{pmatrix} \qquad \text{if } \alpha=2,$$

$$V=\begin{pmatrix} 0 & 0 & 1 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \qquad \text{if } \alpha=1,$$

$$V=\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \qquad \text{if } \alpha=0.$$

W_α is $\beta\times 2$; it is composed entirely of ones if $\alpha=2$, its first column is entirely ones and its second column is entirely zeros if $\alpha=1$, and it is composed entirely of zeros if $\alpha=0$. Finally, W_β is the $\beta\times\beta$ matrix composed entirely of ones.

THEOREM 3.1. *Let f, Nf, and A be as above. Then there is a constant $\gamma>0$ such that for each k there is a collection S_k of fixed points of $(Nf)^k$ satisfying*

(a) *S_k contains at least trace (A^k) points, and*

(b) *if x and y are distinct points in S_k, then for some i, $0\le i\le k$, the distance between $(Nf)^i(x)$ and $(Nf)^i(y)$ is greater than γ (so S_k is a (k,γ)-separated set).*

Proof. We describe the proof in the case $\alpha=2$; the other cases are similar. Let c_1 be the smallest critical point of $f(x)$, and c_2 the largest. $\lim_{x\to c_1^-}Nf(x)=\infty$, $\lim_{x\to c_2^+}Nf(x)=-\infty$, and Nf maps each extreme band onto R, so we can choose $x_1<c_1$ and $x_2>c_2$ with $Nf(x_1)<c_1$, $Nf(x_2)>c_2$. Then $[c_1,\infty)\subset Nf((x_1,c_1))$ and $(-\infty,c_2]\subset Nf((c_2,x_2))$. See Fig. 3. Now choose compact subintervals $I_1\subset(x_1,c_1)$ and $I_2\subset(c_2,x_2)$ with $[c_1,x_2]\subset Nf(I_1)$ and $[x_1,c_2]\subset Nf(I_2)$.

The definition of β implies the existence of β disjoint open bands contained in (c_1,c_2), each containing a fixed point of Nf, and each mapped onto R by Nf. Hence

- Points on the line $y=x$ are identified by their first coordinate.
- Here $\beta=1$.

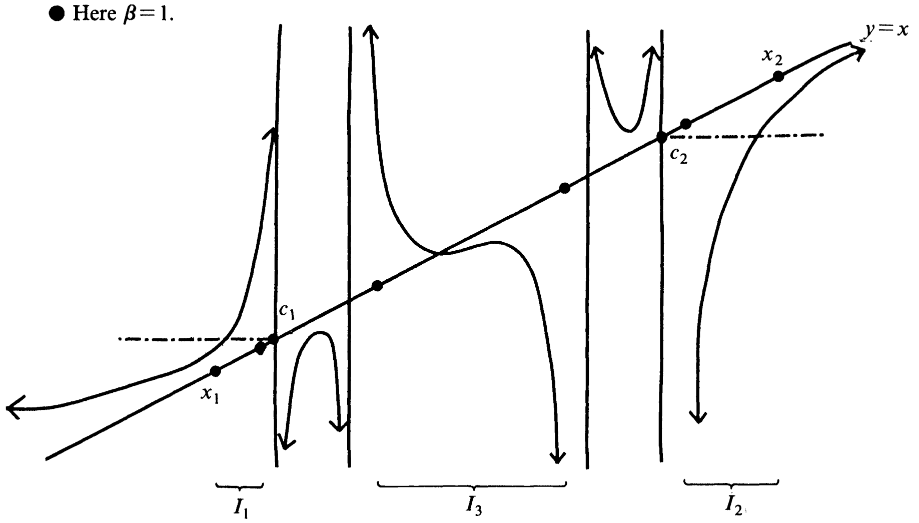


FIG. 3

there are β pairwise disjoint compact intervals $I_3,\cdots,I_{\beta+2}$, each of which is mapped continuously over $[x_1,x_2]$ by Nf. Including the two intervals I_1 and I_2, we now have $\beta+2$ pairwise disjoint compact intervals.

Since a function has a fixed point any time that it maps an interval continuously over itself, the number of fixed points of $(Nf)^k$ is no less than the number of distinct ways any of these $\beta+2$ intervals is mapped over itself by $(Nf)^k$. This is the number of sequences of the form $(j_i)_{i=0}^k$ subject to the conditions

(i) each j_i is an integer in the range $[1,\beta+2]$,
(ii) $j_{i+1}\neq j_i$ whenever j_i is 1 or 2,
(iii) $j_0=j_k$.

Another way of expressing (i) and (ii) is to say that j_{i+1} can be any integer m in the range $[1,\beta+2]$, provided that the entry in row j_i and column m of the matrix A is one. From here it is easy to conclude that the number of such sequences is trace(A^k). (The proof is a standard argument that we do not reproduce in full. To get a feeling of how it goes, note that the (i,i) entry of A^2 is the number of pairs (i,j) such that both the (i,j) and (j,i) entries of A are equal to 1. In this case, we know that Nf maps I_i continuously over I_j, and I_j continuously over I_i. Thus there is a fixed point x_j of A^2 in I_i with $Nf(x_j)$ in I_j.) See [W1], [W2], [DGS], [P] or [Bo2] for more detailed treatments. This establishes Theorem 3.1(a).

Theorem 3.1(b) now follows by choosing $\gamma>0$ to be less than the distance between any two of the intervals I_j, and noting that if x and y are fixed points of $(Nf)^k$ corresponding to the *distinct* sequences $(j_i)_0^k$ and $(j_i^*)_0^k$, then there is an i with $j_i\neq j_i^*$, so that for this value of i, $|(Nf)^i(x)-(Nf)^i(y)|\geq$ the distance between I_{j_i} and $I_{j_i^*}>\gamma$.

COROLLARY 3.2. *The topological entropy of Nf*

(i) $\geq\log(\beta)$ *if* α *is* 0,
(ii) $\geq\log(\beta+(\beta^2+4\beta)^{1/2})/2$ *if* α *is* 1,
(iii) $\geq\log(\beta+1+(\beta^2+6\beta+1)^{1/2})/2$ *if* α *is* 2.

Proof. As noted in part (b) of Theorem 3.1, S_k is a (k,γ)-separated set for each k, and so $s(k,\delta)\geq s(k,\gamma)\geq$trace$(A^k)$ whenever $\delta\leq\gamma$. From the definition it follows that the topological entropy of Nf is bounded below by $\limsup \frac{1}{k} \log \operatorname{trace}(A^k)$, which in turn is equal to log(the largest positive eigenvalue of A) [P]. To compute this eigenvalue, it suffices to find the largest positive eigenvalue of a matrix B where B is related to A as follows: $A = PQ$ and $B = QP$, where P and Q are matrices, not necessarily square, all of whose entries are nonnegative integers [W1], [W2]. We list appropriate choices of P and Q in each of the cases (i)–(iii); from there the result is just a computation. In each case Q has $\beta+2$ columns,

$$\text{(i)} \qquad Q=(0 \;\; 0 \;\; 1 \;\; 1 \;\; \cdots \;\; 1), \quad P=\begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad QP=(\beta);$$

$$\text{(ii)} \quad Q=\begin{pmatrix} 1 & 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 1 & 1 & \cdots & 1 \end{pmatrix}, \quad P=\begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}, \quad QP=\begin{pmatrix} \beta & 1 \\ \beta & 0 \end{pmatrix};$$

$$\text{(iii)} \quad Q=\begin{pmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}, \quad P=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix}, \quad QP=\begin{pmatrix} 0 & 1 & \beta \\ 1 & 0 & \beta \\ 1 & 1 & \beta \end{pmatrix}.$$

COROLLARY 3.3. *If $\alpha \geq 1$, then Nf has positive topological entropy.*

Proof. For definiteness we may assume Nf has an extreme band $B=(-\infty, y)$ with $Nf(B)=R$, so that $Nf(x)>x$ for all x in B. Let C denote the band or extreme band that is immediately to the right of B, so $C=(y,z)$, $y<z\leq\infty$. By Remark 1.6 and the condition on f'' in the definition of $\alpha, Nf(x)<x$ for all x in C. There are two possibilities; either $Nf(C)=R$ or $Nf(C)\subset(-\infty, a]$ for some finite a. If $Nf(C)=R$, then C must be an extreme band and $\alpha=2$, so Corollary 3.3 follows from Corollary 3.2. Thus we assume that $Nf(C)$ is bounded above. There are several cases.

Case 1. See Fig. 4. Assume C is a band and $Nf(C)$ contains the closure of B. Since $Nf(B)=R$, there is a compact interval B_0 in B such that $Nf(B_0)=\text{closure}(C)$. Similarly there is a compact interval B_1 in B such that $Nf(B_1)=B_0$. As in Theorem 2.4 we can choose a compact interval C_0 in C such that $B_0 \cup B_1 \subset Nf(C_0)$, and as in Theorem 3.1 these three intervals give rise to the structure matrix

$$A=\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

As in Corollary 3.2, the entropy of Nf restricted to the union of these 3 intervals is bounded below by the logarithm of the largest eigenvalue of A, which is approximately log(1.325).

Case 2. Assume C is an extreme band with the closure of B contained in $Nf(C)$. In this case we replace C by a bounded subset C^* which also satisfies $\text{closure}(B) \subset Nf(C^*)$. We then define B_0 and B_1 in terms of C^* instead of C, and proceed as in case 1.

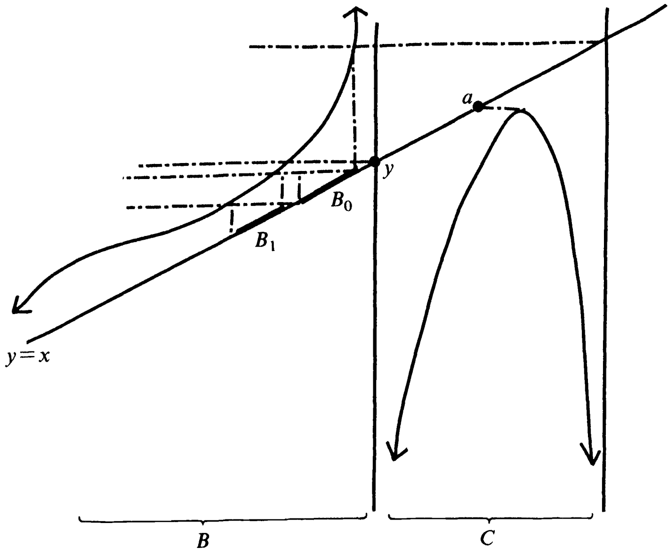


FIG. 4

Case 3. If $Nf(C)$ does not contain the closure of B, then we have to work a little harder. Let B_0 be defined as in the previous cases. Since $Nf(x)>x$ for all x in B and $\liminf_{x\to-\infty} Nf(x)=-\infty$, we can choose a sequence of compact intervals $B_j, j\geq 1$, such that

(i) $B_j \subset B$ for all j,

(ii) $Nf(B_j)=B_{j-1}$,

(iii) if we choose an element x_j of B_j for each j, then $j>i$ if and only if $x_j<x_i$, and $\lim_{j\to\infty} x_j=-\infty$.

Condition (iii) ensures that for some n, $B_n \cup B_{n-1} \subset Nf(C)$. By using the intervals B_0, $B_1, \cdots, B_{n-1}$, B_n, C, we obtain a structure matrix

$$A=\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 1 \end{pmatrix} \qquad \text{(this is an } n+2\times n+2 \text{ matrix).}$$

By expanding $A-xI$ along its first column, it is not hard to check that the characteristic polynomial of A is

$$p(x)=(-1)^n x^{n+2}+(-1)^{n+1}x+(-1)^{n+1}=(-1)^n[x^{n+2}-x-1];$$

the expression in the square brackets is -1 when x is 1, and it tends to infinity as x goes to infinity, so $p(x)$ always has a root λ bigger than 1. Thus the topological entropy of Nf is bounded below by $\log(\lambda)>0$.

COROLLARY 3.4. *Suppose $f(x)$ is a polynomial with n real roots (all distinct), where n is at least* 3. *Then $(Nf)^k$ has at least $(n-2)^k$ fixed points for each k, and the topological entropy of Nf is at least* $\log(n-2)$.

Results like Theorem 3.1 and its corollaries are true generally. If $\alpha=2$ then it follows from Theorem 2.4(1) that Nf can be extended to a continuous map of the circle viewed as the one-point compactification of R. (Saari and Urenko treat this type of Newton function in [SU].) In this situation, a theorem of L. Block states, among other things, that if g is a continuous map of the circle to itself that has a fixed point and a point of some other odd period m, then it has periodic points of all periods k bigger than m. See [B] and [BGMY]. The same references provide estimates of the topological entropy of such maps; however these estimates are not in general as sharp as our estimates in Corollaries 3.2 and 3.3.

Example 3.5. We give an example where Nf has periodic points of all periods greater than 1 even though $\alpha+\beta=0$. The graph of Nf is given in Fig. 5. The essential features of that graph are

(1) $Nf(B_1)$ contains B_2 and B_4,

(2) $Nf(B_2)$ contains B_1,

(3) $Nf(B_4)$ contains B_1 and B_2.

As before, we can choose compact subintervals of B_1, B_2, B_4 that map over each other according to the same scheme as (1)–(3), and so obtain the structure matrix

$$A=\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$

It is not hard to see that the trace of A^k is strictly positive for all k bigger than 1. In fact, by computing eigenvalues one can determine that the topological entropy is at least $\log[(\sqrt{5}+1)/2]$.

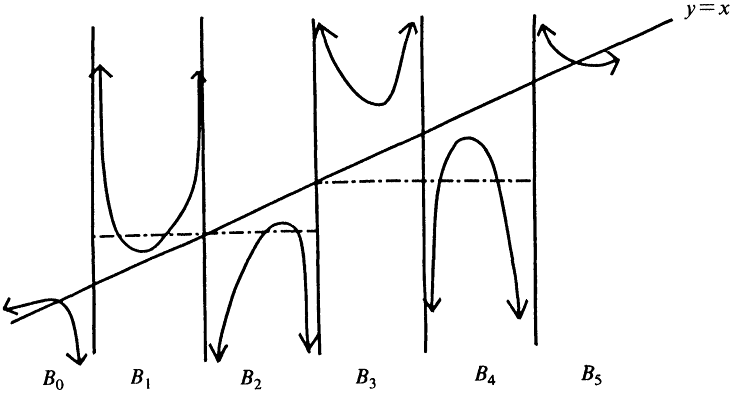


FIG. 5

4. Nonconvergence of $(Nf)^k(x)$—attracting periodic orbits. On one hand the results of the previous two sections are satisfying in that they indicate how for many functions f, Nf has a complicated dynamical structure, including an uncountable number of points x for which $(Nf)^k(x)$ does not converge to a root of f. On the other hand, as we mentioned in Remark 0.6 this uncountable set may be small in the sense of Lebesgue measure, perhaps even measure 0. In such a case one would not expect (in a probabilistic sense) to encounter these nonconvergent points in practice. Consequently, in this section we turn our attention to the existence of attracting periodic orbits (recall that under our definitions a fixed point is not considered to be periodic). If Nf has an attracting periodic orbit then there is an open interval I such that if x is in I, then $(Nf)^k(x)$ does not converge to a root of f; instead it accumulates on the periodic attractor. Thus points in I are well-behaved from a dynamical point of view, but badly behaved from the point of view of Newton's method as a root-finding algorithm.

Probably the oldest and best known result in this area is Barna's theorem that if f is a polynomial with all of its roots real, then Nf has no attracting periodic orbits. The proof relies on Julia's theorem that an attracting periodic orbit of a rational function $r(x)$ of degree at least 2 must attract some critical point of $r(x)$, where $r(x)$ is thought of as acting on the complex plane.

For other types of functions f, attracting periodic orbits of Nf certainly can occur. As an example, in Fig. 6 we have sketched the Newton function of a map with four real roots. The interval AB is mapped inside itself by $(Nf)^2$, and both $|(Nf)'(x)|$ and $|(Nf)'(Nf(x))|$ are less than 1 for all x in AB. Hence Nf has an attracting periodic point in AB of period 2.

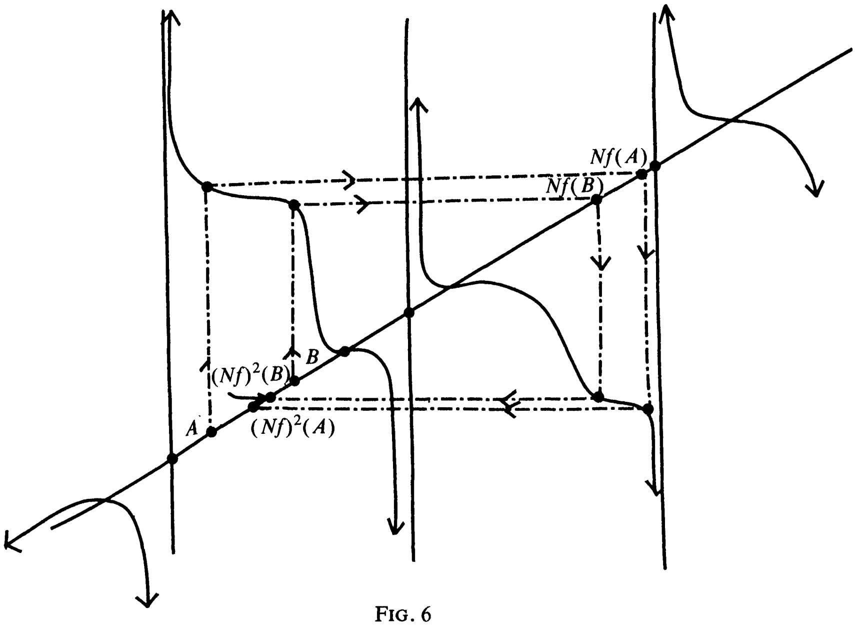


FIG. 6

Similarly one can fashion graphs of Nf with attracting periodic orbits of arbitrary period. That these graphs are the graphs of Nf for some functions f is assured by the following.

PROPOSITION 4.1. *Suppose g is a map from R to R that satisfies*

(1) *g has two continuous derivatives except at a finite number of points $\{c_i\}$ at which g is not defined, and* $\lim_{x\to c_i}|g(x)|=\infty$,

(2) $\lim_{x\to c_i^+} g(x) = -\lim_{x\to c_i^-} g(x)$,

(3) *g has a finite number of fixed points, each of which is a critical point of g, and any two of which are separated by an element of $\{c_i\}$,*

(4) *g has only finitely many critical points.*

Then there is a continuously differentiable function f from R to R such that $g=Nf$.

The proof is elementary, although tedious, and is left to the reader.

PROPOSITION 4.2. *Given $k>1$, there are polynomials whose Newton functions have attracting periodic orbits of period k.*

Proof. Combine Proposition 4.1, the Stone–Weierstrass theorem, and the local stability of attracting periodic points under C^1 perturbations.

Example 4.3. $f(x)=3x^5-10x^3-23x$.

Example 4.4. $f(x)=(x^2-9)(x^2-4)(635x^3-2363x^2+2413x-973)$.

Both of these examples yield Newton functions Nf satisfying $Nf(1)=-1$, $Nf(-1)=1$, and $(Nf)'(1)=0$, so each has $\{1,-1\}$ as an attracting periodic orbit. Example 4.3 has only one root, $x=0$, while the roots of Example 4.4 are ± 2, ± 3, and approximately 2.4070. Barna [Ba3] lists two similar polynomials, $11x^6-34x^4+39x^2$ and x^5-10x^3+69x; each of these has only one real root.

Besides the situation described in Fig. 6 there is a second way for period 2 orbits to occur, when the graph of $(Nf)^2$ within a band that contains a fixed point of Nf is as shown in Fig. 7. Here A and A' form an attracting periodic orbit of period 2. Note that Example 4.4 is of this type.

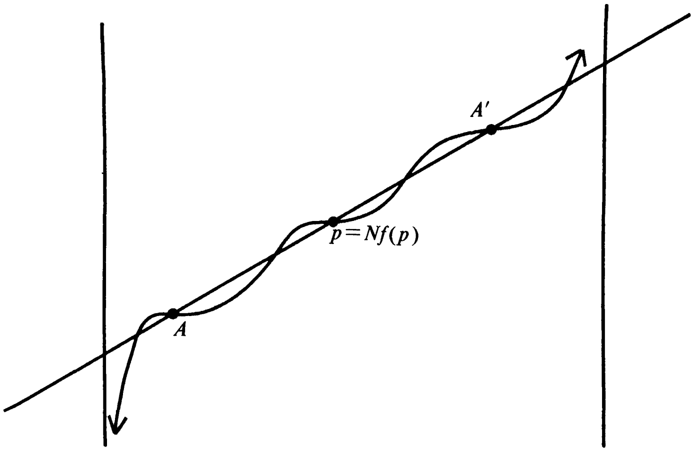


FIG. 7

5. Alternate algorithms.

The discussion in §§2,3 makes clear that chaotic dynamics and the consequent nonconvergence are inherent features of Newton's method. This prompts one to look for algorithms whose dynamics are better behaved. We give below an example that works nonchaotically to find roots of polynomials. It is offered as a motivating example only; we do not suggest that it is a very practical computational device for finding roots of polynomials. We hope that it gives some idea of how chaos-avoiding algorithms can be obstructed.

First consider what we must give up to avoid the complicated dynamics of Newton's method. Suppose f is a smooth map from R to R and that associated to f is a map Tf defined on all but finitely many points of R, smooth where defined, and satisfying $Tf(x)=x$ if and only if $f(x)=0$ and $|Tf(x)|\to\infty$ as x approaches any of the points where Tf is undefined.

LEMMA 5.1. *If each of the fixed points of Tf is attracting, then Tf exhibits chaos similarly to the Newton function Nf.*

Proof. The condition on the fixed points forces the graph of Tf to cross the line $y=x$ from left to right at each fixed point. Hence Tf must have discontinuities, and therefore asymptotes, between any two fixed points. Now arguments analogous to those of §2 serve to establish the existence of chaotic behavior of Tf just as they did for Nf.

This result tells us that to avoid chaos we must not insist that every equilibrium be attracting. Consequently, in order to find all of the roots of f we need a finite set of algorithms $T_1f,\cdots,T_kf$ such that each T_if is nonchaotic and each root of f is an attracting fixed point for at least one of the T_if. An obvious candidate for such a system of algorithms is Euler's method applied to the differential equations $x'=f(x)$ and $x'=-f(x)$. The problem with this approach is that if the step size in Euler's method is too large one might overshoot several roots and thus not detect them. The way to avoid this difficulty is to ensure that the step size is small enough, but it is hard to say how small "small enough" is without some fairly detailed a priori knowledge of the graph of f. For these reasons we present the following alternative which works when f is a polynomial. It also involves the choice of a step size, but this constant can be calculated directly from the coefficients of the polynomial.

THEOREM 5.2. *Suppose $f(x)$ is a polynomial of degree at least 3 and with no multiple roots. For $i=1,2$ define*

$$T_if(x)=x+(-1)^ik\left[\frac{f(x)}{1+f'(x)^2}\right]$$

where k is a constant between 0 and 1 whose choice is described below. Then: (1) *$T_if(x)=x$ if and only if $f(x)=0$, $i=1$ or 2;*

(2) *the fixed points of T_if are alternately attracting and repelling;*

(3) *p is an attracting fixed point of T_1f if and only if it is a repelling fixed point of T_2f;*

(4) *given any x in R, $(T_if)^m(x)$ tends either to $\pm\infty$ or to a root of f as m goes to ∞. In particular, at least one of the two sequences $(T_1f)^m(x)$, $(T_2f)^m(x)$ tends to a root of f, provided that f has a root.*

Proof. Part (1) is clear; (2) and (3) are just direct calculations. To define k and establish (4), note that if f has degree n, then each $(T_if)'$ is a rational function of the form $1+k(-1)^ip(x)/q(x)$, where p and q are polynomials of degree $3n-3$ and $4n-4$ respectively. To identify p and q more precisely, note first that we may assume that the coefficient of x^n in f is $1/n$. Then using the quotient rule we see that $p=f'+(f')^3-2ff'f''$ and $q=[1+(f')^2]^2$, and we can conclude:

(a) $q(x)\geq 1$ for all x,

(b) the leading term of p is $[(2-n)/n]x^{3n-3}$ and the leading term of q is x^{4n-4}. Now let A be the largest absolute value of any coefficient of p, and let B be the largest absolute value of any coefficient of q. By (b) A is at least $1-2/n$ and B is at least 1.

LEMMA 5.3. *$|p(x)/q(x)|<A(A+2B)^{3n-2}$ for all x.*

Let k be $1/[A(A+2B)^{3n-2}]$, so that by the lemma $(T_if)'>0$ for all x. From this it is not hard to conclude that any sequence $(T_if)^k(x)$ is monotonic and so must either converge to a fixed point for T_if or else diverge to plus or minus infinity.

Proof of Lemma 5.3. Several times in the calculations below we use the fact that if $y>2$, then $\Sigma_0^m y^j<y^{m+1}$.

Case 1. If $|x|>A+2B$,

$$|p(x)/q(x)|\leq\left(A\sum_{j=0}^{3n-3}|x|^j\right)\Big/\left(|x|^{4n-4}-B|x|^{4n-5}-B\sum_{j=0}^{4n-6}|x|^j\right)$$

$$<\left(A|x|^{3n-2}\right)/\left(|x|^{4n-4}-B|x|^{4n-5}-B|x|^{4n-5}\right)$$

$$=A/\left(|x|^{n-2}-2B|x|^{n-3}\right)=A/\left[|x|^{n-3}(|x|-2B)\right]$$

$$<A/\left[(A+2B)^{n-3}(A)\right]=1/(A+2B)^{n-3}\leq 1$$

$$<A(A+2B)^{3n-2}\quad\text{by (b).}$$

Case 2. If $|x| \leq A + 2B$,

$$|p(x)/q(x)| \leq |p(x)| \quad \text{(by (a))}$$
$$\leq A \sum_{j=0}^{3n-3} |x|^j \leq A \sum_{0}^{3n-3} (A+2B)^j < A(A+2B)^{3n-2}.$$

Remark. The obvious practical obstacle to implementing this algorithm is that the step size k is quite small even for fairly tame polynomials of reasonably low degree. This has two consequences; first that the algorithm will require a large number of steps to converge to a root (on the order of $1/k$ steps), and secondly that the distance $x - T_i f(x)$ may be small compared to the round-off error involved in computing this quantity.

REFERENCES

[Ba1] B. Barna, *Uber das Newtonsche Verfahren zur Annaherung von wurzeln algebraischen Gleichungen*, Publ. Math. Debrecen, 2 (1951), pp. 50–63.

[Ba2-4] ______, *Uber die Divergenzpunkte des Newtonschen Verfahrens zur bestimmung von wurzeln algebraischen Gleichungen*, I, Publ. Math. Debrecen, 3 (1953), pp. 109–118; II, 4 (1956), pp. 384–397; III, 8 (1961) pp. 193–207.

[B] L. Block, *Periodic orbits of continuous mappings of the circle*, Trans. Amer. Math. Soc., 260 (1980), pp. 553–562.

[BGMY] L. Block, J. Guckenheimer, M. Misiurewicz, and L.-S. Young, *Periodic points and topological entropy of one-dimensional maps*, Global Theory of Dynamical Systems, Lecture Notes in Mathematics 819, Nitecki and Robinson, eds., Springer-Verlag, New York, 1980, pp. 18–34.

[Bo1] R. Bowen, *Topological entropy for noncompact spaces*, Trans. Amer. Math. Soc., 184 (1973), pp. 125–136.

[Bo2] ______, *On Axiom A Diffeomorphisms*, CBMS Regional Conference Series in Applied Mathematics 35, American Mathematical Society, Providence, RI, 1978.

[BBW] J. Baillieul, R. Brockett, and B. Washburn, *Chaotic motion in nonlinear feedback systems*, IEEE Trans. Circuits and Systems, CAS 27 (1980), pp. 990–997.

[CE] P. Collet and J.-P. Eckmann, *Iterated Maps of the Interval as Dynamical Systems*, Birkhauser, Boston, 1980.

[DGS] M. Denker, C. Grillenberger and K. Sigmund, *Ergodic Theory on Compact Spaces*, Lecture Notes in Mathematics 527, Springer-Verlag, New York, 1976.

[G] J. Guckenheimer, *Sensitive dependence on initial conditions for one-dimensional maps*, Comm. Math. Phys., 70 (1979), pp. 133–160.

[LY] T. Li and J. Yorke, *Period three implies chaos*, Amer. Math. Monthly, 82 (1975), pp. 985–992.

[P] W. Parry, *Intrinsic Markov chains*, Trans. Amer. Math. Soc., 112 (1964), pp. 55–66.

[S] S. Smale, *The fundamental theorem of algebra and complexity theory*, Bull. Amer. Math. Soc. (New Series), 4 (1981), pp. 1–36.

[SU] D. Saari and J. Urenko, *Newton's method, circle maps, and chaotic motion*, Amer. Math. Monthly, to appear.

[W1] R. F. Williams, *Classification of symbol spaces of finite type*, Bull. Amer. Math. Soc., 77 (1971), pp. 439–443.

[W2] ______, *Classification of subshifts of finite type*, Ann. Math., 98 (1973), pp. 120–153, *Erratum*, 99 (1974), pp. 380–381.

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1502-0003 $01.25/0

A CONNECTION PROBLEM FOR A REGULAR AND AN IRREGULAR SINGULAR POINT OF COMPLEX ORDINARY DIFFERENTIAL EQUATIONS*

REINHARD SCHÄFKE†

Dedicated to my father Friedrich W. Schäfke on the occasion of his sixtieth birthday

Abstract. In recent papers [SIAM J. Math. Anal., 11 (1980), pp. 848–862, 863–875] D. Schmidt and the author studied the connection problem for two neighboring regular singular points for quite general linear complex ordinary differential equations. In the present paper these results are generalized in the case that one singular point is irregular singular, but of rank 1 and the leading matrix has n distinct eigenvalues; one singular point is regular singular; and there may be several other singular points. The main result is a limit formula for those connection coefficients which are essential in some specified way. An application to the generalized Heun equation is given.

1. Introduction.

An important task in the investigation of the global behaviour of solutions of a linear differential equation in the complex plane consists in finding connection relations between the local solutions. D. Schmidt and the author studied the problem in [12] and [13] for very general differential equations in the case of two neighboring regular singular points. In the present paper we allow one singular point to be irregular, but of rank 1 and not too complicated. Then the method of [12] must be combined with the saddle point method to get comparable results. The Stokes phenomenon for the irregular singular point causes some technical difficulties.

We consider the linear complex differential equations in $\mathbb{C}^n$

$$y'(z)=\left(\frac{1}{z}A_0+\frac{1}{(z-1)^2}B+\frac{1}{z-1}A_1+G(z)\right)y(z) \tag{1.1}$$

where A_0, A_1 and B are complex n by n matrices and G is a matrix-valued function holomorphic in the open disk $\mathfrak{R}=\{z\in\mathbb{C}\,|\,|z|<r\}$, where $1<r<\infty$. We assume that B has n distinct eigenvalues or that without loss of generality

$$B=\operatorname{diag}(\lambda_1,\cdots,\lambda_n),\qquad \lambda_j\neq\lambda_k\quad \text{for} j\neq k. \tag{1.2}$$

Then 0 is a singular point of the first kind of (1.1), 1 is a singular point of the second kind, in general irregular. 1 is neighboring to 0 in the sense that only outside $|z|\le 1$ there can be further (and arbitrary) singular points of (1.1).

It is the main result in this paper, that a pair of singular points can be handled separately from the others. As far as I know, papers ([1],[5],[8],[11],[15], etc.) only study connection relations involving an irregular singular point, if there is at most one other singular point in the whole complex plane (but on the other hand some allow higher ranks than 1). [6] is the first to admit two additional singular points.

The scope of applications of our results is extended largely by transformations of the independent variable. If e.g. for a complex differential equation a simple singular point and one of rank 1 lie within any circle not containing further singular points we achieve the form (1.1) by a transformation $z=(aw+b)/(cw+d)$, $ad-bc\neq 0$. Likewise

* Received by the editors December 14, 1981.

† Universität Essen-GHS, Fachbereich 6-Mathematik, Universitätsstrasse 3, D-4300 Essen 1, Federal Republic of Germany.

that is possible if ∞ is of rank 1 and in an arbitrary halfplane there is only one simple singular point. Section 4 contains an example for this.

Unfortunately not all pairs of singular points can be separated in this way. By conformal transformation, however, we can map any domain containing two singular points onto a circle and apply our results. If, for example, 0 is a singular point of the first kind, 1 is of rank 1 and]0, 1[contains only regular points the form (1.1) for the w-equation can be attained by

$$1-z=\frac{(1+\varepsilon-w)^{1/n}-\varepsilon^{1/n}}{(1+\varepsilon)^{1/n}-\varepsilon^{1/n}},$$

for $\varepsilon>0$ small enough and $n\in\mathbb{N}$ sufficiently large. The disadvantage of such transformations compared with the former one consists in the fact that in most frequent applications, i.e. differential equations with rational coefficients, the recursion formulas for the coefficients of the power series solutions are much more complicated for the w-equations.

The local behaviour of the solutions of (1.1) near 0 is known, there exists a characteristic fundamental system (see [2, p. 120] or [3, Vol. II, pp. 163ff.]). Because of the assumption on B we can characterize in almost every semicircle $H_\theta=\{z\,|\,0<|z-1|<r-1, |\arg(1-z)-\theta|<\pi/2\}$ a fundamental set of solutions by their asymptotic behaviour as $z\to 1$ (see [14, Chap. IV, XI]); the dependence upon θ in general is complicated (see [4]).

Now the problem arises, how the fundamental system near 0 is connected with the one in H_0 (or $H_\delta, |\delta|>0$ small if necessary). In the present paper this problem is solved for the special case where no logarithmic terms appear in the solutions near 0. The general case could be treated then by a reduction method similar to that of [13].

More precisely, we consider a Floquet solution y_0 of (1.1) at 0,

$$y_0(z)=z^\alpha\sum_{k=0}^{\infty}z^k d_k \tag{1.3}$$

where α is an eigenvalue of A_0 and $d_k\in\mathbb{C}^n$. At 1 we have no convergent power series solutions of (1.1), but there are unique formal solutions $\hat{y}_1,\cdots,\hat{y}_n$ of (1.1) of the form

$$\hat{y}_j(z)=\exp\left(\frac{\lambda_j}{1-z}\right)(1-z)^{\alpha_j}\sum_{k=0}^{\infty}(1-z)^k c_j(k)\qquad (j=1,\cdots,n) \tag{1.4}$$

where $\alpha_j\in\mathbb{C}$ and $c_j(k)\in\mathbb{C}^n$, $c_j(0)=e_j$ is the jth unit vector.

Further it is known that there exist solutions $y_1(z),\cdots,y_n(z)$ of (1.1) defined in $H=\{z\,|\,0<|z-1|<r-1, |\arg(1-z)|<\pi/2\}$ which satisfy

$$y_j(z)\sim\hat{y}_j(z)\qquad (H\ni z\to 1, j=1,\cdots,n), \tag{1.5}$$

i.e., for every $m\in\mathbb{N}$ and $\varepsilon,\delta>0$ there exists $\mu>0$ such that for $z\in H$ satisfying $0<|z-1|<\mu$ and $|\arg(z-1)|\le\pi/2<\delta$ the inequality

$$\left|\exp\left(-\frac{\lambda_j}{1-z}\right)(1-z)^{-\alpha_j}y_j(z)-\sum_{k=0}^{m}(1-z)^k c_j(k)\right|\le\varepsilon|1-z|^m$$

holds. These results can be found in [14, Chap. IV], written for a singular point in ∞ instead of 1.

From [4, §6] we conclude that the y_j are uniquely determined by (1.5) if and only if the differences $\lambda_j-\lambda_k$, $j\neq k$ are not real. Section 7 shows, that in every case for sufficiently small positive δ there exist solutions $y_1^+(z),\cdots,y_n^+(z)$ of (1.1) which are defined in $H^+=\{z\,|\,0<|z-1|<r-1,\ -\pi/2<\arg(1-z)<\pi/2+2\delta\}$ and satisfy

$$(1.6)\qquad y_j^+(z)\sim\hat{y}_j(z)\qquad (H^+\ni z\to 1, j=1,\cdots,n).$$

Likewise we have solutions $y_1^-(z),\cdots,y_n^-(z)$ defined in $H^-=\overline{H^+}$ which satisfy

$$(1.7)\qquad y_j^-(z)\sim\hat{y}_j(z)\qquad (H^-\ni z\to 1, j=1,\cdots,n).$$

If no difference $\lambda_j-\lambda_k, j\neq k$ is real, clearly $y_j^-(z)=y_j^+(z)=y_k(z)$ for $z\in H$. In general (cf. [4, p. 76]) we have

$$(1.8)\qquad y_j^+(z)=y_j^-(z)+\sum_{l\prec j}\alpha_{jl}y_l^-(z)\qquad (z\in H, j=1,\cdots,n),$$

where we use the notation

$$(1.9)\qquad l\prec j\quad \text{if and only if}\quad \lambda_j-\lambda_l\in\,]0,\infty[.$$

Now we have a solution at 0 and (even two) fundamental systems near 1, and the connection problem arises. y_0 can be written as a linear combination of the $y_j^\pm$,

$$(1.10)\qquad y_0(z)=\sum_{j=1}^{n}\gamma_j^\pm y_j^\pm(z)\qquad \left(z\in H, |z|<1, |\arg z|<\frac{\pi}{2}\right),$$

where $\gamma_j^\pm\in\mathbb{C}$ are called connection coefficients. Relation (1.10) tells us how the solution $y_0(z)$ known near $z=0$ behaves when z approaches 1. From (1.8) we have

$$(1.11)\qquad \gamma_j^-=\gamma_j^++\sum_{l\succ j}\gamma_l^+\alpha_{lj}.$$

The main result of §2 is an asymptotic formula for the power series coefficients d_k of y_0 which contains the $\gamma_j^\pm$. To obtain this formula we use the idea of [12] together with the saddle point method (see [9]). In §3 we deduce explicit limit formulas for some of the $\gamma_j^\pm$ that are also useful in applications. In §4 we apply our results to the generalized Heun equation

$$(1.12)\quad y''(z)+\left[\frac{1-\mu_0}{z}+\frac{1-\mu_1}{z-1}+\frac{1-\mu_2}{z-a}+\alpha\right]y'(z)+\frac{\beta_0+\beta_1z+\beta_2z^2}{z(z-1)(z-a)}y(z)=0,$$

where $a\in\mathbb{C}\setminus\{0,1\}$ and $\alpha\neq 0$ and μ_j,β_j are complex parameters. This equation with three regular singular points and one irregular singular point at $0,1,a$ and ∞ has been discussed in [12]. There connection relations for pairs of finite singular points have been derived, here we add connection relations for a finite singular point and ∞.

The main results of the paper are Theorems 2.8, 3.1, 3.7 and 4.15. They seem to be new even when restricted to the spheroidal wave equation, where only the complicated connection relations of [7, pp. 295ff.] are known.

2. Asymptotic formula for the power series coefficients d_k. The Cauchy formula yields for $k\in\mathbb{N}$

$$d_k=\frac{1}{2\pi i}\int_{K_\varepsilon(0)} z^{-k-\alpha-1}y_0(z)\,dz.$$

Now $z^{-\alpha}y_0(z)$ can be analytically continued to $\Re\setminus[1,r[$, and we can write

$$d_k=\frac{1}{2\pi i}\int_{\tilde{c}-c} z^{-k-\alpha-1}y_0(z)\,dz, \tag{2.1}$$

where c and $\tilde{c}$ are the two curves of Fig. 1. Here $A=1-r_0e^{-i\psi}$, $E=1-r_0e^{i\psi}$ and $\rho>1$, $0<r_0<r-1$ and $\psi=\pi/2+\eta$, with $0<\eta<\delta$ to be specified later.

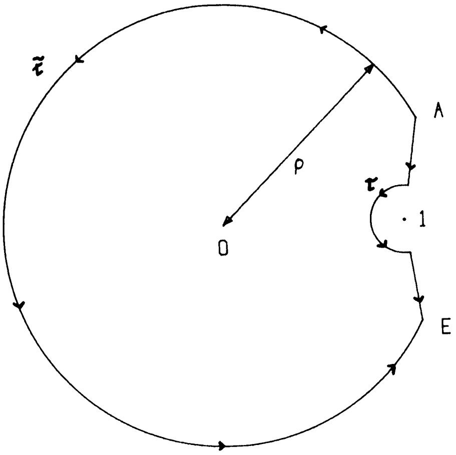


FIG. 1.

Now the y_j^+ can be continued to $0<|z-1|<r-1$, $|\arg(1-z)|<\pi/2+\delta$, but the asymptotic formulas for them in general are only valid in $H^+\ni z\to 1$. Since c lies in the new domain of y_j^+ we can use (1.10) and get

$$d_k=d_k^0+\sum_{j=1}^{n}\gamma_j^+d_k^j \qquad (k\in\mathbb{N}), \tag{2.2}$$

where

$$d_k^0=\frac{1}{2\pi i}\int_{\tilde{c}} z^{-k-\alpha-1}y_0(z)\,dz \qquad (k\in\mathbb{N}) \tag{2.3}$$

and

$$d_k^j=-\frac{1}{2\pi i}\int_{c} z^{-k-\alpha-1}y_j^+(z)\,dz \qquad (k\in\mathbb{N}, j=1,\cdots,n). \tag{2.4}$$

Since $|z|=\rho$ on $\tilde{c}$, we obviously have

$$d_k^0=O(\rho^{-k}) \quad \text{as } k\to\infty. \tag{2.5}$$

The asymptotic formulas for d_k^j are more complicated. They are stated in the following lemmata; their proofs are postponed.

LEMMA 2.6. *Let $\lambda_j\in\mathbb{C}\setminus]-\infty,0]$ and suppose that $r_0>0$ and $\eta>0$ are sufficiently small. Then*

$$d^j_k\sim\exp\left(2\sqrt{\lambda_j}\sqrt{k}\right)k^{-\alpha_j/2-3/4}\sum_{l=0}^{\infty}k^{-l/2}a^j_l\quad \text{as } k\to\infty$$

where $a^j_l\in\mathbb{C}^n$, in particular

$$a^j_0=\frac{1}{2\sqrt{\pi}}e^{\lambda_j/2}\lambda_j^{\alpha_j/2+1/4}e_j.$$

The powers of λ_j are determined by $\arg\lambda_j\in]-\pi,\pi[$.

LEMMA 2.7. *Suppose that $r_0>0$ and $\eta>0$ are sufficiently small. Then for any $\lambda_l\in]-\infty,0]$ there are sequences $(g^{l\pm}_k)_{k\in\mathbb{N}}$ in $\mathbb{C}^n$, such that*

$$d^j_k=g^{j+}_k+g^{j-}_k+\sum_{l<j}\alpha_{jl}g^{l-}_k\qquad(k\in\mathbb{N})$$

for all $j\in\{1,\cdots,n\}$ with $\lambda_j\in]-\infty,0]$, and that the following asymptotic formulas hold:

a) *If $\lambda_l<0$ then*

$$g^{l\pm}_k\sim\exp\left(\pm2i\sqrt{|\lambda_l|}\sqrt{k}\right)k^{-\alpha_l/2-3/4}\sum_{n=0}^{\infty}k^{-m/2}a^{l\pm}_m\quad \text{as } k\to\infty$$

where $a^{l\pm}_m\in\mathbb{C}^n$, in particular $a^{l\pm}_0=\frac{1}{2\sqrt{\pi}}e^{\lambda_l/2}(|\lambda_l|e^{\pm\pi i})^{\alpha_l/2+1/4}e_l$.

b) *If $\lambda_l=0$ then $g^{l-}_k=0$ and*

$$g^{l+}_k\sim\sum_{m=0}^{\infty}\frac{\Gamma(k+\alpha-\alpha_l-m)}{\Gamma(k+\alpha+1)}\frac{1}{\Gamma(-\alpha_l-m)}c_l(m)\quad \text{as } k\to\infty.$$

Together with these lemmata (2.2) immediately yields

THEOREM 2.8. *Suppose that the general assumptions of* §1 *hold. Then for $\lambda_j\in\mathbb{C}\setminus]-\infty,0]$ there exist sequence $(d^j_k)_{k\in\mathbb{N}}$ and for $\lambda_j\in]-\infty,0]$ exist sequences $(g^{j\pm}_k)_{k\in\mathbb{N}}$ such that for $k\in\mathbb{N}$*

$$d_k=\sum_{\lambda_j\notin]-\infty,0]}\gamma^+_j d^j_k+\sum_{\lambda_j\in]-\infty,0]}\gamma^+_j g^{j+}_k+\sum_{\lambda_j\in]-\infty,0[}\tilde{\gamma}_j g^{j-}_k+d^0_k,$$

where $\tilde{\gamma}_j=\gamma^+_j+\Sigma_{l>j,\lambda_l\leq0}\alpha_{lj}\gamma^+_l$ and where asymptotic formulas for d^j_k and $g^{j\pm}_k$ are given in (2.5)–(2.7).

This theorem is unsatisfactory in that it is not symmetric with respect to γ^+_j and γ^-_j and that not all of the summands are essential for the asymptotic representation of d_k. This restriction will be removed in §3.

Proof of Lemma 2.6. We proceed similarly to Perron in [10], who studied the asymptotic formulas for $k\to\infty$ of the confluent hypergeometric functions. Thus in (2.4) we substitute $z=1-s/\sqrt{k}$ and get

$$(2.9)\qquad d^j_k=I_1+I_2+I_3+I_4,$$

$$I_\nu=\frac{1}{2\pi i}\int_{c_\nu}\left(1-\frac{s}{\sqrt{k}}\right)^{-k-\alpha-1}y^+_j\left(1-\frac{s}{\sqrt{k}}\right)\frac{1}{\sqrt{k}}ds,$$

where the curves $\mathfrak{c}_\nu$ are sketched in Fig. 2. Here $\tilde{A}_k=r_0\sqrt{k}\,e^{i(\pi/2+\eta)}$, $A=re^{i(\pi/2+\eta)}$, $C=re^{i(-\pi/2+\eta)}$, $B=e^{-i(\pi/2+\eta)}$ and $\tilde{B}_k=r_0\sqrt{k}\,e^{-i(\pi/2+\eta)}$, r is chosen later. Thus $\mathfrak{c}_3$ and $\mathfrak{c}_4$ depend on k, $\mathfrak{c}_1$ and $\mathfrak{c}_2$ are independent of k.

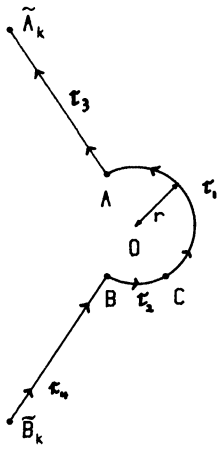


FIG. 2.

In the following I_1 will be evaluated asymptotically using formula (1.6); I_2, I_3 and I_4 will be estimated. First we write

$$I_1=\frac{1}{2\pi i}\sqrt{k}^{\,-\alpha_j-1}\int_{\mathfrak{c}_1}\left(1-\frac{s}{\sqrt{k}}\right)^{-k-\alpha-1}\exp\left(\frac{\lambda_j}{s}\sqrt{k}\right)s^{\alpha_j}f_j^+\left(1-\frac{s}{\sqrt{k}}\right)ds \tag{2.10}$$

where

$$f_j^+(z)\sim\sum_{m=0}^{\infty}(1-z)^m c_j(m)\qquad(H^+\ni z\to 1). \tag{2.11}$$

Hence for s on $\mathfrak{c}_1$, $n\in\mathbb{N}$ and $k\in\mathbb{N}$ we have

$$f_j^+\left(1-\frac{s}{\sqrt{k}}\right)=\sum_{m=0}^{n}\left(\frac{s}{\sqrt{k}}\right)^m c_j(m)+\sqrt{k}^{\,-n-1}r_j(n,s,k) \tag{2.12}$$

where $r_j(n,s,k)$ remains bounded as $k\to\infty$ uniformly for s on $\mathfrak{c}_1$. The power series of the logarithm yields

$$\left(1-\frac{s}{\sqrt{k}}\right)^{-k-\alpha-1}=\exp\left(s\sqrt{k}+\frac{1}{2}s^2\right)\left(1+\sum_{l=1}^{\infty}\beta_l(s)k^{-l/2}\right) \tag{2.13}$$

for $|s|<\sqrt{k}$ and hence for s on $\mathfrak{c}_1$ and k sufficiently large. From (2.11) and (2.13) we obtain for $n\in\mathbb{N}$

$$\begin{aligned}&\left(1-\frac{s}{\sqrt{k}}\right)^{-k-\alpha-1}f_j^+\left(1-\frac{s}{\sqrt{k}}\right)\\&\quad=\exp\left(s\sqrt{k}+\frac{1}{2}s^2\right)\left(\sum_{m=0}^{n}k^{-m/2}b_{jm}(s)+k^{(-n-1)/2}R_j(n,s,k)\right)\end{aligned} \tag{2.14}$$

where $b_{jm}(s)$ are polynomials with $b_{j0}(s)=e_j$ and $R_j(n,s,k)$ remains bounded as $k\to\infty$ uniformly for s on c_1. Insertion in (2.10) yields for $n\in\mathbb{N}$

$$I_1=\frac{1}{2\pi i}k^{(-\alpha_j-1)/2}\left[\sum_{m=0}^{n}k^{-m/2}J_m+k^{(-n-1)/2}\tilde{J}_n\right] \tag{2.15}$$

where

$$J_m=\int_{\mathrm{c}_1}\exp\left(\sqrt{k}\left(s+\frac{\lambda_j}{s}\right)\right)s^{\alpha_j}e^{s^2/2}b_{jm}(s)\,ds, \tag{2.16}$$

$$\tilde{J}_n=\int_{\mathrm{c}_1}\exp\left(\sqrt{k}\left(s+\frac{\lambda_j}{s}\right)\right)s^{\alpha_j}e^{s^2/2}R_j(n,s,k)\,ds. \tag{2.17}$$

Now we obtain the asymptotic representation of J_m from [9, Chap. IV, Thm. 7.1].

Somewhat simplified for our application this theorem says that

$$\int_{\mathrm{c}}e^{-zp(s)}q(s)\,ds\sim 2e^{-zp(s_0)}\sum_{l=0}^{\infty}\Gamma\left(l+\frac{1}{2}\right)a_l z^{-l-1/2}$$

as $z\to\infty$ on $\arg z=\theta$ with $a_0=q(s_0)(2p''(s_0))^{-1/2}$, if

(i) p and q are defined and holomorphic on a neighborhood of the finite smooth path c,

(ii) p' has a simple zero at an interior point s_0 of c,

(iii) $\mathrm{Re}(e^{i\theta}p(s)-e^{i\theta}p(s_0))$ is positive on c except at s_0,

(iv) with a parametrization $s=\varphi(t)$, $t\in[0,1]$ of c satisfying $s_0=\varphi(t_0)$

$$\left|\arg p''(s_0)+\theta+2\lim_{t\searrow t_0}\arg(\varphi(t)-s_0)\right|\leq\frac{\pi}{2}.$$

Now (2.16) can be written

$$J_m=\int_{\mathrm{c}_1}e^{-\sqrt{k}p(s)}q(s)\,ds, \tag{2.18}$$

where $p(s)=-s-\lambda_j/s$ and $q(s)=s^{\alpha_j}e^{s^2/2}b_{jm}(s)$.

The zeros of p' (in $\mathbb{C}\setminus\{0\}$) are $\pm\sqrt{\lambda_j}$. In order that one of them lies on c_1 we now choose $r=\sqrt{|\lambda_j|}$. If $\eta>0$ is sufficiently small then the principal value $s_0=\sqrt{\lambda_j}$ is the only zero of p' on c_1. Then (i) and (ii) are satisfied. Since the limit in (iv) is $\pi/2+\frac{1}{2}\arg\lambda_j$ (principal value), this condition is satisfied if $p''(s_0)=-2\lambda_j^{-1/2}$ takes as argument $-\pi-\frac{1}{2}\arg\lambda_j$.

Towards the essential assumption (iii) on $\mathrm{Re}\,p(s)$ we show

(2.19) *For* $s=re^{i\varphi}$, $\varphi\in[-\pi/2-\eta,\pi/2+\eta]$ *the inequality*

$$\mathrm{Re}\left(s+\frac{\lambda_j}{s}\right)\leq 2\,\mathrm{Re}\,s_0$$

holds. Equality is only possible for $s=s_0$.

If we put $s_0=re^{i\psi}$, $|\psi|<\pi/2$ and thus $\lambda_j=r^2e^{2i\psi}$, we get

$$\mathrm{Re}(s+\lambda_j/s-2s_0)=-2r\cos\psi(1-\cos(\pi-\varphi))$$

and this is negative because of $|\psi|<\pi/2$ if $\psi\neq\varphi$. Statement (2.19) now says that $\operatorname{Re}p(s)>\operatorname{Re}p(s_0)$ for $s\neq s_0$ on $\mathfrak{c}_1$ and that is the key assumption of Olver's theorem. We obtain

$$(2.20)\qquad J_m\sim\exp\left(2\sqrt{\lambda_j}\sqrt{k}\right)\sum_{l=0}^{\infty}2\Gamma\left(l+\frac{1}{2}\right)\sqrt{k}^{-l-1/2}c_l^m \quad \text{as } k\to\infty$$

where

$$(2.21)\qquad c_l^m\in\mathbb{C}^n,\ \text{in particular } c_0^0=\frac{i}{2}e^{\lambda_j/2}\lambda_j^{\alpha_j/2+1/4}e_j,$$

and again all powers take their principal values.

Since $R_j(n,s,k)$ is uniformly bounded on $\mathfrak{c}_1$ as $k\to\infty$, (2.19) immediately yields the estimate

$$(2.22)\qquad \tilde{J}_n=O\left(\exp\left(2\sqrt{\lambda_j}\sqrt{k}\right)\right) \quad \text{as } k\to\infty.$$

(2.15) together with the last two formulas then gives the asymptotic representation of the first part I_1 of d_k^j:

$$(2.23)\qquad I_1\sim\exp\left(2\sqrt{\lambda_j}\sqrt{k}\right)k^{-\alpha_j/2-3/4}\sum_{l=0}^{\infty}k^{-l/2}a_{jl}$$

where

$$(2.24)\qquad a_{jl}\in\mathbb{C}^n,\ \text{in particular } a_{j0}=\frac{1}{2\sqrt{\pi}}e^{\lambda_j/2}\lambda_j^{\alpha_j/2+1/4}e_j.$$

Now we estimate the integrals I_2,I_3,I_4 of (2.9). For the first factor of the integrand we show that for $\varepsilon\in]-\pi/4,\pi/4[$ there exists $r_0>0$ such that

$$(2.25)\qquad \left|\left(1-\frac{s}{\sqrt{k}}\right)^{-k}\right|\leq\left|e^{s\sqrt{k}}\right|$$

for s satisfying $0<|s|\leq r_0\sqrt{k}$ and $\arg s\in]-\pi/2-\varepsilon,-\pi/2+\varepsilon[\cup]\pi/2-\varepsilon,\pi/2+\varepsilon[$. This follows from the fact that $e^{2\operatorname{Re}z}\leq|1+z|^2$ if $|\operatorname{Re}z|\leq\lambda|\operatorname{Im}z|$ with $0<\lambda<1$ and $|z|\leq r_0$ with r_0 depending on λ. Thus if r_0 is appropriately chosen dependent on η we have with a constant C for $\nu=2,3,4$:

$$(2.26)\qquad |I_\nu|\leq\frac{C}{2\pi\sqrt{k}}\int_{\mathfrak{c}_\nu}\left|e^{s\sqrt{k}}\right|\left|y_j^+\left(1-\frac{s}{\sqrt{k}}\right)\right||ds|.$$

For $\nu=3$ we can now use the asymptotic formula (1.6) for y_j^+ and obtain

$$\left|y_j^+\left(1-\frac{s}{\sqrt{k}}\right)\right|\leq\tilde{C}\left|\exp\left(\frac{\lambda_j}{s}\sqrt{k}\right)\left(\frac{s}{\sqrt{k}}\right)^{\alpha_j}\right|$$

for s on $\mathfrak{c}_3$ and $k\in\mathbb{N}$. For $\nu=2,4$ we use the Stokes' relation (1.8) and the asymptotic series (1.7) of the $y^-(z)$ to obtain

$$\left|y_j^+\left(1-\frac{s}{\sqrt{k}}\right)\right|\le\sum_{l\preccurlyeq j}C_l\left|\exp\left(\frac{\lambda_l}{s}\sqrt{k}\right)\left(\frac{s}{\sqrt{k}}\right)^{\alpha_l}\right|$$

for s on $\mathfrak{c}_2$ or $\mathfrak{c}_4$ and $k\in\mathbb{N}$. In all cases we have for $k\in\mathbb{N}$

$$|I_\nu|\le\sum_{l\preccurlyeq j}\tilde{C}_j\sqrt{k}^{\,-\operatorname{Re}\alpha_j-1}\int_{\mathfrak{c}_\nu}\exp\left(\operatorname{Re}\left(s+\frac{\lambda_l}{s}\right)\sqrt{k}\right)|s|^{\operatorname{Re}\alpha_l}|ds|. \tag{2.27}$$

Next we estimate the argument of the exponential term.

(2.28) *Suppose that $\eta>0$ is small enough and $\lambda_j-\lambda_l\in[0,\infty[$. Then for s on $\mathfrak{c}_\nu$, $\nu=2,3$ or 4*

$$\operatorname{Re}\left(s+\frac{\lambda_l}{s}\right)\le\alpha<2\operatorname{Re}\sqrt{\lambda_j}$$

with α independent of ν,s,l.

Proof. First for s,l under consideration we have

$$\operatorname{Re}\left(s+\frac{\lambda_l}{s}\right)\le\operatorname{Re}\left(s+\frac{\lambda_j}{s}\right)+(\lambda_j-\lambda_j)\frac{1}{r}\sin\eta.$$

Hence it suffices to prove (2.28) for $l=j$ and then to reduce $\eta>0$ if necessary. For $l=j$ we put $\sqrt{\lambda_j}=re^{i\psi}$, $|\psi|<\pi/2$. On $\mathfrak{c}_2$ we have $s=re^{i\varphi}$ with $|\varphi+\pi/2|\le\eta$ and (2.19) yields the assertion. On $\mathfrak{c}_3$ we have $s=ti\exp(i\eta)$ with $r\le t\le r_0\sqrt{k}$ and thus

$$\operatorname{Re}\left(s+\frac{\lambda_j}{s}\right)=-t\sin\eta+\frac{r^2}{t}\sin(2\psi-\eta).$$

If $\sin(2\psi-\eta)\le0$ then we can choose $\alpha=0$ else

$$\begin{aligned}\operatorname{Re}\left(s+\frac{\lambda_j}{s}\right)&\le r(-\sin\eta+\sin(2\psi-\eta))\le2r\cos\psi\sin(\psi-\eta)\\&<2r\cos\psi=2\operatorname{Re}\sqrt{\lambda_j}\end{aligned}$$

since $|\psi|<\pi/2$. Similarly (2.28) can be proved for s on $\mathfrak{c}_4$. From (2.27) and (2.28) we get immediately

$$I_\nu=O\left(k^\beta e^{\alpha\sqrt{k}}\right)\quad\text{as }k\to\infty,\quad\nu=2,3,4$$

with some $\beta\in\mathbb{R}$ and $\alpha<2\operatorname{Re}\sqrt{\lambda_j}$. A fortiori for all $n\in\mathbb{N}$

$$\exp\left(-2\sqrt{\lambda_j}\sqrt{k}\right)I_\nu=O(k^{-n})\quad\text{as }k\to\infty,\quad\nu=2,3,4. \tag{2.29}$$

Formulas (2.23) and (2.29) finally yield the assertion of Lemma 2.6.

If we tried to carry out this proof for $\lambda_j\in]-\infty,0]$, we would meet with three difficulties: Both zeros $\pm\sqrt{\lambda_j}$ of p' lie on the imaginary axis and have equal rights, on the circle $|s|=\sqrt{|\lambda_j|}$ we have $\operatorname{Re}p(s)=0$ and an estimate like (2.28) also fails. In the proof of Lemma 2.7 we therefore subdivide the path of integration, deform it in a different way and take the summands from the Stokes' phenomenon into consideration.

Proof of Lemma 2.7. First let $\lambda_j \in]-\infty, 0[$ i.e. $\lambda_j \neq 0$. Then we replace the path $\mathfrak{c}$ of Fig. 1 by the sum of the paths $\mathfrak{c}^+, \mathfrak{c}^-$ outlined in Fig. 3 and obtain

$$d_k^j = -\frac{1}{2\pi i}\int_{\mathfrak{c}^+ + \mathfrak{c}^-} z^{-k-\alpha-1} y_j^+(z)\, dz. \tag{2.30}$$

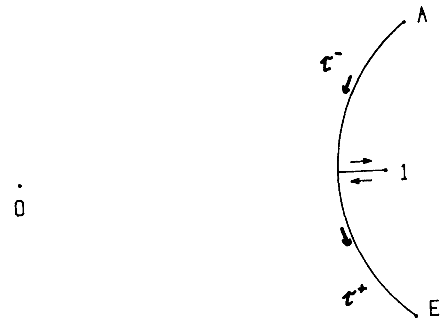


FIG. 3.

We substitute Stokes' relation (1.8) into the integral along $\mathfrak{c}^-$ and get

$$d_k^j = g_k^{j+} + \sum_{l \prec i} \alpha_{jl} g_k^{l-} + g_k^{j-} \tag{2.31}$$

where

$$g_k^{l\pm} = -\frac{1}{2\pi i}\int_{\mathfrak{c}^\pm} z^{-k-\alpha-1} y_l^\pm(z)\, dz. \tag{2.32}$$

By a modification of the proof of Lemma 2.6 we shall show that $g_k^{l\pm}$ have the asymptotic representation given in Lemma 2.7. We only have to do this for g_k^{l+}; the proof for g_k^{l-} is completely analogous.

As in the proof of Lemma 2.6 we put $z = 1 - s/\sqrt{k}$ in (2.32), deform the path of integration and obtain

$$g_k^{l+} = I_1 + I_2, \tag{2.33}$$

$$I_\nu = \frac{1}{2\pi i}\int_{\mathfrak{c}_\nu}\left(1 - \frac{s}{\sqrt{k}}\right)^{-k-\alpha-1} y_l^+\left(1 - \frac{s}{\sqrt{k}}\right)\frac{1}{\sqrt{k}}\, ds,$$

where the paths $\mathfrak{c}_1$ and $\mathfrak{c}_2$ are sketched in Fig. 4. $\mathfrak{c}_1$ is independent of k and parameterized by $s = \rho(\varphi)e^{i\varphi}$, $\varphi \in]0, \pi/2 + \eta]$ where $\rho(\varphi)$ is strictly increasing and $\rho(0+) = 0$, $\rho(\pi/2) = \sqrt{|\lambda_j|}$. $\mathfrak{c}_2$ is a line segment and part of the ray $\arg s = \pi/2 + \eta$, its endpoint is $\tilde{A}_k = r_0\sqrt{k}\, ie^{i\eta}$. We can proceed as in the proof of Lemma 2.6 and only have to show the following three statements:

(2.34) $\operatorname{Re}(s + \lambda_l/s) \leq 0$ *for* s *on* $\mathfrak{c}_1$; *equality holds only for* $s = i\sqrt{|\lambda_j|}$.

(2.35) $\operatorname{Re}(s + \lambda_l/s) \to -\infty$ *as* $s \to 0$ *along* $\mathfrak{c}_1$.

(2.36) *There exists* $\alpha < 0$ *such that* $\operatorname{Re}(s + \lambda_l/s) \leq \alpha$ *for* s *on* $\mathfrak{c}_2$.

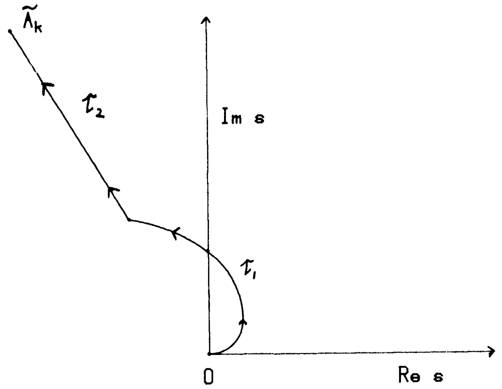


FIG. 4.

The first two assertions follow from

$$\operatorname{Re}\left(s+\frac{\lambda_l}{s}\right)=\left(\rho(\varphi)-\frac{|\lambda_l|}{\rho(\varphi)}\right)\cos\varphi \qquad (s \text{ on } \mathfrak{c}_1)$$

and the choice of $\rho(\varphi)$. For the third we put $s=tie^{i\eta}$ with $t\in[\rho_0, r_0\sqrt{k}]$, $\rho_0=\rho(\pi/2+\eta)$ and obtain

$$\operatorname{Re}\left(s+\frac{\lambda_l}{s}\right)=\left(\frac{|\lambda_l|}{t}-t\right)\sin\eta<\left(\frac{|\lambda_l|}{\rho_0}-\rho_0\right)\sin\eta<0.$$

Now suppose $\lambda_j=0$. Before we can substitute $\mathfrak{c}$ by $\mathfrak{c}^+ + \mathfrak{c}^-$ as in (2.30) we must separate the singular part of y_j^+ to ensure the convergence of the integrals. We choose a sufficiently large $m\in\mathbb{N}$ and put

$$y_j^{\pm}(z)=\sum_{q=0}^{m}(1-z)^{\alpha_j+q}c_j(q)+r_{jm}^{\pm}(z) \tag{2.37}$$

where

$$r_{jm}^{\pm}(z)=o\left((1-z)^{\alpha_j+m}\right) \quad \text{as } H^{\pm}\ni z\to 1. \tag{2.38}$$

By substitution in (2.4) we obtain first

$$d_k^j=-\frac{1}{2\pi i}\sum_{q=0}^{m}\left(\int_{\mathfrak{c}} z^{-k-\alpha-1}(1-z)^{\alpha_j+q}dz\right)c_j(q) \tag{2.39}$$
$$-\frac{1}{2\pi i}\int_{\mathfrak{c}^+ + \mathfrak{c}^-} z^{-k-\alpha-1}r_{jm}^{+}(z)\,dz$$

and then Stokes' relations (1.8) yields as before

$$d_k^j=g_k^{j+}+\sum_{l\prec j}\alpha_{jl}g_k^{l-}, \tag{2.40}$$

where the g_k^{l-} for $l<j$ are given by (2.32) and

$$(2.41)\qquad g_k^{j+} = -\frac{1}{2\pi i}\sum_{q=0}^{m}\int_{\mathfrak{c}} z^{-k-\alpha-1}(1-z)^{\alpha_j+q}dz\,c_j(q)$$

$$-\frac{1}{2\pi i}\int_{\mathfrak{c}^+} z^{-k-\alpha-1}r_{jm}^{+}(z)\,dz-\frac{1}{2\pi i}\int_{\mathfrak{c}^-} z^{-k-\alpha-1}r_{jm}^{-}(z)\,dz.$$

We note that here any sufficiently large $m\in\mathbb{N}$ can be chosen. Completely analogously to the proof of [12, (1.15)] it can be shown that for arbitrary $\alpha,\beta\in\mathbb{C}$

$$(2.42)\quad -\frac{1}{2\pi i}\int_{\mathfrak{c}} z^{-k-\alpha-1}(1-z)^{\beta}dz=\frac{\Gamma(k+\alpha-\beta)}{\Gamma(k+\alpha+1)\Gamma(-\beta)}+O(\rho^{-k})\quad \text{as } k\to\infty$$

and for the proof of the remaining part of Lemma 2.7 we only have to show

$$(2.43)\qquad \int_{\mathfrak{c}^{\pm}} z^{-k-\alpha-1}r_{jm}^{\pm}(z)\,dz=O(k^{-m-\operatorname{Re}\alpha_j-1})\quad \text{as } k\to\infty.$$

The boundedness of $r_{jm}^{\pm}(z)$ first allows us to deform the paths $\mathfrak{c}^+$ and $\mathfrak{c}^-$ into line segments $1E$ and $A1$ and then to estimate $r_{jm}^{\pm}$ there. Insertion of the parametrization of $1A$, $1E$ yields

$$\left|\int_{\mathfrak{c}^{\pm}} z^{-k-\alpha-1}r_{jm}^{\pm}(z)\,dz\right|\le M\int_0^{r_0}\left|1-tie^{i\eta}\right|^{-k-\operatorname{Re}\alpha-1}t^{\operatorname{Re}\alpha_j+m}dt.$$

Since $|1-tie^{i\eta}|\ge 1+t\sin\eta$ the substitution $s=1+t\sin\eta$ gives

$$\left|\int_{\mathfrak{c}^{\pm}} z^{-k-\alpha-1}r_{jm}^{\pm}(z)\,dz\right|\le \tilde{M}\int_1^{\infty}s^{-k-\operatorname{Re}\alpha-1}(s-1)^{\operatorname{Re}\alpha_j+m}ds$$

with $\tilde{M}$ independent of k. Insertion of $s=\tau^{-1}$ yields the first Euler integral and thus

$$\int_{\mathfrak{c}^{\pm}} z^{-k-\alpha-1}r_{jm}^{\pm}(z)\,dz=O\left(\frac{\Gamma(k+\operatorname{Re}\alpha-\operatorname{Re}\alpha_j-m)}{\Gamma(k+\operatorname{Re}\alpha+1)}\right).$$

The Stirling formula finally furnishes the assertion (2.43).

3. Limit formulas for some of the connection coefficients. We begin this section by stating which summands in Theorem 2.8 are essential for the asymptotic behaviour of the d_k. The limit formula for one of the connection coefficients γ_j^+ that we obtain here will be extended such that all connection coefficients which are essential for the behaviour of $y_0(z)$ as $H\ni z\to 1$ can be computed.

From Lemma 2.6 we see that d_k^j grows more rapidly than d_k^l for large k iff $\operatorname{Re}\sqrt{\lambda_j}$ is larger than $\operatorname{Re}\sqrt{\lambda_l}$. The k-powers in the asymptotic formulas are inessential here because $\exp(\alpha\sqrt{k})$ for $\operatorname{Re}\alpha>0$ grows quicker than all powers of k. On the other hand, if $\operatorname{Re}\sqrt{\lambda_l}$ and $\operatorname{Re}\sqrt{\lambda_j}$ are equal then the leading terms only differ by an oscillatory factor $\exp(i\beta\sqrt{k})$ and neither series can be neglected. This leads to the following:

THEOREM 3.1. *If there exists* $j\in\{1,\cdots,n\}$ *such that* $\lambda_j\in\mathbb{C}\setminus]-\infty,0]$ *and* $\gamma_j^+\neq 0$, *then*

$$\alpha:=\max\left\{\operatorname{Re}\sqrt{\lambda_j}\,\middle|\,\gamma_j^+\neq 0\right\}>0,$$

and with

$$\mathfrak{M} := \left\{ j \middle| \gamma_j^+ \neq 0,\ \operatorname{Re}\sqrt{\lambda_j} = \alpha \right\} \neq \varnothing$$

the following assertions hold:

1) $\gamma_j^+ = \gamma_j^-$ *for* $j \in \mathfrak{M}$;

2) *for* $j \in \mathfrak{M}$ *there exist sequences* $(d_k^j)_{k\in\mathbb{N}}$ *such that*

$$d_k = \sum_{j\in\mathfrak{M}} \gamma_j^+ d_k^j \qquad (k \in \mathbb{N}),$$

and

$$d_k^j \sim \exp\left(2\sqrt{\lambda_j}\sqrt{k}\right) k^{-\alpha_j/2-3/4} \sum_{m=0}^{\infty} k^{-m/2} a_m^j \qquad (k\to\infty),$$

where

$$a_m^j = \frac{1}{2\sqrt{\pi}} e^{\lambda_j/2} \lambda_j^{\alpha_j/2+1/4} e_j.$$

Proof. Assertion 1) follows from (1.10) and the fact that $j\in\mathfrak{M}$ and $j \prec l$ imply $\gamma_l^+ = 0$. 2) is an immediate consequence of the preliminary remarks.

If $\gamma_j^+ \neq 0$ in (1.9) only for j with $\lambda_j \in]-\infty, 0]$, then in (2.8) the first sum vanishes and, similar to the above proof, (1.10) yields $\tilde{\gamma}_j = \gamma_j^-$ in the third sum. Thus the assertion becomes symmetric in γ_j^+ and γ_j^- in this case, too.

THEOREM 3.2. *If* $\gamma_j^+ = 0$ *for all* j *such that* $\lambda_j \in \mathbb{C} \setminus]-\infty, 0]$ *then*

$$d_k = \sum_{\lambda_j \in]-\infty,0]} \left(\gamma_j^+ g_k^{j+} + \gamma_j^- g_k^{j-}\right) + d_k^0,$$

where the asymptotic behaviour of the sequences $(g_k^{j\pm})_{k\in\mathbb{N}}$ *and* $(d_k^0)_{k\in\mathbb{N}}$ *is given in Lemma* 2.7 *and* (2.5).

If $\operatorname{Re}\sqrt{\lambda_j}$ attains its maximum only at one j then Theorem 3.1 can be written as a limit formula for γ_j^+.

COROLLARY 3.3. *Suppose that* $\lambda_j \in \mathbb{C} \setminus]-\infty, 0]$ *and*

$$\gamma_l^+ = 0 \quad \textit{or} \quad \operatorname{Re}\sqrt{\lambda_l} < \operatorname{Re}\sqrt{\lambda_j} \qquad (l \neq j).$$

Then $\gamma_j^+ = \gamma_j^-$ *and*

$$\gamma_j^+ e_j = 2\sqrt{\pi} e^{-\lambda_j/2} \lambda_j^{-\alpha_j/2-1/4} \lim_{k\to\infty} \exp\left(-2\sqrt{\lambda_j}\sqrt{k}\right) k^{\alpha_j/2+3/4} d_k.$$

Moreover the above sequence with limit $\gamma_j^+ e_j$ *has an asymptotic series, involving powers of* $1/\sqrt{k}$.

Different from the result of [12] for regular singular points, not all connection coefficients can be computed from the (known) coefficients d_k of the power series $y_0(z)$, in general only one. In order to determine some more of them by limit formulas we investigate how a transformation

$$y(z) = \exp\left(\frac{\lambda}{1-z}\right)\tilde{y}(z)$$

of (1.1) changes Corollary 3.3.

The Floquet solution at 0 is transformed into

$$\tilde{y}_0(z)=\exp\left(-\frac{\lambda}{1-z}\right)y_0(z)=z^{\alpha}\sum_{k=0}^{\infty}z^k\tilde{d}_k(\lambda) \tag{3.4}$$

where the coefficients $\tilde{d}_k(\lambda)$ may be computed in the same way as the coefficients d_k of y_0 by substitution in (1.1), B now replaced by $B-\lambda$. It is easily shown, too, that

$$\tilde{y}_j^{\pm}(z)=\exp\left(-\frac{\lambda}{1-z}\right)y_j^{\pm}(z) \qquad (z\in H^{\pm}) \tag{3.5}$$

are the solutions of the transformed equation having asymptotic series as $H^{\pm}\ni z\to 1$. Thus in (1.4) all λ_j have to be replaced by $\lambda_j-\lambda$. In the transformed form of (1.9), the connection coefficients are the same

$$\tilde{y}_0(z)=\sum_{j=1}^{n}\gamma_j^{\pm}\tilde{y}_j^{\pm}(z) \qquad (z\in H, |z|<1, |\arg z|<\pi/2). \tag{3.6}$$

Now we apply Corollary 3.3 to the transformed equation and get

THEOREM 3.7. *Suppose that the assumptions of* §1 *and* (3.4) *hold. Let* $j\in\{1,\cdots,n\}$. *Assume that there exists* $\lambda\in\mathbb{C}$ *with* $\lambda_j-\lambda\notin\,]-\infty,0]$ *such that*

$$\gamma_l=0 \quad or \quad \mathrm{Re}\sqrt{\lambda_l-\lambda}<\mathrm{Re}\sqrt{\lambda_j-\lambda} \qquad (l\in\{1,\cdots,n\}\setminus\{j\}).$$

Then $\gamma_j^+=\gamma_j^-$ *and*

$$\gamma_j^+e_j=2\sqrt{\pi}\exp\left(-\frac{1}{2}(\lambda_j-\lambda)\right)(\lambda_j-\lambda)^{-\alpha_j/2-1/4}$$
$$\times\lim_{k\to\infty}\exp\left(-2\sqrt{\lambda_j-\lambda}\sqrt{k}\right)k^{\alpha_j/2+3/4}\tilde{d}_k(\lambda).$$

Moreover the above sequence with limit $\gamma_j^+e_j$ *has an asymptotic series, involving powers of* $1/\sqrt{k}$.

Next the question arises which of the γ_j^+ can be determined with the aid of Theorem 3.7, i.e. how the corresponding y_j^+ can be characterized.

PROPOSITION 3.8. *Let* $j\in\{1,\cdots,n\}$. *Then* (1) *and* (2) *are equivalent.*

(1) *There exists* $\lambda\in\mathbb{C}$ *such that*

$$\mathrm{Re}\sqrt{\lambda_j-\lambda}>\max\left\{\mathrm{Re}\sqrt{\lambda_l-\lambda}\,\middle|\,l\neq j,\ \gamma_l^+\neq 0\right\}.$$

(2) *There exists* $\varphi\in\,]-\pi/2,\pi/2[$ *such that*

$$\mathrm{Re}\left(\lambda_je^{-i\varphi}\right)>\max\left\{\mathrm{Re}(\lambda_le^{-i\varphi})\,|\,l\neq j,\ \gamma_l^+\neq 0\right\}.$$

Remark. (2) implies that in a sector

$$S_{\varphi}(\varepsilon)=\{z\,|\,|\arg(1-z)-\varphi|<\varepsilon, |z-1|<r-1\},$$

$\varepsilon>0$ sufficiently small, for $l\neq j$

$$\gamma_l^+=0 \text{ or } \exp\left(-\lambda_j\frac{1}{1-z}\right)y_l^{\pm}(z)=o((1-z)^n) \qquad (S_{\varphi}(\varepsilon)\ni z\to 1, n\in\mathbb{N}).$$

Thus (2) means, that y_j^+ in some subsector or the semicircle $|z-1|<r-1$, $\mathrm{Re}\,z<1$ is dominant among those solutions $y_l^{\pm}$ of (1.1) such that $\gamma_l^+\neq 0$. So aside from some exceptions we can compute $\gamma_j^\pm$ by Lemma 3.7 exactly for j such that the summand $\gamma_j^+ y_j^+$ in (1.9) is relevant for the asymptotic behaviour of $y_0(z)$ as $H \ni z \to 1$. The exceptions only occur if three or more λ_j lie on one straight line.

Proof of Proposition 3.8. By the well-known formula

$$\operatorname{Re}\sqrt{\lambda_j-\lambda}=\sqrt{\frac{1}{2}\left(|\lambda_j-\lambda|+\operatorname{Re}(\lambda_j-\lambda)\right)}\,,$$

(1) is equivalent to

(1′) There exists $\lambda \in \mathbf{C}$ such that

$$|\lambda_j-\lambda|+\operatorname{Re}\lambda_j>\max\left\{|\lambda_l-\lambda|+\operatorname{Re}\lambda_l \mid l\neq j,\ \gamma_l^+\neq 0\right\}.$$

To prove that (1′) implies (2) we put[1]

$$\lambda_j-\lambda=Re^{i\psi} \quad \text{with } R>0,\ \psi\in\,]-\pi,\pi[\tag{3.9}$$

and for $l\neq j$ such that $\gamma_l^+\neq 0$

$$\lambda_j-\lambda_l=r_l e^{i\psi_l} \quad \text{with } r_l>0,\ \psi_l\in[-\pi,\pi[. \tag{3.10}$$

Then for $l\neq j$ with $\gamma_l^+\neq 0$ we get

$$|\lambda_l-\lambda|\geq R-r_l\cos(\psi_l-\psi)$$

and using the addition formula for the cosine (1′) implies

$$0>|\lambda_l-\lambda|+\operatorname{Re}\lambda_l-|\lambda_j-\lambda|-\operatorname{Re}\lambda_j\geq -2r_l\cos\left(\psi_l-\frac{\psi}{2}\right)\cos\frac{\psi}{2}.$$

This means, that for $l\neq j$ and $\gamma_l^+\neq 0$

$$\operatorname{Re}\left((\lambda_j-\lambda_l)e^{-i\psi/2}\right)=r_l\cos\left(\psi_l-\frac{\psi}{2}\right)>0$$

and (2) is proved for $\varphi=\psi/2$.

For the proof that (2) implies (1′) we conversely choose λ such that (3.9) holds for $\psi=2\varphi$. With (3.10) we have here

$$|\lambda_l-\lambda|+\operatorname{Re}\lambda_l-|\lambda_j-\lambda|-\operatorname{Re}\lambda_j=|R-r_l e^{i(\psi_l-2\varphi)}|-R-r_l\cos\psi_l.$$

Now as $R\to\infty$

$$|R-r_l e^{i(\psi_l-2\varphi)}|=R-r_l\cos(\psi_l-2\varphi)+O\left(\frac{1}{R}\right)$$

and thus as before

$$\begin{aligned}|\lambda_l-\lambda|+\operatorname{Re}\lambda_l-|\lambda_j-\lambda|-\operatorname{Re}\lambda_j&=-2r_l\cos\varphi\cos(\psi_l-\varphi)+O\left(\frac{1}{R}\right)\\&=-2\cos\varphi\operatorname{Re}\left[(\lambda_j-\lambda_j)e^{-i\varphi}\right]+O\left(\frac{1}{R}\right).\end{aligned}$$

[1] In the case $\lambda_j-\lambda\in\,]-\infty,0]$ we have $\operatorname{Re}\sqrt{\lambda_j-\lambda}=0$ and because of (1) γ_l^+ vanishes for all $l\neq j$. Hence φ in (2) is arbitrary here.

Because of $\varphi\in]-\pi/2,\pi/2[$ and (2) this implies (1′), if R has been chosen sufficiently large.

4. Application to the generalized Heun equation. If the results are applied to differential equations of second order with rational coefficients, the computation of the connection coefficients is simplified considerably. We demonstrate this for the generalized Heun equation of [12]

$$(4.1)\qquad y''(z)+\left(\frac{1-\mu_0}{z}+\frac{1-\mu_1}{z-1}+\frac{1-\mu_2}{z-a}+\alpha\right)y'(z)+\frac{\beta_0+\beta_1 z+\beta_2 z^2}{z(z-1)(z-a)}y(z)=0$$

with $a\in\mathbb{C}\setminus\{0,1\}$, $\alpha\in\mathbb{C}\setminus\{0\}$ and further complex parameters μ_j,β_j. Formula (4.1) has singular points of the first kind in 0, 1 and a; the corresponding indices are 0 and μ_0, 0 and μ_1, 0 and μ_2, respectively. ∞ is an irregular singular point of (4.1) or rank 1. In the first part we shall transform the connection problem between 1 and ∞ by $z=1/(1-t)$ to achieve the form (1.1) and solve it if $\operatorname{Re}a<\frac{1}{2}$. Then we explain how the connection coefficients between a finite singular point and ∞ can be computed if those between the finite singular points are known from [12, §3].

Because $\alpha\neq 0$ — this is the assumption on (4.1) corresponding to (1.2) – (4.1) has formal solutions belonging to ∞ of the form

$$(4.2)\qquad \hat{y}_1(z)=z^{\mu_{\infty 1}}\left(1+\sum_{k=1}^{\infty}c_1(k)z^{-k}\right),$$

$$\hat{y}_2(z)=e^{-\alpha z}z^{\mu_{\infty 2}}\left(1+\sum_{k=1}^{\infty}c_2(k)z^{-k}\right)$$

and insertion in (4.1) yields

$$(4.3)\qquad \mu_{\infty 1}=-\frac{\beta_2}{\alpha},\qquad \mu_{\infty 2}=\mu_0+\mu_1+\mu_2-3+\frac{\beta_2}{\alpha}.$$

More exactly then in general we know for second order equations like (4.1) (cf. [9, Chap. 7, §2]) that there exist uniquely determined solutions $y_1(z)$ and $y_2(z)$ defined in

$$(4.4)\qquad S_1=\left\{z\,\middle|\,|z|>\max(1,|a|),\ \arg(-\alpha)-\frac{3}{2}\pi<\arg z<\arg(-\alpha)+\frac{3}{2}\pi\right\},$$

$$S_2=\left\{z\,\middle|\,|z|>\max(1,|a|),\ \arg\alpha-\frac{3}{2}\pi<\arg z<\arg\alpha+\frac{3}{2}\pi\right\},$$

where $\arg(\pm\alpha)$ are chosen in $[-\pi,\pi[$, such that

$$(4.5)\qquad y_j(z)\sim\hat{y}_j(z)\qquad (S_j\ni z\to\infty,\ j=1,2).$$

From [12] we know that (4.1) has exactly one solution $y_0(z)$ holomorphic near 1 with $y_0(1)=1$ if $\mu_1\notin\mathbb{N}$, which can be analytically continued to a neighborhood of $[1,\infty[$. Then the following connection problem arises

$$(4.6)\qquad y_0(z)=\gamma_1 y_1(z)+\gamma_2 y_2(z)\qquad (z\in]1,\infty[,\ z\text{ sufficiently large}).$$

With the aid of the transformation $y(z)=(z-1)^{\mu_1}\tilde{y}(z)$ it can be shown that the connection coefficients of the solution to the index μ_1 near 1 can be obtained by computing γ_1 and γ_2 for the problem (4.1)-(4.6) with altered coefficients μ_1,β_0,β_1 and β_2. Analogously the transformation $y(z)=e^{-\alpha z}\tilde{y}(z)$ shows that a limit formula for γ_2 is sufficient because then γ_1 can be obtained by replacing α by $-\alpha$ and β_i by $\tilde{\beta}_i$, which come from the transformed equation. For either transformation the form [12, (3.1)] of the coefficient of y is particularly useful.

In order to apply the results of §3 to the determination of γ_2 we use the transformations

$$z^{-\mu_{\infty 1}}y(z)=v(t),\qquad z=\frac{1}{1-t}. \tag{4.7}$$

The factor $z^{-\mu_{\infty 1}}$ serves to simplify the recursion for the computation of γ_2 to a four-term recursion. v satisfies the differential equation

$$v''(t)+\left[\frac{\alpha}{(t-1)^2}+\frac{\mu_{\infty 2}-\mu_{\infty 1}}{t-1}+\frac{1-\mu_1}{t}+\frac{1-\mu_2}{t-\tilde{a}}\right]v'(t)$$
$$+\frac{\tilde{\beta}_0+\tilde{\beta}_1 t+\tilde{\beta}_2 t^2}{at(t-\tilde{a})(t-1)^2}v(t)=0 \tag{4.8}$$

with $\tilde{a}=1-1/a$ and

$$\begin{aligned}\tilde{\beta}_0&=\beta_0+\beta_1+\beta_2-\mu_{\infty 1}(1-\mu_1)(a-1),\\ \tilde{\beta}_1&=-\beta_0+a\beta_2+\mu_{\infty 1}[(a-1)(\mu_0-\mu_{\infty 1})+a(1-\mu_1)+1-\mu_2],\\ \tilde{\beta}_2&=a\mu_{\infty 1}(\mu_{\infty 1}-\mu_0).\end{aligned} \tag{4.9}$$

The singular points $z=0,1,a,\infty$ are mapped to $t=\infty,0,1-1/a,1$ respectively. 0 and 1 are neighboring, i.e., (4.8), aside from its scalar notation, has the form (1.1) if $|1-1/a|>1$ (i.e. $\operatorname{Re} a<\frac{1}{2}$).

$y_0(z)$ is transformed into the solution $v_0(t)$ of (4.8) holomorphic near 1 with $v_0(1)=1$, which we write

$$v_0(t)=\sum_{k=0}^{\infty} d_k t^k,\qquad d_0=1. \tag{4.10}$$

Insertion in (4.8) supplies a four-term recursion for d_k,

$$\begin{aligned}&\left(1-\frac{1}{a}\right)(k+1)(k+1-\mu_1)d_{k+1}\\ &\quad=\varphi_0(k)d_k+\varphi_1(k-1)d_{k-1}+\varphi_2(k-2)d_{k-2}\qquad (k\in\mathbb{N}),\\ &\qquad\qquad d_0=1,\qquad d_{-1}=d_{-2}:=0,\end{aligned} \tag{4.11}$$

where

$$\begin{aligned}\varphi_0(k)&=\left[\left(3-\frac{2}{a}\right)(k-\mu_1)+\left(1-\frac{1}{a}\right)(\mu_{\infty 2}-\mu_{\infty 1}-\alpha+2)+1-\mu_2\right]k+\tilde{\beta}_0,\\ \varphi_1(k)&=\left[\left(\frac{1}{a}-3\right)(k-\mu_1)+\alpha+\left(\frac{1}{a}-2\right)(\mu_{\infty 2}-\mu_{\infty 1}+2)-2-2\mu_2\right]k+\tilde{\beta}_1,\\ \varphi_2(k)&=(k+\mu_0-\mu_{\infty 1})(k-\mu_{\infty 1}).\end{aligned} \tag{4.12}$$

$y_1(z)$ and $y_2(z)$ pass into local solutions $v_1^+(t)$ and $v_2^+(t)$ of (4.8) at the irregular singular point 1:

$$(4.13)\qquad v_1^+(t)\sim 1+\sum_{k=1}^{\infty} c_1(k)(1-t)^k |\arg(1-t)+\arg(-\alpha)|<\frac{3}{2}\pi, \qquad t\to 1,$$

$$v_2^+(t)\sim \exp\left(-\frac{\alpha}{1-t}\right)(1-t)^{\mu_{\infty 1}-\mu_{\infty 2}}\left(1+\sum_{k=1}^{\infty} c_2(k)(1-t)^k\right)$$

$$|\arg(1-t)+\arg\alpha|<\frac{3}{2}\pi, \qquad t\to 1.$$

Both of the asymptotic sectors contain H^+ of §1; hence v_j^+ is exactly the solution fixed in §1.

Furthermore the connection relation (4.6) corresponds to

$$(4.14)\qquad v_0(z)=\gamma_1 v_1^+(z)+\gamma_2 v_2^+(z) \qquad (z\in]0,1[).$$

Corollary 3.3 immediately applies to (4.8)–(4.14) and yields

THEOREM 4.15. *Let γ_2 be the connection coefficient for the connection problem* (4.1)–(4.6). *Assume that $\mu_1\notin\{1,2,3,\cdots\}$, $\alpha\notin[0,\infty[$ and* Re$a<\frac{1}{2}$. *Let the sequence d_k be determined by the recursion formula* (4.11). *Then*

$$\gamma_2=2\sqrt{\pi}\, e^{\alpha/2}(-\alpha)^{(\mu_{\infty 2}-\mu_{\infty 1}-2)/2}\lim_{k\to\infty}\exp\left(-2\sqrt{-\alpha}\sqrt{k}\right)k^{3/4+(\mu_{\infty 1}-\mu_{\infty 2})/2}d_k.$$

Moreover the above sequence with limit γ_2 has an asymptotic expansion, involving powers of $1/\sqrt{k}$.

Below (4.6) we already explained how all connection coefficients between 1 and ∞ could be computed if a formula for γ_2 were known. Now (4.15) makes this possible if $\mu_1\notin\mathbb{Z}$, $\alpha\notin\mathbb{R}$ and Re$a<\frac{1}{2}$.

With the connection coefficients between the finite singular points already known from [12] we can determine all connection coefficients between 0 and ∞ and between a and ∞, too, if additionally $\mu_0\notin\mathbb{Z}$, $\mu_2\notin\mathbb{Z}$, resp., and $|a-1|\neq 1$.

If Re$a>\frac{1}{2}$, however, we first transform $z=1-\tilde{z}$ and then use the above results. If Re$a=\frac{1}{2}$ and $|a|<1$, then we first transform $z=a\tilde{z}$ and, because of Re $1/a>\frac{1}{2}$, we obtain all connection coefficients, too. More information about these transformations is contained in [12, §3].

Finally we remark that the above conditions (except $a\neq 0,1$ and $\alpha\neq 0$) are not essential. If $\alpha\in\mathbb{R}\setminus\{0\}$ then we use (3.2) and (3.3). For the excluded a we can substitute $z=(1-\delta t)/(1-t)$ in (4.1) and again obtain the connection coefficients, but the recursion formula has more terms. If some $\mu_j\in\mathbb{Z}$ we could use a reduction method similar to [13], but that would be much work. Even the case $\alpha=0$, $\beta_2\neq 0$ in (4.1), where subnormal formal solutions occur, can be treated by $z=t^2$ with success.

REFERENCES

[1] I. BAKKEN, *On the central connection problem for a class of ordinary differential equations, Part* I, II, Funk. Ekvac., 20 (1977), pp. 115–127, 129–156.

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

[3] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1971.

[4] W. B. JURKAT, *Meromorphe Differentialgleichungen*, Springer, Berlin-Heidelberg, 1978.

[5] M. KOHNO, *A Two-point connection problem*, Hiroshima Math. J., 9 (1979), pp. 61–135.

[6] ———, *A multi-point connection problem*, to appear.

[7] J. MEIXNER AND F. W. SCHÄFKE, *Mathieusche Funktionen und Sphäroidfunktionen*, Springer, Berlin, 1954.

[8] K. OBUBO, *A global representation of a fundamental set of solutions and a Stokes phenomenon for a system of linear ordinary differential equations*, J. Math. Soc. Japan, 15 (1963), pp. 268–288.

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

[10] O. PERRON, *Über das Verhalten einer ausgearteten hypergeometrischen Reihe bei unbegrenztem Wachstum eines Parameters*, J. f. Math., 151 (1921), pp. 63–79.

[11] R. SCHÄFKE, *Uber das globale analytische Verhalten der Lösungen der über die Laplacetransformation zusammenhängenden Differentialgleichungen* $tx'=(A+tB)x$ *und* $(s-B)v'=(\rho-A)v$, Dissertation, Essen, 1979.

[12] R. SCHÄFKE AND D. SCHMIDT, *The central connection problem for general linear ordinary differential equations at two regular singular points with applications in the theory of special functions*, this Journal, 11 (1980), pp. 848–862.

[13] ———, *The connection problem for two neighboring regular singular points of general linear complex ordinary differential equations*, this Journal, 11 (1980), pp. 863–875.

[14] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, John Wiley, New York, 1965.

[15] W. WYRWICH, *Eine explizite Lösung des "Central Connection Problem" für eine gewöhnliche lineare Differentialgleichung n-ter Ordnung mit Polynomkoeffizienten*, Dissertation, Dortmund, 1974.

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics

0036-1410/84/1502-0004 $01.25/0

ON BOUNDARY VALUE PROBLEMS FOR HAMILTONIAN SYSTEMS WITH TWO SINGULAR POINTS*

D. B. HINTON† AND J. K. SHAW‡

Abstract. A linear Hamiltonian system of differential equations is considered on an open interval (a,b) where both a and b are singular points. A Green's function is defined by a limit of such functions of regular problems. It is proved that solutions of the differential equations defined by the Green's function satisfy Titchmarsh's λ-dependent boundary conditions at the singular points. A formula linking the Titchmarsh–Weyl matrix m-coefficient to certain square integrable solutions is established for separated boundary conditions.

1. Introduction. Boundary value problems with two singular endpoints occur in many physical problems. An important example is the radial equation for the hydrogen atom which has singular points at 0 and ∞. Recent contributions to the second order scalar case have been given by Krall [15] and Burnap, Greenburg, and Zweifel [3]. We consider two singular endpoint problems for the $2n\times 2n$ Hamiltonian system

$$(1.1)\qquad Jy'=[\lambda A(x)+B(x)]y,\qquad a<x<b,$$

where y is a $2n$-vector and λ is a complex parameter.

It is our purpose here to develop a theory of λ-dependent boundary conditions for (1.1) which parallels the Titchmarsh theory [20] for the second-order scalar equation. In so doing we will show the equivalence of the limit-point definition used in [7] with the usual one (see below) and extend to (1.1) the form of the Green's function and characteristic matrix used in the second order scalar case.

The coefficients A, B, and J satisfy:

(1.2) $A(x)$ and $B(x)$ are $2n\times 2n$ Hermitian matrices of locally Lebesgue integrable functions, $A(x)\geq 0$, and

$$J=\begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix},$$

where I_n is the $n\times n$ identity matrix. A solution of (1.1) is said to be of integrable square if $\int_a^b y^*Ay<\infty$, and we denote this by $y\in\mathcal{L}_A^2(a,b)$ or simply $y\in\mathcal{L}_A^2$. We also assume Atkinson's definiteness condition [1, p. 253], i.e., if y is a nontrivial solution of (1.1), then

$$(1.3)\qquad \int_c^d y^*Ay>0\quad \text{for all } a<c<d<b.$$

We allow the endpoints a and b to be finite or infinite.

The basic theory of (1.1) may be found in Atkinson [1, Chap. 9] and Kogan and Rofe–Beketov [14]. We recall now some of this material. The regular boundary value problems associated with (1.1) are of the form

$$(\mathrm{R})\qquad \begin{aligned} &Jy'=[\lambda A(x)+B(x)]y+f, && c\leq x\leq d,\\ &y(c)=N_1v, && y(d)=N_2v \end{aligned}$$

*Received by the editors February 16, 1982.

†Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996. The research of this author was supported in part by the National Science Foundation under grant MCS-8101712.

‡Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. The research of this author was supported in part by the National Science Foundation under grant MCS-8101536.

where $\mathbf{v}$ is a $2n$-vector and N_1 and N_2 are $2n\times 2n$ matrices such that

$$N_1^*JN_1=N_2^*JN_2, \qquad N_1\mathbf{v}=N_2\mathbf{v}=0\Rightarrow\mathbf{v}=0. \tag{1.4}$$

The number λ is called an eigenvalue of (R) if for $\mathbf{f}=0$ there is a nontrivial $\mathbf{y}$ satisfying (R). The symmetry condition (1.4) implies that all eigenvalues are real.

If λ is not an eigenvalue of (R) and $\mathbf{f}$ is Lebesgue integrable, then (R) has a unique solution $\mathbf{y}$ given by

$$\mathbf{y}(x)=\int_c^d K(x,t,\lambda)\mathbf{f}(t)\,dt \tag{1.5}$$

where

$$K(x,t,\lambda)=\begin{cases} Y(x,\lambda)\big[F(\lambda)J+(1/2)I\big]Y(t,\lambda)^{-1}J^{-1}, & x\le t,\\ Y(x,\lambda)\big[F(\lambda)J-(1/2)I\big]Y(t,\lambda)^{-1}J^{-1}, & x>t,\end{cases} \tag{1.6}$$

$$Y \text{ is the fundamental matrix for (1.1) with } Y(c,\lambda)=I,$$

and

$$F=F(\lambda)=Y(d,\lambda)^{-1}N_2\big[N_1-Y(d,\lambda)^{-1}N_2\big]^{-1}J^{-1}+(1/2)J^{-1}. \tag{1.7}$$

F is the characteristic function of Atkinson. The symmetry condition (1.4) implies for $\operatorname{Im}\lambda\neq 0$ and $\mathbf{f}=A\mathbf{g}$, where $\mathbf{g}\in\mathfrak{L}_A^2(c,d)$, that

$$\int_c^d \mathbf{y}^*A\mathbf{y}\le(\operatorname{Im}\lambda)^{-2}\int_c^d \mathbf{g}^*A\mathbf{g}$$

when $\mathbf{y}$ is given by (1.5). (See the proof of [14, Lemma 2.1].)

The matrix function F lies on the locus

$$\begin{aligned}\big[F+(1/2)J^{-1}\big]^*(J/i)\big[F+(1/2)J^{-1}\big]&\\ =\big[F-(1/2)J^{-1}\big]^*\big(Y(d,\lambda)^*JY(d,\lambda)/i\big)&\big[F-(1/2)J^{-1}\big].\end{aligned} \tag{1.8}$$

To consider two singular endpoints, it is convenient to introduce a fundamental matrix Z of (1.1) where $Z(e,\lambda)=I$ with e fixed as $c\to a$ and $d\to b$. If we define

$$\tilde F=\tilde F(\lambda)=Z(c,\lambda)^{-1}F(\lambda)JZ(c,\lambda)J^{-1}, \tag{1.9}$$

then the formula for K in (1.6) becomes

$$K(x,t,\lambda)=\begin{cases} Z(x,\lambda)\big[\tilde FJ+(1/2)I\big]Z(t,\lambda)^{-1}J^{-1}, & x\le t,\\ Z(x,\lambda)\big[\tilde FJ-(1/2)I\big]Z(t,\lambda)^{-1}J^{-1}, & x>t.\end{cases} \tag{1.10}$$

The matrix function $\tilde F$ satisfies the inequality (with equality holding for $\tilde F$ given by (1.9))

$$\begin{aligned}\big[\tilde F+(1/2)J^{-1}\big]^*\big(Z^*(c,\lambda)^{-1}JZ(c,\lambda)^{-1}/i\big)\big[\tilde F+(1/2)J^{-1}\big]&\\ \ge\big[\tilde F-(1/2)J^{-1}\big]^*\big(Z^*(c,\lambda)^{-1}Z^*(d,\lambda)JZ(d,\lambda)Z(c,\lambda)^{-1}/i\big)&\\ \cdot\big[\tilde F-(1/2)J^{-1}\big].&\end{aligned} \tag{1.11}$$

Further the set $\mathfrak{T}_{cd}$ of matrices $\tilde{F}$ satisfying (1.11) is nested, i.e., $\mathfrak{T}_{c'd'} \subset \mathfrak{T}_{cd}$ if $c' < c < d < d'$ and is bounded independently of c, d, and λ, when λ is restricted to a compact set not intersecting the real axis. This yields that there are functions $\tilde{F}_\infty(\lambda)$, defined and analytic for $\operatorname{Im}\lambda \neq 0$, which are sequential limits of functions $\tilde{F} = \tilde{F}(c, d, N_1, N_2, \lambda)$.

If now K_∞ is defined by replacing $\tilde{F}$ with $\tilde{F}_\infty$ in (1.10), then by the proof of [14, Lemma 2.1] (which treats one singular point)

(i) $$\int_a^b K_\infty^*(x,t,\lambda)A(t)K_\infty(x,t,\lambda)\,dt < \infty;$$

(ii) if $\mathbf{f} \in \mathfrak{L}_A^2(a,b)$ and

$$\mathbf{y}(x) = \int_a^b K_\infty(x,t,\lambda)A(t)\mathbf{f}(t)\,dt, \tag{1.12}$$

then $\mathbf{y} \in \mathfrak{L}_A^2(a,b)$, $\mathbf{y}$ satisfies $J\mathbf{y}' = (\lambda A + B)\mathbf{y} + A\mathbf{f}$, and

$$\int_a^b \mathbf{y}^* A \mathbf{y} \leq (\operatorname{Im}\lambda)^{-2} \int_a^b \mathbf{f}^* A \mathbf{f}. \tag{1.13}$$

We consider now two questions. The first is what boundary conditions does $\mathbf{y}$ defined by (1.12) satisfy at a and b and in what sense is y unique? The second is can the singular structure of $\tilde{F}_\infty$ for separated boundary conditions be determined from the singular structure of two boundary value problems, one on $(a, e]$ and the other on $[e, b)$. In particular, can the formula for the Green's function for the 2nd order scalar equation (cf. [20, p. 42]) be extended. We consider these questions when N_1 and N_2 above represent separated boundary conditions. For the first question we allow only the limit-point or limit-circle case. To define these set

$$N(\lambda) = \dim\{\mathbf{y} \in \mathfrak{L}_A^2 : y \text{ satisfies } (1.1)\}.$$

Then $N(\lambda)$ is constant in $\operatorname{Im}\lambda > 0$ and in $\operatorname{Im}\lambda < 0$ [14]. In analogy to the classical case considered by Weyl [21], we call $N(i) = N(-i) = n$ the *limit-point* case and $N(i) = N(-i) = 2n$ the *limit-circle* case.

In §§2 and 3 below we develop the necessary theory to answer these questions. The boundary value problems are discussed in §4. Theorem 2.1 is the extension to the system (1.1) of [20, Lemma 2.3, p. 26], and Lemma 4.1 shows that $\mathbf{y}$ given by (1.12) satisfies boundary conditions of Titchmarsh's λ-dependent form (cf. [20, p. 31]). Note that even in the limit-point case when no boundary conditions are imposed to obtain a self-adjoint operator, the function $\mathbf{y}$ still satisfies λ-dependent boundary conditions.

Independently of the work of Titchmarsh, Kodaira [13] developed a theory of eigenfunctions expansions for second-order scalar equations based on Hilbert space methods. We mention also that Kim [12] has developed an eigenfunction expansion theory for singular Hamiltonian systems in the limit-point case.

Finally we note a useful identity for (1.1). If

$$J\mathbf{y}' = [\lambda A(x) + B(x)]\mathbf{y} + \mathbf{f}$$

and

$$J\mathbf{z}' = [\mu A(x) + B(x)]\mathbf{z} + \mathbf{g},$$

then

$$(\mathbf{y}^* J \mathbf{z})' = (\mu - \bar{\lambda})\mathbf{y}^* A \mathbf{z} + \mathbf{y}^* \mathbf{g} - \mathbf{f}^* \mathbf{z}. \tag{1.14}$$

2. The Titchmarsh–Weyl coefficient for a singular endpoint. We return to the regular problem (R) with separated boundary conditions. Let α_1, α_2, β_1, β_2 be $n\times n$ matrices such that

$$\text{(2.1)}\qquad \operatorname{rank}[\alpha_1,\alpha_2]=\operatorname{rank}[\beta_1,\beta_2]=n,\quad \alpha_1\alpha_2^*=\alpha_2\alpha_1^*,\quad \beta_1\beta_2^*=\beta_2\beta_1^*.$$

If we define

$$N_1=\begin{pmatrix} 0 & \alpha_2^* \\ 0 & -\alpha_1^* \end{pmatrix},\qquad N_2=\begin{pmatrix} \beta_2^* & 0 \\ -\beta_1^* & 0 \end{pmatrix},$$

then the conditions (1.4) are equivalent to (2.1), and (R) can be written as

$$\text{(R*)}\qquad \begin{gathered} Jy'=[\lambda A(x)+B(x)]\mathbf{y}+\mathbf{f}, \\ [\alpha_1,\alpha_2]\mathbf{y}(c)=0,\qquad [\beta_1,\beta_2]\mathbf{y}(d)=0. \end{gathered}$$

We assume without loss of generality that $\alpha_1\alpha_1^*+\alpha_2\alpha_2^*=I_n$ (since $\alpha_1\alpha_1^*+\alpha_2\alpha_2^*>0$) and define Y_α to be the fundamental matrix of (1.1) satisfying $Y_\alpha(c,\lambda)=E_\alpha$ where

$$E=\begin{pmatrix} \alpha_1^* & -\alpha_2^* \\ \alpha_2^* & \alpha_1^* \end{pmatrix}.$$

Note that $E_\alpha^{-1}=E_\alpha^*$. We decompose Y_α into $n\times n$ blocks by writing

$$Y_\alpha=(\boldsymbol{\theta}_\alpha,\boldsymbol{\Phi}_\alpha)=\begin{pmatrix} \theta_\alpha & \Phi_\alpha \\ \hat{\theta}_\alpha & \hat{\Phi}_\alpha \end{pmatrix}.$$

Then some calculation (cf. [9]) yields that F given by (1.7) satisfies

$$\text{(2.2)}\quad E_\alpha^{-1}[FJ+(1/2)I]E_\alpha=\begin{pmatrix} 0 & 0 \\ -M_\beta & I_n \end{pmatrix},\qquad E_\alpha^{-1}[FJ-(1/2)I]E_\alpha=\begin{pmatrix} -I_n & 0 \\ -M_\beta & 0 \end{pmatrix}$$

where

$$\text{(2.3)}\qquad \begin{aligned} M_\beta=M_\beta(d,\alpha,\lambda)=-[\beta_1\Phi_\alpha(d,\lambda)+\beta_2\hat{\Phi}_\alpha(d,\lambda)]^{-1} \\ \cdot[\beta_1\theta_\alpha(d,\lambda)+\beta_2\hat{\theta}_\alpha(d,\lambda)]. \end{aligned}$$

Note that M_β is the matrix analogue of the Weyl circle at a regular point (cf. [4, p. 226]). Substitution of (2.2) into (1.8) (note: $Y(x,\lambda)=Y_\alpha(x,\lambda)E_\alpha^{-1}$, $E_\alpha^*JE_\alpha=E_\alpha JE_\alpha^*=J$) and additional calculation yields that

$$0=\begin{pmatrix} -I_n & -M_\beta^* \\ 0 & 0 \end{pmatrix}[-iY_\alpha(d,\lambda)^*JY_\alpha(d,\lambda)]\begin{pmatrix} -I & 0 \\ -M_\beta & 0 \end{pmatrix},$$

and hence

$$\text{(2.4)}\qquad 0=[I,M_\beta^*][-iY_\alpha(d,\lambda)^*JY_\alpha(d,\lambda)]\begin{bmatrix} I \\ M_\beta \end{bmatrix}.$$

Alternately, (2.4) follows directly from the definition (2.3).

If we define the solution $\mathbf{X}_\beta$ of (1.1) by $\mathbf{X}_\beta=Y_\alpha\left[\begin{smallmatrix} I \\ M_\beta \end{smallmatrix}\right]$, then (2.4) is equivalent to

$$\mathbf{X}_\beta^*(d)J\mathbf{X}_\beta(d)=0.$$

Further, (2.3) shows that $\mathbf{X}_\beta$ satisfies

$$[\beta_1,\beta_2]\mathbf{X}_\beta(d)=0. \tag{2.5}$$

Suppose now M is a matrix such that

$$0=[I,M^*][-iY_\alpha(d,\lambda)^*JY_\alpha(d,\lambda)]\begin{bmatrix} I \\ M \end{bmatrix}. \tag{2.6}$$

If we define

$$[\beta_1,\beta_2]=[I,M^*]Y_\alpha(d,\lambda)^*J,$$

then $\operatorname{rank}[\beta_1,\beta_2]=n$ and

$$\begin{aligned} -\beta_1\beta_2^*+\beta_2\beta_1^* &= [\beta_1,\beta_2]\begin{bmatrix} -\beta_2^* \\ \beta_1^* \end{bmatrix} \\ &= [\beta_1,\beta_2]J[\beta_1,\beta_2]^* \\ &= [I,M^*]Y_\alpha(d,\lambda)^*JY_\alpha(d,\lambda)\begin{bmatrix} I \\ M \end{bmatrix} \\ &= 0; \end{aligned}$$

thus $\beta_1\beta_2^*=\beta_2\beta_1^*$. Note also

$$[\beta_1,\beta_2]Y_\alpha(d,\lambda)\begin{bmatrix} I \\ M \end{bmatrix}=0,$$

from which it follows that

$$M=-\left[\beta_1\Phi_\alpha(d,\lambda)+\beta_2\hat{\Phi}_\alpha(d,\lambda)\right]^{-1}\left[\beta_1\theta_\alpha(d,\lambda)+\beta_2\hat{\theta}_\alpha(d,\lambda)\right].$$

Hence the M's given by (2.3) are the only ones that satisfy (2.6).

It is convenient to write (2.4) in the matrix-circle form (cf. [2], [19]) and recall some of the basic facts about this representation. Define for $\operatorname{Im}\lambda\neq 0$.

$$\begin{aligned} \begin{pmatrix} \mathfrak{A} & \mathfrak{B}^* \\ \mathfrak{B} & \mathfrak{D} \end{pmatrix} &= \begin{pmatrix} \mathfrak{A} & \mathfrak{B}^* \\ \mathfrak{B} & \mathfrak{D} \end{pmatrix}(d,\lambda) \\ &= \begin{cases} -iY_\alpha^*(d,\lambda)JY_\alpha(d,\lambda), & \operatorname{Im}\lambda>0, \\ iY_\alpha^*(d,\lambda)JY_\alpha(d,\lambda), & \operatorname{Im}\lambda<0, \end{cases} \end{aligned} \tag{2.7}$$

$$E(M)=E_{d,\lambda}(M)=[I,M^*]\begin{pmatrix} \mathfrak{A} & \mathfrak{B}^* \\ \mathfrak{B} & \mathfrak{D} \end{pmatrix}\begin{bmatrix} I \\ M \end{bmatrix}.$$

The above calculations show $E(M)=0$ iff $M=M_\beta$ for some $[\beta_1,\beta_2]$ as in (2.3). We may also write

$$\begin{aligned} E(M) &= M^*\mathfrak{D}M+M^*\mathfrak{B}+\mathfrak{B}^*M+\mathfrak{A} \\ &= (M+\mathfrak{D}^{-1}\mathfrak{B})^*\mathfrak{D}(M+\mathfrak{D}^{-1}\mathfrak{B})+\mathfrak{A}-\mathfrak{B}^*\mathfrak{D}^{-1}\mathfrak{B} \\ &= (M-C)^*R_1^{-2}(M-C)-R_2^2 \end{aligned}$$

where

$$\begin{aligned}C&=C(d,\lambda)=-\mathcal{D}^{-1}\mathcal{B},\\ R_1&=R_1(d,\lambda)=\mathcal{D}^{-1/2},\\ R_2&=R_2(d,\lambda)=[\mathcal{B}^*\mathcal{D}^{-1}\mathcal{B}-\mathcal{A}]^{1/2}.\end{aligned}$$

To see that $\mathcal{D}>0$, we have from (2.7) and (1.14) that

$$(2.8)\qquad \mathcal{D}=-i(\operatorname{sgn}(\operatorname{Im}\lambda))\Phi_\alpha(d,\lambda)^*J\Phi_\alpha(d,\lambda)=2|\operatorname{Im}\lambda|\int_c^d\Phi_\alpha^*A\Phi_\alpha.$$

We now show that

$$(2.9)\qquad \mathcal{B}(d,\lambda)^*\mathcal{D}(d,\lambda)^{-1}\mathcal{B}(d,\lambda)-\mathcal{A}(d,\lambda)=\mathcal{D}(d,\bar{\lambda})^{-1}.$$

To establish (2.9), we follow the argument of McIntosh, Hehenberger and Reyes–Sanchez [17] for the discrete case and [2] for the second-order matrix case. From (1.14) it follows that

$$(2.10)\qquad Y_\alpha(d,\bar{\lambda})^*JY_\alpha(d,\lambda)=J.$$

Reversing the order of the terms and using $J^{-1}=-J$ gives

$$(2.11)\qquad JY_\alpha(d,\lambda)JY_\alpha(d,\bar{\lambda})^*J=-J.$$

Using (2.11) in (2.10) gives

$$\begin{aligned}J&=Y_\alpha(d,\lambda)^*JY_\alpha(d,\bar{\lambda})\\ &=(i)^2Y_\alpha(d,\lambda)^*JY_\alpha(d,\lambda)JY_\alpha(d,\bar{\lambda})^*JY_\alpha(d,\bar{\lambda})\\ &=-\begin{pmatrix}\mathcal{A} & \mathcal{B}^*\\ \mathcal{B} & \mathcal{D}\end{pmatrix}(d,\lambda)J\begin{pmatrix}\mathcal{A} & \mathcal{B}^*\\ \mathcal{B} & \mathcal{D}\end{pmatrix}(d,\bar{\lambda}),\end{aligned}$$

form which we conclude that

$$\begin{aligned}I_n&=-\mathcal{A}(d,\lambda)\mathcal{D}(d,\bar{\lambda})+\mathcal{B}^*(d,\lambda)\mathcal{B}^*(d,\bar{\lambda}),\\ 0&=-\mathcal{B}(d,\lambda)\mathcal{D}(d,\bar{\lambda})+\mathcal{D}(d,\lambda)\mathcal{B}^*(d,\bar{\lambda}).\end{aligned}$$

From these equations we have

$$\begin{aligned}\mathcal{D}(d,\lambda)^{-1}\mathcal{B}(d,\lambda)&=\mathcal{B}^*(d,\bar{\lambda})\mathcal{D}(d,\bar{\lambda})^{-1},\\ \mathcal{D}(d,\bar{\lambda})^{-1}&=-\mathcal{A}(d,\lambda)+\mathcal{B}^*(d,\lambda)\mathcal{B}^*(d,\bar{\lambda})\mathcal{D}(d,\bar{\lambda})^{-1}\end{aligned}$$

from which (2.9) is immediate. We use (2.9) in the form

$$(2.12)\qquad R_2(d,\lambda)=R_1(d,\bar{\lambda}).$$

Note that if $A(x)$ is real, then $Y_\alpha(d,\bar{\lambda})=\overline{Y_\alpha(d,\lambda)}$ and hence $\mathcal{D}(d,\bar{\lambda})=\overline{\mathcal{D}(d,\lambda)}$. Equation (2.8) shows that $\mathcal{D}$ increases as d increases; hence as $d\to b$, $R_1(d,\lambda)$ and $R_2(d,\lambda)$ decrease to nonnegative limits. Further, it can be shown $C(d,\lambda)$ also has a limit [2, 19]. The equation $E(M)=0$ can be written as

$$[R_1^{-1}(M-C)R_2^{-1}]^*[R_1^{-1}(M-C)R_2^{-1}]=I_n$$

so that

$$M = C + R_1 U R_2 \tag{2.13}$$

for some unitary matrix U.

Finally, we note that for $\mathbf{X} = Y_\alpha\left[\begin{smallmatrix} I \\ M \end{smallmatrix}\right]$, it follows from (1.14) that

$$\begin{aligned} E(M) &= -i(\operatorname{sgn}(\operatorname{Im}\lambda))\mathbf{X}(d)^* J\mathbf{X}(d) \\ &= -i(\operatorname{sgn}(\operatorname{Im}\lambda))\mathbf{X}(c)^* J\mathbf{X}(c) + 2|\operatorname{Im}\lambda| \int_c^d \mathbf{X}^* A\mathbf{X}. \end{aligned} \tag{2.14}$$

This relation yields that the sets

$$\mathfrak{S}(d,\lambda) = \{M \colon E(M) \le 0\}$$

are nested, i.e., $\mathfrak{S}(d_2,\lambda) \subset \mathfrak{S}(d_1,\lambda)$ if $d_2 > d_1$. Members M of $\mathfrak{S}(d,\lambda)$ have the representation $M = C + R_1 V R_2$ with $V^* V \le I_n$. This shows $\mathfrak{S}(d,\lambda)$ is compact. If $E(M) \le 0$, then (2.14) yields

$$\int_c^d \mathbf{X}^* A \mathbf{X} \le i[M^* - M]/2\operatorname{Im}\lambda \tag{2.15}$$

since $\mathbf{X}(c)^* J \mathbf{X}(c) = M^* - M$. This inequality can be used to establish the existence of $\mathfrak{L}_A^2(c,b)$ solutions of (1.1). The number of linearly independent such solutions is related to the rank of the limit as $d \to b$ of $R_1(d,\lambda)$ and is discussed in [2].

LEMMA 2.1. *Let $\mu(d)$ be the minimum eigenvalue of $\mathfrak{D}(d,\lambda)$. If the limit-point case holds for* (1.1), *then $\mu(d) \to \infty$ as $d \to b$.*

Proof. Suppose $\mu(d) \le T < \infty$ for all d. Let $\mathbf{v}_d$ be a unit eigenvector of $\mathfrak{D}(d,\lambda)$ corresponding to $\mu(d)$. Set $\mathbf{X}_d = \mathbf{\Phi}_\alpha \mathbf{v}_d$. Application of (1.14) and (2.8) yields

$$2i\operatorname{Im}\lambda \int_c^d \mathbf{X}_d^* A \mathbf{X}_d = \mathbf{v}_d^* \mathbf{\Phi}_\alpha^* J \mathbf{\Phi}_\alpha \mathbf{v}_d \Big|_c^d = i(\operatorname{sgn}\operatorname{Im}\lambda)\mu(d)$$

or

$$\int_c^d \mathbf{X}_d^* A \mathbf{X}_d = \mu(d)/2|\operatorname{Im}\lambda| \le T/2|\operatorname{Im}\lambda|.$$

By considering a convergent subsequence of $\{v_d\}$, we then obtain a solution $\mathbf{X} = \mathbf{\Phi}_\alpha \mathbf{v}$, $\mathbf{v} \ne 0$, such that $\int_c^b \mathbf{X}^* A \mathbf{X} < \infty$. However, the limit point hypothesis yields n linearly independent $\mathfrak{L}_A^2(c,b)$ solutions of the form $Y_\alpha\left[\begin{smallmatrix} I \\ M \end{smallmatrix}\right]$ for appropriate M [7], [9]. Since $\mathbf{\Phi}_\alpha \mathbf{v}$ is not a linear combination of these, a contradiction has been reached.

It is an immediate corollary that

$$\lim_{d \to b} R_1(d,\lambda) = 0 \tag{2.16}$$

if the limit-point hypothesis holds.

LEMMA 2.2. *If $\|\cdot\|$ denotes the euclidean vector norm, then*

$$\left\| \mathfrak{D}(d,\lambda)^{-1/2} \int_c^d \mathbf{\Phi}_\alpha^* A \mathbf{f} \right\| \le \left[\int_c^d \mathbf{f}^* A \mathbf{f} \right]^{1/2} \Big/ [2|\operatorname{Im}\lambda|]^{1/2}$$

for $\mathbf{f} \in \mathfrak{L}_A^2(c,d)$.

Proof. For $\boldsymbol{\eta}$ a unit vector and $\boldsymbol{\mu}=\mathfrak{D}(d,\lambda)^{-1/2}\boldsymbol{\eta}$, application of the Cauchy–Schwarz inequality gives

$$\begin{aligned}\left|\boldsymbol{\eta}^*\mathfrak{D}(d,\lambda)^{-1/2}\int_c^d \boldsymbol{\Phi}^*A\mathbf{f}\right| &= \left|\int_c^d \boldsymbol{\mu}^*\boldsymbol{\Phi}_\alpha^*A\mathbf{f}\right| \\ &= \left|\int_c^d (A^{1/2}\boldsymbol{\Phi}_\alpha\boldsymbol{\mu})^*(A^{1/2}\mathbf{f})\right| \\ &\leq \int_c^d \|A^{1/2}\boldsymbol{\Phi}_\alpha\boldsymbol{\mu}\|\,\|A^{1/2}\mathbf{f}\| \\ &\leq \left[\int_c^d \boldsymbol{\mu}^*\boldsymbol{\Phi}_\alpha^*A\boldsymbol{\Phi}_\alpha\boldsymbol{\mu}\right]^{1/2}\left[\int_c^d \mathbf{f}^*A\mathbf{f}\right]^{1/2} \\ &= \left[\boldsymbol{\mu}^*\mathfrak{D}(d,\lambda)\boldsymbol{\mu}/2|\mathrm{Im}\lambda|\right]^{1/2}\left[\int_c^d \mathbf{f}^*A\mathbf{f}\right]^{1/2} \\ &= \left[1/2|\mathrm{Im}\lambda|\right]^{1/2}\left[\int_c^d \mathbf{f}^*A\mathbf{f}\right]^{1/2}.\end{aligned}$$

The choice $\boldsymbol{\eta}=\mathbf{g}/\|g\|$ where $\mathbf{g}=\mathfrak{D}(d,\lambda)^{-1/2}\int_c^d\boldsymbol{\Phi}_\alpha^*A\mathbf{f}$ completes the proof.

The proof below is a direct generalization of [20, Lemma 2.3].

THEOREM 2.1. *Suppose* (1.1) *is in either the limit-point or limit-circle case at b. Let* $M_\infty(\lambda)$ *be an analytic function on* $\mathrm{Im}\lambda\neq 0$ *determined by a sequential limit, i.e.*,

$$M_\infty(\lambda)=\lim_{n\to\infty} M_{\beta(n)}(d_n,\lambda)$$

for some $d_n\to b$. *Then for all* λ,μ *not real*,

$$\lim_{n\to\infty}\left[I,M_\infty(\lambda)^*\right]Y_\alpha(d_n,\lambda)^*JY_\alpha(d_n,\mu)\begin{bmatrix} I \\ M_\infty(\mu)\end{bmatrix}=0.$$

Proof. We consider first the limit point case. Set $\beta(n)=[\beta_{1n},\beta_{2n}]$ and let

$$\begin{aligned}M_{1n}&=-\left[\beta_{1n}\Phi_\alpha(d_n,\lambda)+\beta_{2n}\hat{\Phi}_\alpha(d_n,\lambda)\right]^{-1}\left[\beta_{1n}\theta_\alpha(d_n,\lambda)+\beta_{2n}\hat{\theta}_\alpha(d_n,\lambda)\right],\\ M_{2n}&=-\left[\beta_{1n}\Phi_\alpha(d_n,\mu)+\beta_{2n}\hat{\Phi}_\alpha(d_n,\mu)\right]^{-1}\left[\beta_{1n}\theta_\alpha(d_n,\mu)+\beta_{2n}\hat{\theta}_\alpha(d_n,\mu)\right].\end{aligned}$$

Then by (2.5),

$$[\beta_{1n},\beta_{2n}]\left[\boldsymbol{\theta}_\alpha(d_n,\lambda)+\boldsymbol{\Phi}_\alpha(d_n,\lambda)M_{1n}\right]=0,$$

which implies [9, p. 223] for some Γ_1,

$$\boldsymbol{\theta}_\alpha(d_n,\lambda)+\boldsymbol{\Phi}_\alpha(d_n,\lambda)M_{1n}=\begin{bmatrix}\beta_{2n}^*\\ -\beta_{1n}^*\end{bmatrix}\Gamma_1.$$

Similarly,

$$\boldsymbol{\theta}_\alpha(d_n,\mu)+\boldsymbol{\Phi}_\alpha(d_n,\mu)M_{2n}=\begin{bmatrix}\beta_{2n}^*\\ -\beta_{1n}^*\end{bmatrix}\Gamma_2,$$

and hence

$$(2.17)\qquad \left[\boldsymbol{\theta}_\alpha(d_n,\lambda)+\boldsymbol{\Phi}_\alpha(d_n,\lambda)M_{1n}\right]^*J\left[\boldsymbol{\theta}_\alpha(d_n,\mu)+\boldsymbol{\Phi}_\alpha(d_n,\mu)M_{2n}\right]=0.$$

Equation (2.17) may be written as

$$\begin{aligned}(2.18)\qquad 0=&[\boldsymbol{\theta}_\alpha(d_n,\lambda)+\boldsymbol{\Phi}_\alpha(d_n,\lambda)M_\infty(\lambda)]^*J[\boldsymbol{\theta}_\alpha(d_n,\mu)+\boldsymbol{\Phi}_\alpha(d_n,\mu)M_\infty(\mu)]\\&+[\boldsymbol{\theta}_\alpha(d_n,\lambda)+\boldsymbol{\Phi}_\alpha(d_n,\lambda)M_\infty(\lambda)]^*J\boldsymbol{\Phi}_\alpha(d_n,\mu)(M_{2n}-M_\infty(\mu))\\&+(M_{1n}-M_\infty(\lambda))^*\boldsymbol{\Phi}_\alpha(d_n,\lambda)^*J[\boldsymbol{\theta}_\alpha(d_n,\mu)+\boldsymbol{\Phi}_\alpha(d_n,\mu)M_{2n}].\end{aligned}$$

The proof will be complete if we show the second and third terms of this equation tend to zero as $n\to\infty$. To consider the third term we use (2.13) to write

$$M_{1n}=C_n+R_{1n}U_nR_{2n},\qquad M_\infty(\lambda)=C_n+R_{1n}V_nR_{2n}$$

where $U_n^*U_n=I$, $V_n^*V_n\le I$, $C_n=-\mathfrak{D}(d_n,\lambda)^{-1}\mathfrak{B}(d_n,\lambda)$, $R_{1n}=R_1(d_n,\lambda)$, and $R_{2n}=R_2(d_n,\lambda)$. Hence

$$\begin{aligned}(2.19)\qquad &(M_{1n}-M_\infty(\lambda))^*\boldsymbol{\Phi}_\alpha(d_n,\lambda)^*J[\boldsymbol{\theta}_\alpha(d_n,\mu)+\boldsymbol{\Phi}_\alpha(d_n,\mu)M_{2n}]\\&=R_{2n}(U_n^*-V_n^*)\mathfrak{D}(d_n,\lambda)^{-1/2}\\&\quad\cdot\left\{[-\alpha_2,\alpha_1]JE_\alpha\begin{bmatrix}I\\M_{2n}\end{bmatrix}+(\mu-\bar\lambda)\int_c^{d_n}\boldsymbol{\Phi}_\alpha^*A[\boldsymbol{\theta}_\alpha+\boldsymbol{\Phi}_\alpha M_{2n}]\right\}\end{aligned}$$

where we have used (1.14). By (2.15) we have a bound on

$$\int_c^{d_n}[\boldsymbol{\theta}_\alpha+\boldsymbol{\Phi}_\alpha M_{2n}]^*A[\boldsymbol{\theta}_\alpha+\boldsymbol{\Phi}_\alpha M_{2n}]$$

which is independent of n; thus by Lemma 2.2 and the fact that $R_{2n}=R_1(d_n,\bar\lambda)\to 0$ as $n\to\infty$, we see that the right-hand side of (2.19) tends to zero as $n\to\infty$. Similar considerations apply to the second term of (2.18) and the proof is complete.

The limit-circle case is much easier. We need only use the facts $M_{1n}\to M_\infty(\lambda)$ as $n\to\infty$, $M_{2n}\to M_\infty(\mu)$ as $n\to\infty$, and (1.14) to represent the other terms of (2.18) in terms of integrals of $\mathfrak{L}_A^2$ functions.

COROLLARY 2.1. *Let M_∞ be as in Theorem 2.1 and set $\mathbf{X}_\infty=Y_\alpha\left[\begin{smallmatrix}I\\M_\infty\end{smallmatrix}\right]$. Then*

$$(2.20)\qquad \lim_{n\to\infty}\mathbf{X}_\infty(d_n)^*J\mathbf{X}_\infty(d_n)=0.$$

Proof. Set $\mu=\lambda$ in Theorem 2.1.

COROLLARY 2.2. *Let M_∞ and $\mathbf{X}_\infty$ be as in Corollary 2.1. Then*

$$(2.21)\qquad \int_c^b\mathbf{X}_\infty^*A\mathbf{X}_\infty=i[M_\infty(\lambda)^*-M_\infty(\lambda)]/2\operatorname{Im}\lambda.$$

Proof. This relation follows by application of Corollary 2.1 to (1.14) with $\mathbf{y}=\mathbf{z}=\mathbf{X}_\infty$.

Equation (2.21) plays an important role in relating the singular structure of $M_\infty(\lambda)$ to the spectrum of differential operators [5], [8], [9].

COROLLARY 2.3. *Equation (1.1) is in the limit-point case at b iff*

$$\lim_{d\to b}\mathbf{X}_1(d)^*J\mathbf{X}_2(d)=0$$

for all $\mathfrak{L}_A^2(c,b)$ solutions $\mathbf{X}_1$ and $\mathbf{X}_2$ of (1.1) for $\lambda=\lambda_1$ and $\lambda=\lambda_2$ respectively, where λ_1 and λ_2 are arbitrary except for $\operatorname{Im}\lambda_1\neq 0$, $\operatorname{Im}\lambda_2\neq 0$.

Proof. The necessity is proved in Theorem 2.1 since in the limit-point case $M_\infty(\lambda)$ is the limit as $d\to b$ of (2.3) rather than being a sequential limit. (Recall $R_1(d,\lambda)\to 0$ as $d\to b$.) The sufficiency is proved in [7]; in fact only $\lambda_2=\bar\lambda_1$ need be considered.

Corollary 2.3 shows the equivalence of the limit-point definition (for even-order systems) used in [7] and that used here. Since the quantity $\mathbf{X}^*J\mathbf{X}$ is the Lagrange bilinear form for self-adjoint scalar equations when put in system form [18, Chap. V], Corollary 2.3 provides an alternate proof (without using the theory of maximal and minimal operators) of this well-known fact for scalar equations [18, §18.3], [11, p. 19].

3. A formula for $\tilde{F}(\lambda)$. Let N_1, N_2, E_α be as in §2; let Z be as in §1, and let $Z_c = Z(c,\lambda)$, $Z_d = Z(d,\lambda)$. Then from (1.7) and (1.9) we have (note that $Y = ZZ_c^{-1}$)

$$\begin{aligned}(3.1)\qquad \tilde{F}J-(1/2)I &= Z_c^{-1}FJZ_c-(1/2)I\\ &= Z_d^{-1}N_2\left[N_1-Z_cZ_d^{-1}N_2\right]^{-1}Z_c\\ &= Z_d^{-1}N_2\left[E_\alpha^{-1}N_1-E_\alpha^{-1}Z_cZ_d^{-1}N_2\right]^{-1}E_\alpha Z_c.\end{aligned}$$

We write

$$Z=\begin{pmatrix}\theta & \Phi\\ \hat{\theta} & \hat{\Phi}\end{pmatrix}$$

and from (1.14), $Z(x,\lambda)^*JZ(x,\bar{\lambda})\equiv J$; thus we compute

$$Z(x,\lambda)^{-1}=\begin{pmatrix}\hat{\Phi}(x,\bar{\lambda})^* & -\Phi(x,\bar{\lambda})^*\\ -\hat{\theta}\,(x,\bar{\lambda})^* & \theta(x,\bar{\lambda})^*\end{pmatrix}.$$

Performing the indicated computations we find that

$$E_\alpha^{-1}N_1-E_\alpha^{-1}Z_cZ_d^{-1}N_2=\begin{pmatrix}-\Delta_1 & 0\\ ** & -I_n\end{pmatrix}$$

where

$$\begin{aligned}\Delta_1 &= \left[\alpha_1\theta(c,\lambda)+\alpha_2\hat{\theta}\,(c,\lambda)\right]\left[\Phi(d,\bar{\lambda})^*\beta_1^*+\hat{\Phi}(d,\bar{\lambda})\beta_2^*\right]\\ &\quad-\left[\alpha_1\Phi(c,\lambda)+\alpha_2\hat{\Phi}(c,\lambda)\right]\left[\theta(d,\bar{\lambda})^*\beta_1^*+\hat{\theta}\,(d,\bar{\lambda})^*\beta_2^*\right].\end{aligned}$$

Since

$$\begin{pmatrix}-\Delta_1 & 0\\ ** & -I_n\end{pmatrix}^{-1}=\begin{pmatrix}-\Delta_1^{-1} & 0\\ ** & -I_n\end{pmatrix},$$

further calculations (the ** terms are unimportant) yields

$$\begin{aligned}(3.2)\quad \tilde{F}J-(1/2)I &= \begin{pmatrix}\Phi^*\beta_1^*+\hat{\Phi}^*\beta_2^*\\ \theta^*\beta_1^*+\hat{\theta}^*\beta_2^*\end{pmatrix}(d,\bar{\lambda})\left(-\Delta_1^{-1}\right)\left(\alpha_1\theta+\alpha_2\hat{\theta},\alpha_1\Phi+\alpha_2\hat{\Phi}\right)(c,\lambda)\\ &= \begin{pmatrix}u\\ v\end{pmatrix}\left(-\Delta_1^{-1}\right)(r,s)\end{aligned}$$

where u,v,r,s are the terms of (3.2). Consider now the $n\times n$ block

$$\begin{aligned}\left(\tilde{F}J-(1/2)I\right)_{11} &= -u\Delta_1^{-1}r = -u[ru-sv]^{-1}r\\ &= -\left[s^{-1}r-vu^{-1}\right]^{-1}s^{-1}r\\ &= -\left[M_\alpha(\lambda)-M_\beta(\bar{\lambda})^*\right]^{-1}M_\alpha(\lambda)\end{aligned}$$

where following the notation of §2,

$$M_\beta(\lambda) = -\left[\beta_1\Phi(d,\lambda)+\beta_2\hat{\Phi}(d,\lambda)\right]^{-1}\left[\beta_1\theta(d,\lambda)+\beta_2\hat{\theta}\,(d,\lambda)\right],$$

$$M_\alpha(\lambda) = -\left[\alpha_1\Phi(c,\lambda)+\alpha_2\hat{\Phi}(c,\lambda)\right]^{-1}\left[\alpha_1\theta(c,\lambda)+\alpha_2\hat{\theta}\,(c,\lambda)\right].$$

Recalling the property $M_\beta(\bar{\lambda})^* = M_\beta(\lambda)$ [7], we have

$$(\tilde{F}J-(1/2)I)_{11} = -\left[M_\alpha(\lambda)-M_\beta(\lambda)\right]^{-1}M_\alpha(\lambda).$$

Similar calculations give (with λ's suppressed)

$$\tilde{F}J-(1/2)I = \begin{pmatrix} [M_\beta-M_\alpha]^{-1}M_\alpha & -[M_\beta-M_\alpha]^{-1} \\ M_\beta[M_\beta-M_\alpha]^{-1}M_\alpha & -M_\beta[M_\beta-M_\alpha]^{-1} \end{pmatrix} \tag{3.3}$$

$$= \begin{pmatrix} I \\ M_\beta \end{pmatrix}(M_\beta-M_\alpha)^{-1}(M_\alpha,-I).$$

From (3.3) it follows readily that

$$\tilde{F}J+(1/2)I = \begin{pmatrix} I \\ M_\alpha \end{pmatrix}(M_\beta-M_\alpha)^{-1}(M_\beta,-I). \tag{3.4}$$

Note that the boundedness of $\tilde{F}$, independent of c and d, ensures that $\det(M_\beta-M_\alpha)$ is bounded away from 0, independent of c and d. Use of the limiting values of (3.3) and (3.4) yields a representation of the Green's function (see (4.4) below) which is the same as the second order scalar case. Further, solving (3.3) for $\tilde{F}$ yields a form of the characteristic function which reduces to a known formula in the second order scalar case (compare with [4, pp. 251] and [13, p. 926]).

4. Singular boundary value problems. Suppose now c_n, d_n, $N_1(n)$, $N_2(n)$ are such that $c_n \to a$, $d_n \to b$ as $n \to \infty$, and for $\operatorname{Im}\lambda \neq 0$,

$$\tilde{F}_\infty(\lambda) = \lim_{n\to\infty} \tilde{F}(c_n,d_n,N_1(n),N_2(n),\lambda) \tag{4.1}$$

is analytic on $\operatorname{Im}\lambda \neq 0$. Suppose (using the notation of §3)

$$M^+(\lambda) = \lim_{n\to\infty} M_\beta(d_n,\lambda) \tag{4.2}$$

and

$$M^-(\lambda) = \lim_{n\to\infty} M_\alpha(c_n,\lambda). \tag{4.3}$$

Then by (3.3) and (3.4),

$$\tilde{F}_\infty J-(1/2)I = \begin{pmatrix} I \\ M^+ \end{pmatrix}(M^+-M^-)^{-1}(M^-,-I)$$

and

$$\tilde{F}_\infty J+(1/2)I = \begin{pmatrix} I \\ M^- \end{pmatrix}(M^+-M^-)^{-1}(M^+,-I).$$

If we allow each endpoint to be limit-point or limit-circle, there are four possible cases. We consider here only the case where a is limit-circle and b is limit-point. The other cases are similar. Using the formulas for Z given in §3, we see from (1.10) that K_∞ may be written as

$$K_\infty(x,t,\lambda)=\begin{cases}\mathbf{\Phi}_a(x,\lambda)[M^+(\lambda)-M^-(\lambda)]^{-1}\mathbf{\Phi}_b(t,\bar{\lambda})^*, & x\le t,\\ \mathbf{\Phi}_b(x,\lambda)[M^+(\lambda)-M^-(\lambda)]^{-1}\mathbf{\Phi}_a(t,\bar{\lambda})^*, & x>t,\end{cases} \tag{4.4}$$

where

$$\mathbf{\Phi}_a(x,\lambda)=Z(x,\lambda)\begin{pmatrix} I\\ M^-(\lambda)\end{pmatrix},\qquad \mathbf{\Phi}_b(x,\lambda)=Z(x,\lambda)\begin{pmatrix} I\\ M^+(\lambda)\end{pmatrix}. \tag{4.5}$$

Note that $\mathbf{\Phi}_b\in\mathcal{L}_A^2(e,b)$.

LEMMA 4.1. *Suppose each of a and b is in the limit-point or limit-circle case*, $\operatorname{Im}\lambda\neq 0$, $\mathbf{f}\in\mathcal{L}_A^2$, *and*

$$\mathbf{y}(x,\lambda)=\int_a^b K_\infty(x,t,\lambda)A(t)\mathbf{f}(t)\,dt. \tag{4.6}$$

Then $\mathbf{y}\in\mathcal{L}_A^2$ *and for all* μ *with* $\operatorname{Im}\mu\neq 0$,

$$\lim_{x\to b}\mathbf{y}(x,\lambda)^*J\mathbf{\Phi}_b(x,\mu)=0=\lim_{x\to a}\mathbf{y}(x,\lambda)^*J\mathbf{\Phi}_a(x,\mu).$$

Proof. The property $\mathbf{y}\in\mathcal{L}_A^2$ is stated in (1.12). Consider first the limit-point case at b. Let $\mathbf{f}_d$ be $\mathbf{f}$ restricted to $(a,d]$, and let $\mathbf{y}_d$ be given by (4.6) for $\mathbf{f}$ replaced by $\mathbf{f}_d$. Then for $x>d$,

$$\mathbf{y}_d(x,\lambda)=\mathbf{\Phi}_b(x,\lambda)[M^+(\lambda)-M^-(\lambda)]^{-1}\int_a^d\mathbf{\Phi}_a(t,\bar{\lambda})^*A(t)\mathbf{f}(t)\,dt.$$

Hence by Theorem 2.1 (note that $Z=Y_\alpha$ if $c=e$ and $E_\alpha=I$),

$$\lim_{x\to b}\mathbf{y}_d(x,\lambda)^*J\mathbf{\Phi}_b(x,\mu)=0. \tag{4.7}$$

From (1.14), with $\mathbf{\Phi}_b=\mathbf{\Phi}_b(\cdot,\mu)$,

$$\mathbf{y}_d^*J\mathbf{\Phi}_b\big|_e^x=(\mu-\bar{\lambda})\int_e^x\mathbf{y}_d^*A\mathbf{\Phi}_b-\int_e^x\mathbf{f}_d^*A\mathbf{\Phi}_b. \tag{4.8}$$

Now by (1.13),

$$\int_a^b(\mathbf{y}-\mathbf{y}_d)^*A(\mathbf{y}-\mathbf{y}_d)\le(\operatorname{Im}\lambda)^{-2}\int_d^b\mathbf{f}^*A\mathbf{f} \tag{4.9}$$

since $\mathbf{f}-\mathbf{f}_d=\mathbf{f}$ on $[d,b)$. Further, the formula (4.4) shows $\mathbf{y}_d$ converges uniformly to $\mathbf{y}$ on compact sets as $d\to b$. Now let $x\to b$ in (4.8). By (4.7) we obtain

$$-\mathbf{y}_d(e,\lambda)^*J\mathbf{\Phi}_b(e,\mu)=(\mu-\bar{\lambda})\int_e^b\mathbf{y}_d^*A\mathbf{\Phi}_b-\int_e^d\mathbf{f}^*A\mathbf{\Phi}_b. \tag{4.10}$$

Application of (4.9) to (4.10) yields

$$-\mathbf{y}(e,\lambda)^*J\mathbf{\Phi}_b(e,\mu)=(\mu-\bar{\lambda})\int_e^b\mathbf{y}^*A\mathbf{\Phi}_b-\int_e^b\mathbf{f}^*A\mathbf{\Phi}_b. \tag{4.11}$$

On the other hand, if $d\to b$ in (4.8) we have

$$\text{(4.12)}\quad \mathbf{y}(x,\lambda)^*J\mathbf{\Phi}_b(x,\mu)-\mathbf{y}(e,\lambda)^*J\mathbf{\Phi}_b(e,\mu)=(\mu-\bar{\lambda})\int_e^x \mathbf{y}^*A\mathbf{\Phi}_b-\int_e^x \mathbf{f}^*A\mathbf{\Phi}_b.$$

By letting $x\to b$ in (4.12) and comparing with (4.11), this completes the proof.

The above argument is not necessary in the limit-circle case since all functions are $\mathfrak{L}_A^2$. The representation,

$$\mathbf{y}(x,\lambda)=\mathbf{\Phi}_b(x,\lambda)[M^+(\lambda)-M^-(\lambda)]^{-1}\int_a^x \mathbf{\Phi}_a(t,\bar{\lambda})^*A(t)\mathbf{f}(t)\,dt$$
$$+\mathbf{\Phi}_a(x,\lambda)[M^+(\lambda)-M^-(\lambda)]^{-1}\int_x^b \mathbf{\Phi}_b(t,\bar{\lambda})^*A(t)\mathbf{f}(t)\,dt,$$

permits a direct application of Corollary 2.1. The proofs for $x\to a$ are similar.

THEOREM 4.1. *Suppose* (1.1)–(1.3), (4.1)–(4.5) *hold*, $\operatorname{Im}\lambda\neq 0$, $\mathbf{L}_a$ *is an n-vector, and* $\mathbf{f}\in\mathfrak{L}_A^2$. *Let* (1.1) *be limit-circle at a and limit-point at b. Then there is a unique* $\mathbf{y}\in\mathfrak{L}_A^2$ *such that*

$$\text{(4.13)}\quad J\mathbf{y}'=[\lambda A(x)+B(x)]\mathbf{y}+A(x)\mathbf{f}$$

and

$$\text{(4.14)}\quad \lim_{x\to a}\mathbf{y}(x)^*J\mathbf{\Phi}_a(x,\bar{\lambda})=\mathbf{L}_a^*.$$

Moreover, y is given by

$$\text{(4.15)}\quad \mathbf{y}(x,\lambda)=\mathbf{\Phi}_b(x,\lambda)[M^+(\lambda)-M^-(\lambda)]^{-1}\mathbf{L}_a+\int_a^b K_\infty(x,t,\lambda)A(t)\mathbf{f}(t)\,dt,$$

and for all μ *with* $\operatorname{Im}\mu\neq 0$,

$$\text{(4.16)}\quad \lim_{x\to b}\mathbf{y}(x,\lambda)^*J\mathbf{\Phi}_b(x,\mu)=0.$$

Proof. Except for (4.14), the other properties of (4.15) have already been established. By (1.14)

$$\text{(4.17)}\quad \mathbf{\Phi}_b(x,\lambda)^*J\mathbf{\Phi}_a(x,\bar{\lambda})\equiv\mathbf{\Phi}_b(e,\lambda)^*J\mathbf{\Phi}_a(e,\bar{\lambda})$$
$$=M^+(\lambda)^*-M^-(\bar{\lambda})=M^+(\lambda)^*-M^-(\lambda)^*$$

since $M(\lambda)^*=M(\bar{\lambda})$ [7]. This matrix is nonsingular by construction of F.

If $\mathbf{y}_1$ and $\mathbf{y}_2$ are $\mathfrak{L}_A^2$ solutions satisfying (4.13) and (4.14), then $\mathbf{y}_3=\mathbf{y}_2-\mathbf{y}_1$ satisfies (1.1). Since (1.1) is limit-point at b, $\mathbf{y}_3(x)=\mathbf{\Phi}_b(x,\lambda)\mathbf{v}$ for some n-vector $\mathbf{v}$. By (4.14), $\mathbf{y}_3$ satisfies

$$\lim_{x\to a}\mathbf{y}_3(x)^*J\mathbf{\Phi}_a(x,\bar{\lambda})=0,$$

which is contrary to (4.17) unless $\mathbf{v}=0$.

In the limit-point case at a, the boundary condition (4.14) is dropped. Uniqueness follows from the fact that the difference $\mathbf{y}_3$ of two solutions $\mathbf{y}_1$ and $\mathbf{y}_2$ has the representations $\mathbf{y}_3(x)=\mathbf{\Phi}_b(x,\lambda)\mathbf{v}$ and $\mathbf{y}_3(x)=\mathbf{\Phi}_a(x,\lambda)\mathbf{u}$. Invertibility of $M^+(\lambda)-M^-(\lambda)$ implies $\mathbf{u}=\mathbf{v}=0$. In the limit-circle case at b we may impose also a boundary condition,

$$\lim_{x\to b}\mathbf{y}(x)^*J\mathbf{\Phi}_b(x,\bar{\lambda})=\mathbf{L}_b^*.$$

In the limit-circle case we may also consider a nonhomogeneous boundary condition

$$\lim_{x\to a} \mathbf{y}(x)^* J\mathbf{\Phi}_a(x,\lambda) = \mathbf{L}_a^*. \tag{4.18}$$

To impose the boundary condition (4.18) in place of (4.14) we need that the matrix Γ defined by

$$\lim_{x\to a} \mathbf{\Phi}_b(x,\lambda)^* J\mathbf{\Phi}_a(x,\lambda) = \Gamma$$

is nonsingular. Now $W = (\mathbf{\Phi}_a, \mathbf{\Phi}_b)$ is a fundamental matrix since $M^+(\lambda) - M^-(\lambda)$ is nonsingular. Thus by Corollary 2.1,

$$\lim_{x\to a} W(x)^* J W(x) = \begin{pmatrix} 0 & \Gamma^* \\ \Gamma & ** \end{pmatrix} \equiv \Delta. \tag{4.19}$$

Since W satisfies (1.1),

$$(\det W)' = \left(\operatorname{trace}[J^{-1}(\lambda A + B)]\right)\det W.$$

Now trace $J^{-1}[(\operatorname{Re}\lambda)A + B]$ is pure imaginary since $A = A^*$ and $B = B^*$; hence if A is real, $|\det W|$ is constant and Δ given by (4.19) is nonsingular. The unique solution $\mathbf{y} \in \mathcal{L}_A^2$ of (4.13) and (4.18) is then given by

$$\mathbf{y}(x,\lambda) = \mathbf{\Phi}_b(x,\lambda)\Gamma^{*-1}\mathbf{L}_a + \int_a^b K_\infty(x,t,\lambda)A(t)\mathbf{f}(t)\,dt.$$

The boundary condition (4.16) is a direct generalization of the λ-dependent condition considered by Titchmarsh [20, p. 31]. In the limit-circle case Fulton [6] has given an equivalent formulation of the boundary conditions which is λ-independent. The authors have recently extended Fulton's formulation to the case of a Hamiltonian system [10]. However, for our purposes here it is more convenient to consider the Titchmarsh form.

Finally we develop the analogue of Corollary 2.2 for two singular endpoints. From Corollary 2.2, we have for $\mathbf{X}_\infty = \mathbf{\Phi}_b$,

$$\int_e^b \left(Z\begin{bmatrix} I \\ M^+ \end{bmatrix}\right)^* A\left(Z\begin{bmatrix} I \\ M^+ \end{bmatrix}\right) = \frac{M^+(\lambda)^* - M^+(\lambda)}{2\operatorname{Im}\lambda},$$

and thus by multiplication of appropriate factors,

(4.20)

$$\{[M^+ - M^-]^{-1}(I, M^-)\}^* \frac{\begin{pmatrix} I \\ M^+ \end{pmatrix}^* J\begin{pmatrix} I \\ M^+ \end{pmatrix}}{2\operatorname{Im}\lambda}\{[M^+ - M^-]^{-1}(I, M^-)\}$$

$$= \int_e^b \left\{Z\begin{pmatrix} I \\ M^+ \end{pmatrix}(M^+ - M^-)^{-1}(I, M^-)\right\}^* A\left\{Z\begin{pmatrix} I \\ M^+ \end{pmatrix}(M^+ - M^-)^{-1}(I, M^-)\right\}.$$

However, we may rewrite (4.20) as

$$\{\tilde{F}_\infty - (1/2)J^{-1}\}^* J\{\tilde{F}_\infty - (1/2)J^{-1}\}/2\operatorname{Im}\lambda$$

$$= \int_e^b \{Z(\tilde{F}_\infty - (1/2)J^{-1})\}^* A\{Z(\tilde{F}_\infty - (1/2)J^{-1})\}.$$

Similarly, it follows that

$$-\{\tilde{F}_\infty+(1/2)J^{-1}\}^*J\{\tilde{F}_\infty+(1/2)J^{-1}\}/2\,\mathrm{Im}\,\lambda$$
$$=\int_a^e\{Z(\tilde{F}_\infty+(1/2)J^{-1})\}^*A\{Z(\tilde{F}_\infty+(1/2)J^{-1})\}.$$

Adding these two equations and simplifying yields the analogue of Corollary 2.2,

$$\frac{\tilde{F}_\infty(\lambda)-\tilde{F}_\infty(\lambda)^*}{2\,\mathrm{Im}\,\lambda}=\int_a^e\{Z(\tilde{F}_\infty+(1/2)J^{-1})\}^*A\{Z(\tilde{F}_\infty+(1/2)J^{-1})\}$$
$$+\int_e^b\{Z(\tilde{F}_\infty-(1/2)J^{-1})\}^*A\{Z(\tilde{F}_\infty-(1/2)J^{-1})\}.$$

REFERENCES

[1] F. V. Atkinson, *Discrete and Continuous Boundary Problems*, Academic Press, New York, 1964.

[2] E. S. Birger and G. A. Kalyalin, *The theory of Weyl limit-circle in the case of non-self-adjoint second-order differential-equation systems*, Differentsial'nye Uravneniya, 12 (1976), pp. 1531–1540; Differential Equations, 12 (1977), pp. 1077–1084.

[3] C. Burnap, W. Greenberg, P. F. Zweifel, *Eigenvalue problems for singular potentials*, Nuovo. Cim., 50 (1979), pp. 457–465.

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

[5] J. Chandhuri and W. N. Everitt, *On the spectrum of ordinary second order differential operators*, Proc. Royal Soc. Edin., 68A (1967-68), pp. 95–119.

[6] C. T. Fulton, *Parametrizations of Titchmarsh's $m(\lambda)$-functions in the limit-circle case*, Trans. Amer. Math. Soc., 229 (1977), pp. 51–63.

[7] D. B. Hinton and J. K. Shaw, *On Titchmarsh–Weyl $M(\lambda)$-functions for linear Hamiltonian systems*, J. Differential Eqs., 40 (1981), pp. 316–342.

[8] ______, *On the spectrum of a singular Hamiltonian system*, Quaestiones Mathematicae, to appear.

[9] ______, *Titchmarsh–Weyl theory for Hamiltonian systems*, in Spectral Theory of Differential Operators, I. W. Knowles and R. T. Lewis, eds., North-Holland, Amsterdam, 1981, pp. 219–231.

[10] ______, *Parameterization of the $M(\lambda)$-function for a Hamiltonian system of limit circle type*, submitted.

[11] R. M. Kauffman, T. T. Read, and A. Zettl, *The Deficiency Index Problem for Powers of Ordinary Differential Expressions*, Lecture Notes in Mathematics, 621, Springer-Verlag, Berlin, 1977.

[12] C. S. Kim, *Spectral theory for a singular linear system of Hamiltonian differential equations*, J. Differential Equations, 10 (1971), pp. 538–567.

[13] K. Kodaira, *The eigenvalue problem for ordinary differential equations of the second order and Heisenberg's theory of S-matrices*, Amer. J. Math., 71 (1949), pp. 921–945.

[14] V. I. Kogan and F. S. Rofe-Beketov, *On square-integrable solutions of symmetric systems of differential equations of arbitrary order*, Proc. Royal Soc. Edin., 74A (1974), pp. 5–40.

[15] A. M. Krall, *Boundary values for an eigenvalue problem with a singular potential*, J. Differential Equations, to appear.

[16] B. M. Levitan and I. S. Sargsjan, *Introduction to Spectral Theory: Self-Adjoint Ordinary Differential Operators*, Trans. Math. Monographs 39, American Mathematical Society, Providence, RI, 1975.

[17] H. V. McIntosh, M. Hehenberger, and R. Reyes-Sanchez, *Lattice dynamics with second-neighbor interactions*, III, Internat. J. Quant. Chem., 11 (1977), pp. 189–211.

[18] M. A. Naimark, *Linear Differential Operators, Part* II, Ungar, New York, 1968.

[19] S. A. Orlov, *Nested matrix disks analytically depending on a parameter, and theorems on the invariance of ranks of radii of limiting disks*, Izv. Akad. Nauk SSSR, 40 (1970), 593–644; Math. USSR Izv., 10 (1976), pp. 565–613.

[20] E. C. Titchmarsh, *Eigenfunction Expansions Associated with Second-order Equations*, I, Oxford Univ. Press, London, 1962.

[21] H. Weyl, *Über genöhnliche Differentialglleichungen mit Singularitäten und die Zugenhörigen Entwicklungen*, Math. Ann., 68 (1910), pp. 220—269.

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics

0036-1410/84/1502-0005 $01.25/0

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

EWA WEINMÜLLER†

Abstract. Analytical properties like existence, uniqueness and smoothness of continuous solutions of nonlinear boundary value problems are considered. Fredholm theory for linear boundary value problems is established. The results are applied to two practical examples from the theory of spherical shells.

1. Introduction. We investigate the nonlinear boundary value problem

$$y''(t)-\frac{A_1}{t}y'(t)-\frac{A_0}{t^2}y(t)=f(t,y(t)), \qquad 0<t\leq 1, \tag{1.1a}$$

$$B(y(0),y'(0);y(1),y'(1))=0, \tag{1.1b}$$

where y, f are vector-valued functions of dimension n, B is a vector-valued function of dimension $m\leq n$ and A_0, A_1 are constant $n\times n$ matrices.

The numerical solution by difference methods of the scalar problem has been examined by different authors; see Jamet [5], Natterer [10], Russel and Shampine [14]. Brabston [1] and de Hoog and Weiss [2] have considered first order systems of ordinary differential equations with a singularity of the type considered here. In our analysis we shall rely heavily on the techniques developed in [2].

The present paper provides a study of basic analytic properties of (1.1) like existence, smoothness and uniqueness of the solutions. Particular attention is paid to the structure and properties of the boundary conditions (1.1b) ensuring the existence of such a solution. We establish a Fredholm theory for the case when (1.1) is linear.

An outline of the paper is as follows: §3 deals with analytic questions for the linear problem (1.1) with constant coefficients. In §4 we consider linear problems where the matrices A_0, A_1 depend on t and $A_0(t)$, $A_1(t)$ are continuous on [0, 1]. In §5 we study the nonlinear problem (1.1) and extend the results to the case where f is a function of the form $f(t,y(t)/t)$. This kind of right-hand side occurs in problems from the nonlinear theory of spherical shells; see Keller and Wolfe [7]. Finally, we apply the theory to two examples given in Keller and Wolfe [7] and Rentrop [13].

2. Preliminaries. The following notation will be used. We denote by X^n the space of complex-valued vectors of dimension n. We use $|\cdot|$ to denote the maximum norm in X^n,

$$|x|=\left|(x_1,x_2,\cdots,x_n)^T\right|=\max_{1\leq i\leq n}|x_i|.$$

$C_n^p[0,1]$ is the space of complex n-vector-valued functions which are p times continously differentiable on [0, 1], and $C_n^p(0,1]$ is defined similarly. For each $y\in C_n^0[0,1]$ we define the norm

$$\|y\|=\max_{0\leq t\leq 1}|y(t)|.$$

*Received by the editors March 8, 1982, and in revised form December 16, 1982.

†Institute for Applied and Numerical Mathematics, Technical University, Vienna, Austria.

Occasionally, we use the following norm on $[0,\delta]$, $\delta>0$;

$$\|y\|_\delta=\max_{0\le t\le\delta}|y(t)|.$$

$C^p_{n\times n}[0,1]$ is the space of complex-valued $n\times n$ matrices with columns from $C^p_n[0,1]$. For $A\in C^0_{n\times n}[0,1]$, $\|A\|$ is the induced norm. When there is no confusion, we shall delete the subscripts n and call $C=C[0,1]=C^0[0,1]$, $C(0,1]=C^0(0,1]$. Let G be a $2n\times 2n$ matrix. We denote by G_1 the $n\times 2n$ matrix consisting of the n first rows of G and by G_2 the $n\times 2n$ matrix consisting of the n last rows of G.

3. Analytic results for the linear problem with constant coefficient-matrices A_0 and A_1. Here we consider the boundary value problem of the form

$$y''(t)-\frac{A_1}{t}y'(t)-\frac{A_0}{t^2}y(t)=f(t),\qquad 0<t\le 1,\tag{3.1a}$$

$$B_0Y(0)+B_1Y(1)=\beta,\tag{3.1b}$$

where $Y(t)=(y(t),y'(t))^T$, $y\in C[0,1]\cap C^2(0,1]$. A_0, A_1 are constant $n\times n$ matrices, B_0, B_1 are constant $m\times 2n$ matrices, β is an m-vector and $f\in C$. The number m of rows of B_0 and B_1, that is necessary for (3.1) to define a well-posed boundary value problem, will be specified later.

As a first step in the analysis of (3.1) we consider the linear system

$$y''(t)-\frac{A_1}{t}y'(t)-\frac{A_0}{t^2}y(t)=f(t),\qquad 0<t\le 1.\tag{3.2}$$

The linear transformation $z_1(t)=y(t)$, $z_2(t)=ty'(t)$ applied to (3.2) yields to the first order system of size $2n$ for the vector $z=(z_1,z_2)^T$

$$z'(t)=\frac{1}{t}Mz(t)+t\mathring{f}(t),\qquad 0<t\le 1,\tag{3.3}$$

where

$$M=\begin{bmatrix}0 & I\\ A_0 & I+A_1\end{bmatrix},\qquad \mathring{f}(t)=\begin{bmatrix}0\\ f(t)\end{bmatrix}.\tag{3.4}$$

The fact that the structure of the general solution of (3.3) depends on the eigenvalues of M, suggests representing this matrix in its Jordan canonical form. Let E be the matrix of (generalized) eigenvectors of M such that $M=EJE^{-1}$. Then $\varphi(t)=E^{-1}z(t)$ satisfies the equation

$$\varphi'(t)=\frac{1}{t}J\varphi(t)+tg(t),\qquad 0<t\le 1,\tag{3.5}$$

with $g(t)=E^{-1}\mathring{f}(t)$. The general solution of (3.5) has the form

$$\varphi(t)=\varphi_p(t)+\Phi(t)c=t^J\int_1^t s^{-J}sg(s)\,ds+t^Jc,\tag{3.6}$$

where c is a constant $2n$-vector, $\varphi_p(t)$ is the solution of (3.5) with $\varphi_p(1)=0$ and

$$\Phi(t)=t^J=\exp(J\ln t),\qquad 0\le t\le 1,$$

is the fundamental solution matrix which satisfies

$$\Phi'(t)=\frac{1}{t}J\Phi(t),\quad 0<t\leq 1,\qquad \Phi(1)=I,$$

cf. [2, Lemma 3.1]. To analyze (3.5) we first assume that J consists of only one box of the form

$$J=\begin{bmatrix}\lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1\\ & & & \lambda\end{bmatrix},\qquad \lambda=\sigma+i\beta$$

and consider separately the three cases: $\sigma\leq 0$, $\lambda=0$, $\sigma>0$. Since the matrix t^J will frequently occur, we note that it has the form

$$(3.7)\qquad t^J=t^\lambda\begin{bmatrix}1 & \ln t & \frac{(\ln t)^2}{2!} & \cdots & \frac{(\ln t)^{2n-1}}{(2n-1)!}\\ & & & & \vdots\\ & & \ddots & & \\ & & & & \ln t\\ & & & & 1\end{bmatrix},\qquad 0\leq t\leq 1.$$

Case 1. $\sigma\leq 0$.

LEMMA 3.1. *For every* $g\in C^p[0,1]$, $p\geq 0$, *there exists a unique solution* φ *of* (3.5). *Furthermore*, $y=E_1\varphi$ *is a solution of* (3.2), $y\in C^{p+2}$ *and the following estimates hold*:

$$|y(t)|\leq t^2D\|g\|,\quad |y'(t)|\leq tD\|g\|,\quad |y''(t)|\leq D\|g\|,\qquad 0\leq t\leq 1,\quad D=\text{const.}$$

Proof. Let $g\in C$. According to (3.6) the unique, continuous solution φ of (3.5) is

$$(3.8)\qquad \varphi(t)=t^J\int_0^t s^{-J}sg(s)\,ds=t^2\int_0^1 s^{-J}sg(ts)\,ds.$$

We substitute (3.8) into (3.5) to obtain

$$\varphi'(t)=Jt\int_0^1 s^{-J}sg(ts)\,ds+tg(t).$$

The estimates for $y(t)$ and $y'(t)$ hold now due to $z^{(k)}(t)=E\varphi^{(k)}(t)$, $k=0$, 1 and the estimate for $y''(t)$ follows from (3.2).If $g\in C^p$, $p\geq 1$, we can differentiate the last equation, so

$$\varphi''(t)=(J-I)J\int_0^1 s^{-J}sg(ts)\,ds+G(t),$$

where $G(t)=(J+I)g(t)+tg'(t)$ and $y''(t)\equiv\xi(t)$, where $\xi\in C^p$ if $g\in C^p$. □

Case 2. $\lambda=0$.

LEMMA 3.2. *For every* $g\in C^p[0,1]$, $p\geq 0$, *there exists a unique solution* φ *of* (3.5) *subject to the terminal condition* $\varphi_1(1)=\kappa$. *Furthermore*, $y=E_1\varphi$ *is a solution of* (3.2), $y\in C^{p+2}$ *and the following estimates hold*:

$$|y(t)|\leq D\{t^2\|g\|+|\kappa|\},\quad |y'(t)|\leq Dt\|g\|,\quad |y''(t)|\leq D\|g\|,\qquad 0\leq t\leq 1, D=\text{const.}$$

Proof. The general solution of (3.5) is

$$\varphi(t)=t^J\left\{\int_0^t s^{-J}sg(s)\,ds+\left[c-\int_0^1 s^{-J}sg(s)\,ds\right]\right\}.$$

Hence, $\varphi\in C$ iff

$$J\left[c-\int_0^1 s^{-J}sg(s)\,ds\right]=0$$

and this implies the following form of any continuous solution φ of (3.5):

$$\varphi(t)=t^2\int_0^1 s^{-J}sg(ts)\,ds+\eta\begin{bmatrix}1\\0\\ \vdots\\0\end{bmatrix},\qquad \eta=\text{const}, \tag{3.9}$$

which yields $\varphi_k(0)=0$, $k=2,\cdots,2n$ and

$$\eta=\varphi_1(1)-\left(\int_0^1 s^{-J}sg(s)\,ds\right)_1.$$

The result follows now as in Case 1, on noting that

$$\varphi'(t)=tJ\int_0^1 s^{-J}sg(ts)\,ds+tg(t). \qquad \square$$

Remark 3.1. When dealing with the solution z of (3.3) we do not use formula (3.9) directly, but first we rewrite it in an equivalent form. To make this clear, we consider the system (3.3) and assume, that all eigenvalues of M are equal to zero. Let X_0 be the eigenspace of M and R be a projection onto X_0. Let us define

$$H=I-R.$$

For simplicity we select a basis in which M is reduced to the Jordan form and use it to construct the projections R and H. Then the following formulas are equivalent:

$$z(t)=t^2\int_0^1 s^{-M}s\mathring{f}(ts)\,ds+\Phi(t)R\gamma,\qquad R\gamma=Rz(1)-\int_0^1 Rs^{-M}s\mathring{f}(s)\,ds$$

and

$$z(t)=t^2\int_0^1 Hs^{-M}s\mathring{f}(ts)\,ds+t^M\int_1^t Rs^{-M}s\mathring{f}(s)\,ds+\Phi(t)R\eta,\qquad R\eta=Rz(1)$$

where $\Phi(t)=t^MR$. This follows immediately on noting that

$$t^M\int_1^t Rs^{-M}s\mathring{f}(s)\,ds=t^2\int_0^1 Rs^{-M}s\mathring{f}(ts)\,ds-\Phi(t)\int_0^1 Rs^{-M}s\mathring{f}(s)\,ds.$$

Case 3. $\sigma>0$.

LEMMA 3.3. *For every $g\in C^p[0,1]$, $p\geq 0$, there exists a unique solution φ of* (3.5) *subject to the terminal condition $\varphi(1)=\eta$. Furthermore, if $y=E_1\varphi$ then $y(0)=0$, $y\in C\cap C^{p+2}(0,1]$ and the following estimates hold*:

(i) *For* $0<\sigma<2$,

$$\left.\begin{aligned}&|y(t)|\leq t^{\sigma}\left(1+|\ln t|^{2n-1}\right)D(\|g\|+|\eta|),\\&|y'(t)|\leq t^{\sigma-1}\left(1+|\ln t|^{2n-1}\right)D(\|g\|+|\eta|),\\&|y''(t)|\leq t^{\sigma-2}\left(1+|\ln t|^{2n-1}\right)D(\|g\|+|\eta|),\end{aligned}\right.\qquad 0\leq t\leq 1,\quad D=\text{const.}$$

(ii) *For* $\sigma=2$,

$$\left.\begin{aligned}&|y(t)|\leq t^{2}\left(1+|\ln t|^{2n}\right)D(\|g\|+|\eta|),\\&|y'(t)|\leq t\left(1+|\ln t|^{2n}\right)D(\|g\|+|\eta|),\\&|y''(t)|\leq \left(1+|\ln t|^{2n}\right)D(\|g\|+|\eta|),\end{aligned}\right.\qquad 0\leq t\leq 1,\quad D=\text{const.}$$

(iii) *For* $\sigma>2$,

$$\left.\begin{aligned}&|y(t)|\leq t^{2}D(\|g\|+|\eta|),\\&|y'(t)|\leq tD(\|g\|+|\eta|),\\&|y''(t)|\leq D(\|g\|+|\eta|),\end{aligned}\right.\qquad 0\leq t\leq 1,\quad D=\text{const.}$$

(iv) (a) *If* $1\leq p<\sigma<p+1$, *then*

$$|y^{(p+1)}(t)|\leq\text{const.}\,t^{\sigma-p-1}\left(1+|\ln t|^{2n-1}\right),\qquad 0<t\leq 1.$$

(b) *If* $\sigma=p+1$, *then*

$$|y^{(p+1)}(t)|\leq\text{const.}\left(1+|\ln t|^{2n}\right),\qquad 0<t\leq 1.$$

(c) *If* $\sigma>p+1$, *then* $y\in C^{p+1}[0,1]$.

(v) (a) *If* $p+1<\sigma<p+2$, *then*

$$|y^{(p+2)}(t)|\leq\text{const.}\,t^{\sigma-p-2}\left(1+|\ln t|^{2n-1}\right),\qquad 0<t\leq 1.$$

(b) *If* $\sigma=p+2$, *then*

$$|y^{(p+2)}(t)|\leq\text{const.}\left(1+|\ln t|^{2n}\right),\qquad 0<t\leq 1.$$

(c) *If* $\sigma>p+2$, *then* $y\in C^{p+2}[0,1]$.

Proof. According to (3.6) the general solution of (3.5), which satisfies the terminal condition, is

$$\varphi(t)=t^{J}\int_{1}^{t}s^{-J}sg(s)\,ds+t^{J}\eta \tag{3.10}$$

and $\varphi\in C\cap C^{p+1}(0,1]$ yields $y\in C\cap C^{p+2}(0,1]$. Clearly,

$$|\varphi(t)|\leq\int_{t}^{1}\left|\left(\frac{t}{s}\right)^{J}\right|s\,ds\|g\|+Dt^{\sigma}\left(1+|\ln t|^{2n-1}\right)|\eta|,\qquad 0\leq t\leq 1,\quad D=\text{const.}$$

For the integral term we have,

$$\int_t^1 \left| \left(\frac{t}{s} \right)^J \right| s\,ds \leq \text{const.} \{ |t^J| + t^2 \}, \quad \text{for } \sigma \neq 2,$$

$$\int_t^1 \left| \left(\frac{t}{s} \right)^J \right| s\,ds \leq \text{const.}\, t^\sigma \{ |\ln t| + |\ln t|^{2n} \}, \quad \text{for } \sigma = 2.$$

This yields the estimates (i), (ii) and (iii).

We now establish (iv) and (v). For the case $p=1$, $g \in C^1[0,1]$ and (iv) (a), (b) and (c) are contained in (i), (ii) and (iii). Since $\varphi'(t)$ is a solution of

$$\varphi''(t) = \frac{1}{t}(J-I)\varphi'(t) + G(t),$$

where $G(t) = 2g(t) + tg'(t) \in C$, the estimate for $y'''(t)$ can be obtained as the estimate for $y''(t)$ before. This process can be continued for $p > 1$. □

We now consider the problem (3.3)

$$z'(t) = \frac{1}{t} Mz(t) + t\mathring{f}(t), \qquad 0 < t \leq 1.$$

Remark 3.2. The special structure of the matrix M yields to the following dependancy between the upper and lower rows of the matrix E:

$$E_2 = E_1 J.$$

This follows immediately from

$$ME = EJ \Leftrightarrow \begin{bmatrix} 0 & I \\ A_0 & I+A_1 \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} J.$$

We use this fact to write the solution y of (3.2) and its first derivative in the form

(3.11a) $\qquad y(t) = I_1 z(t) = E_1 \varphi(t),$

(3.11b) $\qquad y'(t) = I_2(z(t)/t) = E_1 J(\varphi(t)/t)$

or equivalently

(3.11c) $\qquad y'(t) = I_1 z'(t) = E_1 \varphi'(t),$

where I_1 and I_2 are $n \times 2n$ matrices consisting of the appropriate rows of I.

Before we discuss the problem (3.3) we introduce following notation.

Let $d(\lambda)$ be the dimension of the largest Jordan box of M having eigenvalues λ. Let σ_+ be the smallest of the positive real parts of the eigenvalues of M and d_+ be the dimension of the largest Jordan box of M which is associated with an eigenvalue whose real part is σ_+.

Let X_0 be the eigenspace of M corresponding to the eigenvalue $\lambda = 0$ and X_+ the invariant subspace associated with the eigenvalues with positive real part. Let P be a projection onto $X_0 \oplus X_+$. Define

$$P = R + S,$$

where R and S are the projections onto X_0 and X_+, respectively. In addition, let

$$Q = I - P.$$

Let X_+^1 be the eigenspace of M corresponding to the eigenvalue $\lambda=1$ and X_+^2 the invariant subspace associated with the eigenvalues whose real parts are grater than one. Let U and V be the projections onto X_+^1 and X_+^2, respectively and define

$$T=S-U-V.$$

As before (in Remark 3.1, following Lemma 3.2), we select a basis in which M reduces to Jordan form and use this basis to construct the projections.

The following two lemmas, stated without proofs, are consequences of Lemmas 3.1, 3.2, 3.3 and Remark 3.2.

LEMMA 3.4. *Let $z\in C$ be a solution of* (3.3) *with $f\in C$. Then*

$$Qz(0)=0 \quad \textit{and} \quad Sz(0)=0.$$

LEMMA 3.5. *For every $f\in C$ and a constant vector γ, there is a unique solution $z\in C$ satisfying* (3.3) *and the terminal condition $Pz(1)=P\gamma$. This solution has the form*

$$z(t)=(Hf)(t)+\Phi(t)P\gamma,$$

where H: $C\rightarrow C$ is a linear, bounded operator and

$$(Hf)(t)=t^2\int_0^1 Qs^{-M}s\mathring{f}(ts)\,ds+t^M\int_1^t Ps^{-M}s\mathring{f}(s)\,ds,$$
$$\Phi(t)=t^MP.$$

The corresponding two lemmas for the solution y of (3.2) are:

LEMMA 3.6. *Let $y(t)=z_1(t)\in C$ be a solution of* (3.2) *with $f\in C$. Then*

(3.12a) $$y_Q(0)=Q_1z(0)=0, \qquad y_Q'(0)=Q_1z'(0)=0,$$

(3.12b) $$\begin{gathered} y_S(0)=S_1z(0)=0,\\ y_S'(0)=S_1z'(0)=\lim_{t\to 0}T_1z'(t)+U_1z'(0)+V_1z'(0), \end{gathered}$$

where

$$\lim_{t\to 0}T_1z'(t)=\infty, \qquad V_1z'(0)=0.$$

LEMMA 3.7. *For every $f\in C$ and a constant vector γ, there is a unique solution $y\in C$ satisfying* (3.2) *and the terminal condition*

$$PY(1)=P(y(1),y'(1))^T=P\gamma.$$

This solution has the form

$$y(t)=I_1z(t),$$

where z is the solution of (3.3) *defined in Lemma* 3.5.

Let us consider the boundary value problem

(3.13a) $$z'(t)=\frac{1}{t}Mz(t)+t\mathring{f}(t), \qquad 0<t\leq 1,$$

(3.13b) $$B_0Y(0)+B_1Y(1)=\beta.$$

We shall now look for conditions which are necessary for (3.13) to define a well posed boundary value problem. Before we discuss this question, we note, that by (3.12b) $T\neq 0$

implies the discontinuity of the first derivative y' at $t=0$, and therefore, we have to find an additional condition which yields

$$B_{01}T_1z'(0)=0,$$

where B_{01} is the $m\times n$ matrix consisting of the last columns of B_0. This is done in the following remark.

Remark 3.3. We denote by T_1^E the $n\times i$ matrix consisting of nonzero columns of T_1E, say $t_{1,j}^E, j=1,\cdots,i$, where $i=\operatorname{rank}[T]$. Then, it follows immediately, that

$$B_{01}T_1z'(0)=0\Leftrightarrow t_{1,j}^E\in \operatorname{Ker}[B_{01}], \qquad j=1,\cdots,i.$$

For the subsequent analysis we make the following assumption.

A.3.1. If $T\neq 0$, then $t_{1,j}^E\in\operatorname{Ker}[B_{01}], j=1,\cdots,i$, $i=\operatorname{rank}[T]$.

We can now give the condition which is necessary for the uniqueness of a solution $y\in C$ of (3.13), i.e. to (3.1). For any projection matrix W, let us denote by $\tilde{W}$ the $2n\times i$ matrix consisting of the linearly independent columns of W, where $i=\operatorname{rank}[W]$. Then we have the following theorem.

THEOREM 3.1. *Let $y\in C$ be a solution of* (3.2) *defined by Lemma* 3.7. *Then y is a solution of* (3.1) *iff the $m\times m$ matrix*

$$\left[B_0\begin{bmatrix}R_1\\U_1M\end{bmatrix}\tilde{P}+B_1\tilde{P}\right]$$

is nonsingular.

Proof. Since, for every $P\gamma$ there exists a unique vector $\alpha\in X^m$ such that $P\gamma=\tilde{P}\alpha$, we can write the solution z of (3.3) in the following form:

$$z(t)=(Hf)(t)+(t^MP)\tilde{P}\alpha=\tilde{z}(t)+Z(t)\alpha. \tag{3.14}$$

Then we have

$$\begin{aligned}
&Pz(1)=PY(1)=\tilde{P}\alpha,\\
&Qz(1)=QY(1)=Q\tilde{z}(1)=\beta_Q(f),\\
&y_P(0)=P_1z(0)=R_1z(0)=R_1\tilde{z}(0)+R_1\tilde{P}\alpha=\beta_{R_1}(f)+R_1\tilde{P}\alpha,\\
&y_Q(0)=0,\\
&y_P'(0)=P_1z'(0)=S_1z'(0)=T_1z'(0)+U_1z'(0)\\
&\qquad =T_1z'(0)+U_1\tilde{z}'(0)+U_1Z'(0)\alpha=T_1z'(0)+\beta_{U_1}(f)+U_1M\tilde{P}\alpha,\\
&y_Q'(0)=0.
\end{aligned}$$

By substitution into (3.1b) we obtain

$$\left[B_0\begin{bmatrix}R_1\\U_1M\end{bmatrix}\tilde{P}+B_1\tilde{P}\right]\alpha=\beta-B_0\begin{bmatrix}\beta_{R_1}(f)\\\beta_{U_1}(f)\end{bmatrix}-B_1\beta_Q(f)$$

and the result follows. □

Remark 3.4. Note that Theorem 3.1 implies $m=\operatorname{rank}[P]$, i.e., B_0 and B_1 have $\operatorname{rank}[P]$ rows.

Finally, we shall derive the most general boundary condition of the form (3.1b) which yield a Fredholm alternative for the problem (3.1).

We define the differential expression

$$l_0(z)=z'(t)-\frac{1}{t}Mz(t), \qquad 0<t\leq 1$$

and associate with it the operator L_0 defined by

$$L_0 z=l_0(z), \qquad z=(z_1,tz_1')^T$$

if

$$z\in D(L_0)=\left\{z\in C \,\middle|\, l_0(z)=t\mathring{f}(t), f\in C, B_0\begin{bmatrix} z_1(0) \\ z_1'(0)\end{bmatrix}+B_1 z(1)=0\right\}.$$

THEOREM 3.2. *If the assumption* A.3.1 *holds and*

$$\text{(3.15)} \qquad \operatorname{rank}\left[B_0\begin{bmatrix} R_1 \\ U_1 M\end{bmatrix}, B_1\right]=k,$$

then L_0 is Fredholm with index rank $[P]-k$.

Proof. [2, Thm. 3.1] and Theorem 3.1. □

Clearly, if we assume that $m=k$, then L_0 is Fredholm with index zero iff $m=\operatorname{rank}[P]$.

4. Analytic results for the linear problem with variable coefficient-matrices $A_0(t)$ and $A_1(t)$. Here we study boundary value problem of the form

$$\text{(4.1a)} \qquad y''(t)-\frac{1}{t}A_1(t)y'(t)-\frac{1}{t^2}A_0(t)y(t)=f(t), \qquad 0<t\leq 1,$$

$$\text{(4.1b)} \qquad B_0 Y(0)+B_1 Y(1)=\beta,$$

where $y\in C\cap C^2(0,1]$ and B_0, B_1, β, f are defined as before in §3. Section 4.1 deals with the case when $A_0(t)$ and $A_1(t)\in C\cap C^1(0,1]$ and in §4.2 we assume $A_0(t)$ and $A_1(t)$ to be in $C^1[0,1]$. Finally, we consider two special cases, which have to be studied, as a first step in the analysis of the nonlinear problems. Before we construct the general continuous solution y of (4.1a), we prove the following lemma.

LEMMA 4.1. *Given $\delta>0$, $\nu>0$ and $f\in C[0,\delta]$, consider the linear system*

$$\text{(4.2a)} \qquad u'(t)=\frac{1}{t}Mu(t)+t^{\nu-1}\mathring{f}(t), \qquad 0<t\leq\delta, \quad u\in C[0,\delta],$$

$$\text{(4.2b)} \qquad Pu(\delta)=P\gamma.$$

Then

$$u(t)=(Kf)(t)+U_\delta(t)c,$$

where K: $C[0,\delta]\to C[0,\delta]$, is a bounded linear operator, $c\in X^m$, $m=$ rank$[P]$, *and*

$$\|K\|_\delta\leq \text{const.}\,\delta^\nu.$$

Proof. Consider

$$u'(t)=\frac{1}{t}Ju(t)+t^{\nu-1}g(t).$$

Case 1. $\sigma \leq 0$. For $u \in C[0,\delta]$ we have

$$u(t) = t^{\nu} \int_0^1 s^{-J} s^{\nu-1} g(ts)\, ds,$$

and hence

$$|u(t)| \leq t^{\nu} \max_{0 \leq t \leq \delta} |g(t)| \cdot \text{const.} = t^{\nu} \|g\|_{\delta} \cdot \text{const.}$$

Case 2. $\lambda = 0$. For $u \in C[0,\delta]$ it follows that

$$u(t) = t^{\nu} \int_0^1 s^{-J} s^{\nu-1} g(ts)\, ds + \kappa \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

and analogously

$$|u(t)| \leq t^{\nu} \|g\|_{\delta} \cdot \text{const.} + |\kappa|.$$

Case 3. $\sigma > 0$. The continuous solution u, with $u(\delta) = \eta$ has the form

$$u(t) = t^{J} \int_{\delta}^{t} s^{-J} s^{\nu-1} g(s)\, ds + \left(\frac{t}{\delta}\right)^{J} \eta = \tilde{u}(t) + \left(\frac{t}{\delta}\right)^{J} \eta$$

and in a similar way, as in Lemma 3.3, we obtain the following estimates for $\tilde{u}(t)$:

$$\text{if } \sigma \neq \nu, \text{ then } |\tilde{u}(t)| \leq \left(t^{\nu} + \delta^{\nu} \left| \left(\frac{t}{\delta}\right)^{J} \right| \right) \|g\|_{\delta} \cdot D, \quad D = \text{const.},$$

$$\text{if } \sigma = \nu, \text{ then } |\tilde{u}(t)| \leq \delta^{\nu} \left| \ln\left(\frac{t}{\delta}\right) \right| \left| \left(\frac{t}{\delta}\right)^{J} \right| \|g\|_{\delta} \cdot D, \quad D = \text{const.},$$

and hence

$$\|u\|_{\delta} \leq (\delta^{\nu} \|g\|_{\delta} + |\eta|) \cdot \text{const.}$$

The general continuous solution u of (4.2) has the form

$$\begin{aligned} u(t) &= t^{\nu} \int_0^1 Q s^{-M} s^{\nu-1} \mathring{f}(ts)\, ds + t^{M} \int_{\delta}^{t} P s^{-M} s^{\nu-1} \mathring{f}(s)\, ds + \left[\left(\frac{t}{\delta}\right)^{M} P \right] \tilde{P} c \\ &= (Kf)(t) + U_{\delta}(t) c, \end{aligned}$$

since for each $Pu(\delta) = P\gamma$, there exists a unique $c \in X^m$, such that $P\gamma = \tilde{P}c$, and we have

$$\|Kf\|_{\delta} \leq \text{const.}\, \delta^{\nu} \|f\|_{\delta}. \qquad \square$$

We shall now construct the general continuous solution of (4.1a).

4.1. Linear problem with $A_0, A_1 \in C \cap C^1(0,1]$. The matrices $A_0(t)$ and $A_1(t)$ are chosen to have the form

$$(4.3) \qquad A_i(t) = A_i + t^{\nu} C_i(t), \qquad C_i \in C[0,1], \qquad i = 0, 1.$$

If $0 < \nu < 1$, this is a simple characterization of the fact, that $A_i \in C \cap C^1(0,1]$.

Let $0<\nu<1$ and consider the system (4.1a) with (4.3). Applying the linear transformation $z(t)=(y(t),ty'(t))^T$ we obtain

$$z'(t)=\frac{1}{t}M(t)z(t)+t\mathring{f}(t), \qquad 0<t\leq 1, \tag{4.4}$$

where

$$M(t)=\begin{bmatrix} 0 & I \\ A_0(t) & I+A_1(t) \end{bmatrix}=M+t^{\nu}\begin{bmatrix} 0 & 0 \\ C_0(t) & C_1(t) \end{bmatrix}.$$

Hence (4.4) is equivalent to

$$z'(t)=\frac{1}{t}Mz(t)+t^{\nu-1}\mathring{C}(t)z(t)+t\mathring{f}(t), \qquad 0<t\leq 1, \tag{4.5}$$

where

$$\mathring{C}(t)=\begin{bmatrix} 0 & 0 \\ C_0(t) & C_1(t) \end{bmatrix}, \qquad \mathring{f}(t)=\begin{bmatrix} 0 \\ f(t) \end{bmatrix}. \tag{4.6}$$

LEMMA 4.2. *For every $f\in C$ and $\mathring{C}\in C$ there exists a unique, continuous solution of* (4.5) *subject to the terminal condition $Pz(1)=P\eta$. This solution satisfies $Qz(0)=0$ and $z\in C^1(0,1]$.*

Proof. According to Lemmas 3.5 and 4.1 the general continuous solution z of (4.5) satisfies for $0\leq t\leq\delta$,

$$z(t)=(KCz)(t)+(Hf)(t)+\Phi_\delta(t)P\eta, \tag{4.7}$$

where

$$(KCz)(t)=t^{\nu}\int_0^1 Qs^{-M}s^{\nu-1}\mathring{C}(ts)z(ts)\,ds+t^M\int_\delta^1 Ps^{-M}s^{\nu-1}\mathring{C}(s)z(s)\,ds$$

and by Lemma 4.1, $KC\colon C[0,\delta]\to C[0,\delta]$ is a bounded linear operator with

$$\|KCz\|_\delta\leq D\cdot\delta^{\nu}\|z\|_\delta, \qquad D=\text{const}.$$

Hence, for $\delta<\delta_0\leq(1/2D)^{1/\nu}$ the operator KC is contracting and $z\in C[0,\delta]$,

$$z(t)=(I-KC)^{-1}((Hf)(t)+\Phi_\delta(t)P\eta)=\tilde{z}(t)+Z_\delta(t)P\eta.$$

We note, that for δ small enough, $\operatorname{rank}[Z_\delta]=\operatorname{rank}[\Phi_\delta]$, since

$$(I-KC)^{-1}=I+KC\sum_{n=0}^{\infty}(KC)^n$$

and $\|KC\|_\delta\leq\text{const.}\,\delta^{\nu}$.

This solution can be continued uniquely to $t=1$. Clearly, $Qz(0)=0$ and by substitution of z into (4.5) we have $z\in C^1(0,1]$, which completes the proof. □

We now consider the boundary value problem

$$z'(t)=\frac{1}{t}Mz(t)+t^{\nu-1}\mathring{C}(t)z(t)+t\mathring{f}(t), \qquad 0<t\leq 1, \tag{4.8a}$$

$$B_0Y(0)+B_1Y(1)=\beta. \tag{4.8b}$$

Before we formulate a criterion for the existence of a unique solution $z \in C$ of (4.8) we need the following result.

LEMMA 4.3. *The general solution of* (4.8*a*) *has the form*

$$z(t) = \tilde{z}(t) + Z(t)\alpha, \qquad 0 \le t \le 1$$

where $\tilde{z}(t)$ *is the unique continuous solution of*

$$\tilde{z}'(t) - \frac{1}{t}M\tilde{z}(t) - t^{\nu-1}\mathring{C}(t)\tilde{z}(t) = t\mathring{f}(t), \qquad P\tilde{z}(1) = 0 \tag{4.9}$$

and $Z(t)$ *is the unique continuous* $2n \times m$ *matrix solution of*

$$Z'(t) - \frac{1}{t}MZ(t) - t^{\nu-1}\mathring{C}(t)Z(t) = 0, \qquad PZ(1) = \tilde{P}. \tag{4.10}$$

Proof. By Lemma 4.2, $\tilde{z}(t)$ and $Z(t)$ are solutions of (4.9) and (4.10) respectively iff

$$\begin{aligned} \tilde{z}(t) = {} & t^{\nu}\int_0^1 Qs^{-M}s^{\nu-1}\mathring{C}(ts)\tilde{z}(ts)\,ds + t^M\int_1^t Ps^{-M}s^{\nu-1}\mathring{C}(s)\tilde{z}(s)\,ds \\ & + t^2\int_0^1 Qs^{-M}s\mathring{f}(ts)\,ds + t^M\int_1^t Ps^{-M}s\mathring{f}(s)\,ds \\ = {} & (KC\tilde{z})(t) + (Hf)(t) \end{aligned} \tag{4.11}$$

and

$$Z(t) = t^{\nu}\int_0^1 Qs^{-M}s^{\nu-1}\mathring{C}(ts)Z(ts)\,ds + t^M\int_1^t Ps^{-M}s^{\nu-1}\mathring{C}(s)Z(s)\,ds + (t^MP)\tilde{P}. \tag{4.12}$$

The existence of continuous unique solutions of (4.11) and (4.12) is obvious by Lemma 4.2. The result follows now, since z must satisfy

$$z(t) = (KCz)(t) + (Hf)(t) + (t^MP)Pz(1)$$

and for every $Pz(1)$ there exists a unique $\alpha \in X^m$ such that $Pz(1) = P(\tilde{z}(1) + Z(1)\alpha) = \tilde{P}\alpha$. □

Remark 4.1. By Lemma 4.2 and (3.11b), $y'(t)$ can be written as

$$y'(t) = t^{\nu-1}\sum_{i=1}^{n} e_i\xi_i(t),$$

where $\{e_i\}$, $i = 1, \cdots, n$ are linearly independent and $\xi_i(t) \in C$, $i = 1, \cdots, n$. So we have to assume, $B_{01} = 0$. With this assumption we have the following theorem.

THEOREM 4.1. *Let* $y(t) = z_1(t)$. *Let* $z \in C$ *be a solution of* (4.8a) *defined by Lemma* 4.3. *Then* y *is a solution of* (4.8) *iff the* $m \times m$ *matrix* $[B_{00}R_1Z(0) + B_1[QZ(1) + \tilde{P}]]$ *is nonsingular.*

Proof. From Lemma 4.3 and (3.11) we have

$$\begin{aligned} Pz(1) &= PY(1) = \tilde{P}\alpha, \\ Qz(1) &= QY(1) = Q\tilde{z}(1) + QZ(1)\alpha, \\ y_P(0) &= P_1z(0) = R_1\tilde{z}(0) + R_1Z(0)\alpha, \\ y_Q(0) &= Q_1z(0) = 0. \end{aligned}$$

The substitution into (4.8b) gives

$$\left[B_{00}R_1Z(0)+B_1\left[QZ(1)+\tilde{P}\right]\right]\alpha=\beta-B_{00}R_1\tilde{z}(0)-B_1Q\tilde{z}(1)$$

and the result follows. □

As in §3, we see from Theorem 4.1, that $m=\operatorname{rank}[P]$ is the necessary and sufficient condition for the uniqueness of a continuous solution y of (4.8), where P is a projection onto the eigenspace of $M(0)$ corresponding to the eigenvalue zero and the invariant subspace of $M(0)$ associated with the eigenvalues with positive real parts, cf. (4.5).

4.2. Linear problem with $A_0, A_1\in C^1[0,1]$. Using Taylor's theorem we can write $A_0(t)$ and $A_1(t)$ in the form

$$A_i(t)=A_i+tC_i(t),\qquad i=0,1,$$

where $C_0, C_1\in C[0,1]$. This is equivalent to (4.3) with $\nu=1$. The corresponding system of the first order is

$$z'(t)=\frac{1}{t}Mz(t)+\mathring{C}(t)z(t)+t\mathring{f}(t),\qquad 0<t\leq 1 \tag{4.13}$$

and by Lemma 4.2 we have the following result.

LEMMA 4.4. *For every $f\in C$ and $\mathring{C}\in C$ there exists a unique, continuous solution of* (4.13) *subject to the terminal condition $Pz(1)=P\gamma$. This solution satisfies $Qz(0)=0$.*

We now investigate the smoothness properties of the solution z of (4.13). Let $f\in C^p[0,1]$, and consider separately three cases $\sigma\leq 0$, $\lambda=0$, $\sigma>0$.

Case 1. $\sigma\leq 0$. Let $\mathring{C}\in C^p[0,1]$. According to Lemma 4.1, the solution z can be written as

$$z(t)=t\xi(t),\qquad \xi(t)\in C^p[0,1].$$

Hence, the system (4.13) is equivalent to

$$z'(t)=\frac{1}{t}Mz(t)+t\left[\mathring{C}(t)\xi(t)+\mathring{f}(t)\right]$$

and by Lemma 3.1, $y\in C^{p+2}[0,1]$.

Case 2. $\lambda=0$. Let $\mathring{C}\in C^p[0,1]$. Then z has the form

$$z(t)=t\xi(t)+R\gamma,\qquad \xi(t)\in C^p[0,1],$$

where $t\xi(t)$ is the solution of (4.13) and $R\gamma$ is the solution of

$$z'(t)=\frac{1}{t}Mz(t).$$

Hence, (4.13) can be written as

$$z'(t)=\frac{1}{t}Mz(t)+t\left[\mathring{C}(t)\xi(t)+\mathring{f}(t)\right]+\begin{bmatrix}0\\ C_0(t)R_1\gamma\end{bmatrix}$$

and by Lemma 3.2, $y\in C^{p+2}[0,1]$, if $C_0(t)\in C^{p+1}[0,1]$.

Case 3. $\sigma>0$. From Lemma 4.1 we have for $\mathring{C}\in C^p[0,1]$,

$$z(t)=t\eta(t)+t^{\sigma_+}\left(1+(\ln t)^{d_+}\right)\xi(t),\qquad \eta,\xi\in C^p[0,1].$$

If $0<\sigma\leq 1$, then $y\in C\cap C^{p+1}(0,1]$. If $\sigma>1$, then $z=t\xi(t)$, and $\xi(t)\in C^p[0,1]$. Hence, (4.13) can be written as

$$z'(t)=\frac{1}{t}Mz(t)+t[\mathring{C}(t)\xi(t)+\mathring{f}(t)]$$

and by Lemma 3.3 we have

$$\begin{aligned}&\text{if } 1<\sigma\leq 2, \text{ then } y\in C^{p+1}\cap C^{p+2}(0,1],\\ &\text{if } \sigma>2, \text{ then } y\in C^{p+2}.\end{aligned}$$

Hence, the following result is obvious.

LEMMA 4.5. *For every $\mathring{C},f\in C^p[0,1]$, $p\geq 0$ there exists a unique solution $y\in C$ of* (4.13) *subject to the terminal condition $PY(1)=P\gamma$ and*

(i) *if $\sigma_+>p+2$ and $C_0\in C^{p+1}$, then $y\in C^{p+2}$,*
(ii) *if $p+1<\sigma_+\leq p+2$, then $y\in C^{p+1}\cap C^{p+2}(0,1]$,*
(iii) *if $p<\sigma_+\leq p+1$, then $y\in C^p\cap C^{p+1}(0,1]$.*

Consider the boundary value problem

$$z'(t)=\frac{1}{t}Mz(t)+\mathring{C}(t)z(t)+t\mathring{f}(t),\qquad 0<t\leq 1, \tag{4.14a}$$

$$B_0Y(0)+B_1Y(1)=\beta. \tag{4.14b}$$

By Lemma 4.3 we have immediately

LEMMA 4.6. *The general solution of* (4.14a) *is*

$$z(t)=\tilde{z}(t)+Z(t)\alpha,\qquad 0\leq t\leq 1,$$

where $\tilde{z}(t)$ is the unique continuous solution of

$$\tilde{z}'(t)-\frac{1}{t}M\tilde{z}(t)-\mathring{C}(t)\tilde{z}(t)=t\mathring{f}(t),\qquad P\tilde{z}(1)=0 \tag{4.15}$$

and $Z(t)$ is the unique continuous $2n\times m$ matrix solution of

$$Z'(t)-\frac{1}{t}MZ(t)-\mathring{C}(t)Z(t)=0,\qquad PZ(1)=\tilde{P}. \tag{4.16}$$

From (4.11) and (4.12) we have for $\tilde{z}$ and Z

$$\tilde{z}(t)=t\int_0^1 Qs^{-M}\mathring{C}(ts)\tilde{z}(ts)\,ds+t^M\int_1^t Ps^{-M}\mathring{C}(s)\tilde{z}(s)\,ds+(Hf)(t), \tag{4.17}$$

$$Z(t)=t\int_0^1 Qs^{-M}\mathring{C}(ts)Z(ts)\,ds+t^M\int_1^t Ps^{-M}\mathring{C}(s)Z(s)\,ds+(t^MP)\tilde{P}. \tag{4.18}$$

This yields

$$Pz(1)=PY(1)=\tilde{P}\alpha, \tag{4.19a}$$

$$Qz(1)=QY(1)=Q\tilde{z}(1)+QZ(1)\alpha, \tag{4.19b}$$

$$y_P(0)=P_1z(0)=R_1\tilde{z}(0)+R_1Z(0)\alpha. \tag{4.20a}$$

Using (4.14a) and Lemma 4.5 we have finally

$$y_P'(0)=T_1z'(0)+U_1M\tilde{z}(0)+U_1MZ(0)\alpha, \tag{4.20b}$$

$$y_Q'(0)=0. \tag{4.20c}$$

The next result is a consequence of A.3.1, (4.19) and (4.20).

THEOREM 4.2. *Let* $y(t)=z_1(t)$ *and* $z(t)\in C$ *be a solution of* (4.14a) *defined by Lemma* 4.6. *Then* y *is a solution of* (4.14) *iff the* $m\times m$ *matrix*

$$\left[B_0\begin{bmatrix} R_1 \\ U_1M \end{bmatrix}Z(0)+B_1[QZ(1)+\tilde{P}]\right]$$

is nonsingular.

Finally, we define the operator L with domain $D(L)=D(L_0)$ as

$$L=L_0+\mathring{C},$$

where $\mathring{C}$ is a bounded linear operator given by

$$(\mathring{C}z)(t)=\mathring{C}(t)z(t), \qquad z(t)=(z_1(t),tz_1'(t))^T.$$

Now we extend the result of Theorem 3.2 to L.

THEOREM 4.3. *If the assumption* A.3.1 *holds and*

$$\operatorname{rank}\left[B_0\begin{bmatrix} R_1 \\ U_1M \end{bmatrix},B_1\right]=k, \tag{4.21}$$

then L *is Fredholm with index* $\operatorname{rank}[P]-k$.

Proof. From Lemma 3.5, L_0 is compact. Since $\mathring{C}$ is bounded, $\mathring{C}$ is L_0 compact and the result follows from the second stability theorem for Fredholm operators [6]. □

At the end of this section two special cases will be discussed. For the case when

$$A_1(t)\equiv A_1 \quad \text{and} \quad A_0(t)=A_0+tC_0(t) \tag{4.22}$$

where $C_0\in C^1[0,1]$, the above theory can be applied.

We have a similar situation in the second case when

$$A_1(t)\equiv A_1 \quad \text{and} \quad A_0(t)=A_0+t^2C_0(t), \tag{4.23}$$

where $C_0\in C[0,1]$. Now the corresponding system of the first order is equivalent to

$$z'(t)=\frac{1}{t}Mz(t)+t[\mathring{C}(t)z(t)+\mathring{f}(t)], \tag{4.24}$$

and it is clear that all results from this section are also valid.

5. Nonlinear problem. We now consider the nonlinear boundary value problem

$$y''(t)-\frac{A_1}{t}y'(t)-\frac{A_0}{t^2}y(t)=f(t,y(t)), \qquad 0<t\leq 1, \tag{5.1a}$$

$$B(y(0),y'(0);y(1),y'(1))=0. \tag{5.1b}$$

Because of the problems that appear in practice, we consider the case when y is in $C^1\cap C^2(0,1]$ and furthermore make the following assumptions.

A.5.1. $T\equiv 0$.

A.5.2. $f\colon D_1 \to X^n$ and $B\colon D_2 \to X^m$ are nonlinear mappings and $D_1 \subset [0,1] \times X^n$, $D_2 \subset X^n \times X^n \times X^n \times X^n$ are appropriate open sets. As in linear case, we take $m = \operatorname{rank}[P]$.

A.5.3. Problem (5.1) has a solution $y \in C^1 \cap C^2(0,1]$. For this solution we define the spheres

$$S_{\rho_1}(y(t)) = \{v \in X^n \mid |v - y(t)| \le \rho_1,\ \rho_1 > 0\},$$
$$S_{\rho_2}(y'(t)) = \{w \in X^n \mid |w - y'(t)| \le \rho_2,\ \rho_2 > 0\}$$

and the tube

$$T_\rho = \{(t,v) \mid 0 \le t \le 1,\ v \in S_{\rho_1}(y(t))\}.$$

A.5.4. $f(t,v)$ is continuously differentiable with respect to v and $f_v(t,v)$ is continuous on T_ρ. $B(u_1,u_2;\ u_3 u_4)$ is continuously differentiable on $S_{\rho_1}(y(0)) \times S_{\rho_2}(y'(0)) \times S_{\rho_1}(y(1)) \times S_{\rho_2}(y'(1))$.

A.5.5. The solution y of (5.1) is isolated. This means that

$$u''(t) - \frac{A_1}{t}u'(t) - \frac{1}{t^2}(A_0 + t^2 C_0(t))u(t) = 0, \qquad 0 < t \le 1, \quad u \in C, \tag{5.2a}$$

$$B_{00}u(0) + B_{01}u'(0) + B_{10}u(1) + B_{11}u'(1) = 0, \tag{5.2b}$$

where

$$C_0(t) = f_v(t, y(t)),$$
$$B_{00} = B_{u_1}(y(0), y'(0); y(1), y'(1)), \quad B_{01} = B_{u_2}(y(0), y'(0); y(1), y'(1)),$$
$$B_{10} = B_{u_3}(y(0), y'(0); y(1), y'(1)), \quad B_{11} = B_{u_4}(y(0), y'(0); y(1), y'(1)),$$

has only the trivial solution.

The main aim of this section is to show that if the assumptions A.5.1–A.5.5 hold, then the solution of (5.1) is stable as defined in [8]. To prove it we rewrite the problem (5.1) as follows.

Let $\hat{P}$ be the unique $m \times 2n$ matrix such that $\tilde{P}\hat{P} = P$. Then, as in Lemma 3.5, any continuous solution of (5.1) satisfies

$$y(t) = I_1\{(Hf(\cdot, y(\cdot)))(t) + \Phi(t)\tilde{P}\alpha\},$$
$$\alpha = \alpha - B(y(0), y'(0); y(1), y'(1)),$$

where $\alpha = \hat{P}Y(1)$, $\alpha \in X^m$. Equivalently, we write

$$x = N(x),$$

where $x = (y, \alpha)$, $N\colon U_{\rho_1} \times X^m \to C^m$ is a compact nonlinear operator, $U_{\rho_1} = \{u \in C \mid u(t) \in S_{\rho_1}(y(t)),\ 0 \le t \le 1\}$ and $C^m = C \times X^m$. For C^m we define $\|x\| = \max\{\|y\|, |\alpha|\}$, and hence C^m is a Banach space. Let denote by F^* the Fréchet derivative of N at $x^* = (y, \hat{P}Y(1))$, then it follows by A.5.5, that $(I - F^*)^{-1}$ exists. This together with A.5.1 and A.5.4 yields the result, cf. [8, Thm. 2.6].

The smoothness results for the solution y of (5.1) follow in a straightforward way from Lemmas 3.1, 3.2 and 3.3. Let by $f \in C^p[T_\rho]$ denote, that $f(t,v)$ is p times continuously differentiable on T_ρ. Then we have

LEMMA 5.1. *Let $f \in C^p[T_\rho]$, $p \ge 0$. Then*

(i) $y \in C^{p+2}(0,1]$.

(ii) *If all eigenvalues of M are nonpositive, then $y \in C^{p+2}[0,1]$.*
(iii) *If $p+1<\sigma_+<p+2$, then*

$$|y^{(p+2)}(t)| \leq \text{const.}\, t^{\sigma_+ p-2}\left(1+|\ln t|^{d_+-1}\right), \qquad 0<t\leq 1.$$

If $\sigma_+=p+2$, then

$$|y^{(p+2)}(t)| \leq \text{const.}\left(1+|\ln t|^{d_+}\right), \qquad 0<t\leq 1.$$

If $\sigma_+>p+2$, then $y \in C^{p+2}$.

We now consider the first example [13], $n=2$,

$$y''(t)+\frac{3I}{t}y'(t)=F(t,y(t)), \qquad 0<t\leq 1, \tag{5.3a}$$

$$y'(0)=0, \qquad B_1Y(1)=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2/3 & 0 & 1 \end{bmatrix}Y(1)=0, \tag{5.3b}$$

where
$$F(t,y(t))=\begin{bmatrix} y_1(t)y_2(t)-\mu^2y_2(t)-2\gamma \\ -\frac{1}{2}y_1^2(t)+\mu^2y_1(t) \end{bmatrix},$$

and μ, γ are problem parameters.

Since $A_0\equiv 0$ and $A_1=-3I$ we have

$$M=\begin{bmatrix} 0 & I \\ 0 & -2I \end{bmatrix}, \quad J=\begin{bmatrix} 0 & & & \\ & 0 & & \\ & & -2 & \\ & & & -2 \end{bmatrix}, \quad E=\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

We consider now the Fréchet derivative of the nonlinear operator defined by (5.3) and show that it is Fredholm with index zero. Let $y \in C$ be a solution of (5.3). Then

$$u''(t)+\frac{3I}{t}u'(t)-C_0(t)u(t)=0, \qquad 0<t\leq 1, \tag{5.4a}$$

$$u'(0)=0, \qquad B_1U(1)=0, \tag{5.4b}$$

where
$$C_0(t)=\begin{bmatrix} y_2(t) & y_1(t)-\mu^2 \\ -y_1(t)+\mu^2 & 0 \end{bmatrix} \in C, \qquad U(t)=\begin{bmatrix} u(t) \\ u'(t) \end{bmatrix}.$$

Defining $v(t)=(u(t),tu'(t))^T$ we have

$$v'(t)=\frac{1}{t}Mv(t)+t\mathring{C}(t)v(t), \qquad 0<t\leq 1, \tag{5.5a}$$

$$u'(0)=0, \qquad B_1U(1)=0, \tag{5.5b}$$

where
$$\mathring{C}(t)=\begin{bmatrix} 0 & 0 \\ C_0(t) & 0 \end{bmatrix}.$$

Since $S\equiv 0$, any continuous solution u of (5.5a) can be written as

$$u(t)=I_1t^2\int_0^1(Q+R)s^{-M}s\mathring{C}(ts)v(ts)\,ds+c_1\begin{bmatrix} 1 \\ 0 \end{bmatrix}+c_2\begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{5.6}$$

and

$$u'(t)=I_1 Mt\int_0^1 (Q+R)s^{-M}s\mathring{C}(ts)v(ts)\,ds, \qquad u'(0)=0. \tag{5.7}$$

Hence, we recognize the condition $u'(0)=0$ to be necessary for u to be continuous. Furthermore $u\in C^2$. Finally, $T\equiv 0$, $B_0\equiv 0$ and

$$\operatorname{rank}[B_1]=m=2=\operatorname{rank}[R].$$

The result follows now by Theorem 4.3 and the nonlinear problem (5.3) is well posed.

Since in this case M has no eigenvalues which are positive, $f(t,y)$ is continuous with respect to t for $0\le t\le 1$ and all $y\in C$, we can reduce the boundary value problem (5.3) to an initial value problem. Hence the existence of a continuously differentiable solution of (5.3) follows by a contracting argument. Since $Q+R\equiv I$ we have

$$y(t)=I_1 t^2\int_0^1 s^{-M}s\mathring{F}(ts,y(ts))\,ds+\begin{bmatrix} c_1\\ c_2\end{bmatrix},$$

where

$$y(0)=\begin{bmatrix} c_1\\ c_2\end{bmatrix}=\hat{R}Y(1)-B_1Y(1)=\begin{bmatrix} \frac{1}{2}y_1'(1)\\ \frac{1}{3}y_2(1)-\frac{1}{2}y_2'(1)\end{bmatrix}.$$

Furthermore, by Theorem 5.1, $y\in C^\infty[0, 1]$.

We shall now consider the case when

$$y''(t)-\frac{A_1}{t}y'(t)-\frac{A_0}{t^2}y(t)=f\left(t, \frac{y(t)}{t}\right), \qquad 0<t\le 1, \quad y\in C^1, \tag{5.8a}$$

$$B_0(y(0),y'(0);y(1),y'(1))=0. \tag{5.8b}$$

Since the proof of the main result is based on exactly the same arguments, if the assumptions are chosen properly, we shall be satisfied with their formulation and treat the example from [7] in detail.

A.5.3. We have to change the definition of $S_{\rho_1}(y(t))$ to

$$S_{\rho_1}\left(\frac{y(t)}{t}\right)=\left\{v\in X^n \,\middle|\, \left|\frac{v}{t}-\frac{y(t)}{t}\right|\le\rho_1,\ \rho_1>0\right\}.$$

Clearly if $v\in S_{\rho_1}(y(t)/t)$, then $v\in S_{\rho_1}(y(t))$.

We assume, that $y(t)/t\in C[0, 1]$, $y(0)=0$ and define

$$T_\rho=\left\{(t,v)\,\middle|\, 0\le t\le 1,\ v\in S_{\rho_1}\left(\frac{y(t)}{t}\right)\right\}.$$

A.5.5. The only change is the structure of the term in parentheses, which now has the form

$$A_0+tC_0(t), \qquad C_0(t)=f_v\left(t, \frac{y(t)}{t}\right).$$

The other assumptions remain valid with reference to the above definitions. Note, that if $y(0)=0$, then $y(t)/t\in C$ is equivalent to $y\in C^1$.

Consider the following problem [7], $n=2$.

$$y''(t)+\frac{1}{t}y'(t)-\frac{1}{t^2}y(t)=F\left(t,\frac{y(t)}{t}\right),\qquad 0<t\leq 1,\quad y\in C^1, \tag{5.9a}$$

$$y(0)=0,\qquad B_1Y(1)=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & -1/3 & 0 & 1\end{bmatrix}Y(1)=\begin{bmatrix}1\\0\end{bmatrix}, \tag{5.9b}$$

where

$$F\left(t,\frac{y(t)}{t}\right)=\alpha\begin{bmatrix}t\left(\beta+\frac{y_1(t)}{t}\cdot\frac{y_2(t)}{t}\right)\\ t\left(1-\left(\frac{y_1(t)}{t}\right)^2\right)\end{bmatrix}$$

and α, β are parameters.

$A_0=I$ and $A_1=-I$, so we have

$$M=\begin{bmatrix}0 & I\\ I & 0\end{bmatrix},\quad J=\begin{bmatrix}1 & & & \\ & 1 & & \\ & & -1 & \\ & & & -1\end{bmatrix},\quad E=\begin{bmatrix}0 & 1 & 0 & 1\\ 1 & 0 & 1 & 0\\ 0 & 1 & 0 & -1\\ 1 & 0 & -1 & 0\end{bmatrix}.$$

The system corresponding to (5.5) is

$$v'(t)=\frac{1}{t}Mv(t)+\mathring{C}(t)v(t),\qquad 0<t\leq 1, \tag{5.10a}$$

$$u(0)=0,\qquad B_1U(1)=0, \tag{5.10b}$$

where $\mathring{C}\in C$,

$$\mathring{C}(t)=\begin{bmatrix}0 & 0\\ C_0(t) & 0\end{bmatrix},\qquad C_0(t)=\alpha\begin{bmatrix}y_2(t) & y_1(t)\\ 2y_1(t) & 0\end{bmatrix}.$$

Since $R=0$ and $S=U$, any continuous solution of (5.10) has the form

$$u(t)=I_1t\int_0^1 Qs^{-M}\mathring{C}(ts)v(ts)\,ds+I_1t^M\int_1^t Us^{-M}\mathring{C}(s)v(s)\,ds+c_1t\begin{bmatrix}1\\0\end{bmatrix}+c_2t\begin{bmatrix}0\\1\end{bmatrix},$$
$$u(0)=0.$$

Hence, the condition $u(0)=0$ is the regularity condition for the solution of the problem. Furthermore, $u\in C^1$. Since $\operatorname{rank}[B_1]=2=\operatorname{rank}[U]$, the problem is well posed.

The transformation of (5.9) to the first order system yields

$$z'(t)=\frac{1}{t}Mz(t)+t\mathring{F}\left(t,\frac{z_1(t)}{t}\right), \tag{5.11a}$$

$$B_1Y(1)=\begin{bmatrix}1\\0\end{bmatrix}, \tag{5.11b}$$

and any continuous solution of this problem can be written in the following form

$$y(t)=I_1t^2\int_0^1 Qs^{-M}s\mathring{F}\left(ts,\frac{y(ts)}{ts}\right)ds+I_1t^M\int_1^t Us^{-M}s\mathring{F}\left(s,\frac{y(s)}{s}\right)ds+t\begin{bmatrix}c_1\\c_2\end{bmatrix}, \tag{5.12}$$

where

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = (\hat{U} - B_1)Y(1) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} y_1'(1)+1 \\ \frac{2}{3}y_2(1) \end{bmatrix}.$$

The existence of a solution of (5.12) can be shown by contraction and $y \in C^1$. Since $z''(t)$ is the solution of

$$z''(t) = \frac{1}{t}(M-I)z'(t) + G(t),$$

where

$$G(t) = 2\mathring{F}\left(t, \frac{z_1(t)}{t}\right) + t\left[\mathring{F}\left(t, \frac{z_1(t)}{t}\right)\right]'$$

and since $z_1 = y \in C^1$ implies $\mathring{F} \in C^1$, it follows from Lemmas 3.1 and 3.2 that $y \in C^2$. These lemmas are valid, because the eigenvalues of $M - I$ are nonpositive. Straightforward application of this argument yields $y \in C^\infty$.

Finally, we extend the results of Lemma 5.1 to the solution of (5.8).

LEMMA 5.2. *Let* $f \in C^p[T_\rho]$, $p \geq 0$. *Then*

(i) $y \in C^{p+2}(0, 1]$.

(ii) *If all eigenvalues of* M *are nonpositive then* $y \in C^{p+2}[0, 1]$.

(iii) *If* $p+1 < \sigma_+ < p+2$, *then*

$$|y^{(p+2)}(t)| \leq \text{const.}\, t^{\sigma_+ - p - 2}\left(1 + |\ln t|^{d_+ - 1}\right), \qquad 0 < t \leq 1.$$

If $\sigma_+ = p+2$, *then*

$$|y^{(p+2)}(t)| \leq \text{const.}\left(1 + |\ln t|^{d_+}\right), \qquad 0 < t \leq 1.$$

If $\sigma_+ > p+2$, *then* $y \in C^{p+2}$.

Acknowledgment. I am indebted to Richard Weiss for bringing these problems to my attention and for several helpful discussions.

REFERENCES

[1] D. C. BRABSTON, JR., *Numerical solution of singular endpoint boundary value problems*, Ph. D. thesis, part II, Applied Mathematics, California Institute of Technology, Pasadena, 1974.

[2] F. R. DE HOOG AND R. WEISS, *Difference methods for boundary value problems with a singularity of the first kind*, SIAM J. Numer. Anal., 13 (1976), pp. 775–813.

[3] F. R. DE HOOG AND R. WEISS, *On the boundary value problem for systems of ordinary differential equations with a singularity of the second kind*, SIAM J. Numer. Anal., 11 (1980), pp. 41–60.

[4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, *Part* I, Interscience, New York, 1967.

[5] P. JAMET, *On the convergence of finite difference approximations to one-dimensional singular boundary-value problems*, Numer. Math., 14 (1970), pp. 355–378.

[6] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[7] H. B. KELLER AND A. W. WOLFE, *On the nonunique equilibrium states and buckling mechanism of spherical shells*, J. Soc. Indust. Applied Math., 13 (1965), pp. 674–705.

[8] ———, *Approximation methods for nonlinear problems with application to two-point boundary value problems*, Math. Comp., 29 (1975), pp. 464–474.

[9] F. NATTERER, *A generalized spline method for singular boundary value problems in ordinary differential equations*, Linear Algebra and Appl., 7 (1973), pp. 189–216.

[10] ______, *Das Differenzenverfahren für singuläre Rand-Eigenwertaufgaben gewöhnlicher Differentialgleichungen*, Numer. Math., 23 (1975), pp. 387–409.

[11] S. V. PARTER, M. L. STEIN AND P. R. STEIN, *On the multiplicity of solutions of a differential equation arising in chemical reactor theory*, Tech. Rep. 194, Dept. Computer Sciences, Univ. of Wisconsin, Madison, 1973.

[12] S. V. PARTER, *A-posteriori error estimates*, Tech. Rep. 214, Dept. Computer Sciences, Univ. of Wisconsin, Madison, 1974.

[13] P. RENTROP, *Eine Taylorreihenmethode zur numerischen Lösung von Zwei-Punkt Randwertproblemen mit Anwendung auf singuläre Probleme der nichtlinearen Schalentheorie*, TUM, Institut für Mathematik, München, 1977.

[14] R. D. RUSSELL AND L. F. SHAMPINE, *Numerical methods for singular boundary value problems*, SIAM J. Numer. Anal., 12 (1975), pp. 13–35.

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1502-0006 $01.25/0

OSCILLATION THEOREMS FOR DAMPED DIFFERENTIAL EQUATIONS OF EVEN ORDER WITH DEVIATING ARGUMENTS*

S. R. GRACE† AND B.S. LALLI†

Abstract. New oscillation criteria for the damped differential equations with deviating arguments of the form $x^{(n)}(t)+p(t)x^{(n-1)}(t)+q(t)f(x[g_1(t)],x[g_2(t)],\cdots,x[g_m(t)])=0$ are established.

1. Introduction. In this paper we are dealing with the oscillatory behavior of the nth order differential equations with deviating arguments of the form

$$x^{(n)}(t)+p(t)x^{(n-1)}(t)+q(t)f(x[g_1(t)],x[g_2(t)],\cdots,x[g_m(t)])=0, \tag{1}$$

where n is even, $f\in C[R^m, R=(-\infty,\infty)]$, $g_i,p,q\in C[[t_0,\infty),R]$ and $i=1,2,\cdots,m$, and such that:

(i) the functions p,q are nonnegative on the interval $[t_0,\infty)$ and $q(t)$ is not identically zero on any ray $[t^*,\infty)$,

(ii) the function f is nondecreasing on the set Y, where $Y=\{(y_1,y_2,\cdots,y_m)\colon y_i\in R$ and either $y_i>0$ or $y_i<0$, $i=1,2,\cdots,m\}$, i.e. for every $i=1,2,\cdots,m$, $y_i\le z_i$ implies $f(y_1,y_2,\cdots,y_m)\le f(z_1,z_2,\cdots,z_m)$ and

$$f(y_1,y_2,\cdots,y_m)>0 \quad \text{if } y_i>0 \quad \text{for all } i,$$
$$f(y_1,y_2,\cdots,y_m)<0 \quad \text{if } y_i<0 \quad \text{for all } i,$$

(iii) for every $i=1,2,\cdots,m$,

$$g_i(t)\to\infty \quad \text{as } t\to\infty.$$

Without further mention we will assume throughout that every solution $x(t)$ of (1) that is under consideration here is continuable to the right and is nontrivial, i.e. $x(t)$ is defined on some ray $[T_x,\infty)$ and $\sup\{|x(t)|\colon t\ge T\}>0$ for every $T\ge T_x$. Such a solution will be called oscillatory if its set of zeros is unbounded and will be called nonoscillatory otherwise.

Recently C. C. Yeh [9] gave some oscillation criteria for the second order differential equation

$$\ddot{x}+p(t)\dot{x}+q(t)f(x)=0, \qquad \left(\cdot=\frac{d}{dt}\right), \tag{2}$$

where $f\in C[R,R]$, $p,q\in C[[t_0,\infty),R]$ and $xf(x)>0$ for $x\ne 0$. His results improve previous oscillation criteria obtained in [8] for the undamped case ($p=0$) and in [4], [7] for the undamped linear case ($p=0, f(x)=x$). On the other hand, these theorems fail to describe the oscillatory behavior of solutions of the equation

$$\ddot{x}+\frac{k^2}{t^\alpha}x=0 \tag{3}$$

* Received by the editors June 15, 1982, and in revised form November 29, 1982.

† Department of Mathematics, University of Saskatchewan, Saskatoon, Saskatchewan, S7N 0W0 Canada.

according to the values of the constants α, k and $\alpha>1$. So we offer here the following two theorems which unify and improve the results in [9] and can be used to investigate (3). The proofs are similar to those of [9, Thms. 1, 2 and 3] and hence we shall not include them.

THEOREM A. *Let $f'(x)$ exist and $f'(x)\geq k_1>0$ for $x\neq 0$, $('=d/dx)$. If*

$$(\alpha_1)\quad \limsup_{t\to\infty} t^{1-m}\int_{t_0}^{t}(t-u)^{m-3}u^{l}\cdot\left[(t-u)^2q(u)-\frac{1}{4k_1}\left((t-u)\left(p(u)-\frac{l}{u}\right)+m-1\right)^2\right]du=\infty$$

for some integer $m\geq 3$ and some constant l, then every solution of (2) *is oscillatory.*

THEOREM B. *Let $q(t)\geq 0$ and $f(x)/x\geq k_2>0$ for $x\neq 0$. If*

$$(\alpha_2)\quad \limsup_{t\to\infty} t^{1-m}\int_{t_0}^{t}(t-u)^{m-3}u^{l}\cdot\left[k_2(t-u)^2q(u)-\frac{1}{4}\left((t-u)\left(p(u)-\frac{l}{u}\right)+m-1\right)^2\right]du=\infty$$

for some integer $m\geq 3$ and some constant l, then every solution of (2) *is oscillatory.*

We have combined conditions (C_1) and (C_2) of [9, Thm. 1] into one condition, namely (α_1). In (3), if we let $k=1$ and $\alpha=2$, then condition (α_1) is satisfied for $l=1$, but (C_2) of [9, Thm. 1] is not. A similar remark holds for [9, Thms. 2 and 3].

In this paper we are interested in extending Theorems A and B to (1) with the restriction that the functions p and q satisfy (i). In fact nothing much of significance is known regarding (1) when q has a variable sign (see an open problem XIII in [5] and an example in [2]).

The following three lemmas will be needed in the proofs of our results. The first two can be found in [3], [6] and the third appeared in [5].

LEMMA 1. *Let u be a positive and n-times differentiable function on an interval $[t_0,\infty)$. If $u^{(n)}$ is of constant sign and not identically zero on any interval of the form $[t^*,\infty)$, then there exist a $t_4\geq t_0$ and an integer l, $0\leq l\leq n$ with $n+l$ even for $u^{(n)}$ nonnegative or $n+l$ odd for $u^{(n)}$ nonpositive and such that $l>0$ implies $u^{(k)}(t)>0$ for $t\geq t_4$ $(k=0,1,\cdots,l-1)$, and $l\leq n-1$ implies $(-1)^{l+k}u^{(k)}(t)>0$ for $t\geq t_4$ $(k=l,l+1,\cdots,n-1)$.*

LEMMA 2. *If the function u is as in Lemma* 1 *and*

$$u^{(n-1)}(t)u^{(n)}(t)\leq 0 \quad \text{for every } t\geq t_u,$$

then for every λ, $0<\lambda<1$, we have

$$(4)\qquad u(\lambda t)\geq\frac{2^{1-n}}{(n-1)!}\left[\frac{1}{2}-\left|\lambda-\frac{1}{2}\right|\right]^{n-1}t^{n-1}\left|u^{(n-1)}(t)\right| \quad \text{for all large } t.$$

LEMMA 3. *Let*

$$(5)\qquad \lim_{t\to\infty}\int_{\bar{t}}^{t}\exp\left(-\int_{\bar{t}}^{s}p(\tau)\,d\tau\right)ds=\infty \quad \text{for every } \bar{t}\geq t_0.$$

Then if $x(t)$ is a nonoscillatory solution of (1), *we have*

$$x(t)x^{(n-1)}(t)>0 \quad \textit{for all large } t.$$

2. Main results. In the sequel we assume that there exist real valued functions $\sigma_i \in C^1[[t_0,\infty),(0,\infty)]$ for $i=1,2,\cdots,m$ such that

$$\begin{aligned} &\sigma_i(t)=\inf_{s\geq t}(\min\{s,g_i(s)\}),\\ &\dot{\sigma}_i(t)>0,\\ &\sigma_i(t)\to\infty \quad \text{as } t\to\infty. \end{aligned} \tag{6}$$

THEOREM 1. *Let conditions* (i)–(iii), (5) *and* (6) *hold,*

$$\frac{\partial f(y_1,y_2,\cdots,y_m)}{\partial y_i} \ \textit{exist}, \qquad i=1,2,\cdots,m,$$

and

$$\frac{\partial f(y_1,y_2,\cdots,y_m)}{\partial y_i}\geq\alpha_i>0 \quad \textit{for } y_i\neq 0, \quad i=1,2,\cdots,m. \tag{7}$$

If

$$\limsup_{t\to\infty} t^{1-m}\int_{t_0}^{t}(t-u)^{m-3}u^l \cdot\left[(t-u)^2q(u)-2^{2n-3}(n-1)!\,\frac{[(t-u)(p(u)-l/u)+m-1]^2}{\sum_{i=1}^m\alpha_i\sigma_i^{n-2}(u)\dot{\sigma}_i(u)}\right]du=\infty \tag{8}$$

for some integer $m\geq 3$ and some constant l, then every solution of (1) *is oscillatory.*

Proof. Let $x(t)$ be a nonoscillatory solution of (1), say $x(t)>0$ for $t\geq t_1\geq t_0>0$. Then there exists $t_2\geq t_1$ so that $x[\sigma_i(t)]>0$ for $t\geq t_2$, $i=1,2,\cdots,m$. By Lemma 3, there is $t_3\geq t_2$ such that

$$x^{(n-1)}(t)>0 \quad \text{for } t\geq t_3.$$

From (1) and (ii) we obtain

$$x^{(n)}(t)\leq 0 \quad \text{for } t\geq t_3.$$

Moreover $q(t)\not\equiv 0$ on any ray $[t^*,\infty)$ ensures that $x^{(n)}(t)$ also has this property. Notice next that the hypotheses of Lemma 1 are satisfied on $[t_3,\infty)$, which implies that there exists $t_4\geq t_3$ so that

$$\dot{x}(t)>0 \quad \text{and} \quad x^{(n-1)}(t)>0 \quad \text{for } t\geq t_4.$$

It is easy to check that we can apply Lemma 2 for $u=\dot{x}$, $\lambda=\frac{1}{2}$ and conclude that there is a $t_5\geq t_4$ such that

$$\dot{x}\left[\frac{1}{2}\sigma_i(t)\right]\geq\frac{2^{2-2n}}{(n-1)!}\sigma_i^{n-2}(t)x^{(n-1)}[\sigma_i(t)]$$

for every $t \geq t_5$ and $i = 1, 2, \cdots, m$. Notice also that

$$\left(\left(x^{(n-1)}(t)\right)^2\right)^{\cdot} = 2x^{(n-1)}(t)x^{(n)}(t) \leq 0$$

for $t \geq t_3$, so the function $x^{(n-1)}(t)$ is nonincreasing on $[t_3, \infty)$. Therefore for every $t \geq t_3$ we have for each $i = 1, 2, \cdots, m$ that

$$\dot{x}\left[\frac{1}{2}\sigma_i(t)\right] \geq \frac{2^{2-2n}}{(n-1)!}\sigma_i^{n-2}(t)x^{(n-1)}(t). \tag{9}$$

Define

$$w(t) = \frac{t^l x^{(n-1)}(t)}{f\left(x\left[\frac{1}{2}\sigma_1(t)\right], \cdots, x\left[\frac{1}{2}\sigma_m(t)\right]\right)}.$$

Then $w(t)$ satisfies

$$\dot{w}(t) = -t^l q(t)\frac{f\left(x[g_1(t)], \cdots, x[g_m(t)]\right)}{f\left(x\left[\frac{1}{2}\sigma_1(t)\right], \cdots, x\left[\frac{1}{2}\sigma_m(t)\right]\right)} - \left[p(u) - \frac{l}{t}\right]w(t)$$
$$-\frac{1}{2}\frac{w(t)}{f\left(x\left[\frac{1}{2}\sigma_1(t)\right], \cdots, x\left[\frac{1}{2}\sigma_m(t)\right]\right)}\sum_{i=1}^{m}\frac{\partial f\left(x\left[\frac{1}{2}\sigma_1(t)\right], \cdots, x\left[\frac{1}{2}\sigma_m(t)\right]\right)}{\partial y_i}\dot{x}\left[\frac{1}{2}\sigma_i(t)\right]\dot{\sigma}_i(t).$$

Using (ii), (7), (9) and the fact that x is a nondecreasing function, we obtain

$$\dot{w}(t) \leq -t^l q(t) - \left[p(t) - \frac{l}{t}\right]w(t) - \frac{2^{1-2n}}{(n-1)!}t^{-l}w^2(t)\sum_{i=1}^{m}\alpha_i\sigma_i^{n-2}(t)\dot{\sigma}_i(t). \tag{10}$$

Thus

$$\int_{t_5}^{t}(t-u)^{m-1}\dot{w}(u)\,du \leq -\int_{t_5}^{t}(t-u)^{m-1}u^l q(u)\,du$$
$$-\int_{t_5}^{t}(t-u)^{m-1}\left[p(u) - \frac{l}{u}w(u)\right]du$$
$$-\frac{2^{1-2n}}{(n-1)!}\int_{t_5}^{t}(t-u)^{m-1}u^{-l}\left(\sum_{i=1}^{m}\alpha_i\sigma_i^{n-2}(u)\dot{\sigma}(u)\right)w^2(u)\,du.$$

Since

$$\int_{t_5}^{t}(t-u)^{m-1}\dot{w}(u)\,du = (m-1)\int_{t_5}^{t}(t-u)^{m-2}w(u)\,du - (t-t_5)^{m-1}w(t_5),$$

we get

$$t^{1-m}\int_{t_5}^{t}(t-u)^{m-3}u^l\left[(t-u)^2 q(u)\right.$$
$$\left. -2^{2n-3}(n-1)!\,\frac{\left[(t-u)(p(u)-l/u)+(m-1)\right]^2}{\sum_{i=1}^{m}\alpha_i\sigma_i^{n-2}(t)\dot{\sigma}_i(t)}\right]du$$

$$
\begin{aligned}
&\leq t^{1-m}(t-t_5)^{m-1}w(t_5)\\
&\quad -t^{1-m}\int_{t_5}^{t}\left[\left(\frac{2^{1-2n}}{(n-1)!}(t-u)^{m-1}u^{-l}\sum_{i=1}^{m}\alpha_i\sigma_i^{n-2}(u)\dot{\sigma}_i(u)\right)^{1/2}w(u)\right.\\
&\qquad\left.-\frac{(t-u)^{m-2}((t-u)[p(u)-l/u]+m-1)}{2\left(\frac{2^{1-2n}}{(n-1)!}(t-u)^{m-1}u^{-l}\Sigma_{i=1}^{m}\alpha_i\sigma_i^{n-2}(u)\dot{\sigma}_i(u)\right)^{1/2}}\right]^2 du\\
&\leq t^{1-m}(t-t_5)^{m-1}w(t_5)\\
&\to w(t_5)\equiv \text{a finite number} \quad \text{as } t\to\infty,
\end{aligned}
$$

which contradicts condition (8). This proves the theorem. □

COROLLARY 1. *Let condition* (8) *in Theorem* 1 *be replaced by*

$$
\limsup_{t\to\infty} t^{1-m}\int_{t_0}^{t}(t-u)^{m-1}u^{l}q(u)\,du=\infty \tag{11}
$$

and

$$
\lim_{t\to\infty} t^{1-m}\int_{t_0}^{t}(t-u)^{m-3}u^{l}\frac{[(t-u)[p(u)-l/u]+m-1]^2}{\Sigma_{i=1}^{m}\alpha_i\sigma_i^{n-2}(u)\dot{\sigma}_i(u)}\,du<\infty \tag{12}
$$

for some integer $m\geq 3$ *and some constant* l. *Then the conclusion of Theorem* 1 *holds.*

THEOREM 2. *Let conditions* (i)–(iii), (5), (6) *and* (7) *hold. If*

$$
\limsup_{t\to\infty}\int_{t_0}^{t}u^{l}\left[q(u)-2^{2n-3}(n-1)!\,\frac{[p(u)-l/u]^2}{\Sigma_{i=1}^{m}\alpha_i\sigma_i^{n-2}(u)\dot{\sigma}_i(u)}\right]du=\infty \tag{13}
$$

for some constant l, *then every solution of* (1) *is oscillatory.*

Proof. Suppose that $x(t)$ is a nonoscillatory solution of (1), say $x(t)>0$ for $t\geq t_1\geq t_0>0$. Proceeding as in the proof of Theorem 1, we get (10). Thus for $t\geq t_5$ we have

(14)

$$
\begin{aligned}
\dot{w}(t)&\leq -t^{l}q(t)+2^{2n-3}(n-1)!\,t^{l}\frac{[p(t)-l/t]^2}{\Sigma_{i=1}^{m}\alpha_i\sigma_i^{n-2}(t)\dot{\sigma}_i(t)}\\
&\quad-\left[\left(\frac{2^{1-2n}}{(n-1)!}t^{-l}\sum_{i=1}^{m}\alpha_i\sigma_i^{n-2}(t)\dot{\sigma}_i(t)\right)^{1/2}w(t)-\frac{p(t)-l/t}{2\left(\frac{2^{1-2n}}{(n-1)!}t^{-l}\Sigma_{i=1}^{m}\alpha_i\sigma_i^{n-2}(t)\dot{\sigma}_i(t)\right)^{1/2}}\right]^2\\
&\leq -t^{l}q(t)+2^{2n-3}(n-1)!\,t^{l}\frac{[p(t)-l/t]^2}{\Sigma_{i=1}^{m}\alpha_i\sigma_i^{n-2}(t)\dot{\sigma}_i(t)}.
\end{aligned}
$$

Integrating (14) from t_5 to t we obtain

$$\int_{t_5}^{t} u^l\left[q(u)-2^{2n-3}(n-1)!\,\frac{(p(u)-l/u)^2}{\sum_{i=1}^{m}\alpha_i\sigma_i^{n-2}(u)\dot{\sigma}_i(u)}\right]du\leq w(t_5)-w(t)$$
$$\leq w(t_5),$$

which contradicts condition (13). Thus our proof is complete. □

In order to obtain our next results we assume that there exists a real valued function $\sigma\in C^1[[t_0,\infty),(0,\infty)]$ such that

$$\begin{aligned}(15)\qquad &\sigma(t)=\inf_{s\geq t}\{(\min s,g_1(s),\cdots,g_m(s))\},\\ &\dot{\sigma}(t)>0,\\ &\sigma(t)\to\infty\quad\text{as } t\to\infty.\end{aligned}$$

THEOREM 3. *In addition to conditions* (i)–(iii) *and* (15), *assume that there exists a real valued function* $\phi\in C[R,R]$ *such that* $x\phi(x)>0$ *for* $x\neq 0$, $\phi'(x)$ *exists*, $\phi'(x)\geq\alpha>0$ *for* $x\neq 0$ *and*

$$(16)\qquad |f(y,y,\cdots,y)|\geq|\phi(y)|\quad \textit{for all } (t,y)\in[t_0,\infty)\times R-\{0\}.$$

If

$$(17)\qquad \limsup_{t\to\infty} t^{1-m}\int_{t_0}^{t}(t-u)^{m-3}u^l\left[(t-u)^2q(u)-2^{2n-3}(n-1)!\cdot\frac{[(t-u)[p(u)-l/u]+(m-1)]^2}{\alpha\sigma^{n-2}(u)\dot{\sigma}(u)}\right]du=\infty$$

for some integer $m\geq 3$ *and some constant* l, *then every solution of* (1) *is oscillatory.*

Proof. Let $x(t)$ be a nonoscillatory solution of (1) with $x(t)>0$ for $t\geq t_1\geq t_0>0$. It follows as in the proof of Theorem 1 that there exists $t_5\geq t_0$ such that $\dot{x}(t)>0$, $x^{(n-1)}(t)>0$ and

$$(18)\qquad \dot{x}\left[\frac{1}{2}\sigma(t)\right]\geq\frac{2^{2-2n}}{(n-1)!}\sigma^{n-2}(t)x^{(n-1)}(t),\qquad t\geq t_5.$$

Letting

$$w(t)=\frac{t^l x^{(n-1)}(t)}{\phi(x[\frac{1}{2}\sigma(t)])},$$

we have

$$\dot{w}(t)=-t^lq(t)\frac{f(x[g_1(t)],\cdots,x[g_m(t)])}{\phi(x[\frac{1}{2}\sigma(t)])}-\left[p(t)-\frac{l}{t}\right]w(t)$$
$$-\frac{1}{2}\dot{x}\left[\frac{1}{2}\sigma(t)\right]\dot{\sigma}(t)\phi'\left(x\left[\frac{1}{2}\sigma(t)\right]\right)\frac{w(t)}{\phi(x[\frac{1}{2}\sigma(t)])}.$$

Using the hypotheses of the theorem and (18), we obtain

$$\dot{w}(t) \leq -t^l q(t) - \left[p(t) - \frac{l}{t}\right] w(t) - \frac{2^{1-2n}}{(n-1)!} \sigma^{n-2}(t) \dot{\sigma}(t) t^{-l} w^2(t).$$

Thus

$$\begin{aligned}\int_{t_5}^{t} (t-u)^{m-1} \dot{w}(u)\, du \leq & -\int_{t_5}^{t} (t-u)^{m-1} u^l q(u)\, du \\ & -\int_{t_5}^{t} (t-u)^{m-1} \left[p(u) - \frac{l}{u}\right] w(u)\, du \\ & -\frac{2^{1-2n}}{(n-1)!} \int_{t_5}^{t} (t-u)^{m-1} u^{-l} \sigma^{n-2}(u) \dot{\sigma}(u) w^2(u)\, du.\end{aligned}$$

As in the proof of Theorem 1, we have

$$\begin{aligned} & t^{1-m} \int_{t_5}^{t} (t-u)^{m-3} u^l \left[(t-u)^2 q(u) - 2^{2n-3}(n-1)!\, \frac{[(t-u)[p(u)-l/u]+m-1]^2}{\alpha \sigma^{n-2}(u) \dot{\sigma}(u)} \right] du \\ & \leq \left(1 - \frac{t_5}{t}\right)^{m-1} w(t_5) - t^{1-m} \int_{t_5}^{t} \left(\left[\frac{2^{1-2n}}{(n-1)!} \alpha \sigma^{n-2}(u) u^{-l} (t-u)^{m-1} \right]^{1/2} w(u) \right. \\ & \qquad \left. - \frac{(t-u)^{m-3/2} [(t-u)[p(u)-l/u]+m-1]}{2 \left[\frac{2^{1-2n}}{(n-1)!} \alpha \sigma^{n-2}(u) \dot{\sigma}(u) u^{-l} \right]^{1/2}} \right)^2 du \\ & \leq \left(1 - \frac{t_5}{t}\right)^{m-1} w(t_5) \to w(t_5) \equiv \text{a finite number} \end{aligned}$$

as $t \to \infty$, which contradicts condition (17). Thus our proof is complete. □

The following theorem concerns the case when

$$f(x[g_1(t)], \cdots, x[g_m(t)]) = \sum_{i=1}^{m} f_i(x[g_i(t)]),$$

where $f_i \in C[R, R]$, $x f_i(x) > 0$ for $x \neq 0$, $i = 1, 2, \cdots, m$ and f_i, $i = 1, 2, \cdots, m$ are not required to be differentiable.

THEOREM 4. *Let conditions* (i), (iii) *and* (15) *hold and*

$$(19) \qquad \frac{f_i(x)}{x} \geq c_i > 0 \quad \textit{for } x \neq 0, \quad i = 1, 2, \cdots, m.$$

If

(20)

$$\limsup_{t\to\infty} t^{1-m}\int_{t_0}^{t}(t-u)^{m-3}u^{l}\left[c(t-u)^2q(u) - 2^{n-3}(n-1)!\,\frac{[(t-u)[p(u)-l/u]+m-1]^2}{\sigma^{n-2}(u)\dot\sigma(u)}\right]du=\infty$$

for some integer $m\geq 3$, *and some constant* l *and* $c=\sum_{i=1}^{m}c_i$, *then every solution of* (1) *is oscillatory.*

Proof. Let $x(t)$ be a nonoscillatory solution of (1) with $x(t)>0$ for $t\geq t_1\geq t_0>0$. It follows as in the proof of Theorem 3 that there exists $t_5>t_0$ such that $\dot x(t)>0$, $x^{(n-1)}(t)>0$ and

$$\dot x\left[\frac{1}{2}\sigma(t)\right]\geq\frac{2^{2-2n}}{(n-1)!}\sigma^{n-2}(t)x^{(n-1)}(t),\qquad t\geq t_5.$$

Letting

$$w(t)=\frac{t^{l}x^{(n-1)}(t)}{x[\frac{1}{2}\sigma(t)]},$$

we have

$$\dot w(t)=-t^{l}q(t)\sum_{i=1}^{m}\frac{f_i(x[g_i(t)])}{x[\frac{1}{2}\sigma(t)]}-\left[p(t)-\frac{l}{t}\right]w(t)-\frac{1}{2}\dot x\left[\frac{1}{2}\sigma(t)\right]\dot\sigma(t)\frac{w(t)}{x[\frac{1}{2}\sigma(t)]}.$$

Thus

$$\dot w(t)\leq -t^{l}q(t)\sum_{i=1}^{m}c_i-\left[p(t)-\frac{l}{t}\right]w(t)-\frac{2^{1-2n}}{(n-1)!}\sigma^{n-2}(t)\dot\sigma(t)w^2(t).$$

The rest of the proof follows exactly that of Theorem 3 and is omitted. □

Remarks.

1. The results of the present paper are presented in a form which is essentially new. We also mention that we do not stipulate that the functions $g_i(i=1,2,\cdots,m)$ in (1) be either retarded, advanced or mixed type. Hence our theorems may hold for ordinary, retarded, advanced and mixed type equations (see example below).

2. It is obvious that Theorems A and B include the results of Yeh [9], [8, Thm. 2], Wintner [7] and Kamenev [4].

3. Theorems and corollaries similar to Theorem 2 and Corollary 1 can be easily obtained for Theorems 3 and 4. Hence we omit the details.

To illustrate our results we consider the following example.

Example. Consider the mixed equations

(21)

$$x^{(n)}+t^{-1}x^{(n-1)}+2^{2n}(n+1)!t^{-n}\cdot\left[\sinh x[t+\sin t]+x\left[\frac{t}{2}\right]\exp\left(x^2\left[\frac{t}{2}\right]\right)+x[t]\cosh x[t]+x[t^2]\log(e+x^2[t^2])\right]=0$$

and

$$x^{(n)}+t^{-1}x^{(n-1)}+2^{2n}(n+1)!t^{-n} \cdot\left[x\left[\frac{t}{2}\right]\exp\left(\sin x\left[\frac{t}{2}\right]\right)+x[2t](1+\cos^2 x[2t]) + x[t+\cos t]\log(2e+\sin^2 x[t+\cos t])\right]=0, \tag{22}$$

where n is even, $t\geq 1$. It is easy to check that (21) is oscillatory by Theorem 1 for $l=n$ and $m=3$ and that (22) is oscillatory by Theorem 4 for $l=n$ and $m=3$. We may note that the oscillatory character of (21) and (22) are not deducible from any other known oscillation criteria.

REFERENCES

[1] S. R. Grace and B. S. Lalli, *Oscillation theorems for certain second order perturbed nonlinear differential equations*, J. Math. Anal. Appl., 77 (1980), pp. 205–214.

[2] ______, *Oscillation theorems for nth order delay differential equations*, J. Math. Anal. Appl., 91 (1983), pp. 352–366.

[3] M. K. Grammatikopolous, Y. G. Sficas and V. A. Staikas, *Oscillatory properties of strongly superlinear differential equations with deviating arguments*, J. Math. Anal. Appl., 67 (1979), pp. 171–187.

[4] I. V. Kamenev, *An integral criterion for oscillation of linear differential equations of second order*, Math. Zametki, 23 (1978), pp. 249–251 (In Russian.)

[5] A. G. Kartsatos, *Recent results on oscillation of solutions of forced and perturbed nonlinear differential equations of even order*, Stability of Dynamical Systems: Theory and Applications, Lecture Notes in Pure and Applied Mathematics, 28, Springer, New York, 1977, pp. 17–72.

[6] V. A. Staikos, *Basic results on oscillation for differential equations with deviating arguments*, Hiroshima Math. J., 10 (1980), pp. 495–516.

[7] A. Wintner, *A criterion of oscillatory stability*, Quart. Appl. Math., 7 (1949), pp. 115–117.

[8] C. C. Yeh, *An oscillation criterion for second order nonlinear differential equations with functional arguments*, J. Math. Anal. Appl., 76 (1980), pp. 72–76.

[9] ______, *Oscillation theorems for nonlinear second order differential equations with damped term*, Proc. Amer. Math. Soc., 84 (1982), pp. 397–402.

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1502-0007 $01.25/0

BOUNDARY AND CORNER LAYER BEHAVIOR IN SINGULARLY PERTURBED SEMILINEAR SYSTEMS OF BOUNDARY VALUE PROBLEMS*

MARK A. O'DONNELL[†]

Abstract. The existence and asymptotic behavior as $\varepsilon\to 0^+$ of solutions of the boundary value problem $\varepsilon \mathbf{y}''=\mathbf{h}(t,\mathbf{y})$, $a<t<b$, $\mathbf{y}(a)$ and $\mathbf{y}(b)$ prescribed, are studied in the case where $\partial h_i/\partial y_i>m_i>0$ in the domain of interest (m_i a constant, $i=1,\cdots,n$). A mild assumption on the reduced solution essentially decouples the system and allows the application of the scalar theory of singularly perturbed boundary value problems to each component of the system. The components of solutions are shown to exhibit essentially two types of asymptotic behavior:

(i) boundary layer behavior when the reduced solution is smooth and/or

(ii) corner layer behavior when the reduced solution has a discontinuous first derivative in (a,b). Several illustrative examples of both types of behavior are discussed. The results are established by using the theory of differential inequalities for systems of second order boundary value problems.

1. Introduction. We consider in this paper nonlinear boundary value problems of the form

$$\varepsilon \mathbf{y}''=\mathbf{h}(t,\mathbf{y}),\qquad \mathbf{y}(a)=\mathbf{A},\quad \mathbf{y}(b)=\mathbf{B},\tag{1.1}$$

where $\mathbf{y},\mathbf{h},\mathbf{A}$ and $\mathbf{B}$ are n-vectors and $\varepsilon>0$ is a small parameter. The objective is to give sufficient conditions for the existence of solutions of (1.1) and to study the boundary layer and corner layer behavior of these solutions as $\varepsilon\to 0^+$.

The principal assumptions are that the reduced system

$$\mathbf{0}=\mathbf{h}(t,\mathbf{y})$$

has at least one solution $\mathbf{u}=(u_1(t),\cdots,u_n(t))$ which satisfies

$$0=h_i(t,y_1,\cdots,y_{i-1},u_i,y_{i+1},\cdots,y_n),\qquad i=1,\cdots,n,\tag{1.2}$$

for all y_j in some region of interest $\mathfrak{D}_j$, $j\neq i,t$ in $[a,b]$ and that the continuous partials $\partial h_i/\partial y_i$ satisfy

$$\frac{\partial h_i}{\partial y_i}(t,\mathbf{y})>m_i>0,\qquad i=1,\cdots,n,\tag{1.3}$$

in the region $[a,b]\times\mathfrak{D}_1\times\cdots\times\mathfrak{D}_n$. Condition (1.2) on the reduced solutions allows us to use the scalar theory due to Brish [2] to give componentwise estimates for solutions using only (1.3), without having to examine the off-diagonal terms of the Jacobian matrix $(\partial h_i/\partial y_j)$.

Several authors have studied problems of this form. In particular, Howes [6] and Kelley [9] have shown existence under slightly weaker conditions and they have given estimates on the behavior of $\|\mathbf{y}\|$ as $\varepsilon\to 0^+$. However, when (1.2) obtains, componentwise estimates may not only provide more insight into the behavior of the system (1.1) but also extend and improve the norm estimates. (See Example 4.1 below.)

Note that condition (1.2) obtains for the rather general class of systems of the form

$$\varepsilon y_i''=\prod_{j=1}^{n} f_{ij}(t,y_j),\qquad i=1,\cdots,n,$$

* Received by the editors June 11, 1982, and in revised form November 1, 1982.

[†] Department of Mathematics, University of California at Davis, Davis, California 95616.

with

$$0=f_{ii}(t,u_i),\qquad i=1,\cdots,n.$$

Several examples of such systems are given in §4.

2. Preliminaries. Our primary tool will be the following theorem which extends the scalar result of Habets and Laloy [3]. It is proved in [11].

THEOREM 2.1. *Consider the system*

$$y''=\mathbf{h}(t,\mathbf{y}),\qquad \mathbf{y}(a)=\mathbf{A},\quad \mathbf{y}(b)=\mathbf{B}. \tag{2.1}$$

Suppose there exist n bounding pairs (α_i,β_i) *of piecewise* $C^{(2)}$ *functions on* $[a,b]$, *i.e., there are n partitions* $\{t_i^j\}_{j=0}^{m_i}$ *of* $[a,b]$ *with* $a=t_i^0<t_i^1<\cdots<t_i^{m_i}=b$ *such that on each subinterval* $[t_i^{j-1},t_i^j]$ *the bounding functions* α_i *and* β_i *are twice continuously differentiable.* (*At the partition points,* t_i^{j-1} *and* t_i^j, *the derivatives are right-hand and left-hand derivatives, respectively.*) *Suppose that*

$$\alpha_i(a)\le A_i\le\beta_i(a),\qquad \alpha_i(b)\le B_i\le\beta_i(b),\qquad i=1,\cdots,n \tag{2.1.1}$$

and

$$\alpha_i(t)\le\beta_i(t),\qquad t \text{ in } [a,b],\quad i=1,\cdots,n \tag{2.1.2}$$

and that on each subinterval $[t_i^{j-1},t_i^j], j=1,\cdots,m_i$

$$\begin{aligned}&\alpha_i''\ge h_i(t,y_1,\cdots,\alpha_i,\cdots,y_n),\\ &\qquad\qquad\qquad\qquad\qquad \text{for all } y_j \text{ in } \left[\alpha_j(t),\beta_j(t)\right],\quad j\ne i.\\ &\beta_i''\le h_i(t,y_1,\cdots,\beta_i,\cdots,y_n),\end{aligned} \tag{2.1.3}$$

Further suppose that for each t in $[a,b]$,

$$D_l\alpha_i(t)\le D_r\alpha_i(t)\quad\text{and}\quad D_l\beta_i(t)\ge D_r\beta_i(t) \tag{2.1.4}$$

where D_l *and* D_r *denote left-hand and right-hand differentiation, respectively. Finally suppose that* $\mathbf{h}$ *is continuous in the region* $\mathfrak{D}=[a,b]\times\Pi_{i=1}^n[\alpha_i,\beta_i]$.

Then (2.1) *has a solution* $y=(y_1(t),\cdots,y_n(t))$ *of class* $C^2[a,b]$ *with* $\alpha_i(t)\le y_i(t)\le\beta_i(t)$ *for t in* $[a,b]$ *and* $i=1,\cdots,n$.

Note that if the bounding functions α_i and β_i are of class $C^2[a,b]$, then condition (2.1.4) is satisfied and Theorem 2.1 reduces to the more standard existence theorem proved in [1, Chapter 1]. Using this result, the problem of studying the existence and asymptotic behavior of (1.1) reduces to the construction of appropriate bounding pairs (α_i,β_i).

3. Boundary layer and corner layer phenomena. We begin with a result which guarantees the existence of a solution of the boundary value problem

$$\varepsilon\mathbf{y}''=\mathbf{h}(t,\mathbf{y}),\qquad \mathbf{y}(a)=\mathbf{A},\quad \mathbf{y}(b)=\mathbf{B}, \tag{3.1}$$

where $\mathbf{h},\mathbf{y},\mathbf{A}$ and $\mathbf{B}$ are in $\mathbb{R}^n$, which exhibits boundary layer behavior as $\varepsilon\to0^+$. The distinguishing characteristic here is the existence of a reduced solution $\mathbf{u}=(u_1(t),\cdots,u_n(t))$ of class $C^{(2)}[a,b]$ which satisfies condition (1.2).

THEOREM 3.1. *Assume that*

(1) *the reduced system has a solution* $\mathbf{u}=(u_1(t),\cdots,u_n(t))$ *of class* $C^{(2)}[a,b]$ *which satisfies* $0=h_i(t,y_1,\cdots,u_i,\cdots,y_n)$ *for* $i=1,\cdots,n$ *and for all* $(t,y_1,\cdots,y_{i-1},y_{i+1},\cdots,y_n)$ *in* $[a,b]\times\Pi_{j\ne i}\mathfrak{D}_j$ ($\mathfrak{D}_j$ *as defined below*);

(2) *the functions* h_i, $i=1,\cdots,n$, *are of class* $C^{(1)}$ *with respect to* $t,y_1,\cdots,y_n$ *in the region* $\mathfrak{D}=[a,b]\times\Pi_{i=1}^{n}\mathfrak{D}_i$, *where for* $i=1,\cdots,n$ $\mathfrak{D}_i=\{y_i\colon|y_i-u_i(t)|\leq d_i(t)\}$, *for* d_i *a smooth positive function such that* $d_i(t)=|A_i-u_i(a)|+\delta$ *on* $[a,a+\delta/2]$, $d_i(t)=|B_i-u_i(b)|+\delta$ *on* $[b-\delta/2,b]$ *and* $d_i(t)=\delta$ *on* $[a+\delta,b-\delta]$, *for* $\delta>0$ *a small constant* (*see Fig.* 3.1a);

(3) *for all* $(t,y_1,\cdots,y_n)$ *in* $\mathfrak{D}$

$$\frac{\partial h_i}{\partial y_i}(t,y_1,\cdots,y_n)>m_i>0,$$

for a positive constant m_i, $i=1,\cdots,n$.

Then there exists an $\varepsilon_0>0$ *such that the boundary value problem* (3.1) *has a solution* $\mathrm{y}=(y_1(t,\varepsilon),\cdots,y_n(t,\varepsilon))$ *whenever* $0<\varepsilon\leq\varepsilon_0$. *Furthermore, for t in* $[a,b]$ *we have that*

$$y_i(t,\varepsilon)=u_i(t)+(A_i-u_i(a))\exp\left[-\sqrt{\frac{m_i}{\varepsilon}}\,(t-a)\right]$$

$$+(B_i-u_i(b))\exp\left[-\sqrt{\frac{m_i}{\varepsilon}}\,(b-t)\right]+O(\varepsilon),$$

for $i=1,\cdots,n$.

Proof. For definiteness, we construct bounding functions α_i, β_i for the ith component under the assumptions that $u_i(a)<A_i$ and $u_i(b)>B_i$. The other bounding pairs are defined analogously.

Define for t in $[a,b]$ and $\varepsilon>0$ the functions

$$\alpha_i(t,\varepsilon)=u_i(t)+V_i(t,\varepsilon)-\varepsilon_i$$

and

$$\beta_i(t,\varepsilon)=u_i(t)+W_i(t,\varepsilon)+\varepsilon_i,$$

where $V_i(t,\varepsilon)=(B_i-u_i(b))\exp[-\sqrt{m_i/\varepsilon}\,(b-t)]$ and $W_i(t,\varepsilon)=(A_i-u_i(a))\cdot\exp[-\sqrt{m_i/\varepsilon}\,(t-a)]$ and $\varepsilon_i=\varepsilon\gamma_i/m_i$, for γ_i a positive constant greater than $M_i=\max_{[a,b]}|u_i''(t)|$.

Observe that the region between α_i and β_i, that is, the set $\{(t,y_i)\colon t \text{ in } [a,b],\ \alpha_i(t,\varepsilon)\leq y_i\leq\beta_i(t,\varepsilon)\}$ is contained in the region $[a,b]\times\mathfrak{D}_i$ (regardless of the choice of $\delta>0$) when ε is sufficiently small. (See Figs. 3.1a, b.) Similarly, for α_j,β_j defined analogously for $j\neq i$ and ε sufficiently small, we have that

$$[a,b]\times\left[\alpha_j(t,\varepsilon),\beta_j(t,\varepsilon)\right]\subseteq[a,b]\times\mathfrak{D}_j.$$

It is clear that for α_i and β_i defined as above, (2.1.1) and (2.1.2) hold. It remains to show that the conditions (2.1.3) hold. First we consider the differential inequality for α_i:

$$\begin{aligned}\varepsilon\alpha_i''&-h_i(t,y_1,\cdots,\alpha_i,\cdots,y_n)\\&=\varepsilon u_i''+\varepsilon V_i''-h_i(t,y_1,\cdots,\alpha_i,\cdots,y_n)\\&\geq-\varepsilon M_i+m_iV_i-h_i(t,y_1,\cdots,u_i,\cdots,y_n)-\frac{\partial h_i}{\partial y_i}(t,y_1,\cdots,\theta_i,\cdots,y_n)\cdot(V_i-\varepsilon_i)\end{aligned}$$

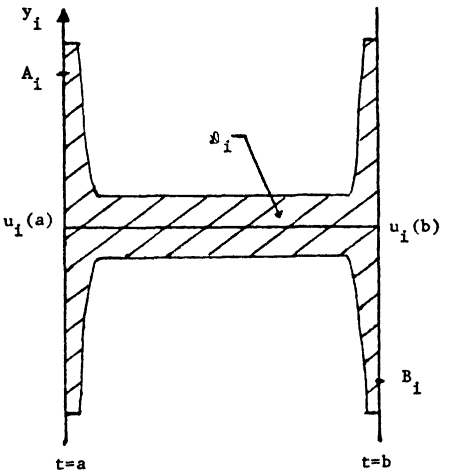


FIG. 3.1a

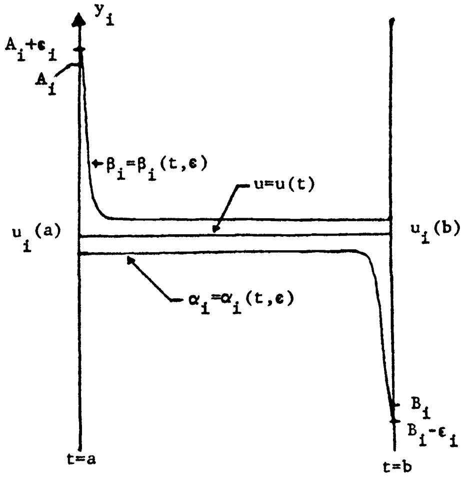


FIG. 3.1b

where $\theta_i = u_i(t) + \theta(V_i(t,\varepsilon) - \varepsilon_i)$, $0 < \theta < 1$. Now since θ_i is between α_i and β_i, it is therefore in $\mathcal{D}_i$ for ε sufficiently small. Thus for $\alpha_j \le y_j \le \beta_j$, $j \ne i$ we have y_j in $\mathcal{D}_j$ for ε sufficiently small, and so

$$\varepsilon\alpha_i'' - h_i(t, y_1, \cdots, \alpha_i, \cdots, y_n) \ge -\varepsilon M_i + m_i V_i(t,\varepsilon) - m_i V_i(t,\varepsilon) + m_i \varepsilon_i$$
$$= \varepsilon(\gamma_i - M_i) > 0$$

since γ_i is greater than M_i.

The argument for β_i is analogous:

$$h_i(t, y_1, \cdots, \beta_i, \cdots, y_n) - \varepsilon\beta_i''$$
$$\ge h_i(t, y_1, \cdots, u_i, \cdots, y_n) + \frac{\partial h_i}{\partial y_i}(t, y_1, \cdots, \eta_i, \cdots, y_n) \cdot (W_i + \varepsilon_i) - \varepsilon M_i - \varepsilon W_i''$$

where $\eta_i = u_i + \eta(W_i + \varepsilon_i)$, $0 < \eta < 1$. Arguing as above, we see that for ε sufficiently small, $\alpha_j \le y_j \le \beta_j$, $j \ne i$, implies that $(t, y_1, \cdots, \eta_i, \cdots, y_n)$ is in $\mathcal{D}$. Hence, it follows that

$$h_i(t, y_1, \cdots, \beta_i, \cdots, y_n) - \varepsilon\beta_i'' \ge m_i W_i + m_i \varepsilon_i - \varepsilon M_i - m_i W_i = \varepsilon(\gamma_i - M_i) > 0.$$

Therefore, we deduce from Theorem 2.1 that the boundary value problem (3.1) has a solution $y = (y_1(t,\varepsilon), \cdots, y_n(t,\varepsilon))$ for $0 < \varepsilon \le \varepsilon_0$, ε_0 sufficiently small. Furthermore, for t in $[a,b]$

$$y_i(t,\varepsilon) = u_i(t) + W_i(t,\varepsilon) + V_i(t,\varepsilon) + O(\varepsilon),$$

for $i = 1, \cdots, n$, since Theorem 2.1 guarantees that $\alpha_i \le y_i \le \beta_i$ for t in $[a,b]$ and $i = 1, \cdots, n$.

In contrast to boundary layer behavior, corner layer behavior in a component is characterized by the existence of two reduced solutions $u_{1i} = u_{1i}(t)$ and $u_{2i} = u_{2i}(t)$ which intersect at a point T_i in (a,b). Different components will in general have corner layers at different points and can exhibit boundary layer behavior simultaneously.

Our main result is the next theorem.

THEOREM 3.2. *Assume that*

(1) *there exist functions* $\mathbf{u}_1=(u_{11}(t),\cdots,u_{1n}(t))$ *and* $\mathbf{u}_2=(u_{21}(t),\cdots,u_{2n}(t))$ *with* u_{1i} *and* u_{2i} *of class* $C^{(2)}$ *on* $[a,T_i]$ *and* $[T_i,b]$ *respectively, satisfying for* $j=1,2$ *and* $i=1,\cdots,n$: $h_i(t,y_1,\cdots,u_{ji},\cdots,y_n)=0$ *for* t *in* $[a,b]$ *and* y_k *in* $\mathcal{D}_k$ *(as given below),* $k\neq i$; *moreover,* $u_{1i}(T_i)=u_{2i}(T_i)$ *and* $u_{1i}'(T_i)<u_{2i}'(T_i)$, T_i *in* (a,b);

(2) *each function* h_i *is of class* $C^{(1)}$ *on the region* $\mathcal{D}=[a,b]\times\mathcal{D}_1\times\cdots\times\mathcal{D}_n$, *where*

$$\mathcal{D}_i=\{y_i\colon |y_i-u_i(t)|\le d_i(t)\} \quad and \quad u_i(t)=\begin{cases}u_{1i}(t), & a\le t\le T_i,\\ u_{2i}(t), & T_i\le t\le b,\end{cases}$$

for d_i *a smooth positive function such that* $d_i(t)=|A_i-u_i(t)|+\delta$ *on* $[a,a+\delta/2]$, $d_i(t)=|B_i-u_i(t)|+\delta$ *on* $[b-\delta/2,b]$ *and* $d_i(t)=\delta$ *on* $[a+\delta,b-\delta]$, *for* $\delta>0$ *a small constant (see Fig.* 3.2a);

(3) *for all* $(t,y_1,\cdots,y_n)$ *in* $\mathcal{D}$

$$\frac{\partial h_i}{\partial y_i}(t,y_1,\cdots,y_n)>m_i>0,$$

for a positive constant m_i, $i=1,\cdots,n$.

Then there exists an $\varepsilon_0>0$ *such that for each* $\varepsilon, 0<\varepsilon\le\varepsilon_0$, *there exists a solution* $y=(y_1(t,\varepsilon),\cdots,y_n(t,\varepsilon))$ *of* (3.1). *Moreover, for* t *in* $[a,b]$

$$|y_i(t,\varepsilon)-u_i(t)|\le u_i(t)+|A_i-u_i(a)|\exp\left[-\sqrt{\frac{m_i}{\varepsilon}}\,(t-a)\right]$$
$$+|B_i-u_i(b)|\exp\left[-\sqrt{\frac{m_i}{\varepsilon}}\,(b-t)\right]+O(\varepsilon^{1/2}),$$

for $i=1,\cdots,n$.

Proof. We define bounding functions α_i,β_i for the ith component. The other bounding pairs are defined analogously. We then verify that the conditions of Theorem 2.1 obtain.

For t in $[a,b]$ and $\varepsilon>0$, define

$$\alpha_i(t,\varepsilon)=\begin{cases}u_{1i}(t)-W_i(t,\varepsilon)-V_i(T_i,\varepsilon)-\varepsilon_i, & a\le t\le T_i,\\ u_{2i}(t)-V_i(t,\varepsilon)-W_i(T_i,\varepsilon)-\varepsilon_i, & T_i\le t\le b,\end{cases}$$

and

$$\beta_i(t,\varepsilon)=\begin{cases}u_{1i}(t)+W_i(t,\varepsilon)+X_i(t,\varepsilon)+\varepsilon_i+(b-t)W_i'(T_i,\varepsilon)\\ \qquad +V_i(T_i\varepsilon)+(b-T_i)V_i'(T_i,\varepsilon), & a\le t\le T_i,\\ u_{2i}(t)+V_i(t,\varepsilon)+X_i(T_i,\varepsilon)+\varepsilon_i+(b-t)V_i'(T_i,\varepsilon)\\ \qquad +W_i(T_i,\varepsilon)+(b-T_i)W_i'(T_i,\varepsilon), & T_i\le t\le b,\end{cases}$$

where

$$W_i(t,\varepsilon)=|A_i-u_i(a)|\exp\left[-\sqrt{\frac{m_i}{\varepsilon}}\,(t-a)\right],$$

$$V_i(t,\varepsilon)=|B_i-u_i(b)|\exp\left[-\sqrt{\frac{m_i}{\varepsilon}}\,(b-t)\right],$$

$$X_i(t,\varepsilon)=\sqrt{\frac{\varepsilon}{m_i}}\,(u_{2i}'(T_i)-u_{1i}'(T_i))\exp\left[\sqrt{\frac{m_i}{\varepsilon}}\,(t-T_i)\right],$$

and where $\varepsilon_i=\varepsilon\gamma_i/m_i$, for γ_i a positive constant such that $\gamma_i>M_i=\max\{\max_{[a,T_i]}|u_{1i}''(t)|, \max_{[T_i,b]}|u_{2i}''(t)|\}$.

We note that the angular behavior of the reduced solution u_i demands these somewhat more complicated bounding functions so that the condition (2.1.4) of Theorem 2.1 will be satisfied. In fact, straightforward calculations show that $\alpha_i(a,\varepsilon)\le A_i\le \beta_i(a,\varepsilon)$, $\alpha_i(b,\varepsilon)\le B_i\le\beta_i(b,\varepsilon)$, $\alpha_i(t,\varepsilon)\le\beta_i(t,\varepsilon)$ for t in $[a,b]$, and that $D_l\alpha_i(T_i,\varepsilon)\le D_r\alpha_i(T_i,\varepsilon)$, $D_l\beta_i(T_i,\varepsilon)\ge D_r\beta_i(T_i,\varepsilon)$ for ε sufficiently small. We verify the differential inequality (2.1.3) for α_i. On $[a,T_i]$ we have that

$$\varepsilon\alpha_i''-h_i(t,y_1,\cdots,\alpha_i,\cdots,y_n)=\varepsilon u_{1i}''-\varepsilon W_i''-h_i(t,y_1,\cdots,u_{1i},\cdots,y_n)$$
$$-\frac{\partial h_i}{\partial y_i}(t,y_1,\cdots,\theta_{1i},\cdots,y_n)\cdot(-W_i-V_i(T_i,\varepsilon)-\varepsilon_i),$$

where $\theta_{1i}=u_{1i}-\theta(W_i+V_i(T_i,\varepsilon)+\varepsilon_i)$, $0<\theta<1$.

Note that for $\varepsilon>0$ sufficiently small, θ_{1i} is in $\mathcal{D}_i$, and if y_k is in $[\alpha_k,\beta_k]$, then y_k is in $\mathcal{D}_k$, for all $k\ne i$. (See Fig. 3.2b.) Thus, for ε small enough

$$\varepsilon\alpha_i''-h_i(t,y_1,\cdots,\alpha_i,\cdots,y_n)\ge-\varepsilon M_i-\varepsilon W_i''+m_iW_i+m_iV_i(T_i,\varepsilon)+m_i\varepsilon_i$$
$$\ge(\gamma_i-M_i)\varepsilon+m_iV_i(T_i,\varepsilon),$$

since $\varepsilon W_i''-m_iW_i=0$. Now observe that $V_i(T_i\varepsilon)=|B_i-u_{2i}(b)|\exp[-\sqrt{m_i/\varepsilon}\,(b-T_i)]$ is transcendentally small, and so since $\gamma_i>M_i$, we have that

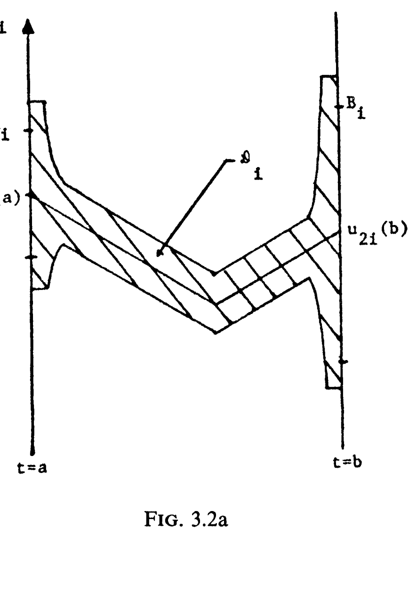


FIG. 3.2a

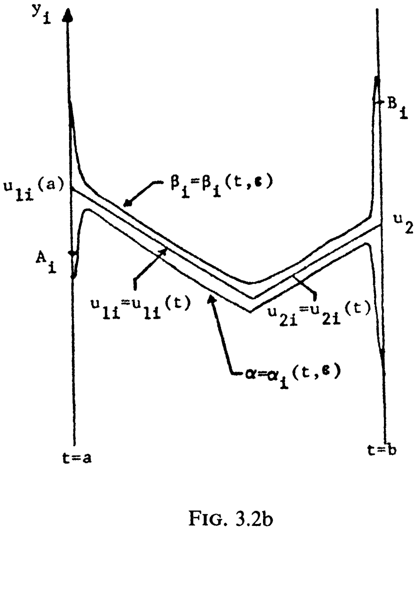


FIG. 3.2b

$$\varepsilon\alpha_i''-h_i(t,y_1,\cdots,\alpha_i,\cdots,y_n)\geq(\gamma_i-M_i)\varepsilon+\text{T.S.T.}>0$$

for y_k in $[\alpha_k,\beta_k]$, $k\neq i$, t in $[a,T_i]$ and $\varepsilon>0$ sufficiently small.

The verification of this differential inequality for $\alpha_i(t,\varepsilon)$ when t is in $[T_i,b]$ is analogous, so we omit it and turn our attention to β_i. On $[a,T_i]$ we have that

$$\begin{aligned}
&h_i(t,y_1,\cdots,\beta_i,\cdots,y_n)-\varepsilon\beta_i''\\
&\quad=h_i(t,y_1,\cdots,u_{1i},\cdots,y_n)+\frac{\partial h_i}{\partial y_i}(t,y_1,\cdots,\eta_{1i},\cdots,y_n)\\
&\qquad\qquad\cdot(W_i+X_i+\varepsilon_i+(b-t)W_i'(T_i,\varepsilon)+V_i(T_i,\varepsilon)\\
&\qquad\qquad\qquad+(b-T_i)V_i'(T_i,\varepsilon))\\
&\qquad-\varepsilon u_{1i}''-\varepsilon W_i''-\varepsilon X_i'',
\end{aligned}$$

where $\eta_{1i}=u_{1i}+\eta(W_i+X_i+\varepsilon_i+(b-t)W_i'(T_i,\varepsilon)+V_i(T_i,t)+(b-T_i)V_i'(T_i,\varepsilon))$, $0<\eta<1$.

As before, for $\varepsilon>0$ sufficiently small, η_{1i} is in $\mathfrak{D}_i$, and y_k in $[\alpha_k,\beta_k]$ implies y_k in $\mathfrak{D}_k$, $k\neq i$. Thus, we continue with the inequality

$$\begin{aligned}
&h_i(t,y_1,\cdots,\beta_i,\cdots,y_n)-\varepsilon\beta_i''\\
&\quad\geq m_iW_i+m_iX_i+m_i\varepsilon_i+m_i(b-t)W_i'(T_i,\varepsilon)\\
&\quad\quad+V_i(T_i,\varepsilon)+(b-T_i)V_i'(T_i,\varepsilon)-\varepsilon M_i-\varepsilon W_i''-\varepsilon X_i''\\
&\quad\geq(\gamma_i-M_i)\varepsilon+m_i(b-t)W_i'(T_i,\varepsilon)+V_i(T_i,\varepsilon)+(b-T_i)V_i'(T_i,\varepsilon)
\end{aligned}$$

since $\varepsilon W_i''-m_iW_i=0$ and $\varepsilon X''_i-m_iX_i=0$. Now observe that the terms $V_i(T_i,\varepsilon)=|B_i-u_i(b)|\exp[-\sqrt{m_i/\varepsilon}(b-T_i)]$, $(b-T_i)V_i'(T_i,\varepsilon)=-\sqrt{m_i/\varepsilon}\,V_i(T_i,\varepsilon)$ and $m_i(b-t)W_i'(T_i,\varepsilon)$ are all transcendentally small on $[a,T_i]$. Hence, for $\varepsilon>0$ sufficiently small, the inequality $\gamma_i>M_i$ implies that

$$h_i(t,y_1,\cdots,\beta_i,\cdots,y_n)-\varepsilon\beta_i''\geq(\gamma_i-M_i)\varepsilon+\text{T.S.T.}>0,$$

for y_k in $[\alpha_k,\beta_k]$, t in $[a,T_i]$ and ε small enough.

The verification of this differential inequality for $\beta_i(t,\varepsilon)$ when t is in $[T_i,b]$ is analogous. Therefore, by Theorem 2.1 we deduce the existence of a solution $y=(y_1(t,\varepsilon),\cdots,y_n(t,\varepsilon))$ for $0<\varepsilon\leq\varepsilon_0$, ε_0 sufficiently small. Furthermore, since $\alpha_i\leq y_i\leq\beta_i$, we have that for t in $[a,b]$ and for $i=1,\cdots,n$

$$|y_i(t,\varepsilon)-u_i(t)|\leq W_i(t,\varepsilon)+V_i(t,\varepsilon)+O(\varepsilon^{1/2}).$$

If some of the derivatives of the functions u_{1i} and u_{2i} satisfy the opposite inequalities $u_{1i}'(T_i)>u_{2i}'(T_i)$, then it is possible to obtain results analogous to Theorem 3.2. (See Fig. 3.3.) We simply make the change of dependent variable $y_i\to-y_i$ and apply Theorem 3.2 to the resulting system. For the sake of clarity we state this result as a theorem.

THEOREM 3.3. *Make the same assumptions as in Theorem* 3.2 *with the exception that in assumption* (1), $u_{1i}'(T_i)\neq u_{2i}'(T_i)$. *Then the conclusions of Theorem* 3.2 *obtain.*

We remark that the three results given above may be combined in the following sense. Some components of the system (3.1) may exhibit the boundary layer behavior of Theorem 3.1, some may exhibit only corner layer behavior and others may exhibit both boundary and corner layers as in Theorems 3.1 and 3.2.

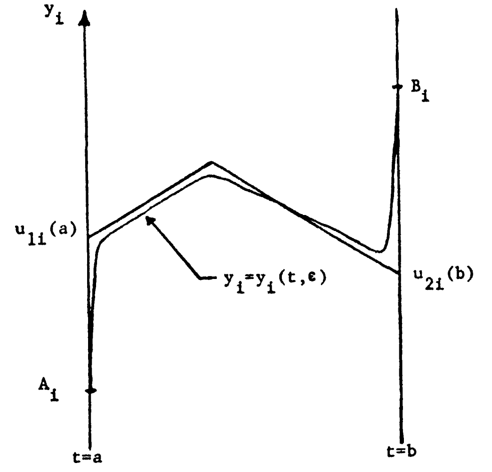


FIG. 3.3

4. Examples. In this section we consider several examples which illustrate the boundary layer and corner layer behavior of the theorems in the preceding section. Throughout this section a "stable reduced solution" will be a reduced solution $\mathbf{u} = (u_1(t), \cdots, u_n(t))$ which satisfies the inequalities

$$\frac{\partial h_i}{\partial y_i}(t, u_1, \cdots, u_n) > m_i > 0,$$

for t in $[a,b]$ and m_i a positive constant, $i = 1, \cdots, n$.

We begin with a simple example of boundary layer behavior.

Example 4.1. The boundary value problem

$$\begin{aligned} &\varepsilon y_1'' = y_1(1-y_2), \quad y_1(a) = A_1, \quad y_1(b) = B_1, \\ &\varepsilon y_2'' = y_2(1-y_1), \quad y_2(a) = A_2, \quad y_2(b) = B_2 \end{aligned} \tag{4.1}$$

has two reduced solutions $u_1 = u_2 = 0$ and $u_1 = u_2 = 1$.Since

$$\frac{\partial h_1}{\partial y_1} = 1 - y_2 \quad \text{and} \quad \frac{\partial h_2}{\partial y_2} = 1 - y_1,$$

$u_1 = u_2 = 0$ is the only stable reduced solution and so from Theorem 3.1, if $A_1, A_2, B_1, B_2 < 1$, the problem (4.1) has a solution $\mathbf{y} = (y_1(t,\varepsilon), y_2(t,\varepsilon))$ such that for t in $[a,b]$

$$y_i(t,\varepsilon) = A_i \exp\left[-\sqrt{\frac{m_i}{\varepsilon}}\,(t-a)\right] + B_i \exp\left[-\sqrt{\frac{m_i}{\varepsilon}}\,(b-t)\right] + O(\varepsilon),$$

where $0 < m_i < \min\{1 - A_i, 1 - B_i\}$, $i = 1, 2$.

Note that the unstable reduced solution $u_1 = u_2 = 1$ is a strict upper bound on the boundary values A_1, A_2, B_1, and B_2 which admit boundary layer behavior.

We now show how our estimates are an improvement over the norm bound results of Howes [6] and Kelley [9]. Their results require the existence of a positive constant m such that the quadratic form

$$Q(\mathbf{y}) = \mathbf{y}^t \cdot (J[\mathbf{h}] - mI)\mathbf{y}$$

is nonnegative definite in $\mathfrak{D}$. Here $J[\mathbf{h}]=(\partial h_i/\partial y_j)$ is the Jacobian matrix, I is the identity matrix and $\mathfrak{D}$ is the set $\{(t,y_1,y_2)$: $a\leq t\leq b$ and $y_1^2+y_2^2\leq d(t)\}$, where $d=d(t)$ is a smooth positive function on $[a,b]$ which is such that $\max\{\|A\|,\|B\|\}\leq d(t)\leq \max\{\|A\|,\|B\|\}+\delta$, for $a\leq t\leq a+\delta/2$ and $b-\delta/2\leq t\leq b$, and $d(t)\leq\delta$ for $a+\delta\leq t\leq b-\delta$, $\delta>0$ a small constant.

Since

$$J[\mathbf{h}]-mI=\begin{bmatrix}(1-m)-y_2 & -y_1\\ -y_2 & (1-m)-y_1\end{bmatrix},$$

we rewrite Q as $Q(\mathbf{y})=\mathbf{y}^t\cdot H\mathbf{y}$ where

$$H=\begin{bmatrix}(1-m)-y_2 & -\dfrac{y_1+y_2}{2}\\ -\dfrac{y_1+y_2}{2} & (1-m)-y_1\end{bmatrix}$$

is a symmetric matrix. Thus Q is nonnegative definite whenever the two conditions

$$(1-m)-y_2\geq 0, \tag{4.1.1}$$

$$\det H=(1-m)^2-(1-m)(y_1+y_2)-\tfrac{1}{4}(y_1-y_2)^2\geq 0 \tag{4.1.2}$$

obtain in the region $\mathfrak{D}$. Conditions (4.1.1) and (4.1.2) imply $1-m>0$ because $u_1=u_2=0$ is the stable reduced solution and so we have that $1>y_2$ and $1>y_1$. Using these conditions it can be shown that condition (4.1.2) implies that $\max\{\|\mathbf{A}\|,\|\mathbf{B}\|\}<2(\sqrt{2}-1)(1-m)$, and since $0<1-m<1$, we have that at best $\max\{\|\mathbf{A}\|,\|\mathbf{B}\|\}<2(\sqrt{2}-1)\approx 0.83$. This condition is stronger than $\max\{|A_1|,|B_1|,|A_2|,|B_2|\}<2(\sqrt{2}-1)$, and so the norm estimate is noticeably cruder than the componentwise estimate, especially for negative boundary conditions. Such improvement is usually the case when assumption (2) of Theorem 3.1 is satisfied.

Example 4.2. Consider the boundary value problem

$$\begin{aligned}\varepsilon y_1''&=\left(\frac{1}{3}y_1^3-y_1\right)(y_2^2+1), & y_1(a)&=A_1, y_1(b)=B_1,\\ \varepsilon y_2''&=\left(\frac{1}{3}y_2^3-y_2\right)(y_1^2+1), & y_2(a)&=A_2, y_2(b)=B_2.\end{aligned}\tag{4.2}$$

Here we have nine reduced solutions:

$$\begin{array}{lll} u_1=0, u_2=0, & u_1=0, u_2=\sqrt{3}, & u_1=0, u_2=-\sqrt{3},\\ u_1=\sqrt{3}, u_2=0, & u_1=\sqrt{3}, u_2=\sqrt{3}, & u_1=\sqrt{3}, u_2=-\sqrt{3},\\ u_1=-\sqrt{3}, u_2=0, & u_1=-\sqrt{3}, u_2=\sqrt{3}, & u_1=-\sqrt{3}, u_2=-\sqrt{3},\end{array}$$

and we have

$$\frac{\partial h_1}{\partial y_1}=(y_1^2-1)(y_2^2+1)>m>0 \quad \text{when } |y_1|>1,$$

$$\frac{\partial h_2}{\partial y_2}=(y_2^2-1)(y_1^2+1)>m>0 \quad \text{when } |y_2|>1.$$

Therefore, among the nine reduced solutions we find four which are stable:

$$u_1=\sqrt{3}\,,u_2=\sqrt{3}\,,\qquad u_1=\sqrt{3}\,,u_2=-\sqrt{3}\,,$$
$$u_1=-\sqrt{3}\,,u_2=-\sqrt{3}\,,\qquad u_1=-\sqrt{3}\,,u_2=\sqrt{3}\,.$$

Since the partials $\partial h_1/\partial y_1$ and $\partial h_2/\partial y_2$ must remain positive, any solution $\mathrm{y}=\mathrm{y}(t,\varepsilon)$ must have components which stay outside the region $|y_i|>1$, $i=1,2$. This will be true if and only if A_i and B_i are both larger than 1 or both smaller than -1, for $i=1,2$. Geometrically, if A_i and B_i are both above or both below the cross-hatched region in Fig. 4.2a, then $u_i=\sqrt{3}$ or respectively $u_i=-\sqrt{3}$ supports boundary layers, $i=1,2$. The precise result is summarized below.

Case 1. $A_1,B_1>1$ and $A_2,B_2>1$. In this case there is a solution of (4.2) satisfying

$$y_i(t,\varepsilon)=\sqrt{3}+\left(A_i-\sqrt{3}\right)\exp\left[-\sqrt{\frac{m_i}{\varepsilon}}\,(t-a)\right]$$
$$+\left(B_i-\sqrt{3}\right)\exp\left[-\sqrt{\frac{m_i}{\varepsilon}}\,(b-t)\right]+O(\varepsilon),$$

where $0<m_i<\min\{A_i^2-1,B_i^2-1\}$, $i=1,2$. (See Fig. 4.2b.)

Case 2. $A_1,B_1>1$ and $A_2,B_2<-1$. For this range of boundary conditions there is a solution of (4.2) satisfying

$$y_1(t,\varepsilon)=\sqrt{3}+\left(A_1-\sqrt{3}\right)\exp\left[-\sqrt{\frac{m_1}{\varepsilon}}\,(t-a)\right]$$
$$+\left(B_1-\sqrt{3}\right)\exp\left[-\sqrt{\frac{m_1}{\varepsilon}}\,(b-t)\right]+O(\varepsilon),$$

and

$$y_2(t,\varepsilon)=-\sqrt{3}+\left(A_2+\sqrt{3}\right)\exp\left[-\sqrt{\frac{m_2}{\varepsilon}}\,(t-a)\right]$$
$$+\left(B_2+\sqrt{3}\right)\exp\left[-\sqrt{\frac{m_2}{\varepsilon}}\,(b-t)\right]+O(\varepsilon),$$

where m_1,m_2 are as above. (See Figs. 4.2b and 4.2c, respectively.)

Case 3. $A_1,B_1<-1$ and $A_2,B_2>1$. This is the reflection of Case 2. There is a solution of (4.2) which satisfies the estimates of Case 2 with the indices 1 and 2 reversed. (See Figs. 4.2c and 4.2b, respectively.)

Case 4. $A_1,B_1<-1$ and $A_2,B_2<-1$. The problem (4.2) has a solution which satisfies

$$y_i(t,\varepsilon)=-\sqrt{3}+\left(A_i+\sqrt{3}\right)\exp\left[-\sqrt{\frac{m_i}{\varepsilon}}\,(t-a)\right]$$
$$+\left(B_i+\sqrt{3}\right)\exp\left[-\sqrt{\frac{m_i}{\varepsilon}}\,(b-t)\right]+O(\varepsilon),$$

where m_i is as above, $i=1,2$. (See Fig. 4.2c.)

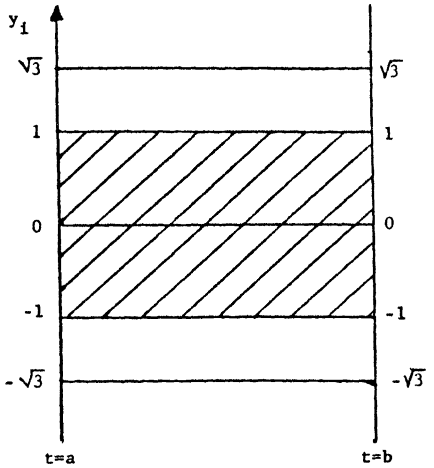


FIG. 4.2a

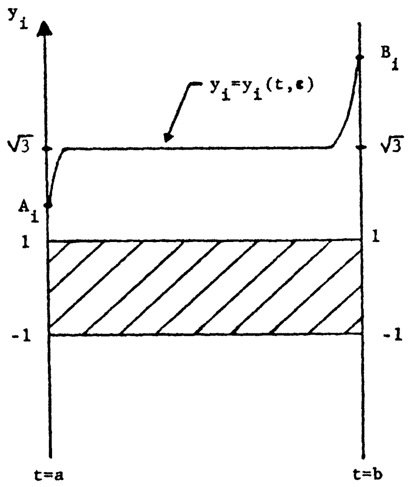


FIG. 4.2b

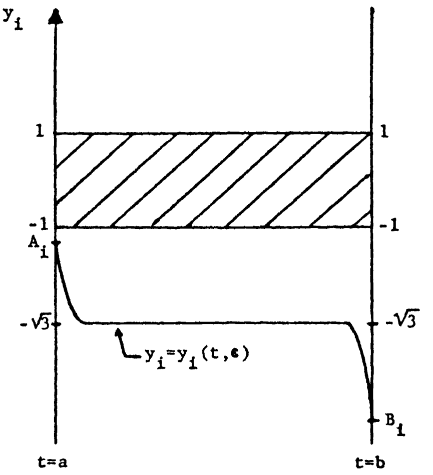


FIG. 4.2c

Next, we consider an example which exhibits both boundary layer behavior and corner layer behavior.

Example 4.3. Consider the boundary value problem

$$(4.3)\qquad \varepsilon y_1'' = \left(y_1 - \left|t-\frac{1}{2}(a+b)\right|\right)(y_2^2+1), \qquad y_1(a)=A_1, \qquad y_1(b)=B_1,$$

$$\varepsilon y_2'' = \left(y_2 + \left|\frac{1}{3}(a+b)-t\right|-1\right)(y_1^2+1), \qquad y_2(a)=A_2, \qquad y_2(b)=B_2.$$

The reduced solution $u_1=|t-(a+b)/2|$, $u_2=1-|(a+b)/3-t|$ (shown in Figs. 4.3a and 4.3b) is stable since

$$\frac{\partial h_1}{\partial y_1}=y_2^2+1\geq 1>0 \quad\text{and}\quad \frac{\partial h_2}{\partial y_2}=y_1^2+1\geq 1>0.$$

Therefore, Theorem 3.3 guarantees the existence of a solution $y=(y_1(t,\varepsilon),y_2(t,\varepsilon))$ for ε sufficiently small which satisfies the componentwise estimates:

$$\left|y_1-\left|t-\frac{1}{2}(a+b)\right|\right|\le\left(A_1-\frac{1}{2}|a-b|\right)\exp\left[-\frac{1}{\sqrt{\varepsilon}}(t-a)\right]$$

$$+\left(B_1-\frac{1}{2}|b-a|\right)\exp\left[-\frac{1}{\sqrt{\varepsilon}}(b-t)\right]+O(\varepsilon^{1/2}),$$

$$\left|y_2+\left|\frac{1}{3}(a+b)-t\right|-1\right|\le\left(A_2+\frac{1}{3}|b-2a|-1\right)\exp\left[-\frac{1}{\sqrt{\varepsilon}}(t-a)\right]$$

$$+\left(B_2+\frac{1}{3}|a-2b|-1\right)\exp\left[-\frac{1}{\sqrt{\varepsilon}}(b-t)\right]+O(\varepsilon^{1/2}),$$

for all possible boundary values A_1, B_1, A_2 and B_2. See Figs. 4.3a and 4.3b.

Finally, we look at an example which combines the results of Theorems 3.1, 3.2 and 3.3.

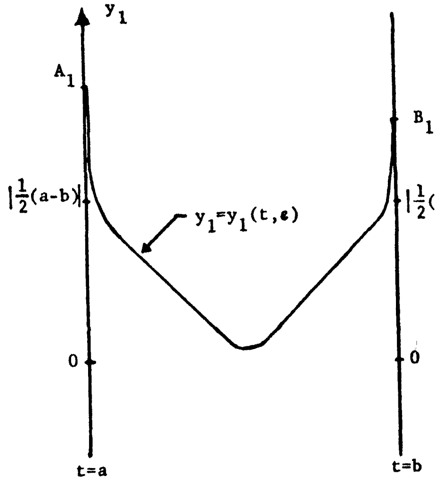


FIG. 4.3a

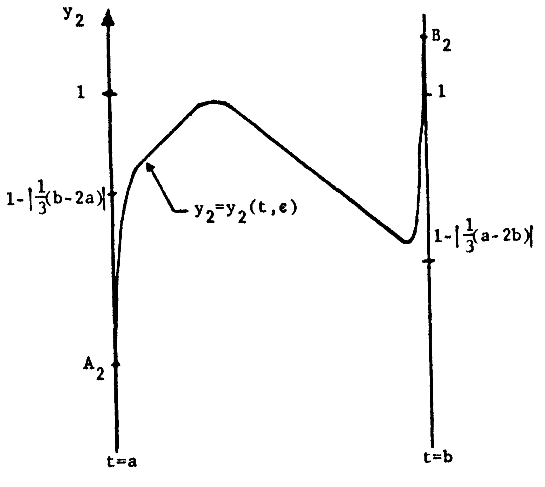


FIG. 4.3b

Example 4.4. Consider the boundary value problem

$$\text{(4.4)}\qquad \begin{aligned}
&\varepsilon y_1''=(y_1)(2-y_2)(y_3^2+1), && y_1(-1)=A_1, && y_1(1)=B_1,\\
&\varepsilon y_2''=(y_2-|t|)(y_1^2+1)(y_3^2+1), && y_2(-1)=A_2, && y_2(1)=B_2,\\
&\varepsilon y_3''=\left(\tfrac{1}{3}y_3^3-y_3\right)(2-y_2)(y_1-K), && y_3(-1)=A_3, && y_3(1)=B_3,\\
&\varepsilon y_4''=(y_4-2+|t|)(2-y_2)^2, && y_4(-1)=A_4, && y_4(1)=B_4.
\end{aligned}$$

The relevant partial derivatives are given by

$$\frac{\partial h_1}{\partial y_1}=(2-y_2)(y_3^2+1), \quad \frac{\partial h_2}{\partial y_2}=(y_1^2+1)(y_3^2+1),$$
$$\frac{\partial h_3}{\partial y_3}=(y_3^2-1)(2-y_2)(y_1-K), \quad \frac{\partial h_4}{\partial y_4}=(2-y_2)^2.$$

There are three reduced solutions

$$u_1=0, \quad u_2=|t|, \quad u_3=0, \quad u_4=2-|t|,$$
$$u_1=0, \quad u_2=|t|, \quad u_3=\sqrt{3}, \quad u_4=2-|t|,$$
$$u_1=0, \quad u_2=|t|, \quad u_3=-\sqrt{3}, \quad u_4=2-|t|,$$

and their stability depends upon the sign of K. It is convenient to divide the discussion into three cases.

Case 1. If $K>0$, the reduced solution $u_1=0$, $u_2=|t|$, $u_3=0$, $u_4=2-|t|$ is stable. For $A_1,B_1<K$, $A_2,B_2<2$ and $-1<A_3,B_3<1$, Theorems 3.1, 3.2 and 3.3 combined imply the existence of a solution of (4.4) satisfying in $[-1,1]$:

$$y_1(t,\varepsilon)=A_1\exp\left[-\sqrt{\frac{m_1}{\varepsilon}}\,(t+1)\right]+B_1\exp\left[-\sqrt{\frac{m_1}{\varepsilon}}\,(1-t)\right]+O(\varepsilon),$$
$$|y_2(t,\varepsilon)-|t||\leq(A_2-1)\exp\left[-\sqrt{\frac{m_2}{\varepsilon}}\,(t+1)\right]+(B_2-1)\exp\left[-\sqrt{\frac{m_2}{\varepsilon}}\,(1-t)\right]$$
$$+O(\varepsilon^{1/2}),$$
$$y_3(t,\varepsilon)=A_3\exp\left[-\sqrt{\frac{m_3}{\varepsilon}}\,(t+1)\right]+B_3\exp\left[-\sqrt{\frac{m_3}{\varepsilon}}\,(1-t)\right]+O(\varepsilon),$$
$$|y_4(t,\varepsilon)-2+|t||\leq(A_4-1)\exp\left[-\sqrt{\frac{m_4}{\varepsilon}}\,(t+1)\right]+(B_4-1)\exp\left[-\sqrt{\frac{m_4}{\varepsilon}}\,(1-t)\right]$$
$$+O(\varepsilon^{1/2}),$$

where

$$0<m_1<\min\{(2-A_2)(A_3^2+1),(2-B_2)(B_3^2+1)\},$$
$$0<m_2<1,$$
$$0<m_3<\min\{(A_3^2-1)(2-A_2)(A_1-K),(B_2^2-1)(2-B_2)(B_1-K)\}, \quad \text{and}$$
$$0<m_4<\min\{(2-A_2)^2,(2-B_2)^2\}.$$

(See Figs. 4.4a–d.) We note that there are no restrictions on the boundary values A_4 and B_4.

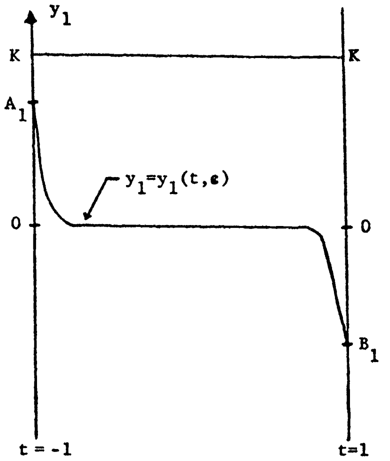


FIG. 4.4a

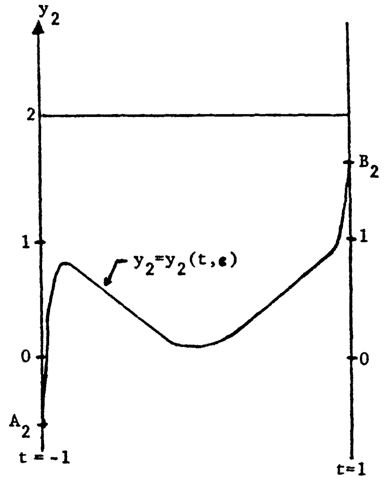


FIG. 4.4b

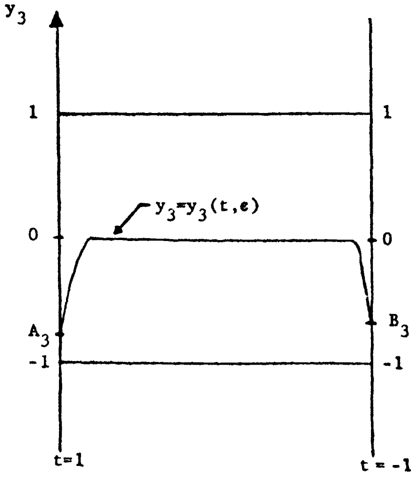


FIG. 4.4c

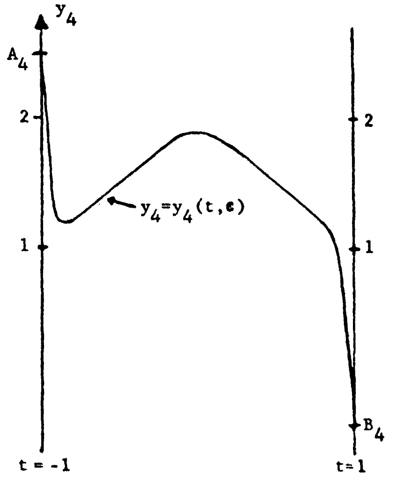


FIG. 4.4d

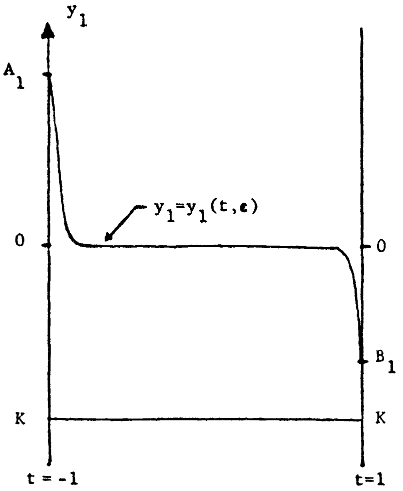


FIG. 4.4e

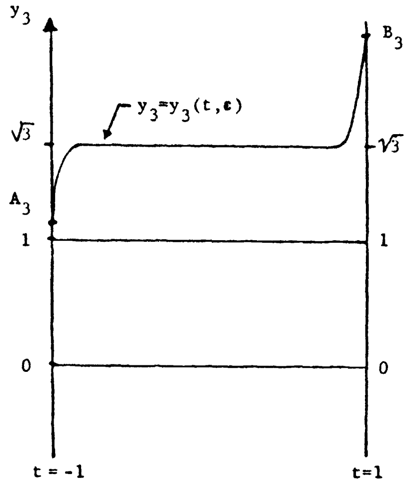


FIG. 4.4f

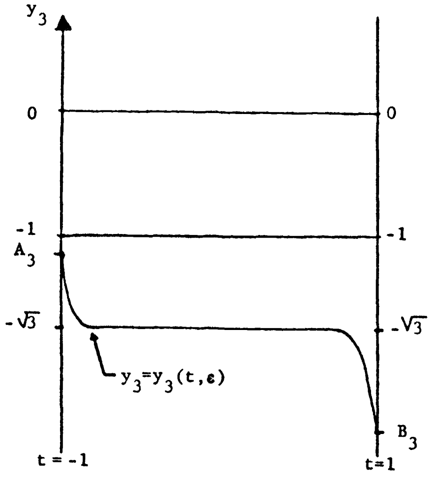


FIG. 4.4g

Case 2. If $K<0$, the reduced solutions $u_1=0$, $u_2=|t|$, $u_3=\sqrt{3}$, $u_4=2-|t|$ and $u_1=0$, $u_2=|t|$, $u_3=-\sqrt{3}$, $u_4=2-|t|$ are both stable. For $A_1,B_1>K$, $A_2,B_2<2$ and $A_3,B_3>1$, we know from Theorems 3.1–3.3 that there is a solution of (4.4) which satisfies the above estimates for y_1,y_2 and y_4, and

$$y_3(t,\varepsilon)=\sqrt{3}+\left(A_3-\sqrt{3}\right)\exp\left[-\sqrt{\frac{m_3}{\varepsilon}}\,(t+1)\right]$$
$$+\left(B_2-\sqrt{3}\right)\exp\left[-\sqrt{\frac{m_3}{\varepsilon}}\,(1-t)\right]+O(\varepsilon),$$

for m_3 as in Case 1. (See Figs. 4.4e, b, f and d, respectively.)

For $A_1,B_1>K$, $A_2,B_2>2$ and $A_3,B_3<-1$, it follows from Theorems 3.1–3.3 that there is a solution of (4.4) which satisfies the estimates of Case 1 for y_1,y_2 and y_4, while

$$y_3(t,\varepsilon)=-\sqrt{3}+\left(A_3+\sqrt{3}\right)\exp\left[-\sqrt{\frac{m_3}{\varepsilon}}\,(t+1)\right]+\left(B_3+\sqrt{3}\right)\exp\left[-\sqrt{\frac{m_3}{\varepsilon}}\,(1-t)\right]$$
$$+O(\varepsilon),$$

for m_3 as above. (See Fig. 4.4g.)

Case 3. If $K=0$, then $\partial h_3/\partial y_3=(y_3^2-1)(2-y_2)y_1=0$ along all three reduced solutions, and our theory is not applicable.

5. Concluding remarks and extensions. As the examples in §4 illustrate, the theory in §3 can be effectively applied to analyze the rich and varied behavior of a general class of systems of nonlinear singularly perturbed boundary value problems. These results may be significantly improved and they may be extended in several directions.

In the case of semilinear systems such as those above, these results have been extended in three ways.

(1) By relaxing the condition $(\partial h_i/\partial y_i)(t,y_1,\cdots,y_n)>m_i>0$ to

$$\frac{\partial^j h_i}{\partial y_i^j}(t,y_1,\cdots,u_i,\cdots,y_n)\equiv 0 \quad \text{for } 0\le j\le 2q,$$

and

$$\frac{\partial^{2q+1}h_i}{\partial y_i^{2q+1}}(t,y_1,\cdots,y_n)>m_i>0,$$

for q a nonnegative integer and $i=1,\cdots,n$, one may obtain results analogous to Theorems 3.1–3.3, the only difference being that the boundary layer terms are no longer exponential functions but rather algebraic functions, as the scalar theory suggests. (See Howes [4], [7].)

(2) By replacing the condition $(\partial h_i/\partial y_i)(t,y_1,\cdots,y_n)>m_i>0$ with $(\partial h_i/\partial y_i)(t,y_1,\cdots,u_i,\cdots,y_n)>m_i>0$ along with certain integral inequalities (cf. Howes [5]) to insure boundary layer stability we may improve the above results dramatically.

(3) Again by assuming that appropriate integral conditions hold, one may analyze interior behavior of the shock layer type for such systems; cf. for example [5], [7].

Similar results may be obtained for both quasilinear and quadratic systems of nonlinear singularly perturbed boundary value problems, including the analysis of turning point and shock layer phenomena in such systems. These extensions of the above results may be anticipated from results in scalar theory. (See, for example, [4], [5] and [7].)

All these results for such nonlinear systems will appear in the author's forthcoming doctoral dissertation [11].

REFERENCES

[1] S. Bernfeld and V. Lakshmikantham, *An Introduction to Nonlinear Boundary Value Problems*, Academic Press, New York, 1974.

[2] N. I. Brish, *On boundary value problems for the equation $\varepsilon y''=f(x,y,y')$ for small ε*, Dokl. Akad. Nauk SSSR, 95 (1954), pp. 429–432. (In Russian.)

[3] P. Habets and M. Laloy, *Étude de problèmes aux limites par la méthode des sur-et sous-solutions*, Lecture notes, Catholic University of Louvain, 1974.

[4] F. A. Howes, *A Class of boundary-value problems whose solutions possess angular limiting behavior*, Rocky Mtn. J. Math., 6 (1976), pp. 591–607.

[5] ———, *Boundary-interior layer interactions in nonlinear singular perturbation theory*, Mem. Amer. Math. Soc., 203, 1978.

[6] ———, *Singularly perturbed semilinear systems*, Stud. Appl. Math., 61 (1979), pp. 185–209.

[7] ———, *Some old and new results on singularly perturbed boundary value problems*, in Singular Perturbations and Asymptotics, Parter and Meyer, eds., Academic Press, New York, 1980, pp. 41–85.

[8] F. A. Howes and R. E. O'Malley, Jr., *Singular perturbations of second-order semilinear systems*, Proc. Conference on O.D.E. and P.D.E. at Dundee, Lecture Notes in Mathematics 827, Springer, Berlin, 1980, pp. 131–150.

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

[10] M. Nagumo, *Über die Differentialgleichung $y''=f(x,y,y')$*, Proc. Phys. Math. Soc. Japan, 19 (1937), pp. 861–866.

[11] M. A. O'Donnell, *Boundary and interior layer behavior in singularly perturbed nonlinear systems*, Doctoral dissertation, Univ. California, Davis, in preparation.

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

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1502-0008 $01.25/0

SINGULARLY PERTURBED HYPERBOLIC EVOLUTION PROBLEMS WITH INFINITE DELAY AND AN APPLICATION TO POLYMER RHEOLOGY*

MICHAEL RENARDY†

Abstract. We prove an existence theorem locally in time for quasilinear hyperbolic equations, in which the coefficients are allowed to depend on the history of the dependent variable. Singular perturbations, which change the type of the equation to parabolic, are included, and continuous dependence of the solutions on the perturbation parameter is shown. It is demonstrated that, for a substantial number of constitutive models suggested in the literature, the stretching of filaments of polymeric liquids is described by equations of the kind under study here.

AMS-MOS subject classification (1980). Primary 35L15, 45K05, 47D05, 76A10

Key words. quasilinear hyperbolic equations, differential delay equations, semigroups, singular perturbations, viscoelastic liquids

1. Introduction. In a recent paper [23] I proved an existence theorem (locally in time) for solutions to a class of quasilinear parabolic differential-delay equations that can be used to model the stretching of filaments of polymeric liquids. Such equations arise, if the constitutive law is such that, besides an "elastic" part which is a functional of the strain history, the stress has also a Newtonian part. For many materials, e.g. molten polyethylene, however, this latter contribution is small. This warrants a theory that can treat the Newtonian contribution as a perturbation rather than as the "leading" term in the equation.

In the present paper, I shall give a partial solution to this problem. Mathematically, we are concerned with differential equations of hyperbolic type with a small perturbation changing the type to parabolic. A mathematical theory applicable to such problems was developed by Kato [12], [14]–[16]. (My results in [23] were based on the theory of Sobolevskii [25].) Since Kato's theory is more easily applicable to pure Cauchy problems than for mixed initial-boundary value problems (some results concerning the latter are in [16]), we confine our attention to the former class of problems here. Physically, this means that rather than a filament pulled at its ends, we will study the deformation of infinite filaments subjected to longitudinal body forces. It is hoped that further research will lead to similar results for the boundary value problem and also for more general (in particular more than one-dimensional) flow geometries. It is clear that the results we obtain apply to other one-dimensional problems in continuum mechanics, e.g., those discussed in [6], [9].

In §2, I quote those results of Kato's theory that are needed in this paper. One of Kato's results will be mildly generalized. In §3, these results are applied to a class of

*Received by the editors June 11, 1982, and in revised form November 18, 1982. This research was sponsored by the United States Army under contract DAAG29-80-C-0041 and Deutsche Forschungsgemeinschaft. The research leading to this paper was started at the Mathematics Research Center, University of Wisconsin-Madison, Madison, Wisconsin, 53705, and concluded at the Department of Aerospace Engineering, University of Minnesota, Minneapolis, Minnesota 55455.

†Mathematics Research Center and Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706.

singularly perturbed quasilinear hyperbolic differential-delay equations, which have the following form

$$(1.1)\qquad \rho u_{\underbrace{tt\cdots t}_{n}}=\eta\cdot\left(f\cdot u_{x\underbrace{t\cdots t}_{n-1}}\right)_x+h\cdot u_{xx\underbrace{t\cdots t}_{n-2}}+k+\phi,\qquad x\in\mathbb{R}.$$

Here η is a small nonnegative constant, ϕ is a given function of x and t, and f, h, k are functionals of the histories of derivatives of u which are of lower order than those displayed. It is assumed that f and h take positive values. Under appropriate assumptions, we prove that the initial history problem associated with (1.1) has a unique solution locally in time and, moreover, that this solution depends continuously on η, even as $\eta\to 0$. In representing differential-delay equations as abstract evolution problems, we follow the method outlined in [23] rather than the classical approach [11].

Section 4 deals with the problem of stretching filaments of polymeric liquids. We use a one-dimensional approximation to this problem based on the thinness of the filament [23]. Various constitutive models suggested in the rheological literature [1]–[5], [8], [10], [13], [17]–[22], [27] are discussed. It is shown that, for all these models, an equation of the form (1.1) is obtained. Section 4 can be read independently of the preceding sections, and physically oriented readers are encouraged to read it first.

The diversity of the models studied here illustrates the fact that—up to now—there is no particular constitutive law describing successfully all the phenomena in polymer rheology. Whether or not one constitutive law can fully describe a substantial number of materials, is not yet known. As pointed out in [26], a "theory of theories" is needed. In [26] particular attention is focussed on special flow geometries, for which the precise nature of the constitutive law is not very important. The situation under study here does not seem to be of such a nature, and we do need some specification of the constitutive law. It turns out that a number of popular rheological models all lead to equations of the form (1.1).

The reason for this common mathematical feature lies basically in smoothing properties: If a function is convoluted with a smooth kernel, then the result is once more often differentiable than the function itself. The same applies if a first order differential equation is solved. Since the strain histories appear in expressions of such a form, we can, after some differentiations, obtain a form in which the highest derivatives occur only by their present values rather than their histories. More precisely, for some $n\in\mathbb{N}$, the nth time derivative of the stress depends linearly on the nth and $(n+1)$st time derivatives of the strain with coefficients depending only on lower order derivatives. One may regard this as a generalization of the old idea that there is a superposition of "elastic" and "viscous" contributions. Instead of a linear superposition as suggested in the oldest models, we have here what may be called a "quasilinear" superposition. It is essential in our development that the integral kernels occurring in the constitutive equation are smooth everywhere, in particular, that they are bounded. This assumption has also been made by other authors [9], [30], [31]. Both molecular theories and experiments suggest, however, that the kernels may have a singularity (see the remark at the end of §4). This leads to pseudo-differential operators of nonintegral order, that is, to terms whose differential order is intermediate between the two terms included in (1.1). Further research needs to be done on such equations.

2. Abstract hyperbolic equations. In this section, I summarize the results from Kato's theory that will be needed in the following. One of Kato's theorems will be generalized.

We study an evolution problem of the form

$$(2.1)\qquad \dot{u}=A(t,u)u+f(t,u),\qquad 0\leq t\leq T,\quad u(0)=\phi,$$

where u takes values in a Banach space X and $A(t,u)$ is a linear operator depending on t and u. Our assumptions will involve further Banach spaces Y and Z such that $Y\subset Z\subset X$ with continuous and dense embeddings. It is assumed that Y, Z and X are reflexive and separable. Let W denote an open set in Y.

First, we quote [12, Thm. I] in a simplified form (with assumption N being obsolete). It is assumed that the following estimates hold for $t,t',\cdots\in[0,T]$ and $w,w',\cdots\in W$ (K denotes a generic constant independent of t and w):

($S1$) There is an isomorphism $S(t,w)$: $Y\rightarrow X$ satisfying

$$\|S(t,w)\|_{Y,X}\leq K,\qquad \|S^{-1}(t,w)\|_{X,Y}\leq K,$$
$$\|S(t',w')-S(t,w)\|_{Y,X}\leq K(|t-t'|+\|w-w'\|_Z).$$

($A1$) $A(t,w)$ generates a quasi-contraction semigroup in X uniformly with respect to t,w: $\|e^{A(t,w)\tau}\|_{X,X}\leq e^{K\tau}$.

($A2$) $S(t,w)A(t,w)S^{-1}(t,w)=A(t,w)+B(t,w)$ where $B(t,w)\in B(X)$, $\|B(t,w)\|_{X,X}\leq K$.

($A3$) $A(t,w)\in B(Y,Z)$ with $\|A(t,w)\|_{Y,Z}\leq K$ and $\|A(t,w')-A(t,w)\|_{Y,X}\leq K\|w'-w\|_X$. The mapping $t\mapsto A(t,w)\in B(Y,X)$ is norm-continuous.

($A4$) There is some $y_0\in W$ such that $A(t,w)y_0\in Y$ and $\|A(t,w)y_0\|_y\leq K$.

($f1$) $f(t,w)\in Y$, $\|f(t,w)\|_Y\leq K$, $\|f(t,w')-f(t,w)\|_X\leq K\|w'-w\|_X$. Moreover, the mapping $t\mapsto f(t,w)\in X$ is continuous

[12, Thm. I] reads as follows.

THEOREM 2.1. *Let* (S), ($A1$)–($A4$) *and* ($f1$) *hold. Then there are a positive* ρ *and a positive* $T'\leq T$ *such that for* $\|\phi-y_0\|_Y\leq\rho$, *equation* (2.1) *has a unique solution* $u\in C^0([0,T'];\ W)\cap C^1([0,T'];\ Z)$. ρ *and* T' *depend only on* K *and the distance of* y_0 *from the boundary of* W.

The second theorem stated in this chapter is a continuous dependence result. It generalizes [12, Thm. II] insofar as it allows S to depend on w. We adopt the following assumptions:

($S2$) There is an open set $W'\subset Z$ such that $W\subset W'$ and the following holds. The definition of $S(t,w)\in B(Y,X)$ can be extended to $w\in W'$. Moreover, we have uniformly on $[0,T]\times W'$:

$$\|S(t,w)\|_{Y,X}\leq K,\quad \|D_wS(t,w)\|_{Z,B(Y,X)}\leq K,\quad \|D_tS(t,w)\|_{Y,X}\leq K,$$
$$\|S(t,w')-S(t,w)\|_{Y,X}\leq K\|w'-w\|_Z,$$
$$\|D_wS(t,w')-D_wS(t,w)\|_{Z,B(Y,X)}\leq K\|w'-w\|_Z,$$
$$\|D_tS(t,w')-D_tS(t,w)\|_{Y,X}\leq K\|w'-w\|_Z.$$

Here D_w and D_t denote the derivatives with respect to t and w.

($A5$) $\|B(t,w')-B(t,w)\|_X\leq K\|w'-w\|_Y$.

$(A6)$ $\|A(t,w)-A(t,w')\|_{Y,Z}\leq K\|w'-w\|_Y$.

$(f2)$ $\|f(t,w')-f(t,w)\|_Y\leq K\|w'-w\|_Y$.

Let us now consider a sequence of evolution problems ($n\in\mathbb{N}$)

$$(2.2)\qquad \dot{u}^n=A(t,u^n)u^n+f(t,u^n),\qquad 0\leq t\leq T,\quad u^n(0)=\phi^n.$$

THEOREM 2.2. *Assume* $(A1)$–$(A6)$, $(f1)$, $(f2)$ *are satisfied uniformly in n and assume* $(S1)$, $(S2)$. (*The operator S shall not depend on n.*) *Moreover, assume* $\|\phi^n-y_0\|_Y<\rho$, $\|\phi-y_0\|_Y<\rho$ *with* ρ *as of Theorem* 2.1. *Finally, assume that for* $t,w\in[0,T]\times W$

$A^n(t,w)\to A(t,w)$ *strongly in* $B(Y,Z)$,

$B^n(t,w)\to B(t,w)$ *strongly in* $B(X)$,

$f^n(t,w)\to f(t,w)$ *in* Y

as $n\to\infty$. *If* $\phi^n\to\phi$ *in the Y-norm as* $n\to\infty$, *then there is a* $T''<T$ *such that* (2.2) *has a solution* $u^n\in C^1([0,T'']; X)\cap C^0([0,T''] W)$ *for any n. Moreover,* $u^n(t)\to u(t)$ *in* Y, *uniformly for* $t\in[0,T'']$', *where u is a solution of* (2.1).

Proof. The proof essentially follows the same line of argument as that in [15]. Theorem 2.1 yields the existence of solutions to (2.2) and the limiting equation (2.1) on some interval $[0,T']$ with T' independent of n. It is moreover proved precisely as in [15] that $u^n\to u$ uniformly in t in the X-norm. To prove convergence in the Y-norm, we rely on [14, Thm., IV]. This involves estimating a number of terms. Most of those estimates go as in [15] or are straightforward, and my exposition will focus only on those terms that present difficulties. As in [15], we use the fact that u^n solves the linear equation

$$(2.3)\qquad \dot{u}^n=A^nu^n+f^n,$$

with $A^n=A^n(t,u^n)$ and $f^n=f^n(t,u^n)$. The limit u solves the linear equation

$$(2.4)\qquad \dot{u}=Au+f$$

with $A=A(t,u), f=f(t,u)$. From [14, Thm. IV], we have the estimate

$$\begin{aligned}\|u^n-u\|_{\infty,Y}\leq K\big(&\|\phi^n-\phi\|_Y+\|f^n-f\|_{1,Y}\big)\\ &+K\big(\|(S^n(0)-S(0))\phi\|_X+\|(S^n-S)f\|_{1,X}+\|(S^n-S)u\|_{\infty,X}\big)\\ &+K\big(\|(B^n-B)Su\|_{1,X}+\|(C^n-C)Su\|_{1,X}\big)\\ &+K\big(\|(U^n-U)(\delta\otimes\psi\oplus g)\|_{\infty,X}\big).\end{aligned}$$

Here S^n denotes $S(t,u^n)$ and B^n denotes $B(t,u^n)$. The symbol C stands for $\dot{S}S^{-1}$. The U, U^n are the evolution operators associated with A, A^n. Finally, ψ denotes $S(0)\phi$, and g stands for $Sf+(C-B)Su$. The indices 1 and ∞ indicate the L^1- and L^∞-norms on the interval $[0,T'']$. On the right-hand side of (2.5), the term $\|\phi^n-\phi\|_Y$ converges to zero by assumption and we are left with seven more contributions. Of these, the first, fifth and seventh have been dealt with in [15], and no change in the argument is needed here. The second, third and fourth terms are estimated in terms of $\|\phi^n-\phi\|_Y$ and $\|u^n-u\|_{\infty,Z}$ by virtue of $(S1)$. Now, note that

$$\begin{aligned}\|u^n-u\|_{\infty,Z}&\le K\big(\|\phi^n-\phi\|_Z+\|\dot u^n-\dot u\|_{1,Z}\big)\\&\le K\big(\|\phi^n-\phi\|_Y+\|A^nu^n+f^n-Au-f\|_{1,Z}\big).\end{aligned}$$

The last term will be estimated below. We may thus focus on the term

$$\begin{aligned}\|(C^n-C)Su\|_{1,X}&=\big\|\big(\dot S^n(S^n)^{-1}-\dot SS^{-1}\big)Su\big\|_{1,X}\\&\le\|(\dot S^n-\dot S)u\|_{1,X}+\big\|\dot S^n\big((S^n)^{-1}-S^{-1}\big)Su\big\|_{1,X}.\end{aligned}$$

By $(S2)$, $\dot S^n$ is bounded in $B(Y,X)$, and by $(S1)$, $\|(S^n)^{-1}-S^{-1}\|_{X,Y}$ can be estimated by $\|u^n-u\|_{\infty,Z}$. This takes care of the second term. For the first term, observe that

$$\begin{aligned}\dot S^n(t)=\frac{d}{dt}S(t,u^n(t))&=\dot S(t,u^n(t))+D_uS(t,u^n(t))\dot u^n(t)\\&=\dot S(t,u^n(t))+D_uS(t,u^n(t))(A^n(t,u^n)u^n+f^n(t,u^n))\end{aligned}$$

and likewise

$$\dot S(t)=\dot S(t,u)+D_uS(t,u)(A(t,u)u+f(t,u)).$$

We have

$$\begin{aligned}\|(\dot S(t,u^n)-\dot S(t,u))u\|_{1,X}&\le\|\dot S(t,u^n)-\dot S(t,u)\|_{1,Y,X}\|u\|_{\infty,Y}\\&\le K\|u^n-u\|_{1,Z}\le K\|u^n-u\|_{1,Y}\end{aligned}$$

and

$$\begin{aligned}&\|(D_uS(t,u^n(t))-D_uS(t,u(t)))(A^n(t,u^n(t))u^n(t)+f^n(t,u^n(t)))\|_{1,B(Y,X)}\\&\qquad\le\|(D_uS(t,u^n(t))-D_uS(t,u(t)))\|_{1,Z,B(Y,X)}\|A^nu^n+f^n\|_{\infty,Z}\\&\qquad\le K\|u^n-u\|_{1,Z}\le K\|u^n-u\|_{1,Y}.\end{aligned}$$

Finally,

$$\begin{aligned}&\|D_uS(t,u(t))(A^nu^n+f^n-Au-f)\|_{1,B(Y,X)}\\&\qquad\le\|D_uS(t,u(t))\|_{\infty,Z,B(Y,X)}\|A^nu^n+f^n-Au-f\|_{1,Z}.\end{aligned}$$

This last term can be estimated by

$$\begin{aligned}&\|A^n(t,u)u-A(t,u)u\|_{1,Z}+\|f^n(t,u)-f(t,u)\|_{1,Z}\\&\quad+\|(A^n(t,u^n)-A^n(t,u))u\|_{1,Z}+\|A^n(t,u^n)(u^n-u)\|_{1,Z}\\&\quad+\|f^n(t,u^n)-f^n(t,u)\|_{1,Z}.\end{aligned}$$

It follows from assumptions ($A6$), ($f2$) that the last three terms can be estimated by $\|u^n-u\|_{1,Y}$. The first two contributions converge to zero because of the assumptions of the theorem. This concludes the proof.

3. Application to delay equations. In this section, we shall apply the preceding results to differential-delay equations of the following form:

$$(3.1)\qquad \begin{aligned} u_{(n-1)}=&\,\eta\big(f(\hat u_x,\hat u_{xt},\cdots,\hat u_{x(n-4)})u_{x(n-2)}\big)_x\\ &+h(\hat u_x,\hat u_{xt},\cdots,\hat u_{x(n-3)};\,\eta)u_{xx(n-3)}\\ &+k(\hat u_x,\cdots,\hat u_{x(n-3)};\,\hat u_{xx},\cdots,\hat u_{xx(n-4)};\,\eta)+\tilde\phi(t,x). \end{aligned}$$

Here, the index (k) stands for k-fold differentiation with respect to the time variable t, and the hat denotes the past history: $\hat u_x(t,x)(S)=u_x(t+S,X)$ for $S\in(-\infty,0]$. The $\tilde f$, $\tilde h$ and $\tilde k$ are smooth functionals on a history space, the topology of which will be specified later.

Differentiating (3.1) with respect to x, we obtain

$$(3.2)\qquad \begin{aligned} u_{x(n-1)}=&\,\eta\big(f(\hat u_x,\hat u_{xt},\cdots,\hat u_{x(n-4)})u_{x(n-2)}\big)_{xx}\\ &+h(\hat u_x,\hat u_{xt},\cdots,\hat u_{x(n-3)};\,\eta)u_{xxx(n-3)}+\frac{\partial h}{\partial x}\cdot u_{xx(n-3)}\\ &+\big(k(\hat u_x,\cdots,\hat u_{x(n-3)};\,\hat u_{xx},\cdots,\hat u_{xx(n-4)};\,\eta)\big)_x+\phi(t,x), \end{aligned}$$

where $\phi=\partial\tilde\phi/\partial x$. In the following, we need only be concerned with equations of the form (3.2). For applying the results of §2, it is convenient to rewrite (3.2) as a system of equations. Let us put $v_k=u_{x(k)}$, $w_k=u_{xx(k)}$. We thus obtain the following system equivalent to (3.2):

$$(3.3)\qquad \begin{aligned} &v_{kt}=v_{k+1},\qquad k=0,1,\cdots,n-3,\\ &w_{kt}=w_{k+1},\qquad k=0,1,\cdots,n-4,\\ &v_{n-2,t}=\eta\big(f(\hat v_0,\cdots,\hat v_{n-4})v_{n-2}\big)_{xx}+h(\hat v_0,\cdots,\hat v_{n-3};\,\eta)w_{n-3,x}+\frac{\partial h}{\partial x}w_{n-3}\\ &\qquad\quad+\big(k(v_0,\cdots,v_{n-3};\,w_0,\cdots,w_{n-4};\,\eta)\big)_x+\phi,\\ &w_{n-3,t}=v_{n-2,x}. \end{aligned}$$

It will be advantageous to make some further substitutions. Let us put $v'_{n-2}=fv_{n-2}$, $w'_{n-3}=f\sqrt{h}\,w_{n-3}+f/\sqrt{h}\cdot k$. Then we obtain the following system:

$$(3.4)\qquad \begin{aligned} &v_{kt}=v_{k+1},\qquad k=0,1,\cdots,n-4,\\ &v_{n-3,t}=\frac{v'_{n-2}}{f},\\ &w_{kt}=w_{k+1},\qquad k=0,1,\cdots,n-5,\\ &w_{n-4,t}=\frac{1}{f\sqrt{h}}w'_{n-3}-\frac{k}{h},\\ &v'_{n-2,t}=\eta fv'_{n-2,xx}+\sqrt{h}\,w'_{n-3,x}\\ &\qquad\quad+\left(\frac{1}{2}f\frac{\partial h}{\partial x}-h\frac{\partial f}{\partial x}\right)\cdot\left(\frac{1}{f\sqrt{h}}w'_{n-3}\right)+\frac{\partial f}{\partial t}\cdot\frac{v'_{n-2}}{f}+\phi\cdot f \end{aligned}$$

$$w'_{n-3,t}=\sqrt{h}\,v'_{n-2,x}-\frac{\sqrt{h}}{f}\frac{\partial f}{\partial x}v'_{n-2}$$
$$+\frac{\partial}{\partial t}\left(f\sqrt{h}\right)\cdot\left(\frac{w'_{n-3}}{f\sqrt{h}}-\frac{k}{h}\right)+\frac{\partial}{\partial t}\left(\frac{fk}{\sqrt{h}}\right).$$

Next, let us define the history spaces, in which (3.4) will be analyzed.

DEFINITION 3.1. For a given Banach space Z, let $\tilde{H}^1(Z)$ denote the space of all functions $(-\infty,0]\to Z$, which are the sum of a constant element $z_0\in Z$ and a function $z(t)$, which is square integrable and has a square integrable derivative (in the Bochner sense). Analogously, let $\tilde{L}^2(z)$ be the space of all functions $(-\infty,0]\to Z$ which are the sum of a constant and a square integrable function.

In particular, $\tilde{H}^1$ shall denote $\tilde{H}^1(\mathbb{R})$ and $\tilde{H}^1_m$ shall denote $\tilde{H}^1(H^m(\mathbb{R}))$ where $H^m(\mathbb{R})$ is the Sobolev space of all functions $\mathbb{R}\to\mathbb{R}$, which have m square integrable derivatives. Analogously, let $\tilde{L}^2_m=\tilde{L}^2(H^m(\mathbb{R}))$.

Remarks. 1. In [23], I used the space $C_b^{\lim}$ (the space of bounded continuous functions having a limit at $-\infty$). The reason why this space cannot be used here is that it is not reflexive as required by Kato's theory.

2. The choice of the space $\tilde{H}^1$ seems to impose rather restrictive conditions on the given history. However, as in [23], one can allow histories in more general spaces, e.g. "fading memory" spaces [7], by reducing the problem to one that has a history in $\tilde{H}^1$, but is equivalent to the given one for $t>0$. The only modification necessitated by this is that f, h, k must be allowed to depend explicitly on x and t. This modification presents no major difficulties.

As in [23], we define a shift operator T_S on $\tilde{H}^1$: $T_S\phi(t)=\phi(t+S)$ for $S\in(-\infty,0]$.

Our assumptions on f, h, k in (3.3) are as follows:

(i) The mappings $f:(\tilde{H}^1)^{n-3}\to\mathbb{R}$, $h:(\tilde{H}^1)^{n-3}\times\mathbb{R}\to\mathbb{R}$ and $k:(\tilde{H}^1)^{2n-3}\times\mathbb{R}\to\mathbb{R}$ are smooth (i.e., sufficiently often continuously differentiable) and the induced operators $\hat{f}$, $\hat{h}$, $\hat{k}$ defined by $\hat{f}(\phi)(S)=f(T_S\phi)$ map into $\tilde{H}^1$ and depend again smoothly on their arguments. Moreover, f and h take strictly positive values: $f\geq\varepsilon>0$, $h\geq\varepsilon>0$. k vanishes if its arguments are zero. Moreover, the Fréchet derivative Df is a linear operator from $(\tilde{L}^2)^{n-3}$ into $\mathbb{R}$ which depends smoothly on the arguments of f (in the topology of $\tilde{H}^1$), and the corresponding operator $D\hat{f}$ maps into $\tilde{L}^2$ and again depends smoothly on the arguments of f. Analogous conditions hold for h and k.

The following lemma is easily proved.

LEMMA 3.2. *If* (i) *holds with a sufficient degree of differentiability, then $\hat{f}$ defines in a natural way* (*acting pointwise in the space variable x*) *a smooth operator from* $(\tilde{H}^1_m)^{n-3}$ *into $\tilde{H}^1_m+\mathbb{R}$. Here m is a given integer greater than or equal to* 1. *The same holds for $\hat{h}$, $\hat{k}$, and an analogous statement also holds for $D\hat{f}$, $D\hat{h}$, $D\hat{k}$* (*regarded as linear operators* $(\tilde{L}^2)^{n-3}$ *or* $(\tilde{L}^2)^{2n-3}\to\tilde{L}^2$).

Remark. Since we are concerned with existence theorems locally in time, it is clearly sufficient that condition (i) holds in a neighborhood of the prescribed initial condition.

Let us also note that

$$\frac{\partial h}{\partial x}=\sum_{i=1}^{n-2}D_ih(\hat{v}_0,\cdots,\hat{v}_{n-3};\eta)\hat{w}_{i-1},$$

where D_i denotes the Fréchet derivative w.r. to the ith argument. Analogous manipulations are possible for the oherr $x-$ and $t-$ derivatives of f, h and k that occur in (3.4).

As in [23], let us assume that the given initial history up to time $t=0$ satisfies the equations (this can be achieved by appropriately changing ϕ). We can then rewrite (3.4) in the abstract form

$$
\begin{aligned}
(3.5)\qquad & \hat{v}_{kt}=\hat{v}_{k+1}, \quad k=0,1,\cdots,n-4,\\
& \hat{w}_{kt}=\hat{w}_{k+1}, \quad k=0,1,\cdots,n-5,\\
& \hat{v}_{n-3,t}=\frac{\hat{v}'_{n-2}}{\hat{f}},\\
& \hat{w}_{n-4,t}=\frac{\hat{w}'_{n-3}}{f\sqrt{\hat{h}}}-\frac{\hat{k}}{\hat{h}},\\
& \hat{v}'_{n-2,t}=\eta\hat{f}\hat{v}'_{n-2,xx}+\sqrt{\hat{h}}\,\hat{w}'_{n-3,x}+\cdots+\hat{\phi}\cdot f,\\
& \hat{w}'_{n-3,t}=\sqrt{\hat{h}}\,\hat{v}'_{n-2,x}+\cdots,
\end{aligned}
$$

where $\hat{f}$, $\hat{h}$ etc. are defined as in Lemma 3.2. Now let the spaces X, Y be as follows: $X=(\tilde{H}_1^1\cap\tilde{L}_4^2)^{2n-5}$, $Y=(\tilde{H}_m^1\cap\tilde{L}_{m+3}^2)^{2n-5}$, where m is an odd number greater or equal to 3. Finally, let $Z=(\tilde{H}_{m-2}^1\cap\tilde{L}_{m+1}^2)^{2n-5}$. When identifying (3.5) with the abstract system (2.1), we incorporate in A only those terms that contain derivatives w.r. to x, everything else is included in f. With this identification, the conditions ($A3$), ($A4$), ($f1$), ($A6$) and ($f2$) are rather obvious, (note that, for $m\geq 3$, $\tilde{H}_m^1\cap\tilde{L}_{m+3}^2$ is a Banach algebra) provided that the following holds:

(ii) $\hat{\phi}$ takes values in $\tilde{H}_m^1$ (it follows automatically that it is continuous into $\tilde{H}_m^1$), and the initial condition at $t=0$ lies in Y.

For verifying the remaining conditions, we have to study the operator $A(\hat{f},\hat{h})$, defined by

$$A(\hat{f},\hat{h})(\hat{v},\hat{w})=\left(\eta\hat{f}\hat{v}_{xx}+\sqrt{\hat{h}}\,\hat{w}_x,\ \sqrt{\hat{h}}\,\hat{v}_x\right).$$

The operator S is defined by $S=(\hat{f}\partial^2/\partial x^2-\lambda)^{(m-1)/2}$ for $\lambda\in\mathbb{R}$ large enough. With this choice of S, conditions ($S1$), ($S2$) are obvious.

Moreover, one sees easily that

$$\left(\left(\hat{f}\frac{\partial^2}{\partial x^2}-\lambda\right)A(\hat{f},\hat{h})-A(\hat{f},\hat{h})\left(\hat{f}\frac{\partial^2}{\partial x^2}-\lambda\right)\right)(\hat{v},\hat{w})$$

yields an expression involving only first and second order derivatives of $\hat{v}$, $\hat{w}$, $\hat{f}$ and $\hat{h}$. From this it is not difficult to conclude ($A2$) and ($A5$). (Note that $T^nA-AT^n=T^{n-1}(TA-AT)+T^{n-2}(TA-AT)T+\cdots$ and apply this with $T=(\hat{f}\partial^2/\partial x^2-\lambda)$). For ($A1$), we have to show that $\operatorname{Re}(A(\hat{v},\hat{w}),(\hat{v},\hat{w}))_{(\tilde{H}_1^1)^2}\leq C((\hat{v},\hat{w}),(\hat{v},\hat{w}))_{(\tilde{H}_1^1)^2}$, which follows from a simple integration by parts in the x-variable.

We have thus proved:

THEOREM 3.3. *Let* (i), (ii) *be satisfied. Then there is a* $T>0$ *such that* (3.5) *has a solution* $\hat{U}=(\hat{v}_0,\hat{v}_1,\cdots,\hat{w}'_{n-2})\in C^1([0,T];\ X)\cap C^0([0,T];\ Y)$. $\hat{U}$ *depends continuously on* $\eta\in[0,\eta_0]$ *in the norm of* Y.

4. Stretching of filaments of viscoelastic liquids. We are studying the motion of an infinitely extended filament of an incompressible viscoelastic liquid under the influence of a longitudinal body force. It was shown in [23] that, if the filament is thin, this problem can be modelled by a one-dimensional approximation, where only longitudinal motions need to be studied. Let $u(x,t)$ denote the position of a fluid particle at time t, which is at the position x in certain reference state. For simplicity, this reference configuration is chosen to be one in which the filament has uniform thickness. I showed in [23] that the evolution of u is governed by the following equation

$$\rho u_{tt}=\frac{\partial}{\partial x}\left(u_x\pi^{11}-u_x^{-2}\pi^{22}\right)+\phi. \tag{4.1}$$

In this equation ρ denotes the density of the fluid (i.e., a constant), ϕ is the given body force, and π^{11}, π^{22} are the longitudinal and transverse components of the convected extra stress tensor (i.e., not including the pressure, which was eliminated in the derivation of (4.1)). The tensor π is related by a constitutive law to the right Cauchy-Green tensor γ (in our notation we follow [22]). In the approximation leading to (4.1), γ is given by

$$\gamma=\begin{pmatrix} u_x^2 & 0 & 0 \\ 0 & u_x^{-1} & 0 \\ 0 & 0 & u_x^{-1} \end{pmatrix}.$$

In the following, we discuss various constitutive laws that have been suggested in the rheology literature and the corresponding equations (4.1) that they lead to. It will be shown that all these equations can be transformed to the form (3.1). In particular, we shall check the positivity of the functions f and h. (It will always be understood that u is the sum of a given function $u_0(x,t)$ and a function tending to zero appropriately as $x\to\pm\infty$, moreover, u_x is always assumed uniformly positive. In comparison to §3, the variable called u there will be identified with $u-u_0(x,t)$ in this section.) The notation used in the original papers cited here is often different from ours, and we have transcribed the constitutive laws appropriately. Tables of some (but not all) constitutive assumptions discussed here can be found in [2] and [22].

a) The rubberlike liquid of Green and Tobolsky [10] and Lodge [19], [20] and modifications of Ward and Jenkins [27] and Lodge [21]. In these theories, the constitutive law has the following form:

$$\pi=-\eta\frac{\partial}{\partial t}(\gamma^{-1})+\int_{-\infty}^{t}a(t-s)\gamma^{-1}(s)\,ds-\int_{-\infty}^{t}b(t-s)\gamma^{-1}(t)\gamma(s)\gamma^{-1}(t)\,ds$$
$$+\int_{-\infty}^{t}c(t-s)\gamma^{-1}(s)\gamma(t)\gamma^{-1}(s)\,ds+\int_{-\infty}^{t}d(t-s)(\gamma(t):\gamma^{-1}(s))\gamma^{-1}(s)\,ds.$$

The first term is a Newtonian contribution, the second is the one given by the rubberlike liquid theory [10], [19], [20]. The third term accounts for a modification suggested by Ward and Jenkins [27], and the last two represent corrections of Lodge [21]. (Lodge finds $c=2d$ from a molecular theory, but we shall make no use of this.)

With this constitutive law, (4.1) assumes the following form:

$$
\begin{aligned}
\rho u_{tt} = {} & 3\eta \frac{\partial^2}{\partial x \partial t}\left(-\frac{1}{u_x}\right) \\
& + \frac{\partial}{\partial x}\left\{ u_x^3(t) \cdot \int_{-\infty}^{t} (c(t-s)+d(t-s)) u_x^{-4}(s)\, ds \right. \\
& \qquad + u_x(t) \cdot \int_{-\infty}^{t} a(t-s) u_x^{-2}(s)\, ds \\
& \qquad + \int_{-\infty}^{t} (b(t-s)+d(t-s)) u_x^{-1}(s)\, ds \\
& \qquad - u_x^{-2}(t) \cdot \int_{-\infty}^{t} a(t-s) u_x(s)\, ds \\
& \qquad \left. - u_x^{-3}(t) \cdot \int_{-\infty}^{t} (b(t-s)+c(t-s)+2d(t-s)) u_x^2(s)\, ds \right\} + \phi.
\end{aligned}
$$

In order to obtain the form (3.1), we differentiate this once with respect to time. This yields

$$
\begin{aligned}
\rho u_{ttt} = {} & 3\eta \frac{1}{u_x^2} \cdot u_{xxtt} - 6\eta \frac{u_{xx}}{u_x^3} u_{xtt} \\
& + u_{xxt} \cdot \left\{ -12\eta \frac{u_{xt}}{u_x^3} \quad + 3u_x^2 \cdot \int_{-\infty}^{t} (c(t-s)+d(t-s)) u_x^{-4}(s)\, ds \right. \\
& \qquad + \int_{-\infty}^{t} a(t-s) u_x^{-2}(s)\, ds \\
& \qquad + 2u_x^{-3} \cdot \int_{-\infty}^{t} a(t-s) u_x(s)\, ds \\
& \qquad \left. + 3u_x^{-4}(t) \cdot \int_{-\infty}^{t} (b(t-s)+c(t-s)+2d(t-s)) u_x^2(s)\, ds \right\} + \cdots.
\end{aligned}
$$

Here, as always in the following, the dots indicate terms involving only lower order derivatives of u and the derivatives of ϕ (ϕ always assumed "smooth enough"). We have assumed that the kernels have derivatives in L^1 so that

$$
\frac{d}{dt} \int_{-\infty}^{t} a(t-s) f(s)\, ds = \int_{-\infty}^{t} a'(t-s) f(s)\, ds + a(0) f(t).
$$

The equation above clearly has the form (3.1), and the coefficient of u_{xxt} is positive if the kernels are positive and η is small enough.

b) The model of Kaye [17] and Bernstein, Kearsley and Zapas [1]. In this model, the constitutive law has the form

$$
\pi = \int_{-\infty}^{t} a(t-s, I_1, I_2) \gamma^{-1}(s)\, ds - \int_{-\infty}^{t} b(t-s, I_1, I_2) \gamma^{-1}(t) \gamma(s) \gamma^{-1}(t)\, ds.
$$

I_1 and I_2 are the invariants of $\gamma^{-1}(t)\gamma(s): I_1=\mathrm{tr}(\gamma(t)\gamma^{-1}(s))$ and $I_2=\mathrm{tr}(\gamma^{-1}(t)\gamma(s))$. In our special problem, we thus have $I_1=u_x^2(t)u_x^{-2}(s)+2u_x^{-1}(t)u_x(s)$ and $I_2=u_x^2(s)u_x^{-2}(t)+2u_x^{-1}(s)u_x(t)$. Both are thus functions of the single variable $I=u_x(t)/u_x(s)$, and we shall use the obvious notation $a(t-s,I)$, $b(t-s,I)$. The dynamic equation (4.1) assumes the form

$$\rho u_{tt}=\frac{\partial}{\partial x}\left\{u_x(t)\int_{-\infty}^{t}a(t-s,I)u_x^{-2}(s)\,ds+\int_{-\infty}^{t}b(t-s,I)u_x^{-1}(s)\,ds\right.$$
$$\left.-u_x^{-2}(t)\int_{-\infty}^{t}a(t-s,I)u_x(s)\,ds-u_x^{-3}(t)\cdot\int_{-\infty}^{t}b(t-s,I)u_x^2(s)\,ds\right\}+\phi.$$

By differentiation with respect to time, we find

$$\rho u_{ttt}=u_{xxt}\left\{\int_{-\infty}^{t}a(t-s,I)u_x^{-2}(s)\,ds+2u_x^{-3}(t)\int_{-\infty}^{t}a(t-s,I)u_x(s)\,ds\right.$$
$$+3u_x^{-4}\int_{-\infty}^{t}b(t-s,I)u_x^2(s)\,ds+u_x(t)\int_{-\infty}^{t}\frac{\partial a}{\partial I}(t-s,I)u_x^{-3}(s)\,ds$$
$$+\int_{-\infty}^{t}\frac{\partial b}{\partial I}(t-s,I)u_x^{-2}(s)\,ds-u_x^{-2}(t)\int_{-\infty}^{t}\frac{\partial a}{\partial I}(t-s,I)\,ds$$
$$\left.-u_x^{-3}(t)\int_{-\infty}^{t}\frac{\partial b}{\partial I}(t-s,I)u_x(s)\,ds\right\}+\cdots.$$

A Newtonian term can be added to this as before. Suppose the kernels a, b are positive. Then the coefficient of u_{xxt} is positive in two cases:

α). If $\partial a/\partial I$, $\partial b/\partial I$ are small, i.e., if the model is considered a perturbation of the Ward-Jenkins model.

β) If $\partial a/\partial I_1$, $\partial a/\partial I_2$, $\partial b/\partial I_1$, $\partial b/\partial I_2$ are positive. It would be interesting if this condition has a physical interpretation.

c) The Bird–Carreau model [3], [5]. In this model, we have

$$\pi=\left(1-\frac{\varepsilon}{2}\right)\int_{-\infty}^{t}a(t-s,I(s))\gamma^{-1}(s)\,ds$$
$$-\frac{\varepsilon}{2}\int_{-\infty}^{t}a(t-s,I(s))\gamma^{-1}(t)\gamma(s)\gamma^{-1}(t)\,ds,$$

where $I(s)=\mathrm{tr}(\dot\gamma(s)\gamma^{-1}(s)\dot\gamma(s)\gamma^{-1}(s))=6u_{xt}^2(s)/u_x^2(s)$.

This leads to the equation

$$\rho u_{tt}=\left(1-\frac{\varepsilon}{2}\right)\frac{\partial}{\partial x}\left\{u_x\int_{-\infty}^{t}a(t-s,I(s))u_x^{-2}(s)\,ds\right.$$
$$\left.-u_x^{-2}(t)\int_{-\infty}^{t}a(t-s,I(s))u_x(s)\,ds\right\}$$
$$+\frac{\varepsilon}{2}\frac{\partial}{\partial x}\left\{\int_{-\infty}^{t}a(t-s,I(s))u_x^{-1}(s)\,ds\right.$$
$$\left.-u_x^{-3}(t)\int_{-\infty}^{t}a(t-s,I(s))u_x^2(s)\,ds\right\}+\phi.$$

Differentiating this with respect to time, we find

$$\rho u_{ttt}=u_{xxt}\Bigg[\left(1-\frac{\varepsilon}{2}\right)\left\{\int_{-\infty}^{t}a(t-s,I(s))u_x^{-2}(s)\,ds\right.$$
$$\left.+2u_x^{-3}(t)\int_{-\infty}^{t}a(t-s,I(s))u_x(s)\,ds\right\}$$
$$+\frac{\varepsilon}{2}\cdot 3u_x^{-4}\int_{-\infty}^{t}a(t-s,I(s))u_x^2(s)\,ds\Bigg]$$
$$+\frac{\partial}{\partial t}\Bigg[\left(1-\frac{\varepsilon}{2}\right)\left\{u_x(t)\int_{-\infty}^{t}\frac{\partial a}{\partial I}(t-s,I(s))I_x(s)u_x^{-2}(s)\,ds\right.$$
$$\left.-u_x^{-2}\int_{-\infty}^{t}\frac{\partial a}{\partial I}(t-s,I(s))I_x(s)u_x(s)\,ds\right\}$$
$$+\frac{\varepsilon}{2}\left\{\int_{-\infty}^{t}\frac{\partial a}{\partial I}(t-s,I(s))I_x(s)u_x^{-1}(s)\,ds\right.$$
$$\left.-u_x^{-3}\int_{-\infty s}^{t}\frac{\partial a}{\partial I}(t-s,I(s))I_x(s)u_x^2(s)\,ds\right\}\Bigg]+\cdots.$$

A second differentiation yields (the kernels are assumed to be twice differentiable with respect to time)

$$\rho u_{tttt}=u_{xxtt}\Bigg[\left(1-\frac{\varepsilon}{2}\right)\left\{\int_{-\infty}^{t}a(t-s,I(s))u_x^{-2}(s)\,ds\right.$$
$$\left.+2u_x^{-3}\int_{-\infty}^{t}a(t-s,I(s))u_x(s)\,ds\right\}$$
$$+\frac{\varepsilon}{2}3u_x^{-4}\int_{-\infty}^{t}a(t-s,I(s))u_x^2(s)\,ds\Bigg]+\cdots.$$

For a positive kernel, the coefficient of u_{xxtt} is positive.

d) The Carreau model B [4]. In this model, it is assumed that

$$\pi=\left(1-\frac{\varepsilon}{2}\right)\int_{-\infty}^{t}\exp\left(-\int_s^t f(I(r))\,dr\right)\gamma^{-1}(s)\,ds$$
$$-\frac{\varepsilon}{2}\int_{-\infty}^{t}\exp\left(-\int_s^t f(I(r))\,ds\right)\gamma^{-1}(t)\gamma(s)\gamma^{-1}(t)\,ds,$$

where I has the same meaning as in the Bird–Carreau model. We thus obtain the following equation:

$$\rho u_{tt}=\frac{\partial}{\partial x}\left\{\int_{-\infty}^{t}\exp\left(-\int_s^t f(I(r))\,dr\right)\right.$$
$$\cdot\left[\left(1-\frac{\varepsilon}{2}\right)\left(u_x(t)u_x^{-2}(s)-u_x^{-2}(t)u_x(s)\right)\right.$$
$$\left.\left.+\frac{\varepsilon}{2}\left(u_x^{-1}(s)-u_x^{-3}(t)u_x^2(s)\right)\right]ds\right\}+\phi.$$

The integral converges, if f takes strictly positive values. Differentiating with respect to time, we obtain

$$\begin{aligned}\rho u_{ttt}=u_{xxt}&\left\{\int_{-\infty}^{t}\exp\left(-\int_s^t f(I(r))\,dr\right)\right.\\&\left.\cdot\left[\left(1-\frac{\varepsilon}{2}\right)\left(u_x^{-2}(s)+2u_x^{-3}(t)u_x(s)\right)+\frac{\varepsilon}{2}3u_x^{-4}(t)u_x^2(s)\right]ds\right\}\\&+\frac{\partial}{\partial t}\left\{-\int_{-\infty}^{t}\exp\left(-\int_s^t f(I(r))\,dr\right)\cdot\int_s^t f'(I(r))I_x(r)\,dr\right.\\&\qquad\cdot\left[\left(1-\frac{\varepsilon}{2}\right)\left(u_x(t)u_x^{-2}(s)-u_x^{-2}(t)u_x(s)\right)\right.\\&\qquad\left.\left.+\frac{\varepsilon}{2}\left(u_x^{-1}(s)-u_x^{-3}(t)u_x^2(s)\right)\right]ds\right\}+\cdots.\end{aligned}$$

The second differentiation with respect to time yields

$$\begin{aligned}\rho u_{tttt}=u_{xxtt}&\left\{\int_{-\infty}^{t}\exp\left(-\int_s^t f(I(r))\,dr\right)\right.\\&\cdot\left[\left(1-\frac{\varepsilon}{2}\right)\left(u_x^{-2}(s)+2u_x^{-3}(t)u_x(s)\right)+\frac{\varepsilon}{2}3u_x^{-4}(t)u_x^2(s)\right]ds\\&-12\frac{u_{xt}(t)}{u_x^2(t)}\int_{-\infty}^{t}\exp\left(-\int_s^t f(I(r))\,dr\right)f'(I(t))\\&\qquad\cdot\left[\left(1-\frac{\varepsilon}{2}\right)\left(u_x(t)u_x^{-2}(s)-u_x^{-2}(t)u_x(s)\right)\right.\\&\qquad\left.\left.+\frac{\varepsilon}{2}\left(u_x^{-1}(s)-u_x^{-3}(t)u_x^2(s)\right)\right]ds\right\}+\cdots.\end{aligned}$$

The coefficient of u_{xxtt} is positive under the restriction that $f'(I)\sqrt{I}$ is not too big.

e). The Leonov model [18]. This model does not explicitly give the stress as a functional of the strain history. Instead it is given by a system of equations as follows:

$$\begin{aligned}\pi=\sum_k W_1^{(k)}(I_{1k},I_{2k})c_k^{-1}-W_2^{(k)}(I_{1k},I_{2k})\gamma^{-1}c_k\gamma^{-1}-\eta W(I_{11},I_{21})\frac{\partial}{\partial t}(\gamma^{-1}),\\ \frac{\partial}{\partial t}(c_k^{-1})=-f_k(I_{1k},I_{2k})\left(c_k^{-1}\gamma c_k^{-1}-\frac{1}{3}I_{1k}c_k^{-1}\right)\\ -g_k(I_{1k},I_{2k})\left(\frac{1}{3}I_{2k}c_k^{-1}-\gamma^{-1}\right),\end{aligned}$$

where $I_{1k}=\operatorname{tr}(c_k^{-1}\gamma)$, $I_{2k}=\operatorname{tr}(\gamma^{-1}c_k)$. The tensors c_k satisfy the restriction $\det c_k=1$ (it can be shown that $\det c_k$ is an invariant of the evolution equation). In Leonov's paper, the analogue of c_k^{-1} is called c_k; we have changed this for consistency of notation. The $W_1^{(k)}$, $W_2^{(k)}$, W, f_k, g_k are positive scalar functions, they are not independent in

Leonov's model. It is convenient to introduce $d_k = c_k^{-1}\gamma$. With this, the constitutive equation becomes

$$(4.2)\quad \pi = \sum_k W_1^{(k)}(I_{1k}, I_{2k})d_k\gamma^{-1} - W_2^{(k)}(I_{1k}, I_{2k})d_k^{-1}\gamma^{-1} - \eta W(I_{11}, I_{21})\frac{\partial}{\partial t}(\gamma^{-1}),$$

$$\frac{\partial}{\partial t}(d_k) = -f_k(I_{1k}, I_{2k})\left(d_k^2 - \frac{1}{3}I_{1k}d_k\right)$$

$$-g_k(I_{1k}, I_{2k})\left(\frac{1}{3}I_{2k}d_k - 1\right) + d_k\gamma^{-1}\dot{\gamma}.$$

I_{1k} and I_{2k} are the first and second invariants of d_k, and we have $\det d_k = 1$. If f_k, g_k have positive values, then, for $\dot{\gamma} = 0$, the solution $d_k = \mathrm{id}$ is an exponentially asymptotically stable solution of (4.2) when this equation is restricted to $\{d_k | \det d_k = 1\}$. Consequently, if $\gamma^{-1}\dot{\gamma} \to 0$ as $t \to -\infty$, then on some interval $(-\infty, T)$ there is a unique solution d_k which converges to the identity as $t \to -\infty$. Whether this solution can be continued up to $t = 0$ depends on the form of f_k, g_k and the history of γ. We shall assume that (4.2) has a solution up to $t = 0$. Then this solution is a smooth functional of the histories of γ and $\dot{\gamma}$: $d_k = F_1(\hat{\gamma}, \hat{\dot{\gamma}})$. From (4.2) or, resp., its differentiated version, we also find functional relationships of the form $\dot{d}_k = F_2(\hat{\gamma}, \hat{\dot{\gamma}})$, $\ddot{d}_k = F_3(\hat{\gamma}, \hat{\dot{\gamma}}) + d_k\gamma^{-1}\ddot{\gamma}$. For the filament problem, γ is a diagonal matrix, and so is d_k. Let us denote the 11- and 22-components of d_k by d_{k1}, d_{k2}. The dynamic equation (4.1) reads now as follows

$$\rho u_{tt} = \frac{\partial}{\partial x}\left\{u_x^{-1}\left[\sum_k W_1^{(k)}(I_{1k}, I_{2k})d_{k1} - W_2^{(k)}(I_{1k}, I_{2k})d_{k1}^{-1}\right.\right.$$

$$\left. - W_1^{(k)}(I_{1k}, I_{2k})d_{k2} + W_2^{(k)}(I_{1k}, I_{2k})d_{k2}^{-1}\right]$$

$$\left. + 3\eta W(I_{11}, I_{21})\frac{\partial}{\partial t}\left(-\frac{1}{u_x}\right)\right\} + \phi.$$

Retaining only terms of the highest differentiation orders, we obtain by two-fold differentiation

$$\rho u_{tttt} = 3\eta W(I_{11}, I_{21})\frac{1}{u_x^2}u_{xxttt} + u_{xttt}O(\eta)$$

$$+ u_{xxtt}\left\{O(\eta) + u_x^{-2}\left(W_1^{(k)}d_{k1} + 3W_2^{(k)}d_{k1}^{-1} + 2W_1^{(k)}d_{k2}\right)\right.$$

$$+ u_x^{-2}\left[2W_{11}^{(k)}(d_{k1} - d_{k2})^2\right.$$

$$\left.\left. + 2W_{22}^{(k)}\left(d_{k2}^{-1} - d_{k1}^{-1}\right)^2 + 2\left(W_{12}^{(k)} + W_{21}^{(k)}\right)(d_{k1} - d_{k2})\left(d_{k2}^{-1} - d_{k1}^{-1}\right)\right]\right\}$$

$$+ \cdots .$$

Here $W_{\mathrm{i}}^{(k)}k_j$ stands for $\partial W_i^{(k)}/\partial I_{jk}$. For small η, the coefficient of u_{xxtt} is positive in particular if $W_1^{(k)}$, $W_2^{(k)}$, $W_{11}^{(k)}$, $W_{22}^{(k)}$ are positive and $(W_{12}^{(k)} + W_{21}^{(k)})^2 \leq 4W_{11}^{(k)}W_{22}^{(k)}$. This corresponds to inequality [1.33] in Leonov's paper.

f). The models of Johnson and Segalman [13] and Chang, Bloch and Tschoegl [28]. This model is described by the following system:

$$\begin{aligned}
&\pi = \int_{-\infty}^{t} a(t-s)G(s,t)\gamma^{-1}(s)G^T(s,t)\,ds,\\
&\frac{\partial G}{\partial t} = -\alpha\gamma^{-1}(t)\dot{\gamma}(t)G,\\
&G(t,t) = \mathrm{id},\\
&\frac{\partial G}{\partial s} = \alpha G\gamma^{-1}(s)\dot{\gamma}(s).
\end{aligned}$$

The parameter α ranges between 0 and $\frac{1}{2}$. For our problem, γ is diagonal, whence $\gamma(t')$, $\gamma(t'')$ commute for any t', t''. The equations for G can therefore be solved as follows.

$$G(s,t) = \exp\left(-\alpha\int_{s}^{t}\gamma^{-1}(r)\dot{\gamma}(r)\,dr\right) = \gamma^{\alpha}(s)\gamma^{-\alpha}(t).$$

This leads to an equation very similar to the ones studied in part a), and the discussion follows closely the one given there. We leave the details to the reader. For α in the range $(0,\frac{1}{4})$, the coefficient of u_{xxtt} turns out to be positive. If α is allowed bigger than $\frac{1}{4}$, the type of the equation may change from hyperbolic to elliptic.

The model of Chang, Bloch and Tschoegl is, for this particular problem, equivalent to that of Johnson and Segalman.

g) The model of Curtiss and Bird [8]. This model proposes the following constitutive law.

$$\begin{aligned}
\pi = &\int_{-\infty}^{t}\int_{\Omega_{\gamma(t)}} a(t-s)\left[1+v^T(\gamma(s)-\gamma(t))v\right]^{-3/2}\frac{vv^T}{\sqrt{v^T\gamma^2(t)v}}\,dv\\
&+\eta\int_{-\infty}^{t}\int_{\Omega_{\gamma(t)}} b(t-s)\left[1+v^T(\gamma(s)-\gamma(t))v\right]^{-3/2}v^T\dot{\gamma}(t)v\cdot\frac{vv^T}{\sqrt{v^T\gamma^2(t)v}}\,dv.
\end{aligned}$$

Here $\Omega_{\gamma(t)}$ is the set $\Omega_{\gamma(t)} = \{v|v^T\gamma(t)v=1\}$. With Ω denoting the unit sphere, this yields for our problem

$$\begin{aligned}
\pi^{11} = &\int_{-\infty}^{t}\int_{\Omega} a(t-s)\big[1+\left(u_x^{-2}(t)u_x^2(s)-1\right)w_1^2\\
&\qquad+\left(u_x(t)u_x^{-1}(s)-1\right)\cdot\left(w_2^2+w_3^2\right)\big]^{-3/2}w_1^2u_x^{-2}(t)\,dw\\
&+\eta\int_{-\infty}^{t}\int_{\Omega} b(t-s)\left[1+\left(u_x^{-2}(t)u_x^2(s)-1\right)w_1^2+\left(u_x(t)u_x^{-1}(s)-1\right)\cdot\left(w_2^2+w_3^2\right)\right]^{-3/2}\\
&\qquad\cdot w_1^2u_x^{-3}(t)\dot{u}_x(t)\left(2w_1^2-w_2^2-w_3^2\right)dw.
\end{aligned}$$

For π^{22}, we have the same expression with $w_1^2u_x^{-2}$ replaced by $w_2^2u_x$. When inserting this into the dynamic equation (4.1), we can again achieve the form (3.1) by differentiating with respect to time. The term involving u_{xxtt} has a positive coefficient proportional to η, the coefficient of u_{xxt} is, up to terms of $O(\eta)$.

$$\begin{aligned}&-\int_{-\infty}^{t}\int_{\Omega}a(t-s)[\cdots]^{-3/2}\left(w_1^2-w_2^2\right)dw\,u_x^{-2}(t)\\&\quad+\frac{3}{2}\int_{-\infty}^{t}\int_{\Omega}a(t-s)[\cdots]^{-5/2}\left(-2u_x^{-4}(t)u_x^2(s)w_1^2\right.\\&\qquad\qquad\qquad\left.+u_x^{-1}(t)u_x^{-1}(s)\left(w_2^2+w_3^2\right)\left(w_2^2-w_1^2\right)\right)dw.\end{aligned}$$

It can easily be checked that this coefficient is positive in a neighborhood of the rest state.

Remark. When we differentiated equations with respect to time, we have always assumed that the integral kernels were sufficiently smooth. Some of the kernels suggested in the literature have singularities at $t=0$ (see e.g. [8], where $a(t)=\Sigma_{\alpha\,\mathrm{odd}}e^{-\alpha^2 t}$). A mathematical theory accommodating such kernels would be of interest. Experimental data on polymer melts (see e.g. [28]) also seem to suggest that the integral kernel may be singular at $t=0$. Kernels with singularities are considered in [32], [33].

Acknowledgments. I have greatly profited from the lectures of M. G. Crandall, by which I was introduced to Kato's theory. I also wish to acknowledge several discussions at MRC, which greatly contributed to motivating this research. In particular, I thank R. B. Bird, C. F. Curtiss, J. D. Ferry, M. E. Gurtin, M. W. Johnson, A. S. Lodge, J. Meissner and J. A. Nohel for participating in these discussions.

REFERENCES

[1] B. Bernstein, E. A. Kearsley and L. J. Zapas, *A study of stress relaxation with finite strain*, Trans. Soc. Rheol., 7 (1963), pp. 391–410.

[2] R. B. Bird et al., *Dynamics of Polymeric Liquids*, 2 volumes, John Wiley, New York, 1977.

[3] R. B. Bird and P. J. Carreau, *A nonlinear viscoelastic model for polymer solutions and melts*, Chem. Eng. Sci., 23 (1968), pp. 427–434.

[4] P. J. Carreau, *Rheological equations from molecular network theories*, Trans. Soc. Rheol., 16 (1972), pp. 99–128.

[5] I. J. Chen and D. C. Bogue, *Time-dependent stress in polymer melts and review of viscoelastic theory*, Trans. Soc. Rheol., 16 (1972), pp. 59–78.

[6] B. D. Coleman, M. E. Gurtin and J. Herrera, *Waves in materials with memory*, Arch. Rat. Mech. Anal., 19 (1965), pp. 1–19, 239–298.

[7] B. D. Coleman and W. Noll, *An approximation theorem for functionals, with applications in continuum mechanics*, Arch. Rat. Mech. Anal., 6 (1960), pp. 355–370.

[8] C. F. Curtiss and R. B. Bird, *A kinetic theory for polymer melts*, J. Chem. Phys., 74 (1981), pp. 2016–2033.

[9] C. M. Dafermos and J. A. Nohel, *A nonlinear hyperbolic Volterra equation in viscoelasticity*, Amer. J. Math., (1981), Supplement, pp. 84–116.

[10] M. S. Green and A. V. Tobolsky, *A new approach to the theory of relaxing polymeric media*, J. Chem. Phys., 14 (1946), pp. 80–100.

[11] J. K. Hale, *Theory of Functional Differential Equations*, Springer, Berlin-Heidelberg-New York, 1977.

[12] T. J. R. Hughes, T. Kato and J. E. Marsden, *Well-posed quasi-linear second-order hyperbolic systems with applications to nonlinear elastodynamics and general relativity*, Arch. Rat. Mech. Anal., 63 (1976), pp. 273–284.

[13] M. W. Johnson and D. Segalman, *A model for viscoelastic fluid behaviour which allows non-affine deformation*, J. Non-Newtonian Fluid Mech., 2 (1977), pp. 255–270.

[14] T. Kato, *Linear evolution equations of "hyperbolic" type* II, J. Math. Soc. Japan, 25 (1973), pp. 648–666.

[15] ______, *Quasi-linear equations of evolution with application to partial differential equations*, in Spectral Theory of Differential Equations, W. N. Everitt, ed., Lecture Notes in Mathematics 448, Springer, Berlin-Heidelberg-New York, 1975, pp. 25–70.

[16] ______, *Linear and quasilinear equations of evolution of hyperbolic type*, in: Hyperbolicity, G. da Prato and G. Geymonat, eds. Centro Internazionale Matematico Estivo, II ciclo, Cortona 1976, pp. 125–191.

[17] A. KAYE, Co A Note No. 134, College of Aeronautics, Cranfield, Bletchley, England, 1962.

[18] A. I. LEONOV, *Nonequilibrium thermodynamics and rheology of viscoelastic polymer media*, Rheol. Acta., 15 (1976), pp. 85–98.

[19] A. S. LODGE, *Some finite strain generalizations to Boltzmann's equation*, Proc. 2nd International Congress on Rheology, 1953, V. G. W. Harrison, ed. 229, Butterworth, London.

[20] ______, *A network theory of flow birefringence and stress in concentrated polymer solutions*, Trans. Faraday Soc., 52 (1956), pp. 120–130.

[21] ______, *Constitutive equations from molecular network theories for polymer solutions*, Rheol. Acta, 7 (1968), pp. 379–392.

[22] ______, *Body Tensor Fields in Continuum Mechanics*, Academic Press, New York-San Francisco-London, 1974.

[23] M. RENARDY, *A class of quasilinear parabolic equations with infinite delay and applications to a problem in viscoelasticity*, J. Differential Equations, 48 (1983), pp. 280–292.

[24] J. C. SAUT AND D. D. JOSEPH, *Fading memory*, Arch. Rat. Mech. Anal., 81 (1983), pp. 53–95.

[25] P. E. SOBOLEVSKII, *Equations of parabolic type in a Banach space*, Amer. Math. Soc. Transl. 49 (1966), pp. 1–62.

[26] C. A. TRUESDELL AND W. NOLL, *The Non-Linear Field Theories of Mechanics*, Handbuch der Physik III/3, Springer, Berlin-Heidelberg-New York, 1965.

[27] A. F. K. WARD AND G. M. JENKINS, *Normal thrust in dynamic torsion for rubberlike materials*, Rheol. Acta, 1 (1958), pp. 110–114.

[28] W. V. CHANG, R. BLOCH AND V. W. TSCHOEGL, *On the theory of viscoelastic behaviour of soft polymers in moderately large deformations*, Rheol. Acta, 15 (1976), pp. 367–378.

[29] H. M. LAUN, *Description of the non-linear shear behaviour of a low density polyethylene melt by means of an experimentally determined strain dependent memory function*, Rheol. Acta, 17 (1978), pp. 1–15.

[30] M. SLEMROD, *A hereditary partial differential equation with applications in the theory of simple fluids*, Arch. Rat. Mech. Anal., 62 (1976), pp. 303–321.

[31] E. F. INFANTE AND J. A. WALKER, *A stability investigation for an incompressible simple fluid with fading memory*, Arch. Rat. Mech. Anal., 72 (1980), pp. 203–218.

[32] B. BERNSTEIN AND R. R. HUILGOL, *On ultrasonic dynamic moduli*, Trans. Soc. Rheol., 18 (1974), pp. 583–590.

[33] M. RENARDY, *Some remarks on the propagation and non-propagation of discontinuities in linearly viscoelastic liquids*, Rheol. Acta, 21 (1982), pp. 251–254.

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1502-0009 $01.25/0

POINTWISE BOUNDS FOR STRONGLY COUPLED TIME DEPENDENT SYSTEMS OF REACTION-DIFFUSION EQUATIONS*

CHRIS COSNER†

Abstract. Sufficient conditions are given for the pointwise boundedness and decay of solutions to time dependent strongly coupled systems of reaction-diffusion equations on spatially bounded domains. The results are obtained via Lyapunov or energy methods.

AMS-MOS subject classification (1980). Primary 35B35, 35K60; secondary 35B40, 35B45

1. Introduction. Because they describe many physical situations, systems of reaction-diffusion equations have been widely studied. A common topic of investigation is the behavior of solutions as the time variable tends to infinity. Many results in that direction have been obtained via the maximum principle and its generalizations. However, maximum principles or invariant set theorems such as those in [4], [7], [8], [9] generally fail if the system is coupled in the highest order terms. Such systems occur in some physical problems; see [3]. In [5], the author presented another method for obtaining pointwise bounds for solutions, but only in the case of time independent coefficients. That restriction on the coefficients limits the applicability of the results.

The object of this paper is to give sufficient conditions for the pointwise boundedness and decay of solutions to strongly coupled systems of semilinear parabolic equations of the form

$$\begin{aligned} &\mathbf{u}_t=\left(a_{ij}^{\mu\nu}(x,t)u_{x_i}^{\nu}\right)_{x_j}+b_i^{\mu\nu}(x,t)u_{x_i}^{\nu}+f^{\mu}(x,t,\mathbf{u}) \quad \text{in } \Omega\times(0,\infty),\\ &\mathbf{u}(x,t)=0 \quad \text{on } \partial\Omega\times(0,\infty),\\ &\mathbf{u}(x,0)=\mathbf{u}_0(x) \quad \text{on } \Omega, \end{aligned} \tag{1.1}$$

where $\mu=1,\cdots,N$, repeated Greek indices are summed from 1 to N, and repeated Roman indices from 1 to n. (This convention is used throughout the paper. We will assume $n\geq 2$.) The domain $\Omega\subseteq\mathbb{R}^n$ is assumed to be bounded, with $\partial\Omega$ of class C^2. The results presented here generalize those of [3] by allowing time dependent coefficients in (1.1) and by weakening the structure conditions on the nonlinearity. Another generalization of the results in [5], requiring time independent coefficients but allowing nonsymmetric coupling in the second order terms, is also derived. Note that the strong coupling in (1.1) generally precludes obtaining pointwise bounds on $\mathbf{u}$ via maximum principle or invariant set arguments; see the discussion in [5]. Since the main results of [5] are based on a Lyapunov or energy method, it is somewhat surprising that they generalize to the time dependent case. See [5] for further discussion and references.

2. Notation. We will assume that $a_{ij}^{\mu\nu}(x,t)$ is C^3 in x, $b_i^{\mu\nu}(x,t)$ is C^2 in x, $f^{\mu}(x,t,\mathbf{u})$ is C^1 in x and $\mathbf{u}$, and all these terms are C^1 in t, with $a_{ij}^{\mu\nu}$, $b_i^{\mu\nu}$, and their first derivatives with respect to t uniformly bounded. The coefficients $a_{ij}^{\mu\nu}$ are assumed to satisfy the symmetry conditions $a_{ij}^{\mu\nu}=a_{ij}^{\nu\mu}$ and $a_{ij}^{\mu\nu}=a_{ji}^{\mu\nu}$. Finally, the following condition is assumed to hold: for some constant $c_0>0$,

$$a_{ij}^{\mu\nu}(x,t)q_i^{\mu}q_j^{\nu}\geq c_0q_i^{\mu}q_i^{\mu} \tag{2.1}$$

*Received by the editors November 2, 1981.

†Mathematics Department, University of Miami, Coral Gables, Florida 33124

for any $\mathbf{q}\in\mathbb{R}^{nN}$. Condition (2.1) implies that the differential operator on the right side of (1.1) is elliptic. The system (1.1) may be rewritten in the form

$$\mathbf{u}_t=A(t)\mathbf{u}+\mathbf{f}(x,t,\mathbf{u}),\qquad \mathbf{u}(0)=\mathbf{u}_0, \tag{2.2}$$

where $A(t)$ represents the differential operator on the right side of (1.1). Equation (2.2) may be regarded as an ordinary differential equation in L^p if $A(t)$ is given the domain $W^{2,p}\cap W_0^{1,p}$, where $W^{k,p}$ is the usual L^p-Sobolev space of functions with k weak derivatives in L^p, and $W_0^{k,p}$ is the completion in $W^{k,p}$ of $C_0^\infty(\Omega)$. Let $\|\mathbf{u}\|_{k,p}$ denote the norm of $\mathbf{u}$ in $W^{k,p}$. The a priori estimates of Agmon, Douglis, and Nirenberg [2] imply that for each $p>1$, there is a constant $C_1(p)$ so that $\|\mathbf{u}\|_{2,p}\leq C_1(p)(\|A(t)\mathbf{u}\|_{0,p}+\|\mathbf{u}\|_{0,p})$ for $\mathbf{u}\in W^{2,p}\cap W_0^{1,p}$. It follows that the operators $A(t)$ are closed. Further, for each p, there exist constants $K(p)$ and $\varepsilon(p)>0$ such that if $\tilde{A}_p(t)=A(t)-k(p)$, the operators $(\tilde{A}_p(t)-\lambda)^{-1}$ exist as bounded operators on L^p and satisfy $\|(\tilde{A}_p(t)-\lambda)^{-1}\|\leq C_2(p)/(1+|\lambda|)$ for $\lambda\in\mathbb{C}$ with $-\pi/2-\varepsilon(p)<\arg\lambda<\pi/2+\varepsilon(p)$. The analysis is much the same as in the case of a single equation as discussed in [6]. It follows that the operators $\tilde{A}_p(t)$ generate analytic semigroups; hence fractional powers of these operators exist. The a priori estimates of [2] combined with standard interpolation theorems for Sobolev spaces imply that for any $\alpha\in(\frac{1}{2},1)$ there exists a constant $C_3(p)$ so that the fractional power $\tilde{A}_p^\alpha(t)$ of $\tilde{A}_p(t)$ satisfies

$$\|\mathbf{u}\|_{1,p}\leq C_3(p)\left\|\tilde{A}_p^\alpha(t)\mathbf{u}\right\|_{0,p} \tag{2.3}$$

for $\mathbf{u}\in W^{2,p}\cap W_0^{1,p}$. Since the coefficients of $\tilde{A}_p(t)$ are smooth in t, $\tilde{A}_p(t)$ generates an operator valued fundamental solution $U_p(s,t)$. Let $\mathbf{g}_p(x,t,\mathbf{u})=\mathbf{f}(x,t,\mathbf{u})+k(p)\mathbf{u}$. Then (2.2) may be rewritten as

$$\mathbf{u}_t=\tilde{A}_p(t)\mathbf{u}+\mathbf{g}_p(t,\mathbf{u}(t)),\qquad \mathbf{u}(0)=\mathbf{u}_0. \tag{2.4}$$

Solutions to (4) in L^p may be represented in the form

$$\mathbf{u}(t)=U_p(t,0)\mathbf{u}_0+\int_0^t U_p(t,s)\mathbf{g}_p(s,\mathbf{u}(s))\,ds. \tag{2.5}$$

Since the operators $\tilde{A}_p(t)$ generate analytic semigroups, $U(s,t)$ satisfies various estimates which will be used later.

3. Analysis. The boundedness conditions on the coefficients of $A(t)$ imply that there exists a finite value $a_0=\inf\{a\colon (\frac{1}{2})|a_{ijt}^{\mu\nu}(x,t)q_i^\mu q_j^\nu|\leq a|q|^2$ for all $\mathbf{q}\in\mathbb{R}^{nN}$, $(x,t)\in\bar{\Omega}\times[0,\infty)\}$. Similarly, let $B^{\mu\nu}=\sup_{x,t}[\Sigma_{i=1}^n(b_i^{\mu\nu})^2]^{1/2}$ and let $b_0=\inf\{b\colon |B^{\mu\nu}u^\mu v^\nu|\leq b|\mathbf{u}|\,|\mathbf{w}|$ for all $\mathbf{u},\mathbf{w}\in\mathbb{R}^N\}$. Assume that $\mathbf{u}$ is a solution of (1) in L^p for all $p\in(1,p_0]$ where $p_0>n$. (Any classical solution has that property.) We have the following:

THEOREM 1. *Suppose that* $\mathbf{u}_0\in W_0^{2,p_0}\cap W^{1,p_0}$ *for some* $p_0>n$ *and that* $\mathbf{f}$ *satisfies the bounds*

$$\mathbf{u}\cdot\mathbf{f}(x,t,\mathbf{u})\leq-\beta_0|\mathbf{u}|^2\quad \textit{for } (x,t)\in\Omega\times(0,\infty) \textit{ and } \mathbf{u}\in\mathbb{R}^N \tag{3.1}$$

and

$$|\mathbf{f}(x,t,\mathbf{u})|\leq f_0(1+|\mathbf{u}|^r), \tag{3.2}$$

where β_0 and f_0 are positive constants, and $r\in[1,\infty)$ for $n=2$, $r\in[1,r_0)$ with $r_0=n/(n-2)$ for $n\geq 3$. Suppose also that the constants b_0, c_0 and β_0 satisfy

$$4\beta_0 c_0 > b_0^2. \tag{3.3}$$

Finally, suppose that there exists a function $G(x,\mathbf{u})$ such that for some positive constants G_0 and σ,

$$G(x,\mathbf{u})\geq -G_0|\mathbf{u}|^2 \quad and \quad \mathbf{u}\cdot\mathbf{f}+\sigma|\mathbf{f}+\nabla_{\mathbf{u}}G|^2\leq 0. \tag{3.4}$$

Then $\sup_{x\in\Omega}|\mathbf{u}|$ is uniformly bounded in terms of $\mathbf{u}_0$. Suppose in addition to the above hypotheses that $\mathbf{f}(x,t,\mathbf{u})$ and $G(x,\mathbf{u})$ are such that for any $R>0$ there exist constants $f_1(R)$, $G_1(R)$ so that

$$|\mathbf{f}(x,t,\mathbf{u})|\leq f_1(R)|\mathbf{u}|, \qquad G(x,\mathbf{u})\leq G_1(R)|\mathbf{u}|^2 \tag{3.5}$$

for all $\mathbf{u}\in\mathbb{R}^N$ with $|\mathbf{u}|<R$. Then $\sup_{x\in\Omega}|\mathbf{u}|$ decays exponentially as $t\to\infty$.

Proof. The proof is similar to that for the case of time independent coefficients discussed in [3]; for more details refer to that article. Let

$$E_0(t)=\int_\Omega\left(u^\mu_{x_i}u^\mu_{x_i}+u^\mu u^\mu\right)dx=\|\mathbf{u}\|^2_{1,2}$$

and

$$E_1(t)=\int_\Omega\left[\frac{1}{2}a^{\mu\nu}_{ij}(x,t)u^\mu_{x_j}u^\nu_{x_i}+G(x,\mathbf{u})+\frac{1}{2}K\mathbf{u}\cdot\mathbf{u}\right]dx.$$

Differentiating $E_1(t)$ with respect to t and integrating by parts yields

$$\begin{aligned}E_1'(t)&=\int_\Omega\left[-\left(a^{\mu\nu}_{ij}u^\nu_{x_i}\right)_{x_j}u^\mu_t+\frac{1}{2}a^{\mu\nu}_{ijt}u^\mu_{x_j}u^\nu_{x_i}+\nabla_{\mathbf{u}}G\mathbf{u}_t+K\mathbf{u}\cdot\mathbf{u}_t\right]dx\\ &=\int_\Omega\left[-|\mathbf{u}_t|^2+b^{\mu\nu}_i u^\nu_{x_i}u^\mu_t+(\mathbf{f}+\nabla_{\mathbf{u}}G)\cdot\mathbf{u}_t+\frac{1}{2}a^{\mu\nu}_{ij}u^\mu_{x_j}u^\nu_{x_i}+K\mathbf{u}\cdot\mathbf{u}_t\right]dx.\end{aligned}$$

Using the bounds on $a^{\mu\nu}_{ijt}$ and $b^{\mu\nu}_i$ and Cauchy's inequality yields

$$E_1'(t)\leq\int_\Omega\left[-\left(1-\frac{b_0\varepsilon_1}{2}\right)|\mathbf{u}_t|^2+(\mathbf{f}+\nabla_{\mathbf{u}}G)\cdot\mathbf{u}_t+\left(\frac{b_0}{2\varepsilon_1}+a_0\right)u^\nu_{x_i}u^\nu_{x_i}+K\mathbf{u}\cdot\mathbf{u}_t\right]dx$$

for any $\varepsilon_1>0$. Choose ε_1 such that $0<d(\varepsilon_1)\equiv 1-b_0\varepsilon_1/2<1$, then complete the square to obtain

$$E_1'(t)\leq\int_\Omega\left[-d(\varepsilon_1)\left|\mathbf{u}_t-\frac{(\mathbf{f}+\nabla_{\mathbf{u}}G)}{2d(\varepsilon_1)}\right|^2+\frac{|\mathbf{f}+\nabla_{\mathbf{u}}G|^2}{4d(\varepsilon_1)}+\left(\frac{b_0}{2\varepsilon_1}+a_0\right)u^\nu_{x_i}u^\nu_{x_i}+K\mathbf{u}\cdot\mathbf{u}_t\right]dx.$$

Dropping the first term, rewriting $\mathbf{u}_t$ via (1.1) in the last term and integrating by parts,

$$\begin{aligned}E_1'(t)\leq\int_\Omega\Bigg[&\frac{|\mathbf{f}+\nabla_{\mathbf{u}}G|^2}{4d(\varepsilon_1)}+\left(\frac{b_0}{2\varepsilon_1}+a_0\right)u^\nu_{x_i}u^\nu_{x_i}\\ &-Ka^{\mu\nu}_{ij}u^\nu_{x_j}u^\nu_{x_i}+Kb^{\mu\nu}_i u^\nu_{x_i}u^\mu+K\mathbf{u}\cdot\mathbf{f}\Bigg]dx.\end{aligned}$$

Using (2), the bounds for $b_i^{\mu\nu}$, and Cauchy's inequality yields

$$E_1'(t)\le K\int_\Omega\left\{\left[-c_0+\frac{b_0\varepsilon_2}{2}+\frac{1}{K}\left(a_0+\frac{b_0}{2\varepsilon_1}\right)\right]u^\nu_{x_i}u^\nu_{x_i}\right.$$
$$\left.+\frac{b_0}{2\varepsilon_2}|\mathbf{u}|^2+\mathbf{u}\cdot\mathbf{f}+\frac{|\mathbf{f}+\nabla_\mathbf{u}G|^2}{4Kd(\varepsilon_1)}\right\}dx \tag{3.6}$$

for any $\varepsilon_2>0$. By (3.3) we may choose $\beta\in(0,\beta_0)$ such that

$$4\beta c_0>b_0^2. \tag{3.7}$$

Let $\varepsilon_3=1-\beta/\beta_0$; then by (3.1),

$$\mathbf{u}\cdot\mathbf{f}\le-\beta|\mathbf{u}|^2+\varepsilon_3\mathbf{u}\cdot\mathbf{f}. \tag{3.8}$$

Let $\varepsilon_2=[-\beta+c_0+\sqrt{(\beta-c_0)^2+b_0^2}\,]/b_0>0$; then $-c_0+b_0\varepsilon_2/2=-\beta+b_0/2\varepsilon_2=-D_1$ where $D_1=[\beta+c_0-\sqrt{(\beta-c_0)^2+b_0^2}\,]/2$. By (3.7), $D_1>0$. Using (3.8) and substituting ε_2 into (3.6), we have

$$E_1'(t)\le K\int_\Omega\left\{\left[-D_1+\frac{1}{K}\left(a_0+\frac{b_0}{2\varepsilon_1}\right)\right]u^\nu_{x_i}u^\nu_{x_i}-D_1|\mathbf{u}|^2+\varepsilon_3\mathbf{u}\cdot\mathbf{f}+\frac{|\mathbf{f}+\nabla_\mathbf{u}G|^2}{4Kd(\varepsilon_1)}\right\}dx.$$

By (3.4) it follows that for any $D\in(0,D_1)$ we may choose K large enough that

$$E_1'(t)\le-KD\int_\Omega\left(u^\nu_{x_i}u^\nu_{x_i}+|\mathbf{u}|^2\right)dx=-KDE_0(t)\le0. \tag{3.9}$$

Thus, $E_1(t)\le E_1(0)$. If we assume $K>G_0+c_0$, then $c_0E_0(t)\le E_1(t)$ by (3.4), so that

$$\|\mathbf{u}\|_{1,2}=[E_0(t)]^{1/2}\le B(u_0)\equiv\left[\frac{1}{c_0}E_1(0)\right]^{1/2}. \tag{3.10}$$

Hence $\mathbf{u}$ is uniformly bounded in $W^{1,2}$. Note that since $p_0>n$, $|\mathbf{u}_0|$ and $|\nabla_x\mathbf{u}_0|$ are bounded by $C\|\mathbf{u}_0\|_{2,p_0}$; hence $B(u_0)$ can be bounded in terms of $\|\mathbf{u}_0\|_{2,p_0}$, independent of the form of $\mathbf{u}_0$. From (3.10) it follows by the Sobolev embedding theorem that $\|\mathbf{u}_0\|_{0,p_1}\le CB(u_0)$ for $p_1=2n/(n-2)$ if $n\ge3$ and any $p_1>2$ if $n=2$. (For a discussion of Sobolev embedding theorems see for example [1].) Consider the case $n\ge3$. It follows from (3.2) that for any fixed k, $\|k\mathbf{u}+\mathbf{f}(x,t,\mathbf{u})\|_{0,p_1/r}\le C(1+B(u_0))^r$. Let $q_1=p_1/r$. Construct $\tilde A_{q_1}(t)$ and $\mathbf{g}_{q_1}(x,t,\mathbf{u})$ as in (2.4); then the representation (2.5) holds with $p=q_1$. Choose α so that (2.3) holds. The following estimates are standard (see [6, §§ II-13–II-16]):

$$\|\tilde A_q^\alpha(t)U_q(t,s)\|\le\|C(t-s)^{-\alpha}e^{-\gamma(t-s)}\|, \tag{3.11}$$
$$\|\tilde A_q^\alpha(t)U_q(t,0)\tilde A_q^{-\alpha}\|\le C,$$

where $\gamma>0$. By (2.5) we may write

$$\tilde A^\alpha_{q_1}(t)\mathbf{u}=\tilde A^\alpha_{q_1}(t)U(t,0)\mathbf{u}_0+\int_0^t\tilde A^\alpha_{q_1}(t)U(t,s)\mathbf{g}_{q_1}(s,\mathbf{u}(s))\,ds. \tag{3.12}$$

The first term on the right side of (3.11) may be written as$\tilde{A}^{\alpha}_{q_1}(t)U(t,0)\tilde{A}^{-\alpha}_{q_1}(t)(A^{\alpha}_{q_1}(t)\mathbf{u}_0)$. Since $\|\mathbf{g}_{q_1}(s,\mathbf{u}(s))\|_{0,q_1}=\|k_{q_1}\mathbf{u}+\mathbf{f}(x,s,\mathbf{u})\|_{0,q_1}\le C(1+B(u_0))^r$, it follows from (2.3), (3.11) and (3.12) that

$$\|\mathbf{u}\|_{1,q_1}\le C\left\|\tilde{A}^{\alpha}_{q_1}(t)\mathbf{u}_0\right\|_{0,q_1}\le C(\mathbf{u}_0). \tag{3.13}$$

If $q_1>n$ then the Sobolev embedding theorem applied to (3.13) yields $\sup_{x\in\Omega}|\mathbf{u}|\le C(\mathbf{u}_0)$. If $q_1=p_1/r<n$, then the Sobolev embedding theorem implies $\|\mathbf{u}\|_{0,p_2}\le C\|\mathbf{u}\|_{1,q_1}\le C(\mathbf{u}_0)$ with $p_2=nq_1/(n-q_1)=(p_1/r)/(1-p_1/nr)$. Since $n(r-1)<n(r_0-1)=p_1$, there exists $\varepsilon>0$ so that $p_1\ge n[(1+\varepsilon)r-1]/(1+\varepsilon)$; thus $1-p_1/nr\le 1/(1+\varepsilon)r$ so $p_2\ge(1+\varepsilon)p_1$. Let $q_2=p_2/r$; then $q_2\ge(1+\varepsilon)q_1$. We may return to (3.11), replacing q_1 and q_2, and repeat the analysis leading to (3.13). This process may be repeated until for some l, $q_l>n$. At that step, we obtain the bound $\sup_{x\in\Omega}|\mathbf{u}|\le C\|\mathbf{u}\|_{1,q_l}\le C(\mathbf{u}_0)$, which is the first conclusion of the theorem. Note that if $\mathbf{u}_0$ is replaced by $\mathbf{v}_0$ such that $\|\mathbf{v}_0\|_{2,p}\le\|\mathbf{u}_0\|_{2,p}$ for $1\le p\le p_0$ and $E_1(0)$ is no larger for $\mathbf{v}_0$ than for $\mathbf{u}_0$, then $C(\mathbf{u}_0)\ge C(\mathbf{v}_0)$. Since $\sup_\Omega|\mathbf{u}|\le R$, hypothesis (3.5) may apply. If so, then for some constant $C_1(R)$, $E_1(t)\le C_1(R)E_0(t)$. Thus (3.10) yields $E_1'(t)\le-(KD/C_1(R))E_1(t)$; thus $E_1(t)$ and hence $\|\mathbf{u}\|_{1,2}$ must decay exponentially. We have $\|\mathbf{u}\|_{1,2}\le C\|\mathbf{u}_0\|_{1,2}e^{-\delta_0 t}$ for some $\delta_0>0$. The process beginning with (3.10) may now be repeated; in fact, by (3.5) we may choose $r=1$. Since $\|\mathbf{u}\|_{1,2}$ decays exponentially, so does $\|\mathbf{u}\|_{0,p_1}$ and hence also $\|\mathbf{g}_{p_1}(s,\mathbf{u}(s))\|_{0,p_1}$. Combining that fact with the bounds of (3.11) yields, analogously with (3.13), that for some $\delta_1>0$, $\|\mathbf{u}\|_{1,p_1}\le C(\mathbf{u}_0)e^{-\delta_1 t}$. Continued iteration leads eventually to the bound $\sup_{x\in\Omega}|\mathbf{u}|\le C\|\mathbf{u}\|_{1,p}\le C(\mathbf{u}_0)e^{-\delta t}$ for some $\delta>0$, where $p>n$. The constant δ depends on δ_0 and the constants γ occurring in (3.11) for the values of q in the iteration. An examination of the way $C(\mathbf{u}_0)$ is obtained shows that $C(\mathbf{u}_0)\to 0$ as $\sup_{1<p\le p_0}\|\mathbf{u}_0\|_{2,p}\to 0$. If $n=2$ then the proof is essentially the same, except that the correct choice of p_1 makes it possible to obtain a bound on $\sup_{x\in\Omega}|\mathbf{u}|$ in one step.

Remarks. In [3], **f** was assumed to be time independent and to have the structure

$$\mathbf{f}(x,\mathbf{u})=M(x)\mathbf{u}+\nabla_{\mathbf{u}}H(x,\mathbf{u}), \tag{3.14}$$

where $M(x)$ is a matrix. In such a case we may choose $G(x,\mathbf{u})=-H(x,\mathbf{u})$; then the second inequality of (3.4) holds provided that (3.1) is satisfied. However, more general forms of the function f can be treated.

Example. Let $\mathbf{u}=(u^1,u^2)$, let D be a positive definite symmetric 2×2 matrix, and let $\mathbf{f}=(f^1(t,\mathbf{u}),f^2(t,\mathbf{u}))$ with $f^1(t,\mathbf{u})=-(2+\cos t)u^1-(u^1)^3+u^1u^2$ and $f^2(t,\mathbf{u})=-(2+\cos t)u^2-(u^2)^3+u^1u^2$. Then Theorem 1 applies to the system $\mathbf{u}_t=(2+\sin t)D\Delta\mathbf{u}+\mathbf{f}(t,\mathbf{u})$ with the choice $G(\mathbf{u})=[(u^1)^4+(u^2)^4]/4$. The function **f** does not satisfy the structure condition (3.14). A key step in the proof of Theorem 1 is the construction of $E_1(t)$; it seems likely that such a construction may be possible in cases not covered by Theorem 1. In particular, the function G may be allowed to depend on t if the proof and hypotheses of Theorem 1 are modified somewhat.

We will now consider the case of nonsymmetric coupling in the second order terms. For simplicity we will assume that the coefficients of the system are time dependent and that the nonlinearity has the structure given in (3.14) where $M(x)$ is a matrix of bounded functions. Without loss of generality we may assume that the matrix coupling the second order terms has been decomposed into symmetric and antisymmetric parts. Thus we consider

$$u^{\mu}_t=\left(a^{\mu\nu}_{ij}(x)u^{\nu}_{x_i}\right)_{x_j}+\left(\alpha^{\mu\nu}_{ij}(x)u^{\nu}_{x_i}\right)_{x_j}+b^{\mu\nu}_i(x)u^{\nu}_{x_i}+f^{\mu}(x,\mathbf{u}). \tag{3.15}$$

All the previous assumptions on coefficients $a_{ij}^{\mu\nu}(x)$, $b_i^{\mu\nu}(x)$ and the nonlinearity remain in force. Also, we require that $\alpha_{ij}^{\mu\nu}(x)$ be bounded for all i,j,μ,ν and that $\alpha_{ij}^{\mu\nu}(x)=\alpha_{ji}^{\mu\nu}(x)$, but $\alpha_{ij}^{\mu\nu}(x)=-\alpha_{ij}^{\nu\mu}(x)$. It follows that

$$\alpha_{ij}^{\mu\nu}(x)p_i^\mu p_j^\nu=0. \tag{3.16}$$

The boundedness of the terms $\alpha_{ij}^{\mu\nu}(x)$ and the entries in $M(x)$ guarantees the existence of constants m_0 and α_0 such that $\alpha_{ij}^{\mu\nu}(x)p_iq_j\le\alpha_0|\mathbf{p}||\mathbf{q}|$ and $M(x)\mathbf{u}\cdot\mathbf{w}\le m_0|\mathbf{u}||\mathbf{w}|$. We will see the notation $\partial f^\mu/\partial u^\nu=f_\nu^\mu$ and assume that

$$f_\nu^\mu u^\nu u^\mu\le f_0|\mathbf{u}|^2. \tag{3.17}$$

THEOREM 2. *Suppose that* $\mathbf{u}_0\in W^{2,p_0}\cap W_0^{1,p_0}$ *for some* $p_0>n$ *and* $\mathbf{u}(x,t)$ *satisfies* (3.15) *with* $\mathbf{u}(x,0)=\mathbf{u}_0$. *Suppose that hypotheses* (3.1), (3.2), (3.3), (3.14) *and* (3.17) *are satisfied, and* $H(x,\mathbf{u})\ge-H_0|\mathbf{u}|^2$. *Then* $\sup_{x\in\Omega}|\mathbf{u}|$ *is uniformly bounded in terms of* $\mathbf{u}_0$. *If in addition* $|\mathbf{f}(x,\mathbf{u})|\le f_1(R)|\mathbf{u}|$ *and* $H(x,\mathbf{u})\le H_1|\mathbf{u}|^2$ *when* $|\mathbf{u}|<R$ *then* $\sup_{x\in\Omega}|\mathbf{u}|$ *decays exponentially as* $t\to\infty$.

Proof. Let

$$E_2(t)=\int_\Omega\left[\frac{1}{2}a_{ij}^{\mu\nu}(x)u_{x_j}^\mu u_{x_i}^\nu+H(x,\mathbf{u})+\frac{1}{2}K_1|\mathbf{u}_t|^2+\frac{1}{2}K_2|\mathbf{u}|^2dx\right].$$

Then we havc

$$\begin{aligned}E_2'(t)=\int_\Omega\Big\{&-u_t^\mu\Big[\big(a_{ij}^{\mu\nu}u_{x_i}^\nu\big)_{x_j}+\big(\alpha_{ij}^{\mu\nu}u_{x_i}^\mu\big)_{x_j}+b_i^{\mu\nu}u_{x_i}^\nu+f^\mu(x,\mathbf{u})\Big]\\&-\alpha_{ij}^{\mu\nu}u_{x_i}^\nu u_{x_jt}^\mu+b_i^{\mu\nu}u_{x_i}^\nu u_t^\mu+M\mathbf{u}\cdot\mathbf{u}_t+K_1\mathbf{u}_t\cdot\mathbf{u}_{tt}+K_2\mathbf{u}\cdot\mathbf{u}_t\Big\}\,dx\\=\int_\Omega\Big\{&-|\mathbf{u}_t|^2-\alpha_{ij}^{\mu\nu}u_{x_i}^\nu u_{x_jt}^\mu+b_i^{\mu\nu}u_{x_i}^\nu u_t^\mu+M\mathbf{u}\cdot\mathbf{u}_t\\&-K_1a_{ij}^{\mu\nu}u_{x_jt}^\nu u_{x_it}^\mu-K_1\alpha_{ij}^{\mu\nu}u_{x_jt}^\mu u_{x_it}^\nu+K_1b_i^{\mu\nu}u_t^\mu u_{x_it}^\nu+K_1f_\nu^\mu u_t^\nu u_t^\mu\\&-K_2a_{ij}^{\mu\nu}u_{x_j}^\mu u_{x_i}^\nu-K_2\alpha_{ij}^{\mu\nu}u_{x_j}^\mu u_{x_i}^\nu+K_2b_i^{\mu\nu}u_{x_i}^\nu u^\mu+K_2\mathbf{u}\cdot\mathbf{f}\Big\}\,dx.\end{aligned}$$

Using (2.1), (3.16), (3.1) and the bounds on the coefficients of the system yields

$$\begin{aligned}E_2'(t)\le\int_\Omega\Big\{&-|\mathbf{u}_t|^2-\alpha_{ij}^{\mu\nu}u_{x_i}^\nu u_{x_jt}^\mu+b_i^{\mu\nu}u_{x_i}^\nu u_t^\mu+m_0|\mathbf{u}||\mathbf{u}_t|\\&-K_1c_0u_{x_jt}^\mu u_{x_jt}^\mu+K_1b_i^{\mu\nu}u_t^\mu u_{x_it}^\nu\\&+K_1f_0|\mathbf{u}_t|^2-K_2c_0u_{x_i}^\mu u_{x_i}^\mu+K_2b_i^{\mu\nu}u_{x_i}^\nu u^\mu-\beta_0|\mathbf{u}|^2\Big\}\,dx.\end{aligned}$$

By Cauchy's inequality and the bounds on $\alpha_{ij}^{\mu\nu}$ and $b_i^{\mu\nu}$ we have

$$\begin{aligned}E_2'(t)\le\int_\Omega\Big\{&\Big[-1+\frac{b_0\varepsilon_2}{2}+\frac{m_0\varepsilon_3}{2}+K_1\Big(\frac{b_0}{2\varepsilon_4}\Big)+K_1f_0\Big]|\mathbf{u}_t|^2\\&+\Big[\frac{\alpha_0\varepsilon_1}{2}+K_1\Big(\frac{b_0\varepsilon_4}{2}-c_0\Big)\Big]u_{x_jt}^\mu u_{x_jt}^\mu\\&+\Big[\frac{\alpha_0}{2\varepsilon_1}+\frac{b_0}{2\varepsilon_2}+K_2\Big(\frac{b_0\varepsilon_5}{2}-c_0\Big)\Big]u_{x_i}^\mu u_{x_i}^\mu+\Big[\frac{m_0}{2\varepsilon_3}+K_2\Big(\Big(\frac{b_0}{2\varepsilon_5}\Big)-\beta\Big)\Big]|\mathbf{u}|^2\Big\}\,dx.\end{aligned}$$

If we choose ε_4 so that $(b_0\varepsilon_4/2-c_0)<0$, K_1 so that $K_1((b_0/2\varepsilon_4)+f_0)<1$, ε_2 and ε_3 so that the coefficient of $|\mathbf{u}_t|^2$ is negative, ε_1 so that the coefficient of $u^\mu_{x_jt}u^\mu_{x_jt}$ is nonpositive, $\varepsilon_5=[-\beta_0+c_0+\sqrt{(\beta_0-c_0)^2+b_0^2}]/b_0>0$, and K_2 sufficiently large, then condition (3.1) implies that the coefficients of $u^\mu_{x_i}u^\mu_{x_i}$ and $|\mathbf{u}|^2$ are negative. Hence we have

$$E_2'(t)\leq -K\int_\Omega\left[|\mathbf{u}_t|^2+u^\mu_{x_i}u^\mu_{x_i}+|\mathbf{u}|^2\right]dx\leq 0$$

for some positive K. From this point the proof is essentially identical to that of Theorem 1, or the results of [5].

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] S. Agmon, A. Douglis and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* II, Comm. Pure Appl. Math., 17 (1964), pp. 35–92.

[3] J. Bell and C. Cosner, *Stability properties of a model of parallel nerves*, J. Differential Equations, 40 (1981), pp. 303–315.

[4] K. Chueh, C. Conley and J. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373–392.

[5] C. Cosner, *Pointwise a priori bounds for strongly coupled semi-linear systems of parabolic partial differential equations*, Indiana Univ. Math. J., 30 (1981), pp. 607–619.

[6] A. Friedman, *Partial Differential Equations*, Krieger, Huntingdon, NY, 1976.

[7] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Prentice Hall, Englewood Cliffs, NJ.

[8] R. Redheffer and W. Walter, *Invariant sets for systems of partial differential equations* I. *Parabolic equations*, Arch. Rational Mech. Anal., 67 (1977), pp. 41–52.

[9] W. Walter, *Differential and Integral Inequalities*, Springer-Verlag, New York, 1970.

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1502-0010 $01.25/0

NONEXISTENCE OF GLOBAL SOLUTIONS TO THE CAUCHY PROBLEM FOR THE DAMPED NONLINEAR SCHRÖDINGER EQUATIONS*

MASAYOSHI TSUTSUMI†

Abstract. Solutions to the Cauchy problem for the equation $iu_t=\Delta u+q(|u|^2)u-(ia/2)u$ ($a>0$, $x\in R^n$, $t>0$); $u(x,0)=u_0(x)$ are considered. Conditions on u_0 and q are given so that solutions do not exist for all $t>0$.

1. Introduction. Consider the Cauchy problem for the equation

$$(1)\qquad iu_t=\Delta u+|u|^{p-1}u-\frac{ia}{2}u,\qquad i=\sqrt{-1},\quad x\in R^n,\quad t>0$$

with

$$(2)\qquad u(x,0)=u_0(x),\qquad x\in R^n,$$

where $p>1$, $a\geq 0$ and $u_0(x)$ is assumed to be smooth and small at infinity.

Equation (1) has various applications in the area of nonlinear optics, plasma physics and fluid mechanics (see [7], [13], [20]). When $p=3$, $n=2$ and $a=0$, (1) shows instability phenomena such as self-focusing of an electromagnetic beam. Global existence theorems and scattering theorems for the problem (1), (2) with $a=0$ have been developed by [3], [10] and [11]. Nonexistence of global solutions of (1), (2) with $a=0$ has been shown first by [21] for radial solutions, the case $p=3$, $n=2$, and then by [5] for more general cases (see also [17]).

We mention the results for (1), (2) with $a=0$ more precisely. When $a=0$, (1) has two conserved quantities:

$$(3)\qquad \begin{aligned} E_0(t)&=\int|u(x,t)|^2dx,\\ E_1(t)&=\int|\nabla u(x,t)|^2dx-\frac{2}{p+1}\int|u(x,t)|^{p+1}dx.\end{aligned}$$

(All integrals are taken over R^n.) Using these quantities it is shown that

i) if $1<p<1+4/n$, there exists a global (weak) solution to (1), (2) for a datum of unrestricted finite energy, i.e., $|E_1(0)|<\infty$;

ii) if $p\geq 1+4/n$, solution with nonpositive energy ($E(0)\leq 0$) cannot exist for all $t>0$.

The first assertion i) is established by Ginibre–Velo [3]. The second assertion ii) is established by Glassey [5] ($p>1+4/n$) and by the author [17] ($p\geq 1+4/n$). In the proofs of ii) one of the crucial hypotheses is that $E(0)\leq 0$ which implies that the solution has nonpositive energy, i.e., $E(t)\leq 0$ for all $t>0$.

For $a>0$ both $E_0(t)$ and $E_1(t)$ are no longer conserved; moreover, nonpositivity of $E(0)$ does not assure nonpositivity of $E(t)$ for all $t>0$. We have $E_0(t)=e^{-at}E_0(0)$, that is, the L^2-norm of solutions decays exponentially if solutions exist globally. Hence,

*Received by the editors February 16, 1982.

†Department of Mathematics, Indiana University, Bloomington, Indiana 47405 and Waseda University, Tokyo, 160 Japan.

there naturally arises the question whether or not the damping term prevents blowing up of solutions.

Our aim of the present paper is to show that even for $a>0$ solutions of (1), (2) cannot exist for all $t>0$ if its initial datum is of nonpositive energy, that is, $E_1(0)\leq 0$.

2. Nonexistence of global solutions and blow-up theorems. The Cauchy problem to be considered is the following:

$$iu_t=\Delta u+q(|u|^2)u-\frac{ia}{2}u, \qquad x\in R^n, \quad t>0, \tag{4}$$

$$u(x,0)=u_0(x), \qquad x\in R^n. \tag{5}$$

Assume that $a\geq 0$ and q is a real valued function defined on $[0,\infty)$. Let

$$Q(v)=\int_0^v q(s)\,ds.$$

DEFINITION. The function $u=u(x,t)$ defined on $[0,T]\times R^n$, $T>0$ is called a *weak solution* of (4) if $u\in C([0,T]:H^1(R^n))$ with $Q(|u|^2)$, $q(|u|^2)u\in C([0,T]:L^1(R^n))$, $|x|u\in C([0,T]:L^2(R^n))$, and satisfies (4) in the sense of distribution and (5) at $t=0$. Here we assume $u_0\in H^1(R^n)$ with $Q(|u_0|^2)$, $q(|u_0|^2)u_0\in L^1(R^n)$ and $|x|u_0\in L^2(R^n)$.

We begin with a lemma giving several identities which will be used in the proof.

LEMMA 1. *Let $b\in R$ and u be a sufficiently smooth solution of* (4) *on* $0\leq t\leq T$. *Then*

(i)

$$\int |u(x,t)|^2dx=e^{-at}\int |u_0(x)|^2dx,$$

(ii)

$$\begin{aligned}e^{bt}\left[\int |\nabla u|^2dx-\int Q(|u|^2)\,dx\right]=&\int |\nabla u_0|^2dx-\int Q(|u_0|^2)\,dx\\&+b\int_0^t e^{bs}\left[\int |\nabla u|^2dx-\int Q(|u|^2)\,dx\right]ds\\&-a\int_0^t e^{bs}\left[\int |\nabla u|^2dx-\int q(|u|^2)|u|^2dx\right]ds,\end{aligned}$$

(iii) $$e^{bt}\int |x|^2|u|^2dx+(a-b)\int_0^t e^{bs}\int |x|^2|u|^2dx\,ds=\int |x|^2|u_0|^2dx+\int_0^t V(s)e^{bs}\,ds,$$

(iv) $$V(t)e^{bt}+(a-b)\int_0^t V(s)e^{bs}\,ds=V(0)+\int_0^t e(s)e^{bs}\,ds,$$

where

$$V(t)=-4\,\mathrm{Im}\int x\cdot\nabla u(x,t)\bar{u}(x,t)\,dx,$$

$$\left(V(0)=-4\,\mathrm{Im}\int x\cdot\nabla u_0(x)\bar{u}_0(x)\,dx\right),$$

and

$$e(t)=8\int |\nabla u|^2dx+4n\int\left[Q(|u|^2)-q(|u|^2)|u|^2\right]dx$$

($\bar{u}$ being the complex conjugate of u).

Proof. Multiplying (4) by $2\bar{u}$ and taking the imaginary part of the result, we get

$$\frac{\partial}{\partial t}|u|^2+a|u|^2=\nabla\cdot(2\,\mathrm{Im}\,\bar{u}\nabla u). \tag{6}$$

Integration of (6) over R^n gives

$$\frac{\partial}{\partial t}\int|u|^2dx+a\int|u|^2dx=0$$

from which (i) follows. We now multiply (6) by $|x|^2$ and integrate over R^n to get

$$\frac{d}{dt}\int|x|^2|u|^2dx+a\int|x|^2|u|^2dx=-4\,\mathrm{Im}\int(x\cdot\nabla u)\bar{u}\,dx\equiv V(t)$$

from which (iii) follows.

To establish (ii) we multiply (4) by $2\bar{u}_t$, integrate over R^n and take the real part of the result to obtain

$$\frac{d}{dt}\left[\int|\nabla u|^2dx-\int Q\big(|u|^2\big)\,dx\right]+a\left[\int|\nabla u|^2dx-\int q\big(|u|^2\big)|u|^2dx\right]=0.$$

Multiplying this identity by e^{bt} and integrating with respect to t, we get (ii). It remains to derive (iv). Integration by parts yields

$$\begin{aligned}\frac{d}{dt}&V(t)+aV(t)\\ &=-4\,\mathrm{Im}\int(x\cdot\nabla u_t)\bar{u}\,dx-4\,\mathrm{Im}\int(x\cdot\nabla u)\bar{u}_t dx+4a\,\mathrm{Im}\int(x\cdot\nabla u)\bar{u}\,dx\\ &=4n\,\mathrm{Im}\int u_t\bar{u}\,dx+8\,\mathrm{Im}\int(x\cdot\nabla\bar{u})u_t dx+4a\,\mathrm{Im}\int(x\cdot\nabla u)\bar{u}\,dx.\end{aligned}$$

The first term can be written as

$$\begin{aligned}4n\,\mathrm{Im}\int u_t\bar{u}\,dx&=-4n\,\mathrm{Re}\int\left[\bar{u}\Delta u+q\big(|u|^2\big)|u|^2\right]dx\\ &=4n\left[\int|\nabla u|^2dx-\int q\big(|u|^2\big)|u|^2dx\right].\end{aligned}$$

The second term can be written as

$$\begin{aligned}8\,\mathrm{Im}&\int(x\cdot\nabla\bar{u})u_t dx\\ &=-8\,\mathrm{Re}\int(x\cdot\nabla\bar{u})\Big(\Delta u+q\big(|u|^2\big)u\Big)\,dx-4a\,\mathrm{Im}\int(x\cdot\nabla\bar{u})u\,dx\\ &=8\int|\nabla u|^2dx-4n\int|\nabla u|^2dx+4n\int Q\big(|u|^2\big)\,dx\\ &\quad-4a\,\mathrm{Im}\int(x\cdot\nabla\bar{u})u\,dx.\end{aligned}$$

Therefore we obtain

$$\frac{d}{dt}V(t)+aV(t)=8\int|\nabla u|^2dx+4n\int\left[Q\left(|u|^2\right)-q\left(|u|^2\right)|u|^2\right]dx$$
$$\equiv e(t)$$

from which we can deduce (iv). Q.E.D.

Our main theorem is given by the following:

THEOREM 1. *Let u be a weak solution satisfying the identities* (i)–(iv). *Assume that*

$$\text{(A)}\qquad E(0)\equiv\int|\nabla u_0|^2dx-\int Q\left(|u_0|^2\right)dx\le 0,$$

(B) *there exist constants* $C_n>1+2/n$, $C_n'>1$ *such that*

$$C_nQ(s)\le sq(s)\le C_n'Q(s)\quad \textit{for all } s>0,$$

and

$$\text{(C)}\qquad \frac{a(C_n'-1)}{d_n-1}\int|x|^2|u_0|^2dx+V(0)\le 0,$$

where $d_n=n(C_n-1)/2>1$.

Then, u cannot exist for all $t>0$.

It is easily seen (see [11], [19]) that the following continuation theorem holds:

THEOREM 2. *Suppose that* $q\in C^\infty([0,\infty))$ *with* $q(0)=0$ *and* $u_0\in H^s(R^n)$, $s>n/2$ (*with* $ru_0\in L^2(R^n)$). *Then we have the following alternatives*:

1) *the* (*smooth*) *solution u exists for all* $t>0$;

2) *there exists a* $T>0$ *such that the* (*smooth*) *solution u exists in* $[0,T)$ *and*

$$\limsup_{t\uparrow T}\|u(t)\|_{H^k(R^n)}=+\infty$$

for any $k\in R$ *satisfying* $n/2<k\le s$.

The solution u satisfies (i)–(iv) *on the maximal time interval of existence.*

Combining Theorem 1 and Theorem 2, we obtain

COROLLARY 1. *In addition to the assumptions in Theorem* 1, *we assume that* $q\in C^\infty([0,\infty))$ *with* $q(0)=0$ *and* $u_0\in H^s(R^n)$, $s>n/2$ *with* $ru_0\in L^2(R^n)$). *Then, there exists a finite time* $T>0$, *estimable from above, such that*

$$\limsup_{t\uparrow T}\|u(t)\|_{H^k(R^n)}=+\infty$$

for any $k\in R$ *satisfying* $n/2<k\le s$.

Ginibre–Velo [3] have shown that

THEOREM 3. *Suppose that* $n\le 3$, $q\in C^1([0,\infty))$ *with* $q(0)=0$ *and* $u_0\in H^1(R^n)\cap L^\infty(R^n)$. *Then the following alternatives hold valid*:

1) *the weak solution u exists for all* $t>0$;

2) *there exists a* $T>0$ *such that the weak solution u exists in* $[0,T]$ *and*

$$\limsup_{t\uparrow T}\|u(t)\|_{L^\infty(R^n)}=+\infty.$$

Therefore we obtain

COROLLARY 3. *In addition to the assumptions in Theorem* 1, *we assume that* $q \in C^1([0,\infty))$ *with* $q(0)=0$ *and* $u_0 \in H^1(R^n) \cap L^\infty(R^n)$ $(n \leq 3)$ *with* $ru_0 \in L^2(R^n)$. *Then, there exists a finite time* $T>0$, *estimable from above, such that*

$$\limsup_{t \uparrow T} \|u(t)\|_{L^\infty(R^n)} = +\infty. \tag{7}$$

If $n=1$, (7) *is equivalent to*

$$\limsup_{t \uparrow T} \|u(t)\|_{H^1(R^n)} = +\infty.$$

The following blow-up theorem shows that for some initial data the singularity of solutions occurs at $x=0$ like the δ-function type, as was suggested by numerical computation of (4) with $a=0$ (see [7], [20]).

THEOREM 4. *Let u be a weak solution satisfying* (i)–(iv). *Assume that* (A), (B) *and* (C) *in Theorem* 1 *hold valid. Let* $[0,T)$ *be the maximal interval of existence for u. If*

$$\lim_{t \uparrow T} \int |x|^2 |u(x,t)|^2 dx = 0,$$

then

$$\lim_{t \uparrow T} \|u(t)\|_{L^p(R^n)} = 0 \quad \textit{for} \max\left(\frac{2n}{n+2}, 1\right) \leq p < 2$$

and for any $\varepsilon > 0$

$$\lim_{t \uparrow T} \|u(t)\|_{L^2(|x|>\varepsilon)} = 0,$$

$$\lim_{t \uparrow T} \|u(t)\|_{L^p(|x|<\varepsilon)} = +\infty \quad \textit{for } 2 < p \leq \infty.$$

Remark. Let $x_0 \in R^n$ be fixed. Multiplying (6) by $|x-x_0|^2$ and integrating over R^n, we get (iii) replacing $|x|^2$ by $|x-x_0|^2$. We also have (iv) replacing $V(t)$ by

$$\hat{V}(t) = -4 \operatorname{Im} \int (x-x_0) \cdot \nabla u(x,t) \bar{u}(x,t)\, dx.$$

Therefore under the assumptions (A), (B) and

$$\text{(C}') \quad \frac{a(C_n'-1)}{d_n - 1s} \int |x-x_0|^2 |u_0|^2 dx - 4 \operatorname{Im} \int (x-x_0) \cdot \nabla u_0(x) \bar{u}_0(x)\, dx \leq 0$$

the singularity of the solution may occur at $x = x_0$ if

$$\lim_{t \uparrow T} \int |x-x_0|^2 |u(x,t)|^2 dx = 0,$$

where $[0,T)$ is the maximal interval of existence for u.

3. Proofs.

Proof of Theorem 1. Suppose that the solution u exists for all $t>0$. Put

$$E(t) = \int |\nabla u|^2 dx - \int Q(|u|^2)\, dx.$$

From (ii) of Lemma 1, we have

$$E(t)e^{bt}=E(0)+\int_0^t e^{bs}g(s)\,ds$$

where

$$g(t)=-(a-b)\int|\nabla u|^2dx-b\int Q(|u|^2)\,dx+a\int q(|u|^2)|u|^2dx.$$

Using hypothesis (B) we find

$$\begin{aligned}g(t)&\le-(a-b)\int|\nabla u|^2dx-b\int Q(|u|^2)\,dx+aC_n'\int Q(|u|^2)\,dx\\&\le-(a-b)\left[\int|\nabla u|^2dx-\frac{n}{2}(C_n-1)\int Q(|u|^2)\,dx\right]\\&\equiv-(a-b)e_1(t),\end{aligned}$$

provided that

$$aC_n'-b\le\frac{n}{2}(a-b)(C_n-1)$$

which is equal to

(8) $$b\le\frac{d_n-C_n'}{d_n-1}a\le a.$$

Therefore we have

$$E(t)e^{bt}\le E(0)-(a-b)\int_0^t e^{bs}e_1(s)\,ds.$$

Hypothesis (B) yields $e_1(t)\le E(t)$ for all $t\ge0$. Hence we get

$$e_1(t)e^{bt}\le E(0)-(a-b)\int_0^t e_1(s)e^{bs}ds$$

which is equivalent to

$$\frac{d}{dt}\left[e^{(a-b)t}\int_0^t e_1(s)e^{bs}ds\right]\le E(0)e^{(a-b)t}.$$

Now $E(0)\le0$ by (A). Hence

$$\int_0^t e_1(s)e^{bs}ds\le0.$$

By (iv) of Lemma 1 and the above inequality we have

$$V(t)e^{bt}+(a-b)\int_0^t V(s)e^{bs}ds\le V(0)$$

from which it follows that

$$\frac{d}{dt}\left[e^{(a-b)t}\int_0^t V(s)e^{bs}ds\right]\le V(0)e^{(a-b)t}.$$

Hence

$$\int_0^t V(s)e^{bs}ds \le \frac{1}{a-b}(1-e^{-(a-b)t})V(0). \tag{9}$$

Put

$$y(t)=e^{bt}\int |x|^2|u(x,t)|^2dx.$$

From (iii) of Lemma 1 and (9) we get

$$y(t)\le\int |x|^2|u_0(x)|^2dx+\frac{1}{a-b}(1-e^{-(a-b)t})V(0). \tag{10}$$

Set

$$\delta=\frac{d_n-C_n'}{d_n-1}a-b.$$

Then by (8) $\delta\ge 0$. Put

$$\begin{aligned} T&=-\frac{1}{a-b}\log\frac{(a-b)\int|x|^2|u_0(x)|^2dx+V(0)}{V(0)}\\ &=-\left(\frac{C_n'-1}{d_n-1}a+\delta\right)^{-1}\log P \end{aligned}$$

where

$$P\equiv\left[\left(\frac{C_n'-1}{d_n-1}a+\delta\right)\int |x|^2|u_0(x)|^2dx+V(0)\right]\Big/V(0).$$

Note that hypothesis (C) implies $P<1$. Hence $T>0$ and $\lim_{t\uparrow T}y(t)=0$. We have

$$\begin{aligned} \int |u|^2dx&=-\int x_i\left(u\frac{\partial \bar{u}}{\partial x_i}+\bar{u}\frac{\partial u}{\partial x_i}\right)dx\\ &\le 2\left(\int x_i^2|u|^2dx\right)^{1/2}\left(\int\left|\frac{\partial u}{\partial x_i}\right|^2dx\right)^{1/2}\\ &\le 2\||x|u\|_{L^2(R^n)}\|\nabla u\|_{L^2(R^n)} \end{aligned}$$

if $x_iu^2\to 0$ as $|x|\to\infty$. Hence from (i) we get

$$\begin{aligned} \|\nabla u(t)\|_{L^2(R^n)}&\ge\|u(t)\|_{L^2(R^n)}/2\||x|u(t)\|_{L^2(R^n)}\\ &=\frac{e^{-(a-b)t}\|u_0\|^2_{L^2(R^n)}}{2y(t)}\\ &\to+\infty \quad \text{as } t\to T, \end{aligned}$$

which leads to a contradiction. This completes the proof.

Proof of Theorem 4. Let $\max(2n/(n+2),1)\le p<2$. Hölder's inequality gives

$$\begin{aligned}\int|u(x,t)|^p dx &= \int_{|x|>R}|u(x,t)|^p dx+\int_{|x|<R}|u(x,t)|^p dx\\ &\le\left(\int_{|x|>R}|x|^{-2p/(2-p)}dx\right)^{(2-p)/2}\left(\int_{|x|>R}|x|^2|u(x,t)|^2dx\right)^{p/2}\\ &\quad+\left(\int_{|x|<R}1\,dx\right)^{(2-p)/2}\left(\int_{|x|<R}|u(x,t)|^2dx\right)^{p/2}\\ &\le W_1(R)\left(\int|x|^2|u(x,t)|^2dx\right)^{p/2}+W_2(R)\left(\int|u_0(x)|^2dx\right)^{p/2}\end{aligned}$$

where

$$W_1(R)=\text{Const.}\,R^{n-(n/2+1)p}\text{ and }W_2(R)=\text{Const.}\,R^{n(2-p)/2}.$$

For arbitrarily small $\mu>0$ we can find $R>0$ so large that

$$W_2(R)\left(\int|u_0(x)|dx\right)^{p/2}<\mu.$$

For such a $R>0$ fixed we obtain

$$\limsup_{t\uparrow T}\int|u(x,t)|^p dx\le\mu.$$

Hence

$$\lim_{t\uparrow T}\|u(t)\|_{L^p(R^n)}=0. \tag{11}$$

We have

$$\int_{|x|>\varepsilon}|u(x,t)|^2dx\le\frac{1}{\varepsilon^2}\int|x|^2|u(x,t)|^2dx\to0 \tag{12}$$

as $t\to T$ for any fixed $\varepsilon>0$. By Hölder's inequality we get

$$\begin{aligned}e^{-at}\|u_0\|^2_{L^2(R^n)}&=\|u(t)\|^2_{L^2(R^n)}\\ &=\|u(t)\|^2_{L^2(|x|<\varepsilon)}+\|u(t)\|^2_{L^2(|x|>\varepsilon)}\\ &\le\|u(t)\|_{L^q(|x|<\varepsilon)}\|u(t)\|_{L^p(|x|<\varepsilon)}+\|u(t)\|^2_{L^2(|x|>\varepsilon)},\end{aligned}$$

where $1/p+1/q=1,\ p,q\ge1$. By (11) and (12) we have

$$\lim_{t\uparrow T}\|u(t)\|_{L^q(|x|<\varepsilon)}=+\infty.$$

Here $2<q<2n/(n-2)$ $(n>2)$ and $2<q<+\infty$ $(n=1,2)$. Since

$$\|u(t)\|^q_{L^q(|x|<\varepsilon)}\le\|u(t)\|_{L^2(R^n)}\|u(t)\|^{q-2}_{L^\infty(|x|<\varepsilon)},$$

we get

$$\lim_{t\uparrow T}\|u(t)\|_{L^\infty(|x|<\varepsilon)}=+\infty.$$

Similarly we obtain

$$\lim_{t\uparrow T}\|u(t)\|_{L^q(|x|<\varepsilon)}=+\infty \quad \text{for all } q\geq\frac{2n}{n-2}.$$

This completes the proof.

4. Final remarks. In recent years, various works have been published on the blowing up of solutions or nonexistence of global solutions of the Cauchy problem and the initial boundary value problem of nonlinear partial differential equations (in particular, nonlinear wave or heat equations). (See [1], [2], [4], [6], [8], [9], [15].) As is shown in Theorems 1 and 4, the nonexistence theorem for the nonlinear Schrödinger equation has a somewhat different character. The L^2-norm of solutions is conserved or even decays exponentially; however solutions may collapse to form point-like foci (see [20], [21]). Analogous instability phenomena may be present in the case of the generalized Korteweg–de Vries equation

$$(*) \qquad u_t+(u^p)_x+u_{xxx}=0, \qquad p=1,2,\cdots,$$

which has three conserved quantities:

$$\int u\,dx, \quad \int u^2dx, \quad \frac{1}{2}\int(u_x)^2dx-\frac{1}{p+1}\int u^{p+1}dx.$$

If $1\leq p\leq 4$, we have a global existence theorem for the Cauchy problem to the equation (*). For $p\geq 5$ we only have a global existence theorem for small Cauchy data (see [12], [16]). The question whether or not the solutions for large data blow up is still unsolved so far as the author knows.

Acknowledgments. The author would like to express his thanks to Professor R. Glassey and Ms. Jacqueline Barab, Indiana University for useful discussions. The author also thanks the referee for valuable advice.

REFERENCES

[1] J. M. Ball, *Remarks on blow-up and nonexistence theorems for nonlinear evolution equations*, Quart. J. Math., Oxford, 28 (1977), pp. 47–486.

[2] H. Fujita, *On the blowing-up of solutions of the Cauchy problem for $u_t=\Delta u+u^{1+\alpha}$*, J. Fac. Sci., Univ. of Tokyo, Sect. IA, 13 (1966), pp. 109–124.

[3] J. Ginibre and G. Velo, *On the class of nonlinear Schrödinger equations* I, II, J. Functional Analysis, 32 (1979), pp. 1–32, 33–71; III, Ann. Inst. Henri Poincaré Sect. A, 28 (1978), pp. 287–316.

[4] R. T. Glassey, *Blow-up theorems for nonlinear wave equations*, Math. Z., 132 (1973), pp. 182–203.

[5] ———, *On the blowing up of solutions to the Cauchy problem for nonlinear Schrödinger equations*, J. Math. Phys., 18 (1977), pp. 1794–1797.

[6] F. John, *Blow up of solutions of nonlinear wave equations in three space dimensions*, Manuscripta Math., 28 (1979), pp. 235–268.

[7] P. L. Kelley, *Self-focusing of optical beams*, Phys. Rev. Lett., 15 (1965), pp. 1005–1008.

[8] H. Levine, *Some nonexistence and instability theorems for solutions of formally parabolic equations of the form $Pu_t=-Au+F(u)$*, Arch. Rat. Mech. Anal., 51 (1973), pp. 371–386.

[9] ———, *Instability and nonexistence of global solutions to nonlinear wave equations of the form $Pu_{tt}=-At+F(u)$*, Trans. Amer. Math. Soc., 192 (1974), pp. 1–21.

[10] J. E. Lin and W. A. Strauss, *Decay and scattering of solutions of a nonlinear Schrödinger equation*, J. Funct. Anal., 30 (1978), pp. 245–263.

[11] H. Pecher and W. von Wahl, *Time dependent nonlinear Schrödinger equations*, Manuscripta Math., 27 (1979), pp. 125–157.

[12] W. Strauss, *Dispersion of low energy waves for two conservative equations*, Arch. Rat. Mech. Anal., 55 (1974), pp. 86–92.

[13] T. Taniuti and H. Washimi, *Self-trapping and instability of hydromagnetic waves along the magnetic field in a cold plasma*, Phys. Rev. Lett., 21 (1968), pp. 209–212.

[14] M. Tsutsumi, *On solutions of semilinear differential equations in a Hilbert space*, Math. Japon., 17 (1972), pp. 173–193.

[15] ———, *Existence and nonexistence of global solutions for nonlinear parabolic equations*, RIMS Kyoto Univ., 8 (1972), pp. 211–229.

[16] ———, *On global solutions of the generalized Korteweg-de Vries equation*, RIMS Kyoto Univ., 7 (1971), pp. 329–344.

[17] ———, *Nonexistence and instability of solutions of nonlinear Schrödinger equations* (unpublished).

[18] ———, *Weighted Sobolev spaces and rapidly decreasing solutions of some nonlinear dispersive wave equations*, J. Differential Equations, 42 (1981), pp. 260–281.

[19] M. Tsutsumi and N. Hayashi, *Classical solutions of nonlinear Schrödinger equations in higher dimensions*, Math. Z., 177 (1981), pp. 217–237.

[20] V. E. Zakharov, V. V. Sobolev and V. C. Synakh, *Behavior of light beams in nonlinear media*, Soviet Phys. JETP, 33 (1971), pp. 77–81.

[21] V. E. Zakharov, *Collapse of Langmuir waves*, Soviet Phys. JETP, 35 (1972), pp. 908–914.

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1502-0011 $01.25/0

A NECESSARY AND SUFFICIENT CONDITION FOR THE COERCIVENESS OF A CLASS OF FUNCTIONALS AND ITS APPLICATIONS*

PATRICK RABIER[†]

Abstract. We show for a family of minimization problems for which a suitable directional growth condition is fulfilled, how the weak sequential lower semicontinuity of the functional is closely related to its coerciveness instead of being a completely separate property. A general example and some of its particular cases, namely applications to semi-linear elliptic problems—for which appropriate results about Nemytskii operators have been established here—are given.

Introduction. We recall that a functional $\mathcal{J}$ defined in a normed space H (norm $\|\cdot\|$) with values in $\mathbb{R}^{\cdot}=\mathbb{R}\cup\{+\infty\}$ is said to be *coercive* if and only if

$$\lim_{\|v\|\to+\infty} \mathcal{J}(v)=+\infty.$$

In this paper, we consider the case when H is a Hilbert space. Denoting by J the "*nonquadratic part*" of $\mathcal{J}$ and assuming

1) an appropriate *directional* growth condition (whose verification is generally obvious in concrete problems),

2) *the weak lower semicontinuity* of J, we show in §1 that the coerciveness of $\mathcal{J}$ is related to a condition (Theorem 1.1, inequality (1.9)) *which generalizes the condition of H-ellipticity for quadratic functionals.*

As in several well-known problems, the possibility for the functional $\mathcal{J}$ to take its values in $\mathbb{R}^{\cdot}$ yields an easy way for solving within a Hilbertian framework some problems that are not naturally defined in a Hilbert space: in §2 we give a general example of application using this method in conjunction with the theoretical results of the first section. Finally, as a particular case of the example of §2, we consider in §3 the typical equation (that can be generalized in many ways)

$$\begin{aligned}&(-1)^m\Delta^m u-\lambda u+\tilde{f}(u)=u^*\in H^{-m}(\Omega),\\&u\in H_0^m(\Omega),\end{aligned}$$

where Ω is a bounded open subset of $\mathbb{R}^N$, λ a given real number and $\tilde{f}$ a Nemytskii operator associated with some given Carathéodory function $f\colon\Omega\times\mathbb{R}\to\mathbb{R}$. Roughly speaking and among other results, we are able to prove the existence of at least one solution of this equation in the space $H_0^m(\Omega)$, even when $\tilde{f}(v)\notin H^{-m}(\Omega)$ for each $v\in H_0^m(\Omega)$. It is interesting to notice in this case that our condition of coerciveness depends upon the fact that the function $y\to f(x,y)/y$ is nondecreasing for $y>0$ and nonincreasing for $y<0$, *for almost all* $x\in\Omega$. To our best knowledge, this type of growth property is usually assumed to hold *uniformly* with respect to $x\in\bar{\Omega}$ and with a function $f(\cdot,y)$ that is continuous on $\bar{\Omega}$ for every $y\in\mathbb{R}$. Finally, let us mention that the results established in §1 are of importance in a general study of nonlinear problems having *exactly three solutions* presented elsewhere (cf. [6], [7]) that generalizes, through a new notion of convexity (cf. [6], [8]) recent works of M. S. Berger or Ambrosetti—Mancini (references in [6], [7]) dealing with particular cases (homogeneous ones). Besides, the

* Received by the editors April 7, 1982, and in revised form September 29, 1982.

[†] Université Pierre et Marie Curie, 4 Place Jussieu, 75230 Paris Cedex 05, France.

existence result of three solutions ($u=0$ and two nontrivial ones) of the variational problems we consider in [6], [7] is given here (Theorem 2.1 and Remark 2.1) *without any assumption of evenness* on the functional under consideration.

1. Abstract results. In the following, we denote by H a real Hilbert space equipped with the inner product $(\cdot, \cdot)$ and the associated norm $\|\cdot\|$. Let

$$J\colon H\to\mathbb{R}^{\cdot}=\mathbb{R}\cup\{+\infty\}$$

be a given functional. We assume for every $v\in H$ that the limit

$$\lim_{t\to+\infty}\frac{J(tv)}{t^2}\in\overline{\mathbb{R}}=\mathbb{R}\cup\{-\infty,+\infty\}$$

exists and we set

$$\omega_\infty(v)=2\lim_{t\to+\infty}\frac{J(tv)}{t^2}. \tag{1.1}$$

In particular we have

$$\omega_\infty(0)=\begin{cases}0 & \text{if } J(0)\in\mathbb{R},\\ +\infty & \text{if } J(0)=+\infty,\end{cases} \tag{1.2}$$

and a straightforward computation shows that

$$\omega_\infty(\lambda v)=\lambda^2\omega_\infty(v) \quad \text{for every } v\in H \text{ and every } \lambda>0. \tag{1.3}$$

Obviously, a sufficient condition for $\omega_\infty(v)$ to be well defined in $\overline{\mathbb{R}}$ for every $v\in H$ is that $J(0)\in\mathbb{R}$ and the function

$$t\in]0,+\infty[\to\frac{J(tv)-J(0)}{t^2}\in\mathbb{R}^{\cdot} \tag{1.4}$$

is nondecreasing. In this case

$$\omega_\infty(v)=2\lim_{t\to+\infty}\frac{J(tv)-J(0)}{t^2}\in\mathbb{R}^{\cdot} \tag{1.5}$$

for every $v\in H$, and

$$\frac{J(tv)-J(0)}{t^2}\le\frac{1}{2}\omega_\infty(v) \tag{1.6}$$

for every $v\in H$ and every $t>0$. From now on, we shall always assume that $J(0)\in\mathbb{R}$ without further mention.

THEOREM 1.1. *Let $J\colon H\to\mathbb{R}^{\cdot}$ be a given functional. We assume that J is weakly sequentially l.s.c. (lower semicontinuous) and for every $v\in H$ that the function*

$$t\in]0,+\infty[\to\frac{J(tv)-J(0)}{t^2}\in\mathbb{R}^{\cdot}, \tag{1.7}$$

is nondecreasing.

On the other hand, let B and L be two linear continuous operators from H into itself such that

(i) *B is selfadjoint and H-elliptic,*

(ii) *L is selfadjoint and compact.*

For every linear continuous form φ *on* H *we denote by* $\mathcal{J}_\varphi\colon H\to\mathbb{R}^{\cdot}$ *the functional*

$$v\in H\to\mathcal{J}_\varphi(v)=\tfrac{1}{2}((B-L)v,v)+J(v)-\varphi(v)\in\mathbb{R}^{\cdot}. \tag{1.8}$$

Then, the following assertions are equivalent:

(a) *For every* $v\in H-\{0\}$

$$(Bv,v)-(Lv,v)+\omega_\infty(v)>0. \tag{1.9}$$

(b) *For every linear continuous form* φ *on* H, *the functional* $\mathcal{J}_\varphi$ (1.8) *is coercive.*

(c) *The functional* $\mathcal{J}_0$ (*i.e.* $\varphi=0$ *in* (1.8)) *is coercive.*

Proof. (a)$\Rightarrow$(b). Let φ be any linear continuous form on H. If the functional $\mathcal{J}_\varphi$ (1.8) is not coercive, there exists a constant $C\in\mathbb{R}$ such that the set $\{v\in H,\mathcal{J}_\varphi(v)\le C\}$ contains an unbounded sequence (v_n). Obviously, we may assume $v_n\neq 0$ for every $n\in\mathbb{N}$ and define

$$t_n=\|v_n\|,\qquad \sigma_n=\frac{v_n}{\|v_n\|},$$

in order that

$$\|\sigma_n\|=1,\qquad \lim t_n=+\infty. \tag{1.10}$$

By definition of (v_n), it follows that $J(v_n)\in\mathbb{R}$ for every $n\in\mathbb{N}$. Thus, after dividing by t_n^2, the condition $\mathcal{J}_\varphi(v_n)\le C$ may be rewritten as

$$\frac{1}{2}(B\sigma_n,\sigma_n)+\frac{J(t_n\sigma_n)}{t_n^2}\le\frac{1}{2}(L\sigma_n,\sigma_n)+\frac{C}{t_n^2}+\frac{1}{t_n}\varphi(\sigma_n), \tag{1.11}$$

for every $n\in\mathbb{N}$. Because of the weak compactness of the unit ball in H, there exists a subsequence of the sequence (σ_n) that tends weakly to a limit $\sigma\in H$. As usual, this subsequence will still be denoted by (σ_n). Let $\tilde t>0$ be given. Then, $\tilde t\sigma_n$ tends weakly to $\tilde t\sigma$ and J being weakly sequentially lower semicontinuous (l.s.c.) so is the functional $J/\tilde t^2$. Thus

$$\underline{\lim}\,\frac{J(\tilde t\sigma_n)}{\tilde t^2}\ge\frac{J(\tilde t\sigma)}{\tilde t^2}. \tag{1.12}$$

Now, since $\lim t_n=+\infty$ (cf. (1.10)), there exists $n_0\in\mathbb{N}$ such that $t_n\ge\tilde t$ for every $n\ge n_0$. For every *given* $n\ge n_0$, the growth property of the function (1.7) with $v=\sigma_n$ shows that

$$\frac{J(t_n\sigma_n)-J(0)}{t_n^2}\ge\frac{J(\tilde t\sigma_n)-J(0)}{\tilde t^2},$$

and then

$$\underline{\lim}\,\frac{J(t_n\sigma_n)-J(0)}{t_n^2}\ge\underline{\lim}\,\frac{J(\tilde t\sigma_n)-J(0)}{\tilde t^2}.$$

But the left-hand side of the above inequality is also $\underline{\lim}\,J(t_n\sigma_n)/t_n^2$ and (1.12) provides

$$\frac{J(\tilde t\sigma)}{\tilde t^2}\le\frac{J(0)}{\tilde t^2}+\underline{\lim}\,\frac{J(t_n\sigma_n)}{t_n^2}.$$

Since this inequality holds for each $\tilde{t}>0$, we deduce by taking the limit when $\tilde{t}$ tends to $+\infty$:

$$\frac{1}{2}\omega_\infty(\sigma)\leq\underline{\lim}\frac{J(t_n\sigma_n)}{t_n^2}. \tag{1.13}$$

On the other hand, the H-ellipticity of the operator B means that there exists a constant $\beta>0$ such that $(Bv,v)\geq\beta\|v\|^2$ for every $v\in H$. Therefore, it follows from (1.11) and $\|\sigma_n\|=1$ that

$$\frac{\beta}{2}+\frac{J(t_n\sigma_n)}{t_n^2}\leq\frac{1}{2}(L\sigma_n,\sigma_n)+\frac{C}{t_n^2}+\frac{1}{t_n}\varphi(\sigma_n),$$

and hence, since L is compact

$$\frac{\beta}{2}+\underline{\lim}\frac{J(t_n\sigma_n)}{t_n^2}\leq\frac{1}{2}(L\sigma,\sigma).$$

From (1.13) we find that

$$\beta+\omega_\infty(\sigma)\leq(L\sigma,\sigma). \tag{1.14}$$

This relation shows that σ cannot be 0. Indeed, if $\sigma=0$, we reach a contradiction with (1.14) since $\beta>0$ (recall that $\omega_\infty(0)=0$).

Coming back to (1.11) and due to the compactness of L we obtain

$$\underline{\lim}\left[\frac{1}{2}(B\sigma_n,\sigma_n)+\frac{J(t_n\sigma_n)}{t_n^2}\right]\leq\frac{1}{2}(L\sigma,\sigma).$$

This inequality yields a fortiori

$$\frac{1}{2}\underline{\lim}(B\sigma_n,\sigma_n)+\underline{\lim}\frac{J(t_n\sigma_n)}{t_n^2}\leq\frac{1}{2}(L\sigma,\sigma).$$

Using (1.13) once again gives

$$\frac{1}{2}\underline{\lim}(B\sigma_n,\sigma_n)+\frac{1}{2}\omega_\infty(\sigma)\leq\frac{1}{2}(L\sigma,\sigma),$$

and finally, since the functional $v\in H\to(Bv,v)$ is weakly l.s.c. (because convex and continuous)

$$(B\sigma,\sigma)-(L\sigma,\sigma)+\omega_\infty(\sigma)\leq 0,$$

which contradicts the assumption (1.9) for we have already proved that σ cannot be 0.

(b)$\Rightarrow$(c) is obvious.

(c)$\Rightarrow$(a) Let us suppose that there exists $v\in H-\{0\}$ such that $(Bv,v)-(Lv,v)+\omega_\infty(v)\leq 0$. From the growth property of the function (1.7) we deduce (cf. (1.6))

$$\frac{J(tv)-J(0)}{t^2}\leq\frac{1}{2}\omega_\infty(v)$$

for every $t>0$. In other words

$$\mathcal{J}_0(tv)-J(0)\leq\frac{t^2}{2}\left[(Bv,v)-(Lv,v)+\omega_\infty(v)\right]\leq 0,$$

for every $t>0$, that is to say,

$$\mathcal{J}_0(tv)\leq J(0)$$

for every $t>0$, and the functional $\mathcal{J}_0$ is not coercive. □

Remark 1.1. The equivalence (a)⇔(b) in Theorem 1.1 shows that the inequality (1.9) is a *necessary and sufficient condition* for the coerciveness of the functionals $\mathcal{J}_\varphi$ altogether. However we emphasize that there may be some linear continuous form $\varphi\neq 0$ for which $\mathcal{J}_\varphi$ is coercive while the inequality (1.9) does *not* hold.

On the other hand, let us point out that the computation of $\omega_\infty(v)$ is quite easy when there is an explicit formulation of the functional J. As a matter of fact, it often happens that $\omega_\infty(v)=+\infty$ for every $v\in H-\{0\}$ and the criterion (1.9) yields the coerciveness of $\mathcal{J}_\varphi$ immediately.

As concerns the minimization problem associated with the functionals $\mathcal{J}_\varphi$, we shall now show that the criterion (1.9) is *optimal.*

COROLLARY 1.1. *Under the same assumptions as in Theorem* 1.1, *a necessary and sufficient condition for the functional* $\mathcal{J}_\varphi$ (1.8) *to have at least one absolute minimizer in the space H for every linear continuous form* φ *on H is that*

$$(Bv,v)-(Lv,v)+\omega_\infty(v)>0,$$

for every $v\in H-\{0\}$.

Proof. The condition is necessary. Indeed, let $v\in H-\{0\}$ be given such that $(Bv,v)-(Lv,v)+\omega_\infty(v)\leq 0$. Using (1.6) we obtain

$$\mathcal{J}_0(tv)-J(0)\leq\frac{t^2}{2}\left[(Bv,v)-(Lv,v)+\omega_\infty(v)\right]\leq 0$$

for every $t>0$. Thus

$$\mathcal{J}_0(tv)\leq J(0) \tag{1.15}$$

for every $t>0$. Since $v\neq 0$, there is a linear continuous form φ on H such that $\varphi(v)>0$. From (1.15)

$$\mathcal{J}_\varphi(tv)=\mathcal{J}_0(tv)-t\varphi(v)\leq J(0)-t\varphi(v),$$

for every $t>0$, and then

$$\lim_{t\to+\infty} J_\varphi(tv)=-\infty,$$

which proves that the functional $\mathcal{J}_\varphi$ has no absolute minimizer in H.

Conversely, it follows from our assumptions that the functional $\mathcal{J}_\varphi$ (1.8) is weakly sequentially l.s.c. (here we again use the weak lower semicontinuity of the functional $v\in H\to(Bv,v)$) for every linear continuous form φ on H. On the other hand, this functional is also coercive from the part "(a)⇒(b)" of Theorem 1.1 and the result follows from classical arguments. □

Theorem 1.1 and Corollary 1.1 can be applied in various situations: for instance, the proof of the existence of minimizers in the von Kármán equations which is given in

[3, Thm. 2.2–1, p. 53] is a simple particular case of this study. Furthermore, the criterion (1.9) is a fundamental theoretical result in a general theory of nonlinear equations with exactly three solutions we have developed elsewhere [6], [7]. Indeed, the same kind of arguments as in Theorem 1.1 or Corollary 1.1 are very useful for proving the existence of *at least two minimizers* of the functional $\mathcal{J}_0$ under suitable assumptions on the operator $B-L$. More precisely, due to the growth property of the function (1.4) for every $v \in H$, we may define

$$\omega_0(v)=2 \lim_{t\to 0_+} \frac{J(tv)-J(0)}{t^2} \in \overline{\mathbb{R}}, \tag{1.16}$$

in order that

$$\frac{J(tv)-J(0)}{t^2} \geq \frac{1}{2}\omega_0(v) \quad \text{for every } t>0 \text{ and every } v\in H. \tag{1.17}$$

It can be immediately verified that the map $\omega_0: H\to\overline{\mathbb{R}}$ is positively homogeneous of degree 2, i.e.,

$$\omega_0(\lambda v)=\lambda^2\omega_0(v) \quad \text{for every } \lambda\geq 0 \text{ and every } v\in H,$$

and that

$$J\geq J(0) \Rightarrow \omega_0\geq 0. \tag{1.18}$$

With these preliminaries, we have

THEOREM 1.2. *Let $J=H\to\mathbb{R}^{\cdot}$ be a functional such that the function*

$$t\in]0,+\infty[\to \frac{J(tv)-J(0)}{t^2} \in \mathbb{R}^{\cdot} \tag{1.19}$$

is nondecreasing for every $v\in H$. On the other hand, let B and L be two linear continuous operators from H into itself such that

(i) *B is selfadjoint and H-elliptic,*

(ii) *L is selfadjoint and compact.*

(a) *We assume either*

$$(Bv,v)-(Lv,v)+\omega_0(v)>0 \tag{1.20}$$

for every $v\in H-\{0\}$, or more generally

$$(Bv,v)-(Lv,v)+\omega_0(v)\geq 0 \tag{1.21}$$

for every $v\in H$ when the function (1.19) *is strictly increasing for every $v\in H-\{0\}$. Then, $0\in H$ is the unique absolute minimizer of the functional*

$$v\in H\to\mathcal{J}_0(v)=\tfrac{1}{2}((B-L)v,v)+J(v)\in\mathbb{R}^{\cdot}. \tag{1.22}$$

(b) *Conversely, we assume that the functional J is* $\geq J(0)$, *weakly sequentially l.s.c., that there exists $v_0\in H$ such that*

$$(Bv_0,v_0)-(Lv_0,v_0)+\omega_0(v_0)<0, \tag{1.23}$$

and that the condition

$$(Bv,v)-(Lv,v)+\omega_\infty(v)>0, \tag{1.24}$$

holds for every $v \in H-\{0\}$. *We assume also that* (Sp *denoting the spectrum*)

$$(1.25) \qquad \mathrm{Sp}(B^{-1/2}LB^{-1/2}) \subset]-\infty, 1[\cup \{\mu_0\},$$

with $\mu_0 > 1$ *a simple eigenvalue. Then, the functional* $\mathcal{J}_0$ (1.22) *has at least one minimizer in the open half-space determined by the hyperplane*

$$(1.26) \qquad B^{-1/2}\left(\mathrm{Ker}\{\mu_0 I - B^{-1/2}LB^{-1/2}\}^{\perp}\right)$$

and containing v_0.

Proof. (a) If we apply (1.17) with $t=1$ we have

$$(1.27) \qquad J(v)-J(0) \geq \tfrac{1}{2}\omega_0(v),$$

for every $v \in H$. In fact, if the function (1.19) is strictly increasing for $v \neq 0$, we see that

$$(1.28) \qquad J(v)-J(0) > \tfrac{1}{2}\omega_0(v),$$

for every $v \in H-\{0\}$. Then, using either (1.20) and (1.27) or (1.21) and (1.28), we find

$$\mathcal{J}_0(v) > J(0) = \mathcal{J}_0(0)$$

for every $v \in H-\{0\}$, so that $0 \in H$ is the unique absolute minimizer of the functional $\mathcal{J}_0$.

(b) Let us first notice that the operators $B^{1/2}$ and $B^{-1/2}$ are well defined, selfadjoint and H-elliptic since the same properties hold for B. It will be convenient to denote by $\mathcal{N}_0$ the one-dimensional null-space

$$\mathcal{N}_0 = \mathrm{Ker}\left(\mu_0 I - B^{-1/2}LB^{-1/2}\right).$$

Then its orthogonal is a hyperplane and so is the space $B^{-1/2}(\mathcal{N}_0^{\perp})$. Let us observe that $B-L$ is $B^{-1/2}(\mathcal{N}_0^{\perp})$-elliptic. Indeed, for $v \in B^{-1/2}(\mathcal{N}_0^{\perp})$ we set $w = B^{1/2}v \in \mathcal{N}_0^{\perp}$ and get

$$((B-L)v, v) = \left((\mu_0 I - B^{-1/2}LB^{-1/2})w, w\right) \geq C\|w\|^2,$$

for some constant $C>0$ since (cf. (1.25)) the eigenvalues of the restriction to $\mathcal{N}_0^{\perp}$ of the compact selfadjoint operator $B^{-1/2}LB^{-1/2}$ are $<\mu_0$. The conclusion follows from the inequality $\|w\| \geq \|v\|/\|B^{-1/2}\|$.

As the functional J is $\geq J(0)$ in H, we have $\omega_0 \geq 0$ in H (cf. (1.18)) and (1.23) one can occur with $((B-L)v_0, v_0) < 0$ only. We conclude that v_0 is not an element of the hyperplane $B^{-1/2}(\mathcal{N}_0^{\perp})$ (because the inner product $((B-L)v_0, v_0)$ would then be ≥ 0) and hence v_0 belongs to one of the two open half-spaces determined by this hyperplane. In what follows, we shall denote by H_{v_0} the half-space in question.

Using the fact that $J \geq J(0)$ in H once again and from the $B^{-1/2}(\mathcal{N}_0^{\perp})$-ellipticity of $B-L$, we find that the functional $\mathcal{J}_0$ (1.22) is $\geq J(0) = \mathcal{J}_0(0)$ in the space $B^{-1/2}(\mathcal{N}_0^{\perp})$. Thus, we shall prove that $\mathcal{J}_0$ has at least one minimizer in H_{v_0} by showing the existence of such a minimizer in the closure $\overline{H}_{v_0}$ while $\mathcal{J}_0$ takes values $<J(0)$ in H_{v_0}. Under our assumptions, the coerciveness of $\mathcal{J}_0$ follows from Theorem 1.1. But $\mathcal{J}_0$ is also weakly sequentially l.s.c. (as noticed in Corollary 1.1) and possesses then at least one minimizer in the (weakly) closed half-space $\overline{H}_{v_0}$. It remains to prove that $\mathcal{J}_0$ assumes values $<J(0)$ in H_{v_0}. By definition of the map ω_0 (cf. (1.16)) and since the condition (1.23) can occur

for $\omega_0(v_0)\in[0,+\infty[$ only (cf. (1.18)), we deduce for every $\varepsilon>0$

$$\frac{J(tv_0)-J(0)}{t^2}\leq\frac{\omega_0(v_0)+\varepsilon}{2},$$

for $t>0$ small enough. Taking $\varepsilon>0$ such that

$$(Bv_0,v_0)-(Lv_0,v_0)+\omega_0(v_0)+\varepsilon<0,$$

we have for $t>0$ small enough

$$\mathcal{J}_0(tv_0)-J(0)=t^2\left[\frac{1}{2}((B-L)v_0,v_0)+\frac{J(tv_0)-J(0)}{t^2}\right]$$

$$\leq\frac{t^2}{2}\left[(Bv_0,v_0)-(Lv_0,v_0)+\omega_0(v_0)+\varepsilon\right]<0,$$

which concludes the proof. □

Remark 1.2. Let us show how Theorem 1.2 can be used for proving the existence of at least two distinct minimizers of the functional $\mathcal{J}_0$. Indeed, under the assumptions of Theorem 1.2, this result is immediate if $\omega_0(-v_0)=\omega_0(v_0)$ because $-v_0$ and v_0 do not belong to the same open half-space determined by the hyperplane (1.26). For instance, when the functional J is defined everywhere around $0\in H$ and is twice differentiable at the origin, it is not difficult to see that $\omega_0(v)=J''(0).\ (v,v)$ for every $v\in H$ and the problem is solved since ω_0 is even (notice that *this property is not related at all to some assumption of eveness of J*). In fact, the relation $\omega_0(-v_0)=\omega_0(v_0)$ holds for some $v_0\neq 0$ as soon as J is twice differentiable at the origin in the direction v_0 and we have $\omega_0(-v_0)=\omega_0(v_0)=(d^2/dt^2)J(tv_0)|_{t=0}$. In this case, J need not be finite everywhere around the origin (only around 0 in the direction v_0) which is useful in important practical applications as we shall see later on.

We do not know (except in the case when J is even) if the assumption (1.25) on the spectrum may be omitted or weakened.

2. A general example. We shall here study in detail a general example (and one of its concrete applications in the next section) in which the assumption that J takes its value $\mathbb{R}^{\bullet}$ is fully employed.

Let W be a real Banach space and H a real Hilbert space. We suppose that both H and W are contained in a locally convex space X with continuous embeddings $H\subset X$ and $W\subset X$. The norms in the spaces H and W will be denoted by $\|\cdot\|_H$ and $\|\cdot\|_W$ respectively. Let

$$\mathbb{F}: W\to\mathbb{R}$$

be a weakly sequentially l.s.c. functional. We define the functional

$$J: v\in H\to J(v)=\begin{cases}\mathbb{F}(v) & \text{if } v\in W,\\ +\infty & \text{if } v\notin W.\end{cases}\tag{2.1}$$

LEMMA 2.1. *We assume that one among the following conditions holds*:

(i) $H\subset W$ (*continuous embedding*),

(ii) *the space W is reflexive and* $\mathbb{F}$ *is coercive on W.*

Then the functional J (2.1) *is weakly sequentially l.s.c. in the space H.*

Proof. The case in which the assumption (i) is fulfilled is obvious since J is the restriction of $\mathbb{F}$ to the space H. Let us suppose that the condition (ii) holds. If (v_n) is a sequence in the space H that converges weakly to $v \in H$ we have

$$v_n \rightharpoonup v \quad \text{in } X, \tag{2.2}$$

because the embedding $H \subset X$ remains continuous when both H and X are equipped with the weak topologies. The inequality

$$J(v) \leq \underline{\lim}\, J(v_n),$$

is obvious if $\lim J(v_n) = +\infty$ and we need only consider the case when the sequence $J(v_n)$ does not tend to $+\infty$. This amounts to saying that

$$\underline{\lim}\, J(v_n) < +\infty,$$

and it is then possible to find a subsequence (v_{n_k}) such that

$$\lim J(v_{n_k}) = \underline{\lim}\, J(v_n) < +\infty. \tag{2.3}$$

From the above relation we may assume $J(v_{n_k}) \in \mathbb{R}$ for every $k \in \mathbb{N}$ and by definition of J

$$v_{n_k} \in W, \qquad J(v_{n_k}) = \mathbb{F}(v_{n_k}), \tag{2.4}$$

for every $k \in \mathbb{N}$. Therefore, the relation (2.3) becomes

$$\lim \mathbb{F}(v_{n_k}) = \underline{\lim}\, J(v_n) < +\infty. \tag{2.5}$$

Under these conditions, the coerciveness of $\mathbb{F}$ on the space W shows that the sequence (v_{n_k}) is bounded in W and then has cluster points in W equipped with the weak topology (since the Banach space W is reflexive). From the continuity of the embedding $W \subset X$ when both W and X are equipped with the weak topologies and since the weak topology in X is separate, it follows that v is the unique cluster point of the sequence (v_{n_k}) in the weak topology of W. On the one hand, this shows that $v \in W$ and on the other we deduce that the whole sequence (v_{n_k}) converges to v in the weak topology of W (following a classical result since the bounded subsets of W are weakly relatively compact). Owing to the fact that the functional $\mathbb{F}$ is weakly sequentially l.s.c. in W we obtain

$$\mathbb{F}(v) \leq \underline{\lim}\, \mathbb{F}(v_{n_k}),$$

and from (2.1) and (2.5), the above inequality is nothing but

$$J(v) \leq \underline{\lim}\, J(v_n),$$

which completes the proof. □

Besides the previous hypotheses on $\mathbb{F}$ we now suppose for every $v \in H \cap W$, $v \neq 0$, that the function

$$t \in]0, +\infty[\to \frac{\mathbb{F}(tv) - \mathbb{F}(0)}{t^2} \in \mathbb{R} \tag{2.6}$$

is nondecreasing. This immediately shows for every $v \in H - \{0\}$ that the function (notice that $J(0) = \mathbb{F}(0) \in \mathbb{R}$ since $0 \in H \cap W$)

$$t \in]0, +\infty[\to \frac{J(tv) - J(0)}{t^2} \in \mathbb{R}^{\cdot},$$

is nondecreasing so that the assumption (1.7) of Theorem 1.1 is satisfied. Moreover, for every $v \in H$

$$\omega_\infty(v) = \begin{cases} 2 \lim\limits_{t \to +\infty} \dfrac{\mathbb{F}(tv)}{t^2} & \text{if } v \in W, \\ +\infty & \text{if } v \notin W. \end{cases} \tag{2.7}$$

Then, considering two linear selfadjoint operators $B \in \mathcal{L}(H)$ and $L \in \mathcal{L}(H)$ such that B is H-elliptic and L compact, it follows from our general results of §1 that the condition (cf. (2.7))

$$(Bv, v) - (Lv, v) + \omega_\infty(v) > 0 \quad \text{for every } v \in H \cap W, \quad v \neq 0, \tag{2.8}$$

ensures for every linear continuous form φ on H that the functional

$$v \in H \to \mathcal{J}_\varphi(v) = \tfrac{1}{2}((B - L)v, v) + J(v) - \varphi(v) \in \mathbb{R}^{\cdot}, \tag{2.9}$$

possesses at least one minimizer $u \in H$ provided one of the conditions (i) or (ii) of Lemma 2.1 is fulfilled. If so, we have

$$u \in H \cap W, \tag{2.10}$$

since the minimum value of $\mathcal{J}_\varphi$ in H cannot be obtained at a point u such that $J(u) = +\infty$ (as $0 \in H \cap W, J$ and $\mathcal{J}_\varphi$ are not constantly equal to $+\infty$). The minimization of $\mathcal{J}_\varphi$ in the space H is then equivalent to its minimization in the space $H \cap W$ on which the functional $\mathcal{J}_\varphi$ has the form

$$v \in H \cap W \to \mathcal{J}(v) = \tfrac{1}{2}((B - L)v, v) + \mathbb{F}(v) - \varphi(v) \in \mathbb{R},$$

and possesses at least one minimizer u (cf. (2.10)).

More particularly, if the functional $\mathbb{F}$ is Gâteaux-differentiable in the space W and if we write

$$\mathcal{J}_\varphi(u) \leq \mathcal{J}_\varphi(u + th) \in \mathbb{R} \quad \text{for every } t \in \mathbb{R} \text{ and every } h \in H \cap W,$$

we find without difficulty that u is a solution of the variational equation

$$((B - L)u, h) + \mathbb{F}'(u) \cdot h = \varphi(h) \quad \text{for every } h \in H \cap W. \tag{2.11}$$

Under the additional assumptions

$$\operatorname{Sp}(B^{-1/2} L B^{-1/2}) \subset]-\infty, 1[\cup \{\mu_0\}, \tag{2.12}$$

with $\mu_0 > 1$ a simple eigenvalue and

$$\mathbb{F} \geq \mathbb{F}(0), \tag{2.13}$$

a quite similar study (using Theorem 1.2 instead of Theorem 1.1) shows that there are at least two solutions

$$u_\alpha \in H \cap W - \{0\}, \qquad \alpha = 1, 2, \tag{2.14}$$

of the variational equation

$$(2.15) \qquad ((B-L)u_\alpha,h)+\mathbb{F}'(u_\alpha)=0 \quad \text{for every } h\in H\cap W,$$

when the conditions

$$(2.16) \qquad (Bv_0,v_0)-(Lv_0,v_0)+\omega_0(v_0)<0, \qquad \omega_0(-v_0)=\omega_0(v_0)$$

hold for some $v_0\in H\cap W$ where, for every $v\in H$ (cf. (1.16) and (2.1))

$$\omega_0(v)=\begin{cases} \lim\limits_{t\to 0_+} \dfrac{\mathbb{F}(tv)-\mathbb{F}(0)}{t^2} & \text{if } v\in W, \\ +\infty & \text{if } v\notin W. \end{cases}$$

When $\mathbb{F}$ is twice Gâteaux-differentiable at the origin, it follows from Remark 1.2 that (2.16) amounts to assuming that there exists $v_0\in H\cap W$ such that

$$(2.17) \qquad ((B-L)v_0,v_0)+\mathbb{F}''(0)\cdot(v_0,v_0)<0.$$

On the other hand, we know that the solutions u_α, $\alpha=1,2$ minimize the functional $\mathcal{J}_0$ (i.e. $\varphi=0$ in (2.9)) in each of the two open half-spaces determined by the hyperplane $B^{-1/2}(\text{Ker}\{\mu_0 I-B^{-1/2}LB^{-1/2}\}^\perp)$ (cf. Theorem 1.2 and Remark 1.2). Hence $u_\alpha\neq 0$, $\alpha=1,2$ and u_1 and u_2 belong to the space $H\cap W$: Indeed, as we have proved in Theorem 2.1, we have $\mathcal{J}_0(u_\alpha)<\mathcal{J}_0(0)=J(0)=\mathbb{F}(0)\in\mathbb{R}$ and in particular $J(u_\alpha)\in\mathbb{R}$, $\alpha=1,2$. By definition of J (cf. (2.1)) we deduce that $u_\alpha\in H\cap W$, $\alpha=1,2$, which proves (2.14).

Remark 2.1. Let us denote by σ_0 a normalized vector (i.e. $\|\sigma_0\|_H=1$) of the space $\text{Ker}(\mu_0 I-B^{-1/2}LB^{-1/2})$ and let ψ_0 be the vector $\psi_0=B^{-1/2}\sigma_0$. Then, the condition (2.16) is satisfied in the following main two cases

1) $\psi_0\in W$ and $\mathbb{F}''(0)\cdot(\psi_0,\psi_0)<\mu_0-1$.

2) $\mathbb{F}''(0)=0$ and the space $H\cap W$ is dense in H.

Indeed, it is immediate that $(\mu_0 B-L)\psi_0=0$ and $(B\psi_0,\psi_0)=1$ so that the first condition is readily equivalent to (2.17) with $v_0=\psi_0$. In the second case, the result follows from the fact that the set $\{v\in H,((B-L)v,v)<0\}$ is open in H and not empty since it contains the vector ψ_0.

A particularly interesting situation is described as follows: let Ω be an open subset of $\mathbb{R}^N$. We take

$$(2.18) \qquad W=L^q(\Omega),$$

for some $1\leq q\leq +\infty$. Now, for some integer $m\geq 0$,

$$(2.19) \qquad H_0^m(\Omega)\subset H\subset H^m(\Omega),$$

denotes a closed subspace of $H^m(\Omega)$ equipped with the inner product $(\cdot,\cdot)$ equivalent to the inner product induced by $H^m(\Omega)$. In this case, the separate locally convex space X may be chosen as

$$X=L_{\text{loc}}^{\min(q,2)}(\Omega).$$

From (2.19), we deduce that the operators B and L in H induce operators

$$\tilde{B}\in\mathcal{L}(H,\mathcal{D}'(\Omega)), \qquad \tilde{L}\in\mathcal{L}(H,\mathcal{D}'(\Omega)),$$

through the formulas

$$\begin{aligned}(2.20)\qquad &\forall v\in H,\quad \langle \tilde{B}v,\theta\rangle=(Bv,\theta)\quad \forall\theta\in\mathcal{D}(\Omega),\\ &\forall v\in H,\quad \langle \tilde{L}v,\theta\rangle=(Lv,\theta)\quad \forall\theta\in\mathcal{D}(\Omega).\end{aligned}$$

Meanwhile, the continuous embedding

$$\mathcal{D}(\Omega)\subset L^q(\Omega)\qquad (1\leq q\leq +\infty)$$

allows us to define the distribution

$$\tilde{f}(v)=\mathbb{F}'(v)|_{\mathcal{D}(\Omega)}\in\mathcal{D}'(\Omega)$$

for every $v\in L^q(\Omega)$.

Then, for $u^*\in H^{-m}(\Omega)=(H_0^m(\Omega))'$ and denoting by φ any linear continuous extension of u^* to the space H, each solution $u\in H\cap L^q(\Omega)$ of the equation (2.11) is also a solution of the equation

$$(2.21)\qquad \tilde{B}u-\tilde{L}u+\tilde{f}(u)=u^*,$$

in the distributional sense. Let us notice that *the elements* $\mathbb{F}'(v)$ *and* $\tilde{f}(v)$ *may be identified* for every $v\in L^q(\Omega)$ when $1\leq q<+\infty$ (because of the denseness of the space $\mathcal{D}(\Omega)$ in the space $L^q(\Omega)$). This remains true for $q=+\infty$ if $\mathbb{F}'(v)$ belongs to the space $L^1(\Omega)$ for every $v\in L^\infty(\Omega)$ (because each element of $L^1(\Omega)$ is entirely determined by its associated distribution).

Remark 2.2. It is easily possible to generalize this situation to the case when W is a closed subspace of $W^{k,q}(\Omega)$ for some integer $k\geq 0$ or a product of such spaces.

We shall now examine particular cases of equations such as (2.21) in the next section.

3. Application to semi-linear elliptic problems. Let $f\colon\Omega\times\mathbb{R}\to\mathbb{R}$ be a Carathéodory function; i.e.,

(3.1) for every $y\in\mathbb{R}, f(\cdot,y)\colon\Omega\to\mathbb{R}$ is measurable,
for almost all $x\in\Omega, f(x,\cdot)\colon\mathbb{R}\to\mathbb{R}$ is continuous.

In addition, we assume for almost all $x\in\Omega$ that

$$(3.2)\qquad f(x,y)\geq 0\quad \forall y\geq 0,\qquad f(x,y)\leq 0\quad \forall y\leq 0$$

(which in particular requires $f(x,0)=0$) and

(3.3) the function $y\to f(x,y)/y$ is nondecreasing for $y>0$ and nonincreasing for $y<0$.

For any real number $q\geq 1$ we call $(\mathcal{P}_{q,q^*})$ the property

$(\mathcal{P}_{q,q^*})$ there exist a function $a\in L^{q^*}(\Omega)(1/q+1/q^*=1)$ and a constant $b>0$ such that for almost all $x\in\Omega$ $|f(x,y)|\leq a(x)+b|y|^{q-1}$, for every $y\in\mathbb{R}$,

and for $q=+\infty$ we call $(\mathcal{P}_{\infty,1})$ the property

$(\mathcal{P}_{\infty,1})$ for every $y\in\mathbb{R}, f(\cdot,y)\in L^1(\Omega)$.

Remark 3.1. It is not difficult to see, when the set Ω is bounded, that for $1 \leq q_1 \leq q_2 \leq +\infty$ one has

$$(\mathcal{P}_{q_1,q^*_1}) \Rightarrow (\mathcal{P}_{q_2,q^*_2}) \Rightarrow (\mathcal{P}_{\infty,1}).$$

From now on, we assume that the function f verifies the property $(\mathcal{P}_{q,q^*})$ for some $2 \leq q \leq +\infty$[1] and for every measurable function v on Ω, we define the measurable function $\tilde{f}(v)$ by

$$\tilde{f}(v)(x) = f(x, v(x)), \tag{3.4}$$

for almost all $x \in \Omega$ (the measurability of $\tilde{f}(v)$ is not obvious at all, cf. [4, Prop. 1.1, p. 218]). We shall denote by $\mathbb{F}$ the functional

$$\mathbb{F}(v) = \int_\Omega F(x, v(x))\, dx, \tag{3.5}$$

where, for almost all $x \in \Omega$

$$F(x,y) = \int_0^y f(x,t)\, dt \tag{3.6}$$

for every $y \in \mathbb{R}$. Then, we have

$$\tilde{f} \in \mathcal{C}^0(L^q(\Omega), L^{q^*}(\Omega)), \tag{3.7}$$

and the functional $\mathbb{F}$ is well defined and Fréchet differentiable in the space $L^q(\Omega)$ with derivative

$$\mathbb{F}' = \tilde{f}. \tag{3.8}$$

For $2 \leq q < +\infty$, these results can be found in Krasnosel'skii [5, Thm. 2.1, Thm. 2.3, Lemma 5.1] and are valid in a more general framework. For $q = +\infty$, they can be derived from the assumption $(\mathcal{P}_{\infty,1})$ and the fact that the function $f(x, \cdot)$ is nondecreasing for almost all $x \in \Omega$ (which follows from (3.2)–(3.3)) through classical theorems in integration theory (see also [8, Thm. 1.3] for a more general result).

We shall now give a simple technical lemma.

LEMMA 3.1. *Let g: $[0, +\infty[\to [0, +\infty[$ be a continuous function such that the function*

$$t \in]0, +\infty[\to \frac{g(t)}{t} \in [0, +\infty[,$$

is nondecreasing (and then $g(0) = 0$). Let G be defined by

$$t \in [0, +\infty[\to G(t) = \int_0^t g(s)\, ds \in [0, +\infty[.$$

Then

(i) *the function*

$$t \in]0, +\infty[\to \frac{G(t)}{t^2} \in [0, +\infty[$$

[1] Clearly, f cannot verify $(\mathcal{P}_{q,q^*})$ for some $1 \leq q < 2$ and (3.3) at the same time, except when $f = 0$.

is nondecreasing,

(ii) *the limits*

$$l_g = \lim_{t\to+\infty} \frac{g(t)}{t}, \qquad l_G = \lim_{t\to+\infty} \frac{G(t)}{t^2}$$

exist (possibly $+\infty$) and verify

$$l_g = 2l_G.$$

Proof. From the properties of g we have for $t>0$

$$G(t) = \int_0^t g(s)\,ds = \int_0^t \left(\frac{g(s)}{s}\right) s\,ds \le \frac{g(t)}{t}\int_0^t s\,ds,$$

hence

$$G(t) \le \tfrac{1}{2}tg(t), \tag{3.9}$$

for every $t>0$. This implies that the derivative of the function $G(t)/t^2$ is ≥ 0 and proves (i).

The existence of the limits l_g and l_G follows from the growth properties of the function $g(t)/t$ and $G(t)/t^2$. Moreover, from (3.9)

$$\frac{G(t)}{t^2} \le \frac{1}{2}\frac{g(t)}{t},$$

for every $t>0$ and consequently

$$l_G \le \frac{l_g}{2}.$$

Let us now show that $l_G \ge l_g/2$. First, we assume that $l_g = +\infty$: for every $M>0$, there exists $t_0>0$ such that

$$t \ge t_0 \Rightarrow \frac{g(t)}{t} \ge M.$$

Then, for $t \ge t_0$

$$G(t) = G(t_0) + \int_{t_0}^t g(s)\,ds \ge G(t_0) + \int_{t_0}^t Ms\,ds,$$

or equivalently

$$\frac{G(t)}{t^2} \ge \frac{G(t_0)}{t^2} + \frac{M}{2}\frac{(t-t_0)^2}{t^2},$$

which proves that $l_G \ge M/2$ and then $l_G = +\infty$.

Now, we assume that $l_g \in \mathbb{R}$: for every $\varepsilon>0$ there exists $t_0>0$ such that

$$t \ge t_0 \Rightarrow \frac{g(t)}{t} \ge l_g - \varepsilon.$$

Thus, for $t \ge t_0$

$$G(t) = G(t_0) + \int_{t_0}^t g(s)\,ds \ge G(t_0) + \int_{t_0}^t (l_g - \varepsilon)s\,ds,$$

or equivalently

$$\frac{G(t)}{t^2} \geq \frac{G(t_0)}{t^2} + \frac{(l_g-\varepsilon)}{2}\frac{(t-t_0)^2}{t^2},$$

which proves that $l_G \geq (l_g-\varepsilon)/2$ and then $l_G \geq l_g/2$ since ε may be taken arbitrarily small. □

As concerns the functional $\mathbb{F}$ (3.5), the previous lemma leads to

PROPOSITION 3.1. *Under our assumptions on the function f,*

(i) *the functional $\mathbb{F}$ is convex,*

(ii) *for every $v \in L^q(\Omega)$ the function*

$$t \in]0, +\infty[\to \frac{\mathbb{F}(tv)}{t^2} \in \mathbb{R}$$

is nondecreasing,

(iii) *for every $v \in L^q(\Omega)$, $\lim_{t\to+\infty} \mathbb{F}(tv)/t^2$ exists (possibly $+\infty$) and*

$$\lim_{t\to+\infty} \frac{1}{t}\int_\Omega \tilde{f}(tv)v = 2\lim_{t\to+\infty} \frac{\mathbb{F}(tv)}{t^2}. \tag{3.10}$$

Proof. Let v_1 and v_2 be two given elements of the space $L^q(\Omega)$ and $0 \leq \lambda \leq 1$ a real number. For almost all $x \in \Omega$, the derivative of the function $F(x, \cdot)$ is the function $f(x, \cdot)$ and it follows from (3.2)–(3.3) that $f(x, \cdot)$ is nondecreasing. Hence $F(x, \cdot)$ is convex. Thus

$$F(x,\lambda v_1(x)+(1-\lambda)v_2(x)) \leq \lambda F(x,v_1(x))+(1-\lambda)F(x,v_2(x)),$$

for almost all $x \in \Omega$. By definition of $\mathbb{F}$, this provides

$$\mathbb{F}(\lambda v_1+(1-\lambda)v_2) \leq \lambda\mathbb{F}(v_1)+(1-\lambda)\mathbb{F}(v_2).$$

Now, let $v \in L^q(\Omega)$ be given. Because of (3.7) we may set

$$g(t) = \int_\Omega \tilde{f}(tv)v, \tag{3.11}$$

for every $t \geq 0$ and according to (3.8)

$$g(t) = \mathbb{F}'(tv)\cdot v,$$

which shows that the function g is continuous and that

$$\mathbb{F}(tv) = \int_0^t g(s)\,ds.$$

From the properties (3.2)–(3.3) we deduce for almost all $x \in \Omega$

$$f(x,tv(x))v(x) \geq 0,$$

and the function

$$t \in]0, +\infty[\to \frac{f(x,tv(x))v(x)}{t} \in \mathbb{R},$$

is nondecreasing. It follows that we may apply Lemma 3.1 to the function g (3.11) and the proof is complete. □

From (3.7)–(3.8) and the results of §2, we conclude that the equation

$$\tilde{B}u-\tilde{L}u+\tilde{f}(u)=u^*, \tag{3.12}$$

posed in a closed subspace H of $H^m(\Omega)$ which contains $H_0^m(\Omega)$ ($m\geq 0$ a given integer) possesses at least one solution $u\in H\cap L^q(\Omega)$ for every $u^*\in H^{-m}(\Omega)$ under the following assumptions.

1) The operators $B\in\mathcal{L}(H)$ and $L\in\mathcal{L}(H)$ inducing the operators $\tilde{B}$ and $\tilde{L}$ as indicated in (2.20) are both selfadjoint, B is H-elliptic and L is compact.

2) One of the conditions (corresponding to the assumptions of Lemma 2.1)

(i) $H\subset L^q(\Omega)$ (continuous embedding),

(ii) $q<+\infty$ and the functional $\mathbb{F}$ is coercive in $L^q(\Omega)$,

is fulfilled.

3) For every $v\in H\cap L^q(\Omega)$, $v\neq 0$

$$(Bv,v)-(Lv,v)+\omega_\infty(v)>0. \tag{3.13}$$

(Notice that the weak lower semicontinuity of the functional $\mathbb{F}$ in $L^q(\Omega)$ follows from its convexity and its continuity.)

We shall give a very simple condition on the function f which ensures that the inequality (3.13) holds. We recall (cf. (2.7) and Proposition 3.1 (iii))

$$\omega_\infty(v)=\lim_{t\to+\infty}\frac{1}{t}\int_\Omega \tilde{f}(tv)v \tag{3.14}$$

for every $v\in H\cap L^q(\Omega)$.

Lemma 3.2. *Under our assumptions on the function f, for every real number l the conditions*

(i) *for almost all $x\in\Omega$*

$$\lim_{y\to\pm\infty}\frac{f(x,y)}{y}\geq l, \tag{3.15}$$

(ii) *for every $v\in L^q(\Omega)$*

$$\lim_{t\to+\infty}\frac{1}{t}\int_\Omega \tilde{f}(tv)v\geq l\|v\|^2_{L^2(\Omega)} \tag{3.16}$$[2]

are equivalent.

Proof. Let us suppose that the condition (i) holds. From the growth property (3.3) it suffices to prove that

$$\lim \frac{1}{n}\int_\Omega \tilde{f}(nv)v\geq l\|v\|^2_{L^2(\Omega)},$$

for every $v\in L^q(\Omega)$. Since (3.16) is obvious for $v=0$, we shall assume $v\neq 0$. Let $\Omega'\subset\Omega$ be the measurable set

$$\Omega'=\{x\in\Omega, v(x)\neq 0\}.$$

It follows from (3.2) and the growth property (3.3) that the sequence $(\tilde{f}(nv)/nv)$ is a nondecreasing sequence of nonnegative measurable functions on the set Ω'. Then, so is the sequence $((1/n)\tilde{f}(nv)v)$ since $(1/n)f(x,nv(x))v(x)=(f(x,nv(x))/nv(x))v^2(x)$ for

[2] where $\|v\|_{L^2(\Omega)}=+\infty$ for every $v\notin L^q(\Omega)\cap L^2(\Omega)$ when Ω is not bounded.

every $x\in\Omega'$. From the definition of the set Ω' and according to (3.15), this last identity shows that

$$\lim \frac{1}{n}\tilde{f}(nv)v\geq lv^2 \quad \text{a.e. in } \Omega'.$$

Thus, from classical theorems in integration theory, we deduce

$$\lim \frac{1}{n}\int_{\Omega'}\tilde{f}(nv)v\geq l\int_{\Omega'}v^2,$$

and (3.16) follows since the integrals over the sets Ω and Ω' are the same by definition of Ω'.

Now, we suppose that (ii) holds. If the set Ω is bounded, for every integers $k\geq 1$ and $n\geq 1$ we define the measurable set

$$\Omega_{k,n}=\left\{x\in\Omega,\ \frac{f(x,n)}{n}\leq l-\frac{1}{k}\right\}, \tag{3.17}$$

in order that the set

$$\Omega_k=\bigcap_{n\geq 1}\Omega_{k,n} \tag{3.18}$$

is measurable. Furthermore, from the growth property (3.3), Ω_k is equivalently defined by

$$\Omega_k=\left\{x\in\Omega,\ \frac{f(x,y)}{y}\leq l-\frac{1}{k}\ \text{for every}\ y>0\right\}. \tag{3.19}$$

Since Ω is bounded, so is Ω_k and it follows from (3.19) for every $y>0$ that the nonnegative measurable function $f(\cdot,y)/y$ is integrable, with

$$\int_{\Omega_k}\frac{f(x,y)}{y}dx\leq\left(l-\frac{1}{k}\right)\operatorname{meas}(\Omega_k),$$

whence

$$\lim_{y\to+\infty}\int_{\Omega_k}\frac{f(x,y)}{y}dx\leq\left(l-\frac{1}{k}\right)\operatorname{meas}(\Omega_k). \tag{3.20}$$

But taking $v=\chi_k$ characteristic function of Ω_k in (3.16) we find

$$\lim_{t\to+\infty}\int_{\Omega_k}\frac{f(x,t)}{t}dx\geq l\operatorname{meas}(\Omega_k).$$

Together with (3.20) we conclude that $\operatorname{meas}(\Omega_k)=0$. Consequently, the measurable set $\bigcup_{k\geq 1}\Omega_k$ is of measure 0. This means for almost all $x\in\Omega$ that x does not belong to Ω_k for any $k\geq 1$. In other words, from (3.17)–(3.18), there exists an integer $n_k\geq 1$ such that

$$\frac{f(x,n_k)}{n_k}>l-\frac{1}{k}.$$

Using the growth property (3.3) once again, we obtain

$$\frac{f(x,y)}{y}>l-\frac{1}{k},$$

for every $y \geq n_k$. Thus

$$\lim_{y \to +\infty} \frac{f(x,y)}{y} > l - \frac{1}{k},$$

for every $k \geq 1$ and finally

$$\lim_{y \to +\infty} \frac{f(x,y)}{y} \geq l.$$

The same process shows that $\lim_{y \to -\infty} f(x,y)/y \geq l$ for almost all $x \in \Omega$ and the condition (i) follows when the set Ω is bounded.

When Ω is not bounded, it is nevertheless the countable union of a family (Ω_n) of open bounded subsets of $\mathbb{R}^N$ such that $\Omega_{n+1} \supset \Omega_n$ for every $n \in \mathbb{N}$. If (ii) holds, we have in particular

$$\lim_{t \to +\infty} \frac{1}{t} \int_{\Omega_n} \tilde{f}(tv)v \geq l \|v\|^2_{L^2(\Omega_n)}$$

for every $v \in L^q(\Omega_n)$. According to the first step, this implies

$$\lim_{y \to \pm\infty} \frac{f(x,y)}{y} \geq l$$

for almost all $x \in \Omega_n$. Since a countable union of sets of measure 0 is a set of measure 0, we conclude

$$\lim_{y \to \pm\infty} \frac{f(x,y)}{y} \geq l$$

for almost all $x \in \Omega$ and the proof is complete. □

Since the function $f(\cdot, y)/y$ is nonnegative for every $y \in \mathbb{R} - \{0\}$ (cf. (3.2)) we can define $\bar{l} \in [0, +\infty]$ by

$$\bar{l} = \sup\left\{ l \in \mathbb{R},\ \lim_{y \to \pm\infty} \frac{f(x,y)}{y} \geq l \text{ a.e.} \right\}. \tag{3.21}$$

With this notation, it is obvious for almost all $x \in \Omega$ that

$$\lim_{y \to \pm\infty} \frac{f(x,y)}{y} \geq \bar{l}, \tag{3.22}$$

and the previous lemma immediately leads to

PROPOSITION 3.2. *Under our assumptions on the function f, we have*

$$\lim_{t \to +\infty} \frac{1}{t} \int_{\Omega} \tilde{f}(tv)v \geq \bar{l} \|v\|^2_{L^2(\Omega)},$$

for every $v \in L^q(\Omega)$ and if $l \in \mathbb{R}$ denotes any real number such that

$$\lim_{t \to +\infty} \frac{1}{t} \int_{\Omega} \tilde{f}(tv)v \geq l \|v\|^2_{L^2(\Omega)}$$

for every $v \in L^q(\Omega)$, then $l \leq \bar{l}$.

It follows from Proposition 3.2 that the inequality (3.13) holds when

$$(Bv,v)-(Lv,v)+\bar{l}\|v\|^2_{L^2(\Omega)}>0,$$

for every $v\in H\cap L^q(\Omega)$, $v\neq 0$. This is in particular the case for any operators B and L when $\bar{l}=+\infty$.

Remark 3.2. Due to the definition of the property $(\mathcal{P}_{q,q^*})$, the case $\bar{l}=+\infty$ requires the restriction $q>2$ for the values of the real number q.

We shall sum up the results we have obtained in the following theorem.

THEOREM 3.1. *Let Ω be an open (bounded or unbounded) subset of $\mathbb{R}^N$ and $m\geq 0$ a given integer. Let*

$$H_0^m(\Omega)\subset H\subset H^m(\Omega)$$

be a closed subspace of $H^m(\Omega)$ equipped with the inner product $(\cdot,\cdot)$ equivalent to the inner product induced by $H^m(\Omega)$. We consider a function f: $\Omega\times\mathbb{R}\to\mathbb{R}$ verifying the properties (3.1)–(3.3) *and $(\mathcal{P}_{q,q^*})$ for some $2\leq q\leq +\infty$ $(1/q+1/q^*=1)$ and we set*

$$\bar{l}=\sup\left\{l\in\mathbb{R},\ \lim_{y\to\pm\infty}\frac{f(x,y)}{y}\geq l \text{ a.e.}\right\}\in[0,+\infty]. \tag{3.23}$$

Then $1\leq q^*\leq 2$ and for every $v\in L^q(\Omega)$ we set

$$\tilde{f}(v)(x)=f(x,v(x)),$$

for almost all $x\in\Omega$. The operator $\tilde{f}$ verifies

$$\tilde{f}\in\mathcal{C}^0(L^q(\Omega),L^{q^*}(\Omega)),$$

and the functional

$$v\in L^q(\Omega)\to\mathbb{F}(v)=\int_\Omega dx\int_0^{v(x)}f(x,s)\,ds \tag{3.24}$$

is well defined, convex and Fréchet differentiable with

$$\mathbb{F}'(v)=\tilde{f}(v)\in L^{q^*}(\Omega), \tag{3.25}$$

for every $v\in L^q(\Omega)$.

Now, we assume that one among the conditions

(i) *$H\subset L^q(\Omega)$ (continuous embedding),*

(ii) *$q\neq+\infty$ and $\mathbb{F}$ is coercive in the space $L^q(\Omega)$,*

holds. We consider two selfadjoint operators $B\in\mathcal{L}(H)$ and $L\in\mathcal{L}(H)$ such that B is H-elliptic and L is compact and assume that the inequality

$$(Bv,v)-(Lv,v)+\bar{l}\|v\|^2_{L^2(\Omega)}>0 \tag{3.26^3}$$

holds for every $v\in H\cap L^q(\Omega)$, $v\neq 0$.

Then, $\tilde{B}\in\mathcal{L}(H,\mathcal{D}'(\Omega))$ and $\tilde{L}\in\mathcal{L}(H,\mathcal{D}'(\Omega))$ being defined from B and L as indicated in (2.20), *the equation*

$$\tilde{B}u-\tilde{L}u+\tilde{f}(u)=u^* \tag{3.27}$$

[3] Let us recall that the case $\bar{l}=+\infty$ is a frequent one.

has at least one solution $u \in H \cap L^q(\Omega)$ *for every* $u^* \in H^{-m}(\Omega)$. *More precisely, if* φ *denotes any linear continuous extension of* u^* *to the space* H, *the functional*

$$v \in H \cap L^q(\Omega) \to \mathcal{J}_\varphi(v) = \tfrac{1}{2}((B-L)v,v) + \mathbb{F}(v) - \varphi(v) \in \mathbb{R} \tag{3.28}$$

has at least one minimizer and each minimizer of $\mathcal{J}_\varphi$ *is a solution of the equation* (3.27).

If the open set Ω is bounded and has a Lipschitzian boundary, the Sobolev's embedding theorems are available to decide whether the continuous embedding $H_0^m(\Omega) \subset L^q(\Omega)$ holds. Since we observed in Remark 3.1 that the property $(\mathcal{P}_{q,q*})$ gets weaker and weaker in proportion as q grows when Ω is bounded, we deduce that the most general form of Theorem 3.1 with Ω bounded corresponds to the values

$$\begin{aligned} &q = +\infty && \text{when } 2m > N, \\ &2 \le q < +\infty && \text{arbitrarily large when } 2m = N, \\ &q = \frac{2N}{N-2m} && \text{when } 2m < N, \end{aligned} \tag{3.29}$$

or else

$$\text{some } \frac{2N}{N-2m} < q < +\infty \quad \begin{array}{l}\text{such that the functional } \mathbb{F} \text{ is coercive in the} \\ \text{space } L^q(\Omega) \text{ when } 2m < N.\end{array} \tag{3.30}$$

As concerns the case $u^* = 0$, namely

$$\tilde{B}u - \tilde{L}u + \tilde{f}(u) = 0, \tag{3.31}$$

we know the existence of at least two solutions $u_\alpha \ne 0$, $\alpha = 1,2$ in the space $H \cap L^q(\Omega)$ under some additional assumptions listed in §2. Let us recall that the first one is concerned with the spectrum of the operator $B^{-1/2}LB^{-1/2}$, namely

$$\mathrm{Sp}(B^{-1/2}LB^{-1/2}) \subset]-\infty, 1[\cup \{\mu_0\}, \tag{3.32}$$

with $\mu_0 > 1$ simple eigenvalues. The condition $\mathbb{F} \ge \mathbb{F}(0)$ is here satisfied and according to Remark 2.1 it remains to assume (for instance) that the operator $\tilde{f}(=\mathbb{F}'$, cf. (3.8)) is Gâteaux-differentiable at the origin of $L^q(\Omega)$ and that there exists $v_0 \in H \cap L^q(\Omega)$ such that

$$(Bv_0, v_0) - (Lv_0, v_0) + \mathbb{F}''(0) \cdot (v_0, v_0) < 0. \tag{3.33}$$

It can be immediately verified that the Gâteaux-differentiability of $\tilde{f}$ at the origin follows from the assumption that the function $f(x, \cdot)$ is differentiable at the origin for almost all $x \in \Omega$, with derivative $f_y(x,0)$ verifying $f_y(\cdot, 0) \in L^{(q/2)*}(\Omega)$ (where we have set $(q/2)^* = q/(q-2)$ for $q > 2$ and $(q/2)^* = 1$ for $q = +\infty$). If so, the condition (3.33) may be rewritten as

$$(Bv_0, v_0) - (Lv_0, v_0) + \int_\Omega f_y(\cdot, 0) v_0^2 < 0.$$

With the notation of Remark 2.1, this is in particular the case when $f_y(\cdot, 0) = 0$ and $H \cap L^q(\Omega)$ is dense in H or else when the generalized eigenvector ψ_0 belongs to $L^q(\Omega)$ with $\int_\Omega f_y(\cdot, 0)\psi_0^2 < \mu_0 - 1$. Since the vector $\psi_0 \in H$ verifies the equation $\mu_0 B\psi_0 = L\psi_0$

and hence the equation $\mu_0\tilde{B}\psi_0=\tilde{L}\psi_0$, it is interesting to notice that the condition $\psi_0\in L^q(\Omega)$ is closely related to a *regularity condition about the operator* $\tilde{B}$ (which has the usual meaning when $\tilde{B}$ is a differential operator).

A classical example is the equation

$$\begin{aligned}(3.34)\qquad &(-1)^m\Delta^m u-\lambda u+\tilde{f}(u)=u^*\in H^{-m}(\Omega),\\ &u\in H_0^m(\Omega),\end{aligned}$$

where λ is a given real number (or a suitable function: for instance, $\lambda\in L^\infty(\Omega)$) and the set Ω is bounded. In this case, we have[4] $B=I$, $L=((-1)^m\Delta^m)^{-1}$ and the space $H_0^m(\Omega)$ is equipped with the inner product $(u,v)=\int_\Omega\nabla^m u\nabla^m v$ where $\nabla^m=\Delta^k$ if $m=2k$, $\nabla^m=\nabla\Delta^k$ if $m=2k+1$. Since $H=H_0^m(\Omega)$, the unique extension φ of u^* to the space $H_0^m(\Omega)$ is u^* itself. In particular, the case $\varphi=0$ (for which we know the existence of at least two solutions $u_\alpha\neq0$, $\alpha=1,2$ in the space $H_0^m(\Omega)\cap L^q(\Omega)$ under the suitable additional assumptions listed above) is equivalent to the case $u^*=0$.

Of course, we may consider more general operators and more general boundary conditions: if the boundary $\partial\Omega$ is divided into two parts Γ_0 and Γ_1 with meas$(\Gamma_0)>0$ and with the choice

$$(3.35)\qquad H=\{v\in H^1(\Omega),\ v=0\text{ on }\Gamma_0\},$$

we solve, for instance,

$$\begin{aligned}(3.36)\qquad &-\partial_j(a_{ij}\partial_i u)-\lambda u+\tilde{f}(u)=u^*,\\ &u=0\quad\text{on }\Gamma_0,\qquad \frac{\partial u}{\partial\nu}=g\quad\text{on }\Gamma_1,\end{aligned}$$

where λ is a given real number (or a suitable function), g a given element of $L^2(\Gamma_1)$ and the operator $-\partial_j(a_{ij}\partial_i)$ fulfills the usual conditions of symmetry ($a_{ij}=a_{ji}$), regularity ($a_{ij}\in L^\infty(\Omega)$) and uniform ellipticity. The space H being equipped with the inner product $(u,v)=\int_\Omega\nabla u\nabla v$, the operator B is defined by

$$(Bu,v)=\int_\Omega a_{ij}\partial_i u\partial_j v,$$

for every pair $(u,v)\in H\times H$, which is a weak form of

$$\begin{aligned}&-\Delta Bu=-\partial_j(a_{ij}\partial_i u),\\ &Bu=0\quad\text{on }\Gamma_0,\qquad \frac{\partial Bu}{\partial\nu}=0\quad\text{on }\Gamma_1,\end{aligned}$$

and the operator L is defined by

$$\begin{aligned}&-\Delta Lu=u,\\ &Lu=0\quad\text{on }\Gamma_0,\qquad \frac{\partial Lu}{\partial\nu}=0\quad\text{on }\Gamma_1.\end{aligned}$$

The right-hand side u^* of (3.36) has to be taken in the space $L^1(\Omega)$ if $N=1$, in $L^{1+\varepsilon}(\Omega)$ for some $\varepsilon>0$ if $N=2$ and in $L^{2N/N+2}(\Omega)$ if $N\geq3$ if we want the element $\partial u/\partial\nu$ to be well defined in the space $H^{-1/2}(\partial\Omega)$. We obtain the existence of at least one solution of

[4] L has to be modified when λ is a function.

(3.36) by minimizing the functional $\mathcal{J}_\varphi$ (3.28) in which the linear extension φ of u^* to the space H is given by

$$\varphi(v)=\int_\Omega u^*v+\int_{\Gamma_1} gv,$$

for every $v\in H$. The case $\varphi=0$ corresponds then to the choices $u^*=0$ *and* $g=0$.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] J. Cea, *Optimisation, théorie et algorithmes*, Dunod, Paris, 1971.

[3] P. G. Ciarlet and P. Rabier, *Les équations de von Kármán*, Lecture Notes in Mathematics 826, Springer-Verlag, Berlin, 1980.

[4] I. Ekeland and R. Temam, *Analyse convexe et problèmes variationels*, Dunod, Gauthier-Villars, Paris, 1974.

[5] M. A. Krasnosel'skii, *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon Press, New York, 1964.

[6] P. Rabier, Doctoral dissertation, Univ. Pierre et Marie Curie, Paris, 1980.

[7] ———, *A general study of nonlinear problems with three solutions in Hilbert spaces*, to appear.

[8] ———, *Definition and properties of a particular notion of convexity*, to appear.

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1502-0012 $01.25/0

INTEGRALS OF HIGH-PASS FUNCTIONS*

B. F. LOGAN†

Abstract. A function $h(x)$ whose Fourier transform vanishes over the interval $(-\Omega,\Omega)$ is called a high-pass function. This paper develops topics in the mathematical theory of high-pass functions. The definition of high-pass functions is extended to functions $h(x)$ which do not have Fourier transforms by requiring that

(i) $\int_x^{x+1}|h(t)|\,dt$ is uniformly bounded for $-\infty<x<\infty$,

(ii) $\int_{-\infty}^{\infty}h(x)g(x)\,dx=0$, for all $g(x)$ in L_1 of the form $g(x)=\int_{-\Omega}^{\Omega}G(\omega)e^{i\omega x}\,d\omega$.

For such $g(x)$, the condition (i) insures that the integral (ii) is absolutely convergent.

Within this extended definition, it is shown that a high-pass function $h(x)$ has a bounded unbiased integral $h^{(-1)}(x)$, which is also a high-pass function, for which $h^{(-1)}(x)-h^{(-1)}(y)=\int_y^x h(t)\,dt$. It is given by convolution of $h(x)$ with any L_1 function $K(x)$ of the form

(iii) $K(x)=\frac{1}{2}\operatorname{sgn}x-g(x)$, where $g(x)$ is the restriction to the real line of an entire function of exponential type Ω,

or, what is shown to be equivalent, by convolution of $h(x)$ with any function in L_1 of the form

(iv) $K(x)=(1/2\pi)\int_{-\infty}^{\infty}\hat{K}(\omega)e^{i\omega x}\,d\omega$ where $\hat{K}(\omega)=1/i\omega$ for $|\omega|\geq\Omega$.

From the convolution representation of $h^{(-1)}(x)$, the following results are obtained. If $h(x)$ is a high-pass function:

(v) The Fourier integral of $h(x)$ is summable $(C,1)$ to zero in the spectral gap $(-\Omega,\Omega)$.

(vi) If $\lim_{x\to\infty}\int_x^{x+1}|h(t)|\,dt=0$, then $\lim_{x\to\infty}\int_0^x h(t)\,dt$ exists.

(vii) $|h^{(-1)}(x)|\leq(\frac{1}{2}+1/2\pi)\sup_t\int_0^{\pi/\Omega}|h(t+u)|\,du$.

1. Introduction. Signals commonly encountered in communication engineering, by virtue of filtering or modulation operations, are for all practical purposes void of low frequencies, i.e. their Fourier spectrum omits an interval about the origin, say $(-\Omega,\Omega)$. Such functions of a real variable are called *high-pass functions* (cf. [2]).

Our first task is to give a suitable definition of such functions. A class of such functions which has reasonable closure properties will naturally include functions which do not have Fourier transforms, so we cannot just define this class of functions to be those functions whose Fourier transforms vanish over the interval $(-\Omega,\Omega)$ or more generally over an open set E. Instead we shall use a general recipe to construct a space of such functions, which is to bound the growth of such functions $h(t)$ and to require them to be orthogonal to a class of test functions $g(t)$, i.e.

$$\int_{-\infty}^{\infty}h(t)\overline{g(t)}dt=0. \tag{1}$$

These test functions must have Fourier transforms supported on E and be of sufficiently rapid decay that the integral (1) is absolutely convergent. Note that if the set E is symmetric with respect to the origin (as it is in the case where $E=(-\Omega,\Omega)$), then the space of test functions can be chosen to consist of real-valued functions $g(t)$, and (1) can then be replaced by

$$\int_{-\infty}^{\infty}h(t)g(t)\,dt=0.$$

We introduce the following spaces of test functions. For $\Omega>0$, $B(\Omega)$ denotes the collection of functions which are restrictions to the real line of entire functions of exponential type $\leq\Omega$. For each p with $1\leq p\leq\infty$, we let $B_p(\Omega)$ denote the subset of

*Received by the editors March 10, 1982, and in revised form August 27, 1982.

†Bell Laboratories, Murray Hill, New Jersey 07974.

$B(\Omega)$ consisting of those functions which are also in $L_p(-\infty,\infty)$. We call the functions in $B_p(\Omega)$ *band-limited* or *low-pass functions.* The smallest of the classes $B_p(\Omega)$ is $B_1(\Omega)$, since we have

$$B_p(\Omega)\subset B_p'(\Omega) \quad \text{if } 1\le p<p'\le\infty,$$

(see Boas [1, Thm. 6.7.18]). Also for $1\le p\le 2$ and $g(t)$ in $B_p(\Omega)$, we have

$$g(t)=\int_{-\Omega}^{\Omega} G(\omega)e^{i\omega t}d\omega$$

for some function G in $L_q(-\Omega,\Omega)$, where $1/p+1/q=1$. (Boas [1, Thm. 6,8,13]), so that the spectrum of $g(t)$ is confined to $[-\Omega,\Omega]$. (The Paley-Wiener theorem asserts that for $p=2$ the converse assertion holds, i.e. all functions $G\in L_2(-\Omega,\Omega)$ give rise to a function $g\in B_2(\Omega)$.)

The growth condition we impose on high-pass functions is that they belong to the class Λ of *uniform locally-*L_1 *functions.* The class Λ consists of all complex-valued measurable functions $h(x)$ defined on $(-\infty,\infty)$ which are locally integrable and for which

$$(2) \qquad M_T(h)=\sup_{-\infty<t<\infty}\int_0^T |h(x+t)|\,dx$$

is finite for all $T<\infty$. Note that if (2) is satisfied for one positive T, then it is satisfied for all finite positive T. Members of the class Λ are not necessarily bounded, but do not grow in the sense that the function $S_h(x)=(1/T)\int_x^{x+T}h(t)\,dt$ is bounded when h is in Λ. Also Λ contains all functions in $L_p(-\infty,\infty)$, $1\le p\le\infty$; and if f_1 is in L_{p_1}, f_2 in L_{p_2}, $1\le p_1,p_2\le\infty$, then f_1+f_2 is in Λ, although f_1+f_2 may not belong to any L_p space.

We now define the class $H(\Omega)$ of *high-pass functions* to consist of all those (complex-valued) functions $h(x)$ in Λ for which[1]

$$\int_{-\infty}^{\infty} h(x)g(x)=0, \quad \text{all } g\in B_1(\Omega).$$

We also define the classes $H_p(\Omega)$ by

$$H_p(\Omega)=H(\Omega)\cap L_p(-\infty,\infty).$$

Note that since $(-\Omega,\Omega)$ is symmetric, the real and imaginary parts of any function in $H(\Omega)$ (resp. $H_p(\Omega)$) are also in $H(\Omega)$ (resp. $H_p(\Omega)$).

We can immediately check that $H(\Omega)$ contains all "reasonable" trigonometric series with all frequencies $|\lambda_k|\ge\Omega$. Indeed the Fourier transform of an L_1-function is continuous, and since the Fourier transform of a function g in $B_1(\Omega)$ vanishes outside $[-\Omega,\Omega]$, it must vanish at $\pm\Omega$. In particular, then, any sum of the form $\Sigma_k a_k e^{i\lambda_k t}$, where λ_k is real and $|\lambda_k|\ge\Omega$, that converges in Λ, represents a function in $H(\Omega)$. The sum *converges in* Λ if there exists h in Λ such that

$$\lim_{n\to\infty}\left(\sup_x\int_x^{x+1}\left|h(t)-\sum_{k=1}^{n}a_k e^{i\lambda_k t}\right|dt\right)=0.$$

[1] The absolute convergence of the integral follows from (4), (5) and (6) below.

2. The unbiased integral. Our first goal is to characterize the operator U that takes a high-pass function h in $H(\Omega)$ to the unique high-pass function $U(h)=h^{(-1)}$ having the property that

$$h^{(-1)}(x_2)-h^{(-1)}(x_1)=\int_{x_1}^{x_2}h(t)\,dt. \tag{3}$$

We call $h^{(-1)}$ the *unbiased integral* of h. For example, the unbiased integral of $\cos\lambda t$ is $(1/\lambda)\sin\lambda t$.

We will show subsequently that $h^{(-1)}$ is a bounded function, i.e. $h^{(-1)}\in H_\infty(\Omega)$. This fact is of considerable importance in the theory of high-pass functions. Indeed we can use it to show that the linear functional

$$\phi_f(h)=\lim_{T\to\infty}\int_{-T}^{T}f(t)h(t)\,dt$$

is a bounded linear functional on $H(\Omega)$ whenever $f(t)$ is a function of total bounded variation on $(-\infty,\infty)$ which tends to 0 at $\pm\infty$, even though the integral defining $\phi_f(h)$ above may be only conditionally convergent. To establish this, we use integration by parts to obtain

$$\begin{aligned}\phi_f(h)&=\lim_{T\to\infty}\left[f(T)h^{(-1)}(T)-f(-T)h^{(-1)}(-T)-\int_{-T}^{T}h^{(-1)}(t)\,df(t)\right]\\&=-\int_{-\infty}^{\infty}h^{(-1)}(t)\,df(t),\end{aligned}$$

which converges absolutely since $h^{(-1)}$ is bounded.

We shall show that $h^{(-1)}$ is given by the absolutely convergent convolution integral

$$h^{(-1)}(t)=\int_{-\infty}^{\infty}h(x)K(t-x)\,dx,$$

where $K(t)$ is any kernel of the form

$$K(t)=\frac{1}{2}\operatorname{sgn}t-g(t),\qquad g\in B(\Omega),$$

such that $K(t)\in L_1(-\infty,\infty)$, i.e. $g(t)$ is an L_1-approximation in $B(\Omega)$ to $\frac{1}{2}\operatorname{sgn}t$.

For convenience in describing our results, we introduce a subclass of $L_1(-\infty,\infty)$ denoted by S_1, which we call the class of *absolutely lattice summable functions*. This consists of all functions $g(x)$ on $(-\infty,\infty)$ that satisfy: For each $T>0$,

$$\Sigma_T(g)\equiv\sum_{n=-\infty}^{\infty}\max_{nT\le x\le(n+1)T}|g(x)|<\infty. \tag{4}$$

Note that if (4) holds for one $T>0$, it holds for all $T>0$. The space S_1 has the property that translates of members of S_1 are also in S_1. Also if h is in Λ and g is in S_1, then

$$\int_{-\infty}^{\infty}|h(x)g(x)|\,dx\le M_T(h)\Sigma_T(g). \tag{5}$$

Finally we note that if g is in $B_1(\Omega)$, then g is in S_1 with

$$\Sigma_T(g) \leq \frac{4}{\pi T} \frac{e^{\Omega T}}{2} \int_{-\infty}^{\infty} |g(x)|\, dx, \tag{6}$$

cf. [1, Thm. 6.7.15].

Our main result is as follows.

THEOREM 1. *Let K be a function in $L_1(-\infty,\infty)$. The following two conditions are equivalent.*

(i) $K(t)=\frac{1}{2}\operatorname{sgn} t-g(t)$, *$g$ in $B(\Omega)$,*

(ii) $K(t)=(1/2\pi)\lim_{A\to\infty}\int_{-A}^{A}\hat{K}(\omega)e^{i\omega t}\,d\omega$ *for all real t, where* $\hat{K}(\omega)=1/i\omega$, $|\omega|\geq\Omega$.

Furthermore any function K in $L_1(-\infty,\infty)$ for which (i) *holds satisfies*:

(iii) *K is in S_1 and hence* $\lim_{t\to\pm\infty}K(t)=0$, *and g in* (i) *belongs to* $B_\infty(\Omega)$.

(iv) *To each h in $H(\Omega)$ there corresponds a function $h^{(-1)}$ in $H_\infty(\Omega)$ given by*

$$h^{(-1)}(t)=\int_{-\infty}^{\infty} h(x)K(t-x)\,dx.$$

This function satisfies

$$h^{(-1)}(b)-h^{(-1)}(a)=\int_a^b h(x)\,dx, \qquad -\infty<a<b<\infty.$$

Proof. We suppose first the existence of a function K_1 in $L_1(-\infty,\infty)$ that is of both forms (i) and (ii) (with corresponding functions g_1 in $B(\Omega)$ and $\hat{K}_1$) and in addition that K_1 is in S_1. We later exhibit such a function.

Now let K_2 be any L_1-function of the form (i): $K_2(t)=\frac{1}{2}\operatorname{sgn} t-g_2(t)$, g_2 in $B(\Omega)$. Then $K_2(t)-K_1(t)=g(t)$, where $g(t)=g_1(t)-g_2(t)$ is in $B(\Omega)\cap L_1(-\infty,\infty)=B_1(\Omega)$. Then g is in S_1, and hence $K_2=K_1+g$ is in S_1. Since g is in $B_1(\Omega)$, it follows that $g(t)=(1/2\pi)\int_{-\Omega}^{\Omega}G(\omega)e^{i\omega t}\,d\omega$, where G is continuous and $G(\omega)=0$, $|\omega|\geq\Omega$. Thus

$$K_2(t)=\frac{1}{2\pi}\lim_{A\to\infty}\int_{-A}^{A}\left[\hat{K}_1(\omega)+G(\omega)\right]e^{i\omega t}\,d\omega,$$

with

$$\hat{K}_1(\omega)+G(\omega)=\frac{1}{i\omega}, \qquad |\omega|\geq\Omega;$$

i.e. K_2 is of form (ii).

On the other hand, if K_2 is any L_1-function of the form (ii):

$$\text{(ii)}\quad K_2(t)=\frac{1}{2\pi}\lim_{A\to\infty}\int_{-A}^{A}\hat{K}_2(\omega)e^{i\omega t}\,d\omega, \qquad \hat{K}_2(\omega)=\frac{1}{i\omega}\ \text{ for } |\omega|\geq\Omega,$$

then

$$K_2(t)-K_1(t)=\frac{1}{2\pi}\lim_{A\to\infty}\int_{-A}^{A}\left[\hat{K}_2(\omega)-\hat{K}_1(\omega)\right]e^{i\omega t}\,d\omega$$

$$=\frac{1}{2\pi}\int_{-\Omega}^{\Omega}\left[\hat{K}_2(\omega)-\hat{K}_1(\omega)\right]e^{i\omega t}\,d\omega,$$

since

$$\hat{K}_2(\omega)=\hat{K}_1(\omega)=\frac{1}{i\omega}, \qquad |\omega|\geq\Omega.$$

It follows that $K_2(t)-K_1(t)=g(t)$ is in $B(\Omega)$, and hence in $B_1(\Omega)$. Again this gives K_2 in S_1; also

$$K_2(t)=K_1(t)+g(t)$$
$$=\frac{1}{2}(\operatorname{sgn} t)-[g_1(t)-g(t)],$$

and $g_1(t)-g(t)$ is in $B(\Omega)$; i.e. K_2 is of the form (i). Now any function of S_1 must be bounded and tend to zero at $\pm\infty$. Hence if g belongs to $B(\Omega)$ and $[\frac{1}{2}\operatorname{sgn} t-g(t)]$ belongs to S_1, it follows that $g(t)$ has limits $\frac{1}{2}$ and $-\frac{1}{2}$ at $+\infty$ and $-\infty$, respectively. In particular, g is bounded, i.e., g belongs to $B_\infty(\Omega)$.

Thus any L_1-function of the form (i) or (ii) is necessarily of both forms and is in S_1, provided there is one function K_1 of both forms and in the class S_1. We now construct K_1. We let

$$\hat{K}_1(\omega)=\begin{cases}\dfrac{1}{i\omega}, & |\omega|\geq\Omega,\\ \dfrac{\omega}{i\Omega^2}, & |\omega|<\Omega,\end{cases}$$

and then

$$K_1(t)=\frac{1}{2\pi}\lim_{A\to\infty}\int_{-A}^{A}\hat{K}_1(\omega)e^{i\omega t}\,d\omega.$$

Since $\hat{K}_1$ is odd,

$$K_1(t)=\frac{i}{\pi}\lim_{A\to\infty}\int_0^A\hat{K}_1(\omega)\sin\omega t\,d\omega.$$

Therefore, $K_1(t)$ is odd, and for $t>0$,

$$K_1(t)=\frac{1}{\pi(\Omega t)^2}\int_0^{\Omega t}x\sin x\,dx+\frac{1}{\pi}\lim_{A\to\infty}\int_{\Omega t}^{A}\frac{\sin x}{x}dx.$$

Integrating $\int_0^{\Omega t}x\sin x\,dx$ by parts gives

$$K_1(t)=\frac{1}{\pi}\left(\frac{\sin\Omega t}{(\Omega t)^2}-\frac{\cos\Omega t}{\Omega t}\right)+\frac{1}{\pi}\lim_{A\to\infty}\int_{\Omega t}^{A}\frac{\sin x}{x}dx,\qquad t>0. \tag{7}$$

Thus

$$K_1(t)=-\frac{1}{\pi\Omega^2}\frac{d}{dt}\left(\frac{\sin\Omega t}{t}\right)+\frac{1}{\pi}\lim_{A\to\infty}\int_{\Omega t}^{A}\frac{\sin x}{x}ds.$$

Since $\lim_{A\to\infty}\int_0^A(\sin x/x)\,dx=\pi/2$, it follows that

$$K_1(t)=\frac{1}{2}-\left[\frac{1}{\pi}\int_0^{\Omega t}\frac{\sin x}{x}dx+\frac{1}{\pi\Omega^2}\frac{d}{dt}\left(\frac{\sin\Omega t}{t}\right)\right],\qquad t>0.$$

Since K_1 is odd, we have

$$K_1(t)=\frac{1}{2}\operatorname{sgn} t-g_1(t),\qquad -\infty<t<\infty, \tag{8}$$

where

$$g_1(t)=\frac{1}{\pi}\int_0^{\Omega t}\frac{\sin x}{x}dx+\frac{1}{\pi\Omega^2}\frac{d}{dt}\left(\frac{\sin\Omega t}{t}\right)$$

is in $B(\Omega)$. Thus K_1 is of the form (i), provided that K_1 is in $L_1(-\infty,\infty)$. To see that K_1 is in $L_1(-\infty,\infty)$, and indeed in S_1, integrate $\int_{\Omega t}^{A}(\sin x/x)\,dx$ by parts to obtain

$$\lim_{A\to\infty}\int_{\Omega t}^{A}\frac{\sin x}{x}dx=\frac{\cos\Omega t}{\Omega t}-\int_{\Omega t}^{\infty}\frac{\cos x}{x^2}dx.$$

It follows from (7) that

$$K_1(t)=\frac{1}{\pi}\frac{\sin\Omega t}{(\Omega t)^2}-\frac{1}{\pi}\int_{\Omega t}^{\infty}\frac{\cos x}{x^2}dx,\qquad t>0.$$

Since K_1 is odd, this gives $K_1(t)=O(1/t^2)$ as $|t|\to\infty$, and since K_1 is bounded (by (8)), we find that K_1 is in S_1 as sought. Thus we have proved the equivalence of the forms (i) and (ii) and the conclusion (iii) of the theorem.

Now let K be an L_1-function of the form (i):

$$K(t)=\frac{1}{2}\operatorname{sgn}t-g(t),\qquad g\text{ in }B(\Omega).$$

For h in $H(\Omega)$, define

$$h^{(-1)}(t)=\int_{-\infty}^{\infty}h(x)K(t-x)\,dx.$$

The integral converges absolutely, since K is in S_1 and $|h^{(-1)}(t)|\le M_T(h)\Sigma_T(K)$. For $-\infty<a<b<\infty$,

$$h^{(-1)}(b)-h^{(-1)}(a)=\int_{-\infty}^{\infty}h(x)[K(b-x)-K(a-x)]\,dx,$$

and

$$K(b-x)-K(a-x)=C_{(a,b)}(x)-[g(b-x)-g(a-x)],$$

where $C_{(a,b)}$ is the characteristic function of the interval (a,b). Hence $g(b-x)-g(a-x)$ is in $B(\Omega)\cap L_1=B_1(\Omega)$. Therefore

$$h^{(-1)}(b)-h^{(-1)}(a)=\int_{-\infty}^{\infty}h(x)C_{(a,b)}(x)\,dx=\int_a^b h(x)\,dx.$$

If f is in $B_1(\Omega)$, then $\int_{-\infty}^{\infty}h^{(-1)}(t)f(t)\,dt=\int_{-\infty}^{\infty}h(t)f^*(t)\,dt$, where

$$f^*(t)=\int_{-\infty}^{\infty}K(x)f(t-x)\,dx,$$

is also in $B_1(\Omega)$. Hence $h^{(-1)}$ is in $H(\Omega)$, and thus in $H_\infty(\Omega)$, since $|h^{(-1)}|$ is bounded. This completes the proof of Theorem 1. □

3. Conditionally convergent integrals and integrals summable $(C,1)$. We now derive a number of results concerning the convergence of integrals involving functions in $H(\Omega)$ as corollaries of Theorem 1. These results justify certain formal operations on high-pass functions that would follow directly if their Fourier transforms existed. In particular we consider the use of the $(C,1)$ summation

$$\lim_{T\to\infty}\int_{-T}^{T}\left(1-\frac{|t|}{T}\right)f(t)\,dt \tag{9}$$

as a replacement for the integral $\int_{-\infty}^{\infty}f(t)\,dt$, which may not exist. In this connection, we note that whenever the limit (9) exists, then the Abel summation

$$\lim_{a\to 0^{+}}\int_{-\infty}^{\infty}e^{-a|t|}f(t)\,dt$$

also exists and agrees with the former limit. In some applications it may be more convenient to use the latter summation method in connection with the following results.

We first show that for $h\in H(\Omega)$ the improper integral $\int_0^{\infty}h(t)\,dt$ is summable $(C,1)$ to $-h^{(-1)}(0)$.

COROLLARY 1. *For h in $H(\Omega)$,*

$$\lim_{A\to\infty}\int_0^{A}\left(1-\frac{t}{A}\right)h(t)\,dt=-h^{(-1)}(0),$$

$$\lim_{A\to\infty}\int_{-A}^{0}\left(1+\frac{t}{A}\right)h(t)\,dt=h^{(-1)}(0),$$

and hence

$$\lim_{A\to\infty}\int_{-A}^{A}\left(1-\frac{|t|}{A}\right)h(t)\,dt=0.$$

Proof. Integrating by parts, we have

$$\int_0^{A}\left(1-\frac{t}{A}\right)h(t)\,dt=-h^{(-1)}(0)+\frac{1}{A}\int_0^{A}h^{(-1)}(t)\,dt.$$

Since $h^{(-1)}$ is in $H(\Omega)$, $\int_0^{A}h^{(-1)}(t)\,dt$ is bounded for all A. The result follows. □

We next specialize Corollary 1 to show that certain low-pass and high-pass functions are "weakly" orthogonal; i.e. the integral of their product is summable $(C,1)$ to zero. In particular, the Fourier integrals of high-pass and bounded low-pass functions are summable $(C,1)$ to zero for points outside their respective spectral supports. So we have come full circle from the orthogonality condition we used as a device to define a class of functions whose "Fourier transforms" vanish over $(-\Omega,\Omega)$, to the conclusion that the Fourier integrals of these functions are actually summable $(C,1)$ to zero in the open interval $(-\Omega,\Omega)$. Thus we could have used the latter condition to define the class $H(\Omega)$, but it would have appeared to be more restrictive than the orthogonality condition. It is easy from this point to show that the two conditions are equivalent, i.e. lead to the same subclass of the space Λ of uniform locally-L^1 functions.

COROLLARY 2. *If g is in $B_{\infty}(\alpha)$, h in $H(\Omega)$, $0\le\alpha<\Omega$, then*

$$\lim_{A\to\infty}\int_{-A}^{A}\left(1-\frac{|t|}{A}\right)g(t)h(t)\,dt=0.$$

In particular

$$\lim_{A\to\infty}\int_{-A}^{A}\left(1-\frac{|t|}{A}\right)g(t)e^{i\omega t}\,dt=0,\qquad \omega>\alpha \text{ or } \omega<-\alpha,$$

and

$$\lim_{A\to\infty}\int_{-A}^{A}\left(1-\frac{|t|}{A}\right)h(t)e^{i\omega t}dt=0, \qquad -\Omega<\omega<\Omega.$$

Proof. $f(t)=g(t)h(t)$ is in Λ, since h is in Λ and g is in L_∞. If g_1 is any function in $B_1(\beta)$, where $\beta=\Omega-\alpha>0$, then

$$\int_{-\infty}^{\infty}f(t)g_1(t)\,dt=\int_{-\infty}^{\infty}h(t)[g(t)g_1(t)]\,dt=0,$$

since $g(t)g_1(t)$ is in L_1 and is the restriction to the real line of an entire function of exponential type $\leq\alpha+\beta=\Omega$. Thus $f(t)$ is in $H(\beta)$, and the result follows from applying Corollary 1 to f. □

We next show that $\lim_{T\to\infty}\int_0^T h(t)\,dt$ exists for h in $H(\Omega)$, provided $\lim_{T\to\infty}\int_T^{T+1}|h(t)|\,dt=0$; i.e. if h tends "weakly" to zero.

COROLLARY 3. *If h is in $H(\Omega)$ and if for some $T>0$,* $\lim_{t\to\infty}$ *or* $\lim_{t\to-\infty}\int_0^T|h(x+t)|\,dx=0$, *then* $\lim h^{(-1)}(t)=0$, *as $t\to\infty$, or $t\to-\infty$, respectively, and accordingly,* $h^{(-1)}(t)=-\lim_{A\to\infty}\int_t^A h(x)\,dx$ *or* $h^{(-1)}(t)=\lim_{A\to\infty}\int_{-A}^{t}h(x)\,dx$. *Hence if* $\lim_{|t|\to\infty}\int_0^T|h(x+t)|\,dx=0$, *then* $\lim_{A,B\to\infty}\int_{-B}^{A}h(x)\,dx=0$. *In particular, if h belongs to $H_p(\Omega)$, $1\leq p<\infty$, all the above assumptions (and conclusions) hold, and $h^{(-1)}$ belongs to $H_p(\Omega)$.*

Proof. From Theorem 1,

$$h^{(-1)}(t)=\int_{-\infty}^{\infty}h(x)K(t-x)\,dx,$$

where K is in S_1. Then

$$\begin{aligned}h^{(-1)}(2A)&=\int_{-\infty}^{A}h(x)K(2A-x)\,dx+\int_{A}^{\infty}h(x)K(2A-x)\,dx\\&=\int_{A}^{\infty}h(2A-x)K(x)\,dx+\int_{A}^{\infty}h(x)K(2A-x)\,dx.\end{aligned}$$

Thus

$$|h^{(-1)}(2A)|\leq M_T(h)\Sigma_T(K_A)+\Sigma_T(K)\sup_{t\geq A}\int_0^T|h(x+t)|\,dx,$$

where

$$K_A(t)=\begin{cases}K(t), & t\geq A,\\ 0, & t<A.\end{cases}$$

Then if $\lim_{t\to\infty}\int_0^T|h(x+t)|\,dx=0$, it follows that $\lim_{A\to\infty}h^{(-1)}(2A)=0$. Similarly, if $\lim_{t\to-\infty}\int_0^T|h(x+t)|\,dx=0$, then $\lim_{A\to\infty}h^{(-1)}(-2A)=0$.

The other conclusions follow from the equations

$$h^{(-1)}(b)-h^{(-1)}(a)=\int_a^b h(x)\,dx.$$

Finally, if h is in $L_p(-\infty,\infty)$, $1\leq p<\infty$, then

$$\lim_{|t|\to\infty}\int_0^T|h(x+t)|^p\,dx=0=\lim_{|t|\to\infty}\int_0^T|h(x+t)|\,dx,$$

Also, since $h^{(-1)}=h*K$,

$$\|h^{(-1)}\|_p\leq\|K\|_1\|h\|_p.$$

Thus for h in $H_p(\Omega)$, $h^{(-1)}$ is also in $H_p(\Omega)$. □

Next we specialize Corollary 3 to show that the integrals of products of certain low-pass and high-pass functions converge conditionally.

COROLLARY 4. *If g is in $B_\infty(\alpha)$ and h in $H(\Omega)$, $0\leq\alpha<\Omega$, and if either*

(i) $\lim_{|t|\to\infty}g(t)=0$, *or*

(ii) $\lim_{|t|\to\infty}\int_0^T|h(x+t)|\,dx=0$, *then*

$$\lim_{A,B\to\infty}\int_{-B}^{A}g(x)h(x)\,dx=0.$$

In particular, if (i) *holds, then*

$$\lim_{A,B\to\infty}\int_{-B}^{A}g(x)e^{i\omega x}dx=0$$

for $\omega>\alpha$, $\omega<-\alpha$. And if (ii) *holds, then*

$$\lim_{A,B\to\infty}\int_{-B}^{A}h(x)e^{i\omega x}dx=0,\qquad -\Omega<\omega<\Omega.$$

Proof. Since g is in $B_\infty(\alpha)$, and h is in $H(\Omega)$, $g(t)h(t)$ is in $H(\Omega-\alpha)$, as shown in the proof of Corollary 2. If either (i) or (ii) holds, then

$$\lim_{|t|\to\infty}\int_0^T|g(x+t)h(x+t)|\,dx=0.$$

The result follows from applying Corollary 3 to $g(t)h(t)$. □

Next we apply Corollary 4 to obtain two results which give sufficient conditions for $\sin\beta t/\pi t$ to be a reproducing kernel for low-pass functions.

COROLLARY 5. *If g is in $B(\Omega)$ and $\lim_{|t|\to\infty}(g(t)/t)=0$, then*

$$g(x)=\lim_{A,B\to\infty}\int_{-B}^{A}g(t)\frac{\sin\beta(x-t)}{\pi(x-t)}dt$$

for any $\beta>\Omega$.

Proof. $f(t)=(g(x)-g(t))/(x-t)$ belongs to $B_\infty(\Omega)$ and $\lim_{|t|\to\infty}f(t)=0$, and $h(t)=(1/\pi)\sin\beta(x-t)$ belongs to $H_\infty(\beta)$. By Corollary 4, $\lim_{A,B\to\infty}\int_{-B}^{A}f(t)h(t)\,dt=0$. Thus

$$\lim_{A,B\to\infty}\int_{-B}^{A}g(t)\frac{\sin\beta(x-t)}{\pi(x-t)}dt=g(x)\lim_{A,B\to\infty}\int_{-B}^{A}\frac{\sin\beta(x-t)}{\pi(x-t)}dt$$

whenever either limit exists. Since

$$\lim_{A,B\to\infty}\int_{-B}^{A}\frac{\sin\beta(x-t)}{\pi(x-t)}dt=1,$$

the result follows. □

COROLLARY 6. *If g is in $B_p(\Omega)$, $1 \le p < \infty$, then*

$$g(x) = \int_{-\infty}^{\infty} g(t) \frac{\sin \beta(x-t)}{\pi(x-t)} dt$$

for any $\beta \ge \Omega$.

Proof. For $0 < \alpha < \Omega$,

$$f(t) = \frac{g(t) - g(x)(\sin \alpha(x-t))/\alpha(x-t)}{x-t}$$

belongs to $B_1(\Omega)$, and hence to $B_1(\beta)$, and $h(t) = (1/\pi)\sin \beta(x-t)$ is in $H_\infty(\beta)$. Thus $\int_{-\infty}^{\infty} f(t) h(t)\, dt = 0$. Hence

$$\int_{-\infty}^{\infty} g(t) \frac{\sin \beta(x-t)}{\pi(x-t)} dt = g(x) \int_{-\infty}^{\infty} \frac{\sin \alpha(x-t)}{\alpha(x-t)} \frac{\sin \beta(x-t)}{\pi(x-t)} dt$$

as both integrals are absolutely convergent, and by Corollary 5,

$$\int_{-\infty}^{\infty} \frac{\sin \alpha(x-t)}{\alpha(x-t)} \frac{\sin \beta(x-t)}{\pi(x-t)} dt = 1. \qquad \square$$

4. Bounds on the unbiased integral. An important characteristic of real-valued functions in $H(\Omega)$ is that they must oscillate infinitely often and in fact fairly regularly. One measure of this is the bound

$$\left| \int_0^x h(t)\, dt \right| \le \frac{13}{4} M_{\pi/\Omega}(h), \tag{10}$$

where

$$M_{\pi/\Omega}(h) = \sup_{-\infty < t < \infty} \int_0^{\pi/\Omega} |h(x+t)|\, dx,$$

valid for all (complex-valued) functions in $H(\Omega)$ ([2, Thm. 6.1.1]).

Our object here is to obtain sharper upper bounds for

$$\|h^{(-1)}\|_\infty = \sup_{-\infty < t < \infty} |h^{(-1)}(x)|$$

in terms of $M_{\pi/\Omega}(h)$ for h in $H(\Omega)$, and in terms of $\|h\|_\infty$ for h in $H_\infty(\Omega)$.

THEOREM 2. *For all $h \in H_\infty(\Omega)$,*

$$\|h^{(-1)}\|_\infty \le \frac{\pi}{2\Omega} \|h\|_\infty. \tag{11}$$

The inequality (11) was first given by Bohr [3] for periodic high-pass functions and later was generalized by Lewitan [4] and Hormander [5].

Proof. We first prove the theorem in the special case $\Omega = \pi$. The general case follows by a change of variable.

In order to obtain the bound (11) we construct a particular L_1-function $K(t) = \frac{1}{2}\operatorname{sgn} t - g(t)$, satisfying condition (i) of Theorem 1 for $\Omega = \pi$. Let

$$g(t) = \lim_{N \to \infty} \sum_{k=1}^{N} \frac{1}{2} \left(\frac{\sin \pi(t-k)}{\pi(t-k)} - \frac{\sin \pi(t+k)}{\pi(t+k)} \right). \tag{12}$$

We note that g is an odd function. Since

$$\frac{\sin \pi x}{\pi x}=\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega x}\,d\omega,$$

it is easily verified that

$$g(t)=\frac{1}{4\pi}\int_{-\pi}^{\pi}\cot\frac{\omega}{2}\sin\omega t\,d\omega, \tag{13}$$

so that g is in the class $B(\pi)$. We show now that $K(t)=\frac{1}{2}\operatorname{sgn} t-g(t)$ is in $L_1(-\infty,\infty)$. Since

$$\lim_{N\to\infty}\sum_{k=-N}^{N}\frac{\sin\pi(t-k)}{\pi(t-k)}=1,$$

we have

$$\begin{aligned}\frac{1}{2}-g(t)&=\lim_{N\to\infty}\frac{1}{2}\left(\sum_{k=-N}^{N}\frac{\sin\pi(t-k)}{\pi(t-k)}\right)-g(t)\\ &=\lim_{N\to\infty}\left(\frac{\sin\pi(t)}{2\pi t}+\sum_{k=1}^{N}\frac{\sin\pi(t+k)}{\pi(t+k)}\right)\\ &=\frac{\sin\pi t}{\pi}\lim_{N\to\infty}\left(\frac{1}{2t}+\sum_{k=1}^{N}\frac{(-1)^k}{t+k}\right).\end{aligned}$$

For $x>0$, $1/x=\int_0^\infty e^{-sx}ds$, and hence for $t>0$,

$$K(t)=\frac{1}{2}-g(t)=\frac{\sin\pi t}{\pi}\lim_{N\to\infty}\int_0^\infty\left(\frac{1}{2}+\sum_{k=1}^{N}(-1)^k e^{-sk}\right)e^{-st}ds.$$

Thus, this particular kernel $K(t)$ has the representation

$$K(t)=\frac{\sin\pi t}{2\pi}\int_0^\infty\left(\frac{1-e^{-s}}{1+e^{-s}}\right)e^{-st}ds,\qquad t>0. \tag{14}$$

Since $(1-e^{-s})/(1+e^{-s})\le s/2$ for $s\ge 0$, we have

$$\int_0^\infty\frac{1-e^{-s}}{1+e^{-s}}e^{-st}ds\le\frac{1}{2}\int_0^\infty se^{-st}ds=\frac{1}{2t^2},\qquad t>0.$$

Hence

$$|K(t)|\le\frac{1}{4\pi t^2},\qquad t>0. \tag{15}$$

Since g is an odd function, K is also odd, and hence (15) holds for all $t\neq 0$. Also $K(t)=\frac{1}{2}\operatorname{sgn} t-g(t)$ is clearly bounded near 0. Consequently (15) shows that K is an L_1-function.

If h is in the class $H_\infty(\pi)$, then

$$h^{(-1)}(x)=\int_{-\infty}^{\infty}h(x-t)K(t)\,dt,$$

since K satisfies condition (i) of Theorem 1 for $\Omega=\pi$. Hence

$$|h^{(-1)}(x)|\leq\left(\int_{-\infty}^{\infty}|K(t)|\,dt\right)\sup_{-\infty<t<\infty}|h(t)|. \tag{16}$$

Since $K(t)$ is an odd function, it follows from (14) that $K(t)$ and $\sin\pi t$ have the same sign for all real t. Thus

$$\int_{-\infty}^{\infty}|K(t)|\,dt=\int_{-\infty}^{\infty}K(t)\operatorname{sgn}(\sin\pi t)\,dt.$$

The function $h_1(t)=\operatorname{sgn}(\sin\pi t)$ has the Fourier series $(4/\pi)\Sigma_0^\infty(\sin(2k+1)\pi t)/(2k+1)$ and hence belongs to the class $H_\infty(\pi)$. Consequently

$$\int_{-\infty}^{\infty}|K(t)|\,dt=-\int_{-\infty}^{\infty}h_1(t)K(-t)\,dt=-h_1^{-1}(0).$$

Now $h_1(t+\frac{1}{2})$ is an even function of t, and since $K(t)$ is odd, it follows that $h^{(-1)}(\frac{1}{2})=0$. Then since

$$\int_0^{1/2}\operatorname{sgn}(\sin\pi t)\,dt=\frac{1}{2},$$

it follows that $h_1^{(-1)}(0)=-\frac{1}{2}$. Thus,

$$\int_{-\infty}^{\infty}|K(t)|\,dt=\frac{1}{2}.$$

Together with (16) this gives

$$\sup_{-\infty<t<\infty}|h^{(-1)}(t)|\leq\frac{1}{2}\sup_{-\infty<t<\infty}|h(t)|,\qquad h\text{ in }H_\infty(\pi).$$

For any $\Omega>0$, the function $K(\Omega t/\pi)=\frac{1}{2}\operatorname{sgn}t-g(\Omega t/\pi)$ is in L_1 with $g(\Omega t/\pi)$ in $B(\Omega)$. Hence for h in $H_\infty(\Omega)$,

$$h^{(-1)}(x)=\int_{-\infty}^{\infty}K\left(\frac{\Omega}{\pi}t\right)h(x-t)\,dt,\qquad -\infty<x<\infty.$$

Since $\int_{-\infty}^{\infty}|K(\Omega t/\pi)|\,dt=\pi/2\Omega$, the inequality (11) follows. □

An immediate consequence of the proof of Theorem 2 is the following corollary.

COROLLARY 7. *The function*

$$g(t)=\frac{1}{4\pi}\int_{-\pi}^{\pi}\cot\frac{\omega}{2}\sin\omega t\,d\omega$$

is the best L_1-approximation to $\frac{1}{2}$ sgn(t) by an entire function of exponential type $\leq\pi$.

Proof. This follows from the inequality (16) applied to the function $h(t)=\operatorname{sgn}(\sin\pi t)$. □

THEOREM 3. *For all $h\in H(\Omega)$,*

$$\|h^{(-1)}\|_\infty\leq\frac{1}{2}\left(1+\frac{1}{\pi}\right)M_{\pi/\Omega}(h). \tag{17}$$

This inequality is not the best possible, but allows us to considerably improve the bound $\frac{13}{4}$ to $1+1/\pi$ in (10). (See (25) below.)

Proof. Again by a change of variable we need only consider the case $\Omega=\pi$. We use the same kernel $K(t)$ as in the proof of Theorem 2, which is defined by (14). We have

$$(18)\quad |h^{(-1)}(x)|\leq\int_{-\infty}^{\infty}|K(t)h(x-t)|\,dt=\int_{-1}^{1}|K(t)h(x-t)|\,dt+\int_{|t|>1}|K(t)h(x-t)|\,dt$$

and estimate each of the integrals on the right separately. By (13),

$$g''(t)=-\frac{1}{4\pi}\int_{-\pi}^{\pi}\omega^2\cot\frac{\omega}{2}\sin\omega t\,d\omega.$$

Hence $g''(t)<0$, $0\leq t\leq 1$. Since $g(0)=0$ and $g(1)=\frac{1}{2}$, we have, for $0\leq t\leq 1$, $g(t)\geq t/2$, and hence $0\leq K(t)=\frac{1}{2}-g(t)\leq\frac{1}{2}-t/2$. Since K is odd,

$$|K(t)|\leq\frac{1-|t|}{2},\qquad |t|\leq 1.$$

Thus

$$(19)\qquad \int_{-1}^{1}|K(t)h(x-t)|\,dt\leq\frac{1}{2}\int_{-1}^{1}(1-|t|)|h(x-t)|\,dt.$$

Let

$$\mu(t)=\begin{cases}1, & |t|\leq\frac{1}{2},\\ 0, & |t|>\frac{1}{2}.\end{cases}$$

Then

$$\mu*\mu(t)=\int_{-\infty}^{\infty}\mu(s)\mu(t-s)\,ds=\begin{cases}1-|t|, & |t|\leq 1,\\ 0, & |t|>1,\end{cases}$$

and hence

$$\int_{-1}^{1}(1-|t|)|h(x-t)|\,dt=\int_{-\infty}^{\infty}\mu*\mu(t)|h(x-t)|\,dt=\int_{-\infty}^{\infty}\mu(x-t)\mu*|h|(t)\,dt.$$

We may assume that $M_1(h)\leq 1$, so that

$$\mu*|h|(t)=\int_{t-1/2}^{t+1/2}|h(s)|\,ds\leq 1,$$

and since

$$\int_{-\infty}^{\infty}\mu(x-t)\,dt=1,$$

we have

$$\int_{-1}^{1}(1-|t|)|h(x-t)|\,dt\leq 1.$$

Then by (19),

$$(20)\qquad \int_{-1}^{1}|K(t)h(x-t)|\,dt\leq\frac{1}{2}.$$

Consider now the second integral on the right in (18). Let

$$K_n = \max_{n \le |t| \le n+1} |K(t)| = \max_{n \le t \le n+1} |K(t)|, \qquad n = 1, 2, \cdots .$$

Then since $\sup_{-\infty < x < \infty} \int_0^1 |h(x-t)|\, dt \le 1$, we have

$$\text{(21)} \qquad \int_{|t|>1} |K(t)h(x-t)|\, dt \le 2 \sum_{n=1}^{\infty} K_n .$$

From (14), we write $K(t) = (\sin \pi t) f(t)$, $t > 0$, where

$$f(t) = \frac{1}{2\pi} \int_0^\infty \frac{1-e^{-s}}{1+e^{-s}} e^{-st}\, ds .$$

Since $K'(t) = \pi \cos \pi t f(t) + \sin \pi t f'(t)$, $K'(t)$ vanishes for exactly those t for which

$$\pi \cot \pi t + \frac{f'(t)}{f(t)} = 0 .$$

Now

$$\frac{d}{dt}\left(\frac{f'(t)}{f(t)}\right) = \frac{\int_0^\infty s^2 \frac{1-e^{-s}}{1+e^{-s}} e^{-st}\, ds}{\int_0^\infty \frac{1-e^{-s}}{1+e^{-s}} e^{-st}\, ds} - \left[\frac{\int_0^\infty s \frac{1+e^{-s}}{1+e^{-s}} e^{-st}\, ds}{\int_0^\infty \frac{1+e^{-s}}{1+e^{-s}} e^{-st}\, ds}\right]^2 ,$$

and thus (being the form of a second moment minus the square of a first moment)

$$\text{(22)} \qquad \frac{d}{dt}\left(\frac{f'(t)}{f(t)}\right) = \min_{-\infty < a < \infty} \left[\frac{\int_0^\infty (s-a)^2 \frac{1-e^{-s}}{1+e^{-s}} e^{-st}\, ds}{\int_0^\infty \frac{1-e^{-s}}{1+e^{-s}} e^{-st}\, ds}\right], \qquad t > 0 .$$

Since $1 - e^{-s/2} \le (1-e^{-s})/(1+e^{-s}) \le s/2$, $s \ge 0$, we have

$$\frac{d}{dt}\left(\frac{f'(t)}{f(t)}\right) \le \frac{\int_0^\infty (s-a)^2 \frac{s}{2} e^{-st}\, ds}{\int_0^\infty (1-e^{-s/2}) e^{-st}\, ds} = \frac{\frac{1}{2}\frac{6}{t^4} - a\frac{2}{t^3} + \frac{a^2}{2}\frac{1}{t^2}}{\frac{1}{t} - \frac{1}{t+1/2}} ,$$

for any a, $-\infty < a < \infty$. Choosing $a = 2/t$, we obtain

$$\frac{d}{dt}\left(\frac{f'(t)}{f(t)}\right) \le \frac{2}{t^2} + \frac{1}{t^3}, \qquad t > 0 .$$

It follows that $(d/dt)(f'(t)/f(t)) \le 3$ for $t \ge 1$. Then

$$\frac{d}{dt}\left[\pi \cot \pi t + \frac{f'(t)}{f(t)}\right] = -\left(\frac{\pi}{\sin \pi t}\right)^2 + \frac{d}{dt}\left(\frac{f'(t)}{f(t)}\right) \le -\pi^2 + 3 < 0$$

for $t>1$, $t\neq$ integer. Consequently, for each integer $n\geq 1$, there is exactly one t_n, $n<t_n<n+1$, for which

$$\pi\cot\pi t_n+\frac{f'(t_n)}{f(t_n)}=0.$$

Then

$$K_n=|K(t_n)|=|\sin\pi t_n|f(t_n)\leq f(t_n).$$

From (22) we see that $-f'(t)/f(t)$ is a decreasing (positive) function. It follows that $\delta_n=n+\frac{1}{2}-t_n$ is a decreasing (positive) sequence. If $0<\theta\leq t_1$, then since $\delta_n\leq\delta_1$, we have $t_n\geq t_1+(n-1)\geq\theta+(n-1)$. Then $K_n\leq f(t_n)\leq f(\theta+n-1)$, since $f(t)$ is decreasing. Hence

$$\sum_{n=1}^{\infty}K_n\leq\sum_{n=0}^{\infty}f(\theta+n),\qquad 0\leq\theta\leq t_1. \tag{23}$$

We shall now show that we can apply (23) with $\theta=\frac{4}{3}$, i.e., show that $t_1>\frac{4}{3}$. We have

$$f(t)=\frac{1}{2\pi}\int_0^\infty\frac{1-e^{-s}}{1+e^{-s}}e^{-st}ds=\frac{1}{2\pi}\int_0^1\frac{u(1-u)^{t-1}}{2-u}du,\qquad t>0.$$

Since

$$\frac{1}{2-u}=\frac{1}{2}\sum_{k=0}^{\infty}\left(\frac{u}{2}\right)^k\geq\frac{1}{2}\left(1+\frac{u}{2}\right)$$

and

$$\int_0^1\left(u+\frac{u^2}{2}\right)(1-u)^{t-1}du=\frac{t+3}{t(t+1)(t+2)},$$

we have

$$2\pi f(t)\geq\frac{t+3}{2t(t+1)(t+2)},\qquad t>0.$$

In particular, $2\pi f(\frac{4}{3})>\frac{117}{560}$. Also

$$f'(t)=-\frac{1}{2\pi}\int_0^\infty\frac{s(1+e)^{-s}}{1+e^{-s}}e^{-st}ds.$$

Since

$$\frac{1}{1+e^{-s}}\leq 1-\frac{e^{-s}}{2},\qquad s>0,$$

and

$$\int_0^\infty s\left(1-\frac{3}{2}e^{-s}+\frac{e^{-2s}}{2}\right)e^{-st}=\frac{1}{t^2}-\frac{3}{2(t+1)^2}+\frac{1}{2(t+2)^2},$$

we have

$$-2\pi f'(t) \leq \frac{1}{t^2} - \frac{3}{2(t+1)^2} + \frac{1}{2(t+2)^2}, \qquad t>0.$$

In particular, $-2\pi f'(\frac{4}{3}) < \frac{27}{80}$. Consequently

$$-\frac{f'(\frac{4}{3})}{f(\frac{4}{3})} < \frac{21}{13} < \frac{5}{3}.$$

Since $\pi \cot \frac{4}{3}\pi = \pi/\sqrt{3} > \sqrt{3} > \frac{5}{3}$, we have $-f'(\frac{4}{3})/f(\frac{4}{3}) < \pi \cot \frac{4}{3}\pi$, and hence $\frac{4}{3} < t_1$.

By (23), then, $\Sigma_{n=1}^{\infty} K_n \leq \Sigma_{n=0}^{\infty} f(\frac{4}{3}+n)$. Since

$$f\left(\frac{4}{3}+n\right) = \frac{1}{2\pi}\int_0^1 \frac{u(1-u)^{1/2+n}}{2-u}\,du,$$

we have

$$\sum_{n=0}^{\infty} f\left(\frac{4}{3}+n\right) = \frac{1}{2\pi}\int_0^1 \frac{(1-u)^{1/3}}{2-u}\,du,$$

and putting $1-u=t^3$:

$$\sum_{n=0}^{\infty} f\left(\frac{4}{3}+n\right) = \frac{1}{2\pi}\int_0^1 \frac{3t^3}{1+t^3}\,dt.$$

This is readily evaluated by partial fractions to give

$$\sum_{n=0}^{\infty} f\left(\frac{4}{3}+n\right) = \frac{1}{2\pi}\left[3 - \left(\log 2 + \frac{\pi}{\sqrt{3}}\right)\right] < \frac{1}{4\pi}.$$

Together with (21), this gives $\int_{|t|>1} |K(t)h(x-t)|\,dt < 1/2\pi$. Then with (18) and (20) we have:

$$|h^{(-1)}(x)| < \frac{1}{2} + \frac{1}{2\pi}, \qquad -\infty < x < \infty,$$

for h in $H(\pi)$ with

$$\sup_{-\infty<t<\infty} \int_0^1 |h(u+t)|\,du \leq 1.$$

Consequently,

$$|h^{(-1)}(x)| \leq \left(\frac{1}{2} + \frac{1}{2\pi}\right) \sup_{-\infty<t<\infty} \int_0^1 |h(u+t)|\,du, \qquad -\infty<x<\infty, \quad h \text{ in } H(\pi).$$

For $\Omega > 0$ arbitrary and h in $H(\Omega)$, let $h_\pi(x) = h(\pi x/\Omega)$. h_π is in the class $H(\pi)$, and hence

$$|h_\pi^{(-1)}(x)| \leq \left(\frac{1}{2} + \frac{1}{2\pi}\right) \sup_{-\infty<t<\infty} \int_0^1 |h_\pi(u+t)|\,du.$$

But

$$h^{(-1)}(x)=\int_{-\infty}^{\infty}K\left(\frac{\Omega}{\pi}t\right)h(x-t)\,dt=\frac{\pi}{\Omega}\int_{-\infty}^{\infty}K(t)h\left(x-\frac{\pi}{\Omega}t\right)dt$$
$$=\frac{\pi}{\Omega}h_{\pi}^{(-1)}\left(\frac{\Omega}{\pi}x\right).$$

Hence

$$\frac{\Omega}{\pi}|h^{(-1)}(x)|\le\left(\frac{1}{2}+\frac{1}{2\pi}\right)\sup_{-\infty<t<\infty}\int_0^1|h_{\pi}(u+t)|\,du$$
$$=\left(\frac{1}{2}+\frac{1}{2\pi}\right)\frac{\Omega}{\pi}\sup_{-\infty<t<\infty}\int_0^{\pi/\Omega}|h(u+t)|\,du.$$

Thus we have the desired inequality (17). □

Lastly, we note that since

$$\int_{-T}^{T}h(t)\,dt=h^{(-1)}(T)-h^{(-1)}(-T),$$

Theorem 2 implies that

$$\left|\int_{-T}^{T}h(t)\,dt\right|1\le\frac{\pi}{\Omega}\|h\|_{\infty},\qquad h\in H_{\infty}(\Omega),\tag{24}$$

and Theorem 3 implies that

$$\left|\int_{-T}^{T}h(t)\,dt\right|\le\left(1+\frac{1}{\pi}\right)M_{\pi/\Omega}(h),\qquad h\in H(\Omega).\tag{25}$$

5. Acknowledgment. The author is indebted to J. C. Lagarias for a revision of this paper.

REFERENCES

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

[2] B. F. LOGAN, JR., *Properties of high-pass functions*, Doctoral thesis, Columbia University, New York, 1965.

[3] H. BOHR, *Ein allgemeiner Satz über die Integration eines trigonomerischen Polynoms*, Prace Matematyczno-Fizyczne, 43 (1935), pp. 273–288.

[4] B. M. LEWITAN, *Uber eine Verallgemeinerung der Ungleichungen vor S. Bernstein und H. Bohr*, Doklady Akad. Nauk., 15 (1937), pp. 169–172.

[5] LARS HORMANDER, *A new proof and a generalization of an inequality of Bohr*, Math. Scand., 2 (1954), pp. 33–45.

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1502-0013 $01.25/0

SOME INFINITE SERIES OF EXPONENTIAL AND HYPERBOLIC FUNCTIONS*

I. J. ZUCKER†

Abstract. The sums of several infinite series of exponential and hyperbolic functions containing a parameter, c, are expressed in closed form in terms of the complete elliptic integral of the first kind and its modulus. Several quadruple sums are evaluated, and from these some triple sums of hyperbolic functions are evaluated. Certain double sums of exponential and hyperbolic functions are also given.

Introduction. Many authors have investigated various infinite sums of hyperbolic functions. Berndt [1], [2] gives many references. The purpose of this note is to gather together several well-known (but not widely known) results in closed form for the sums of several exponential and hyperbolic function series, which depend on a certain parameter, c. When c is the square root of a rational number the sums may be expressed in terms of Γ-functions and other well-known transcendental numbers. Some new results are also presented for some quadruple series and for some double and triple series of exponential and hyperbolic series. The methods employed in obtaining these results have been described by Glasser [3], Zucker [4] and Glasser and Zucker [5]. Considerable use is made of results from the theory of θ-functions and elliptic integrals. The notation and procedures are outlined below.

Notation.

(1) $$K=\int_0^1[(1-x^2)(1-k^2x^2)]^{-1/2}dx, \qquad k^2+k'^2=1, \quad K'=K(k')$$

where k is the modulus of the complete elliptic integral of the first kind, K.

(2) $$K'/K=c, \quad q=e^{-\pi c}, \quad K=K(k)=K[c].$$

(3) $$\theta_2=\sum_{-\infty}^{\infty} q^{(n-1/2)^2}, \qquad \theta_3=\sum_{-\infty}^{\infty} q^{n^2},$$
$$\theta_4=\sum_{-\infty}^{\infty}(-1)^n q^{n^2}, \qquad \theta_5=2\sum_{-\infty}^{\infty}(-1)^n q^{(2n-1/2)^2},$$
$$\theta_1'=\theta_2\theta_3\theta_4=2\sum_{0}^{\infty}(-1)^n(2n+1)q^{(n-1/2)^2}.$$

(4) $$Q_0=\prod_1^{\infty}(1-q^{2n}), \qquad Q_1=\prod_1^{\infty}(1+q^{2n}),$$
$$Q_2=\prod_1^{\infty}(1+q^{2n-1}), \quad Q_3=\prod_1^{\infty}(1-q^{2n-1}).$$

(5) $$\theta_2^2=\frac{2kK}{\pi}=4q^{1/2}Q_0^2Q_1^4, \qquad \theta_5^2(q^{1/2})=\frac{4K\sqrt{kk'}}{\pi}=4q^{1/4}\frac{Q_0^2}{Q_2^2},$$
$$\theta_3^2=\frac{2K}{\pi}=Q_0^2Q_2^4,$$
$$\theta_4^2=\frac{2k'K}{\pi}=Q_0^2Q_3^4, \qquad \theta_1'=\left(\frac{8K^3kk'}{\pi^3}\right)^{1/2}=2q^{1/4}Q_0^3.$$

*Received by the editors March 23, 1982, and in revised form September 8, 1982.

†Department of Physics, University of Surrey, Guildford, England GU2 5XH.

All these results (except those involving θ_5) are standard and may be found in Whittaker and Watson [6]. The following series are referred to:

$$\begin{aligned}
&\zeta(s)=L_1(s)=1+2^{-s}+3^{-s}+4^{-s}+\cdots,\\
&\eta(s)=(1-2^{1-s})\zeta(s)=1-2^{-s}+3^{-s}-4^{-s}+\cdots,\\
&\lambda(s)=(1-2^{-s})\zeta(s)=1+3^{-s}+5^{-s}+\cdots,\\
&\beta(s)=L_{-4}(s)=1-3^{-s}+5^{-s}-7^{-s}+\cdots,\\
&L_{-8}(s)=1+3^{-s}-5^{-s}-7^{-s}-\cdots,\\
&L_{+8}(s)=1-3^{-s}-5^{-s}+7^{-s}+\cdots.
\end{aligned}\tag{6}$$

Let Σ mean the sum over all integer values of all indices which appear and Σ' the same sum excluding only the case where all the integers are simultaneously zero. Denote the Mellin transform M_s where

$$M_s[f(t)]=\Gamma(s)^{-1}\int_0^{\infty}t^{s-1}f(t)\,dt.$$

It is possible to express multiple sums as Mellin transforms of θ-functions of argument e^{-t}, e.g.,

$$\begin{aligned}
&\sum{}'(n^2)^{-s}=M_s[\theta_3(e^{-t})-1],\\
&\sum{}'(-1)^n(n^2)^{-s}=M_s[\theta_4(e^{-t})-1],\\
&\sum\left[(n-\tfrac{1}{2})^2\right]^{-s}=M_s[\theta_2(e^{-t})],\\
&\sum{}'(m^2+n^2)^{-s}=M_s[\theta_3^2-1],
\end{aligned}\tag{7}$$

etc. The property

$$M_s[f(e^{-kt})]=k^{-s}M_s[f(e^{-t})]\tag{8}$$

is also used. Further details may be found in [5] and the references therein.

The results presented here are exhibited in Tables 1–3 and are discussed below.

Deduction of results in Table 1. The method of obtaining the results of Table 1 is illustrated for (T1.1). From (4) and (5) it is easily established that

$$Q_0=q^{-1/12}\left(\frac{2kk'K^3}{\pi^3}\right)^{1/6}.\tag{9}$$

Taking logarithms, the left-hand side of (9) becomes

$$-\log Q_0=-\log\prod_1^{\infty}(1-q^{2n})=-\sum_1^{\infty}\log(1-q^{2n})=\sum_1^{\infty}\sum_1^{\infty}\frac{q^{2nm}}{m}.$$

Reversing the order of summation we have

$$-\log Q_0=\sum_1^{\infty}\frac{q^{2m}}{m(1-q^{2m})}=\sum_1^{\infty}\frac{1}{m(e^{2\pi mc}-1)}=\frac{1}{2}\sum_1^{\infty}\frac{\coth(\pi mc)-1}{m}$$

and thus we get (T1.1). Equations (T1.2)–(T1.9) may be found in a similar fashion. (T1.10) is not so simple to find. Jacobi [7] established it by successive transformations

Table 1

(T1.1)	$-\log Q_0=\sum_1^\infty\frac{1}{n(e^{2\pi nc}-1)}=\frac{1}{2}\sum_1^\infty\frac{\coth(\pi nc)-1}{n}=-\frac{\pi c}{12}-\frac{1}{6}\log\frac{2K^3kk'}{\pi^3}$
(T1.2)	$\begin{aligned}-\log Q_1&=\sum_1^\infty\frac{(-1)^n}{n(e^{2\pi nc}-1)}\\&=\frac{1}{2}\sum_1^\infty\frac{(-1)^n(\coth(\pi nc)-1)}{n}=-\frac{\pi c}{12}-\frac{1}{12}\log\frac{k^2}{16k'}\end{aligned}$
(T1.3)	$-\log Q_2=\frac{1}{2}\sum_1^\infty\frac{(-1)^n\operatorname{cosech}(\pi nc)}{n}=\frac{\pi c}{24}-\frac{1}{12}\log\frac{4}{kk'}$
(T1.4)	$-\log Q_3=\frac{1}{2}\sum_1^\infty\frac{\operatorname{cosech}(\pi nc)}{n}=\frac{\pi c}{24}-\frac{1}{12}\log\frac{4k'^2}{k}$
(T1.5)	$\begin{aligned}\log Q_0Q_1^2&=\log\frac{\theta_2}{2q^{1/4}}=\sum_1^\infty\frac{1}{n(e^{2\pi nc}+1)}\\&=\frac{1}{2}\sum_1^\infty\frac{1-\tanh(\pi nc)}{n}=\frac{\pi c}{4}+\frac{1}{2}\log\frac{kK}{2\pi}\end{aligned}$
(T1.6)	$\frac{1}{2}\log Q_0Q_2^2=\frac{1}{2}\log\theta_3=\sum_1^\infty\frac{1}{(2n-1)[e^{\pi(2n-1)c}+1]}=\frac{1}{4}\log\frac{2K}{\pi}$
(T1.7)	$\begin{aligned}&-\frac{1}{2}\log Q_0Q_3^2=-\frac{1}{2}\log\theta_4\\&=\sum_1^\infty\frac{1}{(2n-1)[e^{\pi(2n-1)c}-1]}=\sum_1^\infty\tanh^{-1}e^{-\pi nc}=-\frac{1}{4}\log\frac{2k'K}{\pi}\end{aligned}$
(T1.8)	$\begin{aligned}\log Q_1(q^2)/Q_2(q^2)&=\sum_1^\infty\frac{(-1)^n}{n(e^{2\pi nc}+1)}\\&=\frac{1}{2}\sum_1^\infty\frac{(-1)^n[1-\tanh(\pi nc)]}{n}=\frac{\pi c}{4}+\frac{1}{2}\log\left(\frac{1-k'}{2k}\right)\end{aligned}$
(T1.9)	$\log Q_1Q_2^2=\sum_1^\infty\frac{\operatorname{cosech}[\pi(2n-1)c]}{2n-1}=-\frac{1}{4}\log k'$
(T1.10)	$\sum_1^\infty(-1)^{n+1}\frac{\operatorname{sech}[(2n-1)\pi c/2]}{2n-1}=2\sum_1^\infty(-1)^{n+1}\tan^{-1}e^{-\pi c(2n-1)/2}=\frac{1}{2}\sin^{-1}k$

of (T1.9). First put $q^{1/2}$ for q; then if this is done k' must be replaced by $(1-k)/(1+k)$ and c by $c/2$. Thus

$$(10)\qquad \sum_1^\infty\frac{\operatorname{cosech}[(2n-1)\pi c/2]}{2n-1}=-\frac{1}{4}\log\left(\frac{1-k}{1+k}\right).$$

Then put $-q$ for q, whence $k\to ik/k'$. Further let $k=\sin\theta$. The right-hand side of (10) becomes (since $k'=\cos\theta$)

$$-\frac{1}{4}\log\left(\frac{k'-ik}{k'+ik}\right)=-\frac{1}{4}\log\left(\frac{\cos\theta-i\sin\theta}{\cos\theta+i\sin\theta}\right)=\frac{i\theta}{2}.$$

The left-hand side of (10) becomes

$$i\sum_{1}^{\infty}(-1)^{n+1}\frac{\operatorname{sech}[(2n-1)\pi c/2]}{(2n-1)};$$

therefore

$$\sum_{1}^{\infty}(-1)^{n+1}\frac{\operatorname{sech}[(2n-1)\pi c/2]}{2n-1}=2\sum_{1}^{\infty}(-1)^{n+1}\tan^{-1}q^{(2n-1)/2}=\frac{\theta}{2}=\frac{1}{2}\sin^{-1}k.$$

Jacobi [7] (who does not actually give the result in terms of hyperbolic functions) refers to this result as "quae inter formulas elegantissimas censeri debit".

Many results similar to (T1.1)–(T1.10) may be found by using the transformations $q^{1/2}$, q^2 or $-q$ for q. We summarize the essential points below.

$$q\to q^{1/2},\quad k'\to\frac{1-k}{1+k},\quad k\to\frac{2k^{1/2}}{1+k},\quad K\to(1+k)K,\quad c\to c/2;$$

$$q\to q^{2},\quad k'\to\frac{2k'^{1/2}}{1+k'},\quad k\to\frac{1-k'}{1+k'},\quad K\to\frac{1+k'}{2}K,\quad c\to 2c;$$

$$q\to -q,\quad k\to\frac{ik}{k'},\quad k'\to\frac{1}{k},\quad K\to k'K;$$

e.g., in (T1.10) put q^2 for q; then

$$\sum_{1}^{\infty}(-1)^{n+1}\frac{\operatorname{sech}[(2n-1)\pi c]}{2n-1}=2\sum_{1}^{\infty}(-1)^{n+1}\tan^{-1}q^{2n-1}=\frac{1}{2}\sin^{-1}\left(\frac{1-k'}{1+k'}\right).$$

Deduction of results in Table 2. The method of obtaining results similar to those in Table 2 has been described elsewhere [4]. The results given here are new. Some follow quite naturally from previous results; e.g., (T2.1) may be found by representing the sum

TABLE 2

Σ implies summation over all integer values of the indices. Σ' implies summation over all integer values of the indices excluding the case where they are simultaneously zero.

(T2.1)	$\sum'(-1)^{m}(m^2+n^2+p^2+r^2)^{-s}=M_s[\theta_4\theta_3^3-1]$ $=4\beta(s)\beta(s-1)-2^{3-2s}\eta(s)\eta(s-1)$
(T2.2)	$\sum'(-1)^{m+n+p}(m^2+n^2+p^2+r^2)^{-s}=M_s[\theta_4^3\theta_3-1]$ $=-4\beta(s)\beta(s-1)-2^{3-2s}\eta(s)\eta(s-1)$
(T2.3)	$\sum(m^2+(n-\frac{1}{2})^2+(p-\frac{1}{2})^2+(r-\frac{1}{2})^2)^{-s}=M_s[\theta_3\theta_2^3]$ $=2^{2s}[\lambda(s)\lambda(s-1)-\beta(s)\beta(s-1)]$
(T2.4)	$\sum(m^2+n^2+p^2+(r-\frac{1}{2})^2)^{-s}=M_s[\theta_3^3\theta_2]$ $=2^{2s}[\lambda(s)\lambda(s-1)+\beta(s)\beta(s-1)]$
(T2.5)	$\sum(-1)^{m+n+p}(m^2+n^2+p^2+(r-\frac{1}{2})^2)^{-s}=M_s[\theta_4^3\theta_2]$ $=2^{2s}[L_{-8}(s)L_{-8}(s-1)+L_8(s)L_8(s-1)]$
(T2.6)	$\sum(-1)^{m}(m^2+(n-\frac{1}{2})^2+(p-\frac{1}{2})^2+(r-\frac{1}{2})^2)^{-s}=M_s[\theta_4\theta_2^3]$ $=2^{2s}[L_{-8}(s)L_{-8}(s-1)-L_8(s)L_8(s-1)]$

as $M_s[\theta_4\theta_3^3-1]$. Using the fact that $\theta_3\theta_4=\theta_4^2(q^2)$ and $\theta_3^2=\theta_3^2(q^2)+\theta_4^2(q^2)$, we have

$$\begin{aligned} M_s[\theta_4\theta_3^3-1] &= M_s[\theta_4^2(q^2)\theta_3^2(q^2)-1+\theta_4^2(q^2)\theta_2^2(q^2)] \\ &= 2^{-s}M_s[\theta_4^2\theta_3^2-1+\theta_4^2\theta_2^2]. \end{aligned}$$

The results for both $M_s[\theta_3^2\theta_4-1]$ and $M_s[\theta_4^2\theta_2^2]$ are given in [4], and hence (T2.1) is found. The results for (T2.5) and (T2.6) are not so easy to establish. We use the relations $\theta_2^2(q^{1/2})=2\theta_2\theta_4$, $\theta_2^2(q^{1/2})=2\theta_2\theta_3$, $[\theta_2/2q^{1/4}](iq)=\theta_5/2q^{1/4}$ and $\theta_3(-q)=\theta_4$.Thus

$$M_s[\theta_2\theta_4^3]=M_s[2^{-1}\theta_5^2(q^{1/2})\theta_4^2]=2^{s-1}M_s[\theta_5^2\theta_4^2(q^2)].$$

Now $\theta_2^2\theta_3^2(q^2)=\frac{1}{2}[\theta_2^2\theta_3^2+\theta_2^2\theta_4^2]$, and from series given by Jacobi [7], this may be written as

$$\frac{\theta_2^2\theta_3^2(q^2)}{q^{1/2}}=4\sum_0^\infty\frac{(4n+1)q^{2n}}{1-q^{8n+2}}+\frac{(4n+3)q^{6n+4}}{1-q^{8n+6}}. \tag{11}$$

Putting iq for q in (11), we have

$$\frac{\theta_5^2\theta_4^2(q^2)}{q^{1/2}}=4\sum_0^\infty\frac{(-1)^n(4n+1)q^{2n}}{1+q^{8n+2}}+\frac{(-1)^n(4n+3)q^{6n+4}}{1+q^{8n+6}};$$

therefore

$$\begin{aligned} M_s[\theta_5^2\theta_4^2(q^2)] &= 4\sum_0^\infty\sum_0^\infty\frac{(-1)^m(-1)^n}{(4n+1)^{s-1}(2m+\frac{1}{2})^s}+\frac{(-1)^m(-1)^n}{(4n+3)^{s-1}(2m+\frac{3}{2})^s} \\ &= 2^{s+1}[L_{-8}(s)L_{-8}(s-1)+L_{+8}(s)L_{+8}(s-1)]. \end{aligned}$$

Finally,

$$\begin{aligned} M_s[\theta_2\theta_4^3] &= 2^{s-1}M_s[\theta_5^2\theta_4^2(q^2)] \\ &= 2^{2s}[L_{-8}(s)L_{-8}(s-1)+L_8(s)L_8(s-1)], \end{aligned}$$

which is (T2.5). (T2.6) is found in a similar fashion.

Deduction of results of Table 3. Whereas the results in Tables 1 and 2 may be derived in a fairly systematic manner, those in Table 3 are just a collection of odd results which appear from a variety of sources. (T3.1) and (T3.2) were obtained by evaluating certain triple sums by two differing ways involving their reduction to double sums. For details see Chaba and Pathria [8] who first gave them. (T3.3) was obtained from a certain dipole sum originally investigated by Nijboer and de Wette [9], and (T3.4) was found by Glasser [5] using contour integration. Both may be obtained by using the identities

$$\sum_{-\infty}^{\infty}\frac{b^2}{m^2+b^2}=b+\frac{2\pi b}{e^{2\pi b}-1}, \tag{12}$$

TABLE 3

(T3.1)	$\sum_1^\infty \sum_1^\infty [m^2+(2n-1)^2]^{-1/2}\{\exp[\pi(m^2+(2n-1)^2)^{1/2}]+(-1)^m\}^{-1}$ $=\frac{4-3\sqrt{2}}{8}\zeta(\frac{1}{2})\beta(\frac{1}{2})-\frac{\eta}{16}$
(T3.2)	$\sum_1^\infty \sum_1^\infty [m^2+4n^2]^{-1/2}\{\exp[\pi(m^2+4n^2)^{1/2}]+(-1)^m\}^{-1}$ $=\frac{9}{16}\log 2-\frac{\pi}{16}+\frac{\sqrt{2}}{8}\zeta(\frac{1}{2})\beta(\frac{1}{2})-\frac{\eta}{16}$ $\eta=2\log\frac{2K[1]}{\pi}=\log\Gamma^4(\frac{1}{4})/4\pi^3$
(T3.3)	$\sum_1^\infty \sum_1^\infty (m^2+n^2)^{1/2}\{\exp[2\pi(m^2+n^2)^{1/2}]-1\}^{-1}=\frac{\zeta(\frac{3}{2})\beta(\frac{3}{2})}{8\pi^2}+\frac{1}{24}\left(\frac{1}{\pi}-1\right)$
(T3.4)	$\sum_1^\infty \sum_1^\infty (-1)^{m+n}(m^2+n^2)^{1/2}\operatorname{cosech}[\pi(m^2+n^2)^{1/2}]=1/12\pi$
(T3.5)	$\sum_1^\infty \sum_1^\infty \frac{\operatorname{cosech}\left\{\pi\left[m^2+\left(n-\frac{1}{2}\right)^2\right]^{1/2}\right\}}{\left[m^2+\left(n-\frac{1}{2}\right)^2\right]^{1/2}}=\frac{1}{2}[1-\log(1+\sqrt{2})]$
(T3.6)	$\sum_1^\infty \sum_1^\infty \frac{\operatorname{cosech}\left\{\pi\left[\left(m-\frac{1}{2}\right)^2+\left(n-\frac{1}{2}\right)^2\right]^{1/2}/\sqrt{2}\right\}}{\left[\left(m-\frac{1}{2}\right)^2+\left(n-\frac{1}{2}\right)^2\right]^{1/2}}=\frac{1}{\sqrt{2}}$
(T3.7)	$\sum_1^\infty \sum_1^\infty \frac{\operatorname{cosech}\left\{\pi\left[\frac{1}{2}m^2+\frac{1}{4}\left(n-\frac{1}{2}\right)^2\right]^{1/2}\right\}}{\left[\frac{1}{2}m^2+\frac{1}{4}\left(n-\frac{1}{2}\right)^2\right]^{1/2}}=\sqrt{8}-\log(\sqrt{2}+1)(2^{1/4}+1)^2$
(T3.8)	$\sum{}' \frac{\operatorname{cosech}\left[\pi\left(m^2+n^2+p^2\right)^{1/2}\right]}{\left(m^2+n^2+p^2\right)^{1/2}}=\frac{\pi}{6}+\frac{1}{2}-\frac{\log 2}{\pi}$
(T3.9)	$\sum_1^\infty \sum_1^\infty \sum_1^\infty \frac{\operatorname{cosech}\left\{\pi\left[(2m-1)^2+(2n-1)^2+(2p-1)^2\right]^{1/2}\right\}}{\left[(2m-1)^2+(2n-1)^2+(2p-1)^2\right]^{1/2}}=\frac{3\log 2}{32\pi}-\frac{1}{64}$
(T3.10)	$\sum_1^\infty \sum_1^\infty \sum_1^\infty \frac{\operatorname{cosech}\left\{\pi\left[\left(m-\frac{1}{2}\right)^2+\left(n-\frac{1}{2}\right)^2+\left(p-\frac{1}{2}\right)^2\right]^{1/2}\right\}}{\left[(2m-1)^2+(2n-1)^2+(2p-1)^2\right]^{1/2}}=\frac{\sqrt{2}}{16}$

(13)
$$\sum_{-\infty}^{\infty}\frac{(-1)^m}{m^2+b^2}=\frac{\pi\operatorname{cosech}(\pi b)}{b}.$$

(T3.5)–(T3.10) may be evaluated using (13). We evaluate (T3.5) as an example. We have that

(14)
$$\sum(-1)^m\left[m^2+n^2+\left(p-\frac{1}{2}\right)^2\right]^{-s}=M_s[\theta_2\theta_3\theta_4]=M_s[\theta_1']=2^{2s+1}\beta(2s-1).$$

For $s=1$ the left-hand side of (14) may be written

$$\sum \frac{\pi \operatorname{cosech}\left\{\pi\left[n^2+\left(p-\frac{1}{2}\right)^2\right]^{1/2}\right\}}{\left[n^2+\left(p-\frac{1}{2}\right)^2\right]^{1/2}}$$

$$=4\pi \sum_1^\infty \sum_1^\infty \frac{\operatorname{cosech}\left\{\pi\left[n^2+\left(p-\frac{1}{2}\right)^2\right]^{1/2}\right\}}{\left[n^2+\left(p-\frac{1}{2}\right)\right]^{1/2}}+4\pi \sum_1^\infty \frac{\operatorname{cosech}[(2p-1)\pi/2]}{2p-1}.$$

But from (T1.9)

$$\sum_1^\infty \frac{\operatorname{cosech}[(2n-1)\pi/2]}{2p-1}=\frac{1}{2}\log\left(\sqrt{2}+1\right),$$

since for $c=\frac{1}{2}$, k is $(\sqrt{2}+1)^{-2}$. For $s=1$ the right-hand side of (14) is equal to 2π, and thus (T3.5) is obtained. (T3.5)–(T3.10) are by no means unique—many other similar formulae are obtainable.

Discussion. The results in Table 1 may be extended in a number of ways. One method is to combine the various formulae of Table 1. For example, Berndt [2] obtains the result

$$(15) \qquad \sum_{n=1}^\infty \left[\frac{1}{n\left[(-1)^n e^{\pi n\sqrt{3}}+1\right]}-\frac{2}{(2n-1)\left[e^{(2n-1)\pi\sqrt{3}}+1\right]}\right]=\frac{\pi\sqrt{3}}{8}-\log 2.$$

However, the left-hand side of (15) may be expressed for general c by the formulae in Table 1. Indeed, for any c it is

$$(16) \qquad \frac{1}{2}(\mathrm{T}1.5)-(\mathrm{T}1.7)-2(\mathrm{T}1.6)=\frac{\pi c}{8}+\frac{1}{4}\log\frac{kk'}{4}.$$

For $c=\sqrt{3}$, $kk'=\frac{1}{4}$ and (15) follows. Just as neat a result then appears for $c=\sqrt{7}$, for then $kk'=\frac{1}{16}$. Hence, if in (15) we replace $\sqrt{3}$ by $\sqrt{7}$, the right-hand side becomes $\pi\sqrt{7}/8-(3\log 2)/2$.

Another way of obtaining further formulae is to differentiate with respect to the parameter c. It may be shown that

$$(17) \qquad \frac{dk}{dc}=-\frac{2kk'^2K^2}{\pi}, \qquad \frac{dk'}{dc}=\frac{2k^2k'K^2}{\pi}.$$

Thus for example differentiating (T1.3) gives

$$(18) \qquad \sum_1^\infty (-1)^{n+1}\operatorname{cosech}(\pi nc)\coth(\pi nc)=\frac{1}{12}-\frac{K^2}{3\pi^2}(2k^2-1).$$

Each formula in Table 1 when differentiated will give such another similar result; many of them have been discussed elsewhere by the author [10]. There it was also explained that when c is a rational number, K may be written in terms of Γ-functions, and k and k' are surds. Further, for special values of c, particularly attractive looking results may be obtained.

For example, if $c=1$, $k=k'=1/\sqrt{2}$, and $(2k^2-1)$ vanishes. Then (18) gives

$$(19)\qquad \sum_1^\infty(-1)^{n+1}\operatorname{cosech}(\pi n)\coth(\pi n)=\frac{1}{12}=2\sum_0^\infty\frac{2n+1}{e^{(2n+1)\pi}+1}.$$

Again for $c=1$ and $k=1/\sqrt{2}$ from (T1.10) we have

$$(20)\qquad \sum_1^\infty(-1)^{n+1}\frac{\operatorname{sech}[(2n-1)\pi/2]}{(2n-1)}=\frac{1}{2}\sin^{-1}\frac{1}{\sqrt{2}}=\frac{\pi}{8}.$$

For $c=\sqrt{3}$, $k^2=(2-\sqrt{3})/4$; i.e., $\sin^{-1}k=\pi/12$, hence

$$(21)\qquad \sum_1^\infty(-1)^{n+1}\frac{\operatorname{sech}\left[(2n-1)\sqrt{3}\,\pi/2\right]}{(2n-1)}=\frac{\pi}{24}.$$

Both (20) and (21) have been derived by Berndt [1], [2] by other methods. Here, however, another result is immediately apparent for $c=1/\sqrt{3}$, for then $k^2=(2+\sqrt{3})/4$ and $\sin^{-1}k=5\pi/12$, hence

$$(22)\qquad \sum_1^\infty(-1)^{n+1}\frac{\operatorname{sech}\left[(2n-1)\pi/2\sqrt{3}\right]}{(2n-1)}=\frac{5\pi}{24}.$$

Equation (21) is of interest as it is a special case of a general formula given by Berndt [2]. This is

$$(23)\qquad \sum_1^\infty(-1)^{n+1}\frac{\operatorname{sech}\left[(2n-1)\pi\sqrt{3}/2\right]}{(2n-1)^{6N+1}}$$
$$=\frac{1}{2}(-1)^{N+1}\pi^{6N+1}\sum_{k=0}^{3N}\frac{E_{2k+1}}{(2k+1)!}\frac{B_{6N-2k}}{(6N-2k)!}\cos\left[(2k+1)\frac{\pi}{3}\right]$$

where E_n and B_n are the Euler and Bernoulli numbers respectively. Berndt [2] attributes this result (in different form) to Cauchy.

REFERENCES

[1] B. C. Berndt, *Modular transformation and generalisations of several formulae of Ramanujan*, Rocky Mountain J. Math., 7 (1977), pp. 147–189.

[2] ______, *Analytic Eisenstein series, theta functions, and series relations in the spirit of Ramanujan*, J. Reine Angew. Math., 303/304 (1978), pp. 332–365.

[3] M. L. Glasser, *The evaluation of lattice sums*, I. *Analytic procedures*, J. Math. Phys., 14 (1973), pp. 409–413.

[4] I. J. Zucker, *Exact results for some lattice sums in* 2, 4, 6 *and* 8 *dimensions*, J. Phys. A, 7 (1974), pp. 1568–1575.

[5] M. L. Glasser and I. J. Zucker, *Lattice sums*, Theoret. Chem. Adv. and Perspectives, 5 (1980), pp. 67–139.

[6] E. T. Whittaker and G. N. Watson, *Modern Analysis*, 4th ed., Cambridge Univ. Press, Cambridge, 1927.

[7] C. G. J. Jacobi, *Fundamenta Nova Theoriae Functionum Ellipticurum*, Königsberg, 1829.

[8] A. N. Chaba and R. K. Pathria, *Evaluation of lattice sums using Poisson's summation formula* III. J. Phys. A, 9 (1976), pp. 1801–1810.

[9] B. R. A. Nijboer and F. W. de Wette, *The internal field in dipole lattices*, Physica, 24 (1958), pp. 422–431.

[10] I. J. Zucker, *The summation of series of hyperbolic functions*, this Journal, 10 (1979), pp. 192–206.

SIAM J. MATH. ANAL.
Vol. 15, No. 2, March 1984

© 1984 Society for Industrial and Applied Mathematics
0036-1410/84/1502-0014 $01.25/0

ON QUADRATIC TRANSFORMATIONS OF BASIC SERIES*

W. A. AL-SALAM[†] AND A. VERMA[‡]

Abstract. We prove the formula

$$\sum_{n\geq 0} A_nB_n(-xw)^n=\sum_{n=0}^{\infty}(-x)^n[hqp^n;q]_{n-1}\sum_{j=0}^{n}\frac{(1-hq^jp^j)w^jA_j}{[p;p]_{n-j}[hqp^n;q]_j}$$

$$\cdot\sum_{k=0}^{\infty}\frac{p^{k(k-1)/2}[hq^np^n;q]_k}{[p;p]_k}x^kB_{n+k}$$

and show that it implies not only the Euler transformation formula and its q-analogue, but also Carlitz' q-analogue of Whipple's transformation, as well as several other new quadratic transformations of basic (or Heine) series.

1. Introduction.

Euler's transformation formula,

$$\sum_{n\geq 0} a_nb_nx^n=\sum_{k=0}^{\infty}(-1)^k\frac{x^k}{k!}f^{(k)}(x)\Delta^ka_0, \tag{1.1}$$

where

$$f(x)=b_0+b_1x+b_2x^2+\cdots$$

and

$$\Delta^ka_0=\sum_{j=0}^{k}(-1)^j\binom{k}{j}a_{k-j},$$

has been known to be useful not only in addressing accelerating convergence questions but also in finding series identities. For example (1.1) implies Pfaff's formula

$$F(a,b;c;z)=(1-z)^{-a}F\left(a,c-b;c;\frac{-z}{1-z}\right).$$

F. H. Jackson [6] obtained the following q-analogue of (1.1):

$$\sum_{n\geq 0} a_nb_nx^n=\sum_{n=0}^{\infty}\frac{x^k}{[q]_k}\{D_q^nf(x)\}\{\Delta_q^na_0\}, \tag{1.2}$$

where $[a;q]_0=1$, $[a;q]_n=(1-a)(1-aq)\cdots(1-aq^{n-1})$ for $n=1,2,3,\cdots$, $D_qf(x)=(f(x)-f(qx))/x$, and

$$\Delta_q^na_0=\sum_{j=0}^{n}(-1)^j\frac{[q;q]_n}{[q:q]_j[q;q]_{n-j}}q^{j(j-1)/2}a_{n-j}.$$

Thus when $q\to 1$ this formula reduces to (1.1).

*Received by the editors November 20, 1981, and in revised form September 13, 1982.

[†] Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1.
[‡] Arizona State University, Tempe, Arizona 85281.

Jackson himself showed that (1.2) implies the transformation formula [6]

$$ {}_2\phi_1\left[\begin{matrix} a,b \\ c \end{matrix};q;x\right]=\frac{(ax;q)_\infty}{(x;q)_\infty}\,{}_2\phi_2\left[\begin{matrix} a,\frac{c}{b} \\ c,ax \end{matrix};q;bx\right], $$

as well as a large number of other q-formulas.

More recently Chaundy [4], and in an equivalent form Niblett [8], stated a formula which is more general than (1.1) and applied it to find series identities (of hypergeometric type). In particular, Niblett obtained several quadratic transformations of hypergeometric series, some of which are rather strange. We shall refer to them below.

Later Fields and Ismail obtained a formula [5,(2.3)] which contain both (1.1) and (1.2). However they did not apply their formula to its full potential and made no attempt to find quadratic transformations for ordinary or basic hypergeometric functions.

In §2 of this paper we shall first give a formula (see (2.1) below) which is different from that of Fields and Ismail and which contains both (1.1) and (1.2), as well as the Chaundy–Niblett formula. We shall then in §3 apply (2.1) to obtain several q-analogues of Niblett's results. We show that it can be used also to obtain Carlitz' q-analogue of Whipples's transformation.

In §4 we shall give further results on quadratic transformations of basic (or q-) series.

In addition to the notation defined above, we shall also use

$$ [a;q]_{-n}=\frac{(-1)^n q^{n(n+1)/2}}{a^n\left[\frac{q}{a};q\right]_n}, \qquad n=1,2,3,\cdots, $$

$$ [a;q]_\infty=\prod_{k=0}^{\infty}(1-aq^k). $$

Throughout this paper we shall always assume that $|q|<1$.

2. q-analogue of an expansion of Chaundy. We begin this section by proving the following q-analogue of an expansion of Chaundy [4]:

(2.1)

$$ \sum_{n=0s\infty} A_nB_n(-xw)^n=\sum_{k=0}^{\infty}(-x)^k[hqp^k;q]_{k-1} $$

$$ \cdot\sum_{n=0}^{k}\frac{(1-hq^np^n)w^nA_n}{[p;p]_{k-n}[hqp^k;q]_n}\sum_{j=0}^{\infty}\frac{p^{j(j-1)/2}[hq^kp^k;q]_jx^j}{[p;p]_j}B_{j+k}. $$

Proof of (2.1). Since the jth q-difference of a polynomial of degree $\leq j-1$ is equal to zero, we have

(2.2)
$$ \left(1-\frac{d}{q}\right)\sum_{k=0}^{j}\frac{(-1)^kp^{(j-k)(j-k-1)/2}}{[p;p]_k[p;p]_{j-k}}[dp^k;q]_{j-1}=\delta_{j,0}. $$

Multiplying both sides by $B_{n+j}(-x)^{j+n}$ and summing from $j=0$ to ∞, interchanging the summation and then replacing k by $k-n$ and x by $-x$, we get

$$(-x)^n B_n=\left(1-\frac{d}{q}\right)\sum_{k=n}^{\infty}\frac{(-x)^k}{[p:p]_{k-n}}\sum_{j=0}^{\infty}\frac{\left[dp^{-n+k};q\right]_{j-n+k-1}p^{j(j-1)/2}x^j}{[p;p]_j}B_{j+k}.$$

Now setting $d=hq^{1+n}p^n$, multiplying both sides by $A_n w^n$ and summing from $n=0$ to ∞ and rearranging the series, we get (2.1).

It might be of interest to note that formula (2.1) is equivalent to the fact that the triangular matrix $H=(h_{nj})$, where

$$h_{nj}=\frac{(-1)^{n+j}[hqp^n;q]_{n-1}(1-hq^jp^j)}{[p;p]_{n-j}[hqp^n;q]_j},$$

is inverse to the triangular matrix $G=(g_{jn})$, where

$$g_{jn}=\frac{[hp^nq^n;q]_{j-n}p^{(j-n)(j-n-1)/2}}{[p;p]_{j-n}}.$$

The expansion (2.1) is a q-analogue of equivalent results in [4], [8] and [12] which contain a number of known results as special cases and in particular the Euler transformation (1.1).

In (2.1), setting $p=1/q$, $w=q$ and letting $h\to\infty$, we get the q-analogue of Euler's transformation (1.2) due to Jackson [6] (see also Bailey [2]).

On the other hand, in (2.1), replacing x and w by $-\beta x$ and $qh/bc\beta$ respectively, and then assuming that h is of the form q^{-m} (so that all the series involved are finite series) and

$$p=q,\qquad A_n=\frac{[h;q]_n[b;q]_n[c;q]_n[\beta;q]_n}{[q;q]_n\left[\frac{qh}{b};q\right]_n\left[\frac{qh}{c};q\right]_n},\qquad B_n=\frac{1}{[\beta;q]_n},$$

we get on letting $\beta\to\infty$

$$ {}_3\phi_2\left[\begin{matrix}h,b,c,\\ \frac{qh}{b},\frac{qh}{c}\end{matrix};q;\frac{xqh}{bc}\right]=\sum_{k=0}^{\infty}\frac{[h;q]_{2k}(-x)^kq^{-k(k-1)/2}}{[hq;q]_k[q;q]_k}\,{}_1\phi_0\left[\begin{matrix}hq^{2k}\\ \overline{}\end{matrix};q;xq^{-k}\right]$$

$$\cdot{}_6\phi_5\left[\begin{matrix}h,q\sqrt{h},-q\sqrt{h},b,c,q^{-k}\\ \sqrt{h},-\sqrt{h},\frac{qh}{b},\frac{qh}{c},hq^{1+k}\end{matrix};q;\frac{hq}{bc}^{1+k}\right].$$

Summing the ${}_1\phi_0[x]$ and ${}_6\phi_5$ [10,(6.2)], we get Carlitz' [3] q-analogue of Whipple's transformation

$$(2.3)\qquad {}_3\phi_2\left[\begin{matrix}h,b,c\\ \frac{qh}{b},\frac{qh}{c}\end{matrix};q;\frac{xqh}{bc}\right]=\frac{[hx;q]_\infty}{[x;q]_\infty}\,{}_5\phi_4\left[\begin{matrix}\sqrt{h},-\sqrt{h},\sqrt{qh},-\sqrt{qh},\frac{qh}{bc}\\ xh,\frac{q}{x},\frac{qh}{b},\frac{qh}{c}\end{matrix};q;q\right],$$

where h is of the form q^{-m}. In fact this result was also given earlier by Sears [10,(4.1)].

3. *q*-analogues of some quadratic transformations of Niblett. Let us consider (2.1) and put

$$A_n=\frac{[h;q]_n[\gamma;q]_n[\delta;q]_n[\varepsilon;q]_n[\theta;q]_n[\beta;q]_n}{[q;q]_n\left[\frac{qh}{\gamma};q\right]_n\left[\frac{qh}{\delta};q\right]_n\left[\frac{qh}{\varepsilon};q\right]_n\left[\frac{qh}{\theta};q\right]_n},$$

$$B_n=\frac{1}{[\beta;q]_n},\quad p=q,\quad w=\frac{h^2q^2}{\gamma\delta\varepsilon\theta\beta};$$

replacing x by $-\beta x$, $h=q^{-m}$ we get on letting $\beta\to\infty$

(3.1)

$$\begin{aligned}
&{}_5\Phi_4\left[\begin{matrix}h,\gamma,\delta,\varepsilon,\theta\\ \frac{qh}{\gamma},\frac{qh}{\delta},\frac{qh}{\varepsilon},\frac{qh}{\theta}\end{matrix};q;\frac{xh^2q^2}{\gamma\delta\varepsilon\theta}\right]\\
&=\frac{[xh;q]_\infty}{[x;q]_\infty}\sum_{k=0}^{\infty}\frac{[h;q]_{2k}q^k}{[q;q]_k[qh;q]_k}\\
&\quad\cdot\frac{1}{[xh;q]_k\left[\frac{q}{x};q\right]_k}\,{}_8\Phi_7\left[\begin{matrix}h,q\sqrt{h},-q\sqrt{h},\gamma,\delta,\varepsilon,\theta,q^{-k}\\ \sqrt{h},-\sqrt{h},\frac{qh}{\gamma},\frac{qh}{\delta},\frac{qh}{\varepsilon},\frac{qh}{\theta},hq^{1+k}\end{matrix};q;\frac{h^2q^{2+k}}{\gamma\delta\varepsilon\theta}\right].
\end{aligned}$$

Transforming the inner well-poised ${}_8\Phi_7$ by Watson's q-analogue of Whipple's transformation [1, 8.5(2)], we get

(3.2)

$$\begin{aligned}
&{}_5\Phi_4\left[\begin{matrix}h,\gamma,\delta,\varepsilon,\theta\\ \frac{qh}{\gamma},\frac{qh}{\delta},\frac{qh}{\varepsilon},\frac{qh}{\theta}\end{matrix};q;\frac{xq^2h^2}{\gamma\delta\varepsilon\theta}\right]\\
&=\frac{[xh;q]_\infty}{[x;q]_\infty}\sum_{k=0}^{\infty}\frac{[h;q]_{2k}\left[\frac{qh}{\varepsilon\theta};q\right]_k}{[q;q]_k\left[\frac{qh}{\varepsilon};q\right]_k}\\
&\quad\cdot\frac{q^k}{\left[\frac{qh}{\theta};q\right]_k[xh;q]_k[q/x;q]_k}\,{}_4\Phi_3\left[\begin{matrix}\frac{qh}{\gamma\delta},\varepsilon,\theta,q^{-k}\\ \frac{\varepsilon\theta}{h}q^{-k},\frac{qh}{\gamma},\frac{qh}{\delta},\end{matrix};q;q\right]\\
&=\frac{[xh;q]_\infty}{[x;q]_\infty}\sum_{k=0}^{\infty}\frac{[h;q]_{2k}q^k}{[q;q]_k\left[\frac{qh}{\varepsilon};q\right]_k\left[\frac{qh}{\theta};q\right]_k[xh;q]_k\left[\frac{q}{x};q\right]_k}\\
&\quad\cdot\sum_{j=0}^{k}\frac{\left[\frac{qh}{\gamma\delta};q\right]_j[\varepsilon;q]_j[\theta;q]_j[q^{-k};q]_j\left[\frac{hq}{\varepsilon\theta};q\right]_{k-j}}{[q;q]_j\left[\frac{qh}{\gamma};q\right]_j\left[\frac{qh}{\delta};q\right]_j}\left(\frac{-h}{\varepsilon\theta}\right)^j q^{kj-j(j-1)/2+j}.
\end{aligned}$$

Now let $\varepsilon\theta = hq^2$ so that $[hq/\varepsilon\theta; q]_{k-j} = [q^{-1}; q]_{k-j}$. Hence, after some calculation, we get that

(3.3)

$$
{}_5\Phi_4\left[\begin{matrix} h,\gamma,\delta,\varepsilon,\theta \\ \dfrac{qh}{\gamma},\dfrac{qh}{\delta},\dfrac{qh}{\varepsilon},\dfrac{qh}{\theta} \end{matrix} ;q;\frac{xh}{\gamma\delta}\right] = \frac{[xh;q]_\infty}{[x;h]_\infty}
$$

$$
\cdot {}_6\Phi_5\left[\begin{matrix} \sqrt{h},-\sqrt{h},\sqrt{qh},-\sqrt{qh},\dfrac{h}{\delta\gamma},qD \\ \dfrac{qh}{\gamma},\dfrac{qh}{\delta},xh,\dfrac{q}{x},D; \end{matrix} ;q;q\right],
$$

where $\varepsilon\theta = hq^2$ and

$$
\left\{1+\frac{h}{\delta}+\frac{h}{\gamma}-\frac{\varepsilon+\theta}{q}-\frac{h}{\gamma\delta}\right\}(1-D) = \left(1-\frac{h}{\gamma\delta}\right)\left(1-\frac{\theta}{q}\right)\left(1-\frac{\varepsilon}{q}\right).
$$

The special case $D\to\infty$ (so that $(\varepsilon+\theta)/q = 1-h/\gamma\delta+h/\gamma+h/\delta$) leads to

(3.4)

$$
{}_5\Phi_4\left[\begin{matrix} h,\gamma,\delta,\varepsilon,\theta \\ \dfrac{qh}{\gamma},\dfrac{qh}{\delta},\dfrac{qh}{\varepsilon},\dfrac{qh}{\theta} \end{matrix} ;q;\frac{xh}{\gamma\delta}\right] = \frac{[xh;q]_\infty}{[x;q]_\infty}\,{}_5\Phi_4\left[\begin{matrix} \sqrt{h},-\sqrt{h},\sqrt{qh},-\sqrt{qh},\dfrac{h}{\gamma\delta} \\ \dfrac{qh}{\gamma},\dfrac{qh}{\delta},xh,\dfrac{q}{x} \end{matrix} ;q;q^2\right],
$$

where ε and θ are roots of the quadratic equation

$$
m^2-q\left(1-\frac{h}{\gamma\delta}+\frac{h}{\gamma}+\frac{h}{\delta}\right)m+hq^2=0.
$$

Formula (3.4) is a q-analogue of another result of Niblett [8,(22)] which could be deduced from (3.4) by replacing $h,\gamma,\delta,\varepsilon,\theta$ by q^{2a}, q^{e-c-1}, q^{1+2a-e}, $q^{1+\phi}$ and $q^{1+\theta}$, where θ and ϕ are now the roots of the quadratic equation

$$
\begin{aligned}
(1-m)^2-\left[(1-q^{e-1})+q^c(1-q^{2a-e+1})\right](1-m)& \\
+q^c(1-q^{2a-e+1})(1-a^{e-c-1})&=0,
\end{aligned}
$$

and letting $q\to 1$.

4. Some further quadratic transformations. Bailey [1, p. 97, ex. 5] pointed out that there is equivalence between Whipple's quadratic transformation [1, p. 97, ex. 4(iv)] for the well-poised ${}_3F_2[x]$ and Whipple's transformation of a nearly-poised ${}_4F_3(1)$ into a Saalschützian ${}_5F_4(1)$ [1, 4.5(1)] (i.e one implies the other).

The q-analogue (2.3) of Whipple's transformation does not seem to imply a transformation between q-series, which, in the limiting case, reduce to Whipple's transformation [1, 4.5(1)] of a nearly-poised ${}_4F_3(1)$ into a Saalschützian ${}_5F_4(1)$. In this section we shall obtain another q-analogue (different from (3.1)) of Whipple's quadratic transformation [1, p. 97, ex. 4(iv)] for ${}_3F_2(x)$ which overcomes this difficulty.

Verma and Jain [13] have obtained the following q-analogue of Whipple's transformation [1, 4.5(1)] of a nearly-poised ${}_4F_3(1)$ into a Saalschützian ${}_5F_4(1)$:

$$(4.1)\quad \sum_{r\geq 0}\frac{[a^2;q^2]_r[-aq^2;q^2]_r[b^2;q^2]_r[c^2;q^2]_r\left[-\frac{aq}{w};q\right]_r[d;q]_r}{[q^2;q^2]_r[-a;q^2]_r\left[\frac{a^2q^2}{b^2};q^2\right]_r\left[\frac{a^2q^2}{c^2};q^2\right]_r[w;q]_r\left[-\frac{aq}{d};q\right]_r}\cdot\left(\frac{awq^2}{b^2c^2d}\right)^r$$

$$=\frac{[-aq;q]_\infty\left[\frac{w}{a};q\right]_\infty\left[-\frac{q}{d};q\right]_\infty\left[\frac{w}{d};q\right]_\infty}{[-q;q]_\infty[w;q]_\infty\left[-\frac{aq}{d};q\right]_\infty\left[\frac{w}{ad};q\right]_\infty}$$

$$\cdot{}_5\phi_4\left[\begin{matrix}\frac{a^2q^2}{b^2c^2},a,aq,\frac{a^2q^2}{w^2},d^2\\ \frac{a^2}{b^2}q^2,\frac{a^2}{c^2}q^2,\frac{aqd}{w},\frac{adq^2}{w}\end{matrix};q^2;q^2\right],$$

where a or d is of the form q^{-m}.

In (4.1), setting $d=q^{-m}$, multiplying both sides by

$$\frac{[w;q]_m[x;q]_m}{[q;q]_m[-aq;q]_m}\left(-\frac{aq}{xw}\right)^m$$

and summing from $m=0$ to ∞, interchanging the order of summations on both sides and summing the resulting inner ${}_2\phi_1$ series by the q-analogue of Gauss' summation theorem, we get on assuming that a is of the form q^{-j}

$$(4.2)\quad \sum_{r=0}^{\infty}\frac{[a^2;q^2]_r[-aq^2;q^2]_r[b^2;q^2]_r[c^2;q^2]_r[x;q]_rq^{r(r+1)/2}}{[q^2;q^2]_r[-a;q^2]_r\left[\frac{a^2q^2}{b^2};q^2\right]_r\left[\frac{a^2q^2}{c^2};q^2\right]_r\left[-\frac{aq}{x};q\right]_r}\cdot\left(\frac{a^2q^2}{b^2c^2x}\right)^r$$

$$=\frac{[-aq;q]_\infty\left[-\frac{q}{x};q\right]_\infty}{[-q;q]_\infty\left[-\frac{aq}{x};q\right]_\infty}\,{}_4\phi_2\left[\begin{matrix}\frac{a^2q^2}{b^2c^2},a,aq,x\\ \frac{a^2q^2}{b^2},\frac{a^2q^2}{c^2}\end{matrix};q^2;\frac{q}{x^2}\right].$$

This is another q-analogue of Whipple's quadratic transformation [1, p. 97, ex. 4(iv)], to which it reduces on replacing a,b,c, by q^a, q^b, q^c respectively, and then letting $q\to 1$ (with x replaced by $(1-x)/(1+x)$).

On the other hand, in (4.1), replacing a by $-a$ and then setting $c^2=-aq$, $d=q^{-m}$, so that the left-hand side reduces to a well-poised ${}_8\phi_7$, which, when transformed by

Watson's q-analogue of Whipple's transformation [1, 8.5(2)], yields (on replacing a by $-a$)

$$(4.3)\qquad {}_4\Phi_3\left[\begin{matrix} a, b, -\frac{w}{b}, q^{-m} \\ \frac{aq}{b}, -bq^{-m}, w \end{matrix} ;q;q\right] = \frac{\left[\frac{w}{a};q\right]_m \left[\frac{-aq}{b};q\right]_m}{[w;q]_m\left[-\frac{q}{b};q\right]_m}$$

$$\cdot {}_4\Phi_3\left[\begin{matrix} \frac{aq}{b^2}, a, \frac{a^2q^2}{w^2}, q^{-2m} \\ \frac{a^2q^2}{b^2}, \frac{aq^{1-m}}{w}, \frac{aq^{2-m}}{w} \end{matrix} ;q^2;q^2\right].$$

(4.3) for $b=1$ yields the summation

$${}_4\Phi_3\left[\begin{matrix} a, aq, \frac{a^2q^2}{w^2}, q^{-2m} \\ a^2q^2, \frac{aq^{1-m}}{w}, \frac{aq^{2-m}}{w} \end{matrix} ;q^2;q^2\right] = \frac{[w;q]_m[-q;q]_m}{\left[\frac{w}{a};q\right]_m[-aq;q]_m}.$$

In (4.3), replacing a, b, w, by q^a, q^b and q^w respectively, and letting $q \to 1$, we get the following special case of Whipple's transformation [1]:

$${}_3F_2\left[\begin{matrix} a, b, -m; \\ 1+a-b, w \end{matrix}\right] = \frac{(w-a)_m}{(w)_m}\,{}_4F_3\left[\begin{matrix} \frac{1+a-2b}{2}, \frac{a}{2}, 1+a-w, -m; \\ 1+a-b, \frac{1}{2}(1+a-m-w), \frac{1}{2}(2+a-m-w) \end{matrix}\right].$$

Furthermore, transforming the left-hand side of (4.3) by the transformation of a ${}_4\phi_3[q]$ Saalschützian series [9, (8.3)], we get the following q-analogue of Gauss' quadratic transformation for ${}_2F_1(z)$ [1, p. 97, ex. 4(III)]:

$${}_4\Phi_3\left[\begin{matrix} \alpha^2, \beta^2, -\frac{\alpha^2 q}{w}, q^{-m} \\ \alpha\beta\sqrt{q}, \frac{\alpha^2q^{1-m}}{w}, -\alpha\beta q^{1/2} \end{matrix} ;q;q\right] = {}_4\Phi_3\left[\begin{matrix} \alpha^2, \beta^2, \frac{\alpha^4q^2}{w^2}, q^{-2m} \\ \alpha^2\beta^2 q, \frac{\alpha^2}{w}q^{1-m}, \frac{\alpha^2}{w}q^{2-m} \end{matrix} ;q^2;q^2\right].$$

This quadratic transformation was first given in a different but equivalent form by Singh [11]. It was rediscovered later by Askey and Wilson.

Acknowledgment. We are grateful to a referee for pointing out that (3.2) can be put in the form (3.3). He also remarked that (3.3) may be obtained by a different method using the fact that a 2-balanced ${}_3\phi_2$ can be written as a sum of two terms.

REFERENCES

[1] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge, Univ. Press, Cambridge, 1935.
[2] ______, *Identities of Rogers-Ramanujan type*, Proc. London Math. Soc., (2) 50 (1949), pp. 1–10.
[3] L. CARLITZ, *Some formulas of F. H. Jackson*, Monatshefte für Math., 73 (1969), pp. 193–198.
[4] T. W. CHAUNDY, *Some hypergeometric identities*, J. London Math. Soc., 26 (1951), pp. 42–44.
[5] J. L. FIELDS AND M. E. H. ISMAIL, *Polynomial expansions*, Math. Comp., 29 (1975), pp. 894–902.

[6] F. H. JACKSON, *Examples of generalizations of Euler's transformation for power series*, Mess. Math., 57 (1927/28), pp. 169–187.

[7] V. K. JAIN, *Some transformations of basic hypergeometric functions. Part* II, this Journal, 12 (1981), pp. 957–961.

[8] J. D. NIBLETT, *Some hypergeometric identities*, Pacific J. Math., 2 (1952), pp. 219–225.

[9] D. B. SEARS, *On the transformation theory of basic hypergeometric functions*, Proc. London Math. Soc., (2) 53 (1951), pp. 158–191.

[10] ______, *Transformations of basic hypergeometric functions of special type*, Proc. London Math. Soc., (2) (1951), pp. 467–483.

[11] V. N. SINGH, *The basic analogue identities of the Cayley–Orr type*, J. London Math. Soc., 34 (1959), pp. 15–22.

[12] A. VERMA, *On generating functions of classical polynomials*, Proc. Amer. Math. Soc., 46 (1974), pp. 73–76.

[13] A. VERMA AND V. K. JAIN, *q-analogue of Whipple's transformation* (to appear).

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics
001

CLASSIFICATION AND UNFOLDING OF SEQUENTIAL BIFURCATIONS*

GERHARD DANGELMAYR† AND IAN STEWART‡

Abstract. Golubitsky and Schaeffer have developed an extensive theory of imperfect bifurcation by adapting the determinacy and unfolding theorems of singularity/catastrophe theory to singularities with a distinguished parameter λ. In this paper we adapt their methods and results to "sequential bifurcations" of the form $a(u,\lambda)=0$, $b(x,u,\lambda)=0$. Here λ is a bifurcation parameter, x represents the final "output" state of the system, and u is interpreted as a "hidden" variable. Such problems arise when a bifurcating process in (u,λ) is coupled in sequence with a u-dependent bifurcation process in (x,λ). Assuming that the first process is independent of x, it is inappropriate to treat the coupled system as a simple bifurcation of (x,u) with λ. Instead, it is necessary to develop a version of the theory which preserves the special "intermediate" role of u —much as the Golubitsky–Schaeffer theory preserves the special role of λ. Such a theory is developed here. As well as finding determinacy and unfolding criteria, we classify all sequential bifurcation problems of codimension 4 or less when x and u are one-dimensional. We exhibit the possible bifurcation diagrams for codimension 2 or less (those of higher codimension are omitted for reasons of space) and give analytic conditions for the occurrence of a given bifurcation in the classification. We discuss applications of these ideas to suitable systems of chemical reactions.

The presence of "hidden" or intermediate variables u has a strong effect on the expected bifurcation phenomenology, and hence on the inferences that may be made from theoretical results: (a) Multiple limit points can occur in a persistent (structurally stable) fashion. (b) Eliminating a hidden variable can change the codimension of a bifurcation problem (because some perturbations of the resulting problem, that contribute to the codimension, may be incompatible with the elimination step). (c) Bifurcation diagrams that are ordinarily considered inequivalent may become equivalent if a hidden variable is present.

The effect of (a) is to introduce new types of persistent diagram. The likelihood of seeing a given diagram in a parametrized family of bifurcation problems is affected by (b). And (c) implies that different ways of eliminating hidden variables from equations may produce apparently different observable consequences.

1. Introduction. In this paper we apply singularity/catastrophe theory to the classification and unfolding of bifurcation problems of the form

$$a(u,\lambda)=0,\qquad b(x,u,\lambda)=0 \tag{1}$$

where $x,u,\lambda\in\mathbb{R}$ and λ is a bifurcation parameter. Equations (1) arise in processes of the type drawn schematically in Fig. 1. For this reason we call (1) a *sequential bifurcation*. The applications we have in mind are particularly to chemical reactions, but include also nonlinear electric circuits and mechanical analogues. The common feature of all physical systems which can be schematized as in Fig. 1 is that process 1 is only *weakly coupled* to process 2, for which some *irreversible mechanism* in the system may be responsible. The bifurcation equations (1) will then, in general, result from a Lyapunov–Schmidt reduction of some nonlinear equations governing the steady-state behaviour of the system.

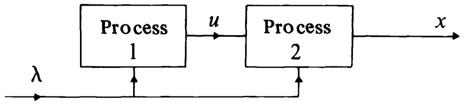


FIG. 1. *Block diagram of a sequential bifurcation with bifurcation parameter* λ, *internal variable* u, *and output* x.

*Received by the editors August 16, 1982, and in revised form January 8, 1983. This research was supported in part by the A. v. Humboldt Foundation and in part by the Volkswagenwerk Foundation.

† Institute for Information Sciences, University of Tübingen, Tübingen, West Germany.

‡ Mathematics Institute, University of Warwick, Coventry, England CV4 7AL.

The crucial feature here is that process 1 is considered to be *independent* of x. In consequence, when unfolding (1) to find all small perturbations, we consider only perturbations of a, that are independent of x. The resulting theory thus *differs* from that which would be obtained if (1) were considered to be a bifurcation problem in (x,u) of the form

$$a(x,u,\lambda)=0, \qquad b(x,u,\lambda)=0 \tag{2}$$

where, by chance, a is independent of x. We expand on this point in §2.

Our approach is closely analogous to the work of Golubitsky and Schaeffer [6] on imperfect bifurcation, which itself is a modification of singularity/catastrophe theory of Thom [14], Zeeman [16], Arnold [1], [2], Mather [9], [10] and others. In particular, we discuss the following problems:

(a) *Recognition*. Under what conditions will (1) be equivalent to a given type of problem?

(b) *Classification*. What types can occur (in low codimension)?

(c) *Unfolding*. What perturbations can occur for a given sequential bifurcation?

In §2 we discuss qualitatively the effect of the first equation of (1) on the bifurcation diagrams in the (x,λ)-plane, thereby considering u as a "hidden variable". In §3 the basic mathematical machinery is set up for sequential bifurcations. We confine ourselves to $(x,u)\in\mathbb{R}\times\mathbb{R}$; generalizations to higher dimensions should be straightforward. The structure of (1) requires a new type of contact equivalence, whose geometrical implications will be discussed in §4. In §5 we present normal forms of sequential bifurcation problems up to some reasonable degree of degeneracy. Perturbed bifurcation diagrams associated with several normal forms of §5 are sketched in §6. Finally, in §7, we describe a chemical reaction whose steady states are governed by sequential bifurcations: two stirred tank reactors coupled in series.

2. The hidden variable. To clarify the special role of u, consider the equations

$$u^2-\lambda=0, \tag{3}$$

$$x^2-u=0. \tag{4}$$

Consider this first as a bifurcation problem in (x,u) over λ. Then we can solve (4) to get $u=x^2$ and substitute in (3) to get the *quartic fold* (Stewart [12]):

$$x^4-\lambda=0. \tag{5}$$

If we analyse (5) by the methods of Golubitsky and Schaeffer [6], we find it has codimension two, with universal unfolding

$$(x^4-\lambda)+\alpha x^2+\beta x=0, \tag{6}$$

giving the structurally stable bifurcation diagrams of Fig. 2. However, as a sequential bifurcation, (3) and (4) have codimension one and a universal unfolding

$$\begin{bmatrix} u^2-\lambda \\ x^2-u+\alpha \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{7}$$

The bifurcation diagrams for x against λ are shown in Fig. 3. Note the *double limit points* in the second diagram. In the Golubitsky–Schaeffer theory these are a codimension-1 phenomenon, hence structurally unstable. But in the setting of sequential bifurcations they are codimension-0, and *structurally stable*.

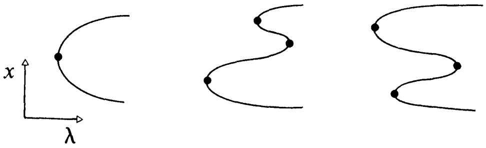


FIG. 2. *Structurally stable bifurcation diagrams in the universal unfolding of* (5), *considered as an ordinary bifurcation problem.*

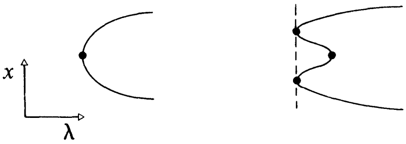


FIG. 3. *Structurally stable bifurcation diagrams in the universal unfolding of* (3), (4), *considered as a sequential bifurcation. Compare with Fig.* 2.

To see why this should be so, note that (3) is a codimension-0 bifurcation of u against λ, and (4) is codimension-0 considered as a bifurcation of x against u. Hence, processes 1 and 2 of Fig. 1 are both structurally stable here. However, there is still the possibility of perturbing the way they are *coupled*, and an origin shift in u in (4) is plausibly the only perturbation of relevance.

Specifically, the solutions of (3) and (4) are ruled paraboloids (Fig. 4a). Perturbations of (3), (4) which respect the fact that process 1 is independent of x will move each of these surfaces independently (but preserve their identity as separate surfaces). If the vertex of the (x,u)-paraboloid B is at a positive value of u, the intersection will be as in Fig. 4b, giving Fig. 3a as projection. If negative, then the curve of intersection "wraps around" the paraboloid A, giving a double limit point (cf. Fig. 3b) along the vertex of A $(\lambda=0)$ as in Fig. 4c. Analytically the appearance of this structurally stable double-limit point is seen at once, if we replace the unfolding $(0,\alpha)$ in (6) by the equivalent unfolding $(\alpha u,0)$ (which also just shifts the origin in u), solve the lower equation of (7) for u and substitute into the upper equation. The result obtained in this way is just (6) with $\beta=0$. In fact, as we shall see below, *the newest phenomenon encountered in sequential bifurcations is the structurally stable occurrence of double* (*and hence multiple*) *limit points.*

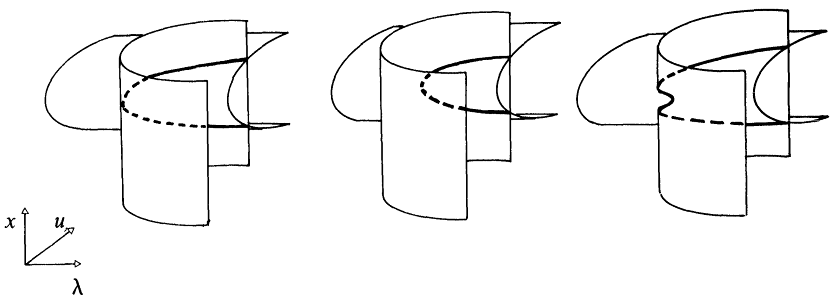


FIG. 4. *Interpretation of* (3), (4) *and their perturbations, as the intersection of two ruled paraboloids.*

We emphasize here the dependence of x on λ, that is, of the "output" of the sequence of processes on its "input". This is because we view the variable u as a *hidden* or *intermediate parameter*, which in general may not be accessible to observation (or, at least, not measured in a given experiment). This means that the three variables x,u,λ are on different "levels" and our equivalence relation is chosen to reflect this. Because information about u is lost when only (x,λ)-projection of the bifurcation curve is drawn, it is necessary to discuss carefully the *interpretation* of features of bifurcation diagrams. We explain this further in §4.

3. Mathematical setting. In this section we develop the main ideas involved in the mathematical treatment. Since this is closely analogous to that of Golubitsky and Schaeffer [6] and subsequent variations [7], [8], we omit routine details.

A *sequential bifurcation problem* (SBP) (in one variable) is an equation

$$g(x,u,\lambda)=\begin{bmatrix} a(u,\lambda) \\ b(x,u,\lambda) \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{8}$$

where $g\in S\equiv \mathcal{E}_{u\lambda}\times\mathcal{E}_{xu\lambda}$, the $\mathcal{E}$'s denoting rings of germs in the usual way (see Golubitsky and Schaeffer [6], Gibson [5], etc.). We further assume that

$$g(0,0,0)=0, \qquad a_u(0,0)=b_x(0,0,0)=0$$

where subscripts u,x denote partial derivatives. This is done to avoid trivial cases where one equation has single-valued solutions and the system reduces to a bifurcation problem (in the usual sense of Golubitsky and Schaeffer) in one variable. Note that we do *not* require $b_u(0,0,0)=0$.

The particular version of contact equivalence that we use is the following. Two germs $g,h\in S$ are (*sequentially contact*) *equivalent* if and only if

$$g(x,u,\lambda)=M(x,u,\lambda)h(X(x,u,\lambda),U(u,\lambda),\Lambda(\lambda)), \tag{9}$$

where

$$M(x,u,\lambda)=\begin{bmatrix} M_{11}(u,\lambda) & 0 \\ M_{12}(x,u,\lambda) & M_{22}(x,u,\lambda) \end{bmatrix}$$

is invertible at 0 (that is $(M_{11}M_{22})|_0\neq 0$) and

$$X_x|_0,\ U_u|_0,\ \Lambda_\lambda|_0>0$$

to preserve orientations. We write $g\sim h$ if g and h are equivalent. A discussion of the intuitive meaning of this equivalence is deferred until §4.

If $g\in S$ then we define the space $\tilde{T}g$ and the *formal tangent space* Tg as follows. Set

$$\tilde{T}_1 g=\mathcal{E}_{xu\lambda}\left\langle \begin{bmatrix} 0 \\ a \end{bmatrix}, \begin{bmatrix} 0 \\ b \end{bmatrix}, \begin{bmatrix} 0 \\ b_x \end{bmatrix} \right\rangle, \tag{10a}$$

$$\tilde{T}_2 g=\mathcal{E}_{u\lambda}\left\langle \begin{bmatrix} a \\ 0 \end{bmatrix}, \begin{bmatrix} a_u \\ b_u \end{bmatrix} \right\rangle. \tag{10b}$$

Then

$$\tilde{T}g=\tilde{T}_1 g+\tilde{T}_2 g, \tag{11a}$$

$$Tg=\tilde{T}g+\mathcal{E}_\lambda\langle g_\lambda\rangle. \tag{11b}$$

The formal tangent space has the standard interpretation in terms of orbits for the equivalence relation $\sim$. We say that g has *finite codimension* if $\dim S/\tilde{T}g$ is finite. If g has finite codimension in this sense, then the *codimension* of g is

$$\operatorname{cod} g=\dim S/Tg. \tag{12}$$

Note that g has finite codimension if and only if

$$\dim \mathcal{E}_{xu\lambda}/\mathcal{E}_{xu\lambda}\langle a,b,b_x\rangle<\infty \tag{13a}$$

and

$$\dim \mathcal{E}_{u\lambda}/\mathcal{E}_{u\lambda}\langle a,a_u\rangle<\infty . \tag{13b}$$

(The proof is not entirely obvious, but uses properties of $\mathcal{E}_{u\lambda}$ and the generators.) The use of $\tilde{T}g$ to define finiteness of the codimension, but Tg to define the codimension itself, lacks elegance, but is required for a proof of the unfolding theorem below. Exactly the same problem arises in Golubitsky and Schaeffer [6]. In their case, however, Damon (then unpublished but mentioned in Golubitsky and Schaeffer [8] and Stewart [12]) has shown that Tg has finite codimension if and only if $\tilde{T}g$ has. (We use here their notation.) Prof. Damon has informed us that his results, to be published in [3], also show that the same phenomenon occurs for sequential bifurcation problems; that is, $\dim S/Tg$ is finite if and only if $\dim S/\tilde{T}g$ is finite. Essentially this is because the equivalence relation used for sequential bifurcations corresponds to what Damon [3] calls a *geometric subgroup* of the group K of contact equivalences. The action of this group satisfies certain algebraic properties on tangent spaces given in [3].

In fact, the general unfolding theorem of [3] applies in our present context. However, none of this affects the main results of this paper and for that reason we shall not provide further discussion.

In order to classify sequential bifurcations and to give conditions for a particular type to occur, we need suitable determinacy statements which are achieved as follows.

Define the *restricted tangent space* RTg to be

$$RTg=\mathcal{E}_{xu\lambda}\left\langle\begin{bmatrix}0\\a\end{bmatrix},\begin{bmatrix}0\\b\end{bmatrix}\right\rangle+\mathfrak{m}_{xu\lambda}\left\langle\begin{bmatrix}0\\b_x\end{bmatrix}\right\rangle+\mathcal{E}_{u\lambda}\left\langle\begin{bmatrix}a\\0\end{bmatrix}\right\rangle+\mathfrak{m}_{u\lambda}\left\langle\begin{bmatrix}a_u\\b_u\end{bmatrix}\right\rangle, \tag{14}$$

with the $\mathfrak{m}$'s in (14) denoting maximal ideals in the usual way. RTg is the tangent space of g with respect to an obvious restriction of the class of transformations defining the equivalence relation $\sim$. Extending the techniques used by Golubitsky and Schaeffer [8] to include sequential bifurcations, it is easily seen that, for given $g,h\in S$, $g+h$ is equivalent to g if either

$$RT(g+th)=RT(g)$$

or

$$RT(g+th)+\mathfrak{m}_\lambda\langle g_\lambda+th_\lambda\rangle=RTg+\mathfrak{m}_\lambda\langle g_\lambda\rangle$$

for all $t\in\mathbb{R}$. We note that in the case of germs f under right equivalence, the restricted tangent space corresponds to $\mathfrak{m}J(f)$ where $J(f)$ is the Jacobian ideal of f.

Following the lines of Golubitsky and Schaeffer [8], we define the space $P(g)$ as

$$P(g)=\{p\in S|RT(h+p)=RTh \text{ for all } h\sim g\}. \tag{15}$$

$P(g)$ contains all monomials which can be ignored in the Taylor expansion of h if one is determining whether h is equivalent to g. In order to compute the space $P(g)$ explicitly for a given $g\in S$ we have to extend the notation of intrinsic ideals due to B. L. Keyfitz (see [8]) to include sequential bifurcations. Since the techniques involved are nearly the same as in bifurcations in the Golubitsky–Schaeffer (G–S) sense, we only state the basic facts. Recall from [8] that an ideal J in $\mathcal{E}_{u\lambda}$ is called intrinsic if whenever

a is in J and a' is contact equivalent to a in the G–S sense, then a' is in J. Moreover, any finitely generated intrinsic ideal has the form

$$J=\mathfrak{m}_{u\lambda}^{n_1}\langle\lambda^{m_1}\rangle+\cdots+\mathfrak{m}_{u\lambda}^{n_s}\langle\lambda^{m_s}\rangle \tag{16}$$

where $0\leq m_1<\cdots<m_s$ and $n_1+m_1>\cdots>n_s+m_s$ ($m_1=0$ if J has finite codimension). Let now I and J be ideals in $\mathcal{E}_{xu\lambda}$ and $\mathcal{E}_{u\lambda}$, respectively, and let V be defined by

$$V=J\left\langle\begin{bmatrix}1\\0\end{bmatrix}\right\rangle+I\left\langle\begin{bmatrix}0\\1\end{bmatrix}\right\rangle. \tag{17}$$

Extending the concept of intrinsic ideals to sequential bifurcations, we say that a subspace V of S as given by (17) is intrinsic if any $h\in S$ which is equivalent to a $g\in V$ is in V. It is not hard to see that V is intrinsic and of finite codimension if and only if J is intrinsic and of finite codimension in $\mathcal{E}_{u\lambda}$, $J\subset I$, and I is given by

$$I=\mathfrak{m}_{xu\lambda}^{k}+\mathfrak{m}_{xu\lambda}^{k_1}J_1+\cdots+\mathfrak{m}_{xu\lambda}^{k_r}J_r$$

where each $J_i\subset\mathcal{E}_{u\lambda}$ has the form of (16) and

$$0<n_{s_1,1}+m_{s_1,1}<\cdots<n_{s_r,r}+m_{s_r,r},\qquad k>k_1+n_{1,1}+m_{1,1}>\cdots>k_r+n_{1,r}+m_{1,r}.$$

If a subspace W of S is the sum of an $\mathcal{E}_{xu\lambda}$- and an $\mathcal{E}_{u\lambda}$-module, then there is a largest intrinsic space contained in W which will be called the *intrinsic part of* W and will be denoted by Itr W. Let now $g\in S$ have finite codimension. The main statement about $P(g)$ is that $P(g)$ is an intrinsic space of finite codimension and

$$\operatorname{Itr}(\mathfrak{m}RTg)\subset P(g)\subset\operatorname{Itr}(RTg)$$

where

$$\mathfrak{m}RTg=\mathcal{E}_{u\lambda}\left\langle\begin{bmatrix}a\\0\end{bmatrix},\begin{bmatrix}ua_u\\ub_u\end{bmatrix},\begin{bmatrix}\lambda a_u\\\lambda b_u\end{bmatrix}\right\rangle+(\mathcal{E}_{xu\lambda}\langle a,b\rangle+\mathfrak{m}_{xu\lambda}\langle b_x\rangle)\left\langle\begin{bmatrix}0\\1\end{bmatrix}\right\rangle.$$

The computation of $P(g)$ is facilitated by the following observation which is an extension of [8, Lemma 1.9]: Let V_1 and V_2 be intrinsic subspaces of S with corresponding ideals I_1, J_1 and I_2, J_2, respectively, and let $V_2\subset P(g)$. Assume that $J_1=J_2+\mathcal{E}_{u\lambda}\langle p\rangle$ and $I_1=I_2+\mathcal{E}_{xu\lambda}\langle q\rangle$ and that

$$RT(g+t(p,0))=RT(g+\tau(0,q))=RTg \tag{18}$$

holds for all $t,\tau\in\mathbb{R}$. Then $V_1\subset P(g)$. In practice, one chooses $V_2=\operatorname{Itr}(\mathfrak{m}RTg)$ and looks for all germs $(p,q)\in S$ such that $V_2+\mathcal{E}_{u\lambda}\langle(p,0)\rangle+\mathcal{E}_{xu\lambda}\langle(0,q)\rangle$ is still intrinsic and (18) holds to obtain $P(g)$. As an example, consider

$$g=\begin{bmatrix}u^2+\lambda\\x^2+u^3\end{bmatrix},$$

which has codimension three. Proceeding as explained before it is easily seen that here

$$P(g)=\operatorname{Itr}(\mathfrak{m}RTg)$$

with the corresponding ideals given by

$$J=\mathfrak{m}_{u\lambda}^4+\mathfrak{m}_{u\lambda}^2\langle\lambda\rangle+\mathfrak{m}_{u\lambda}\langle\lambda^2\rangle,\qquad I=\mathfrak{m}_{xu\lambda}^4+\mathfrak{m}_{xu\lambda}^2\langle\lambda\rangle+\mathfrak{m}_{xu\lambda}\langle\lambda^2\rangle.$$

In the Appendix we tabulate I and J for each normal form of §5 which has codimension less than four.

Finally we deal with unfoldings. For $\alpha \in \mathbb{R}^k$ let

$$S_\alpha = \mathcal{E}_{u\lambda\alpha} \times \mathcal{E}_{xu\lambda\alpha}.$$

We say that $G \in S_\alpha$ is an unfolding of $g \in S$ if $G|_{\alpha=0} = g$. We write

$$G_\alpha(x,u,\lambda) = G(x,u,\lambda,\alpha) = \begin{bmatrix} A(u,\lambda,\alpha) \\ B(x,u,\lambda,\alpha) \end{bmatrix}.$$

Suppose that G_α and H_β are unfoldings of g, with $\alpha \in \mathbb{R}^k$, $\beta \in \mathbb{R}^l$. Then

(1) H_β *factors through* G_α (or G_α induces H_β) if there is a smooth map Φ: $\mathbb{R}^l \to \mathbb{R}^k$ such that $G_{\Phi(\beta)} \sim H_\beta$ for all (sufficiently small) β.

(2) G_α is a *versal unfolding* of g if any unfolding of g factors through G_α.

(3) G_α is a *universal unfolding* of g if it is versal and has minimal possible unfolding dimension k.

Next we state the

UNFOLDING THEOREM. *Let $G \in S_\alpha$ be a k-parameter unfolding of $g \in S$ and let g have finite codimension. Then G is versal if and only if*

$$Tg + \mathbb{R}\left\{ \frac{\partial G}{\partial \alpha_i}\bigg|_{\alpha=0} \right\} = S.$$

The proof of the unfolding theorem follows traditional lines and we only sketch the salient points. The crucial step here is an analogue of [6, Lemma 3.3].

LEMMA. *Let $G \in S_\alpha$ be an unfolding of $g \in S$ and assume that g has finite codimension. Let $H_1, \cdots, H_s$ be in S_α and $h_i = H_i|_{\alpha=0}$. Then*

(a) $$S = Tg + \mathbb{R}\{h_i\}$$

if and only if

(b) $$S_\alpha = \mathcal{E}_{xu\lambda\alpha} \left\langle \begin{bmatrix} 0 \\ A \end{bmatrix}, \begin{bmatrix} 0 \\ B \end{bmatrix}, \begin{bmatrix} 0 \\ B_x \end{bmatrix} \right\rangle + \mathcal{E}_{u\lambda\alpha} \left\langle \begin{bmatrix} A \\ 0 \end{bmatrix}, \begin{bmatrix} A_u \\ B_u \end{bmatrix} \right\rangle + \mathcal{E}_{\lambda\alpha}\langle G_\lambda \rangle + \mathcal{E}_\alpha \langle H_i \rangle.$$

Proof. (b)⇒(a) is clear, so assume that (a) holds. Since g has finite codimension, we can write

$$Tg = \tilde{T}_1 g + \mathcal{E}_{u\lambda} \left\langle \begin{bmatrix} a \\ 0 \end{bmatrix} \right\rangle + \mathfrak{m}^m_{u\lambda} \left\langle \begin{bmatrix} a_u \\ 0 \end{bmatrix} \right\rangle + \mathbb{R}\{r_j g_u, \lambda^l g_\lambda\},$$

for some m and a finite set of powers λ^l and germs $r_j \in \mathcal{E}_{u\lambda}$. Since (a) holds we can select a finite set of germs $p_\nu \in \mathcal{E}_{xu\lambda}$ and $q_\mu \in \mathcal{E}_{u\lambda}$ generating $\mathcal{E}_{xu\lambda}/\mathcal{E}_{xu\lambda}\langle a, b, b_x \rangle$ and $\mathcal{E}_{u\lambda}/\mathcal{E}_{u\lambda}$ $(\langle a \rangle + \mathfrak{m}^m_{u\lambda}\langle a_u \rangle)$ as vector spaces, respectively, such that $(0, p_\nu)$ and $(q_\mu, 0)$ are contained in $\mathbb{R}\{h_i, r_j g_u, \lambda^l g_\lambda\}$. Now apply the Malgrange preparation theorem (see, e.g. [15]) separately to $\mathcal{E}_{xu\lambda\alpha}/\mathcal{E}_{xu\lambda\alpha}\langle A, B, B_x \rangle$ and $\mathcal{E}_{u\lambda\alpha}/\mathcal{E}_{u\lambda\alpha}(\langle A \rangle + \mathfrak{m}^m_{u\lambda}\langle A_u \rangle)$ to obtain (b).

The proof of the unfolding theorem on the basis of the lemma follows the same lines as [6, proof of Thm. 2.4]. If $\{p_1, \cdots, p_k\}$ is a basis of a vector subspace of S complementary to Tg, then

$$G(x,u,\lambda,\alpha) = g(x,u,\lambda) + \sum_{i=1}^{k} \alpha_i p_i(x,u,\lambda)$$

is a universal unfolding of g. We call the p_i's *unfoldings* and the α_i's *unfolding parameters*.

4. Interpreting the bifurcation diagrams. Our version of contact equivalence has the following geometrical meaning (we shall ignore for this discussion the matrix M in (9) since the zeros of g are the same as the zeros of Mg). The space with coordinates (x,u,λ) may be deformed by a diffeomorphism which leaves invariant the set of planes $\lambda=$ constant, and further on each such plane leaves invariant the set of lines $u=$ constant (Fig. 5). However, the vertical lines above points (u,λ) can be deformed differently for different (u,λ).

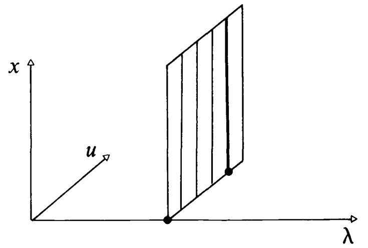


FIG. 5. *Geometric meaning of sequential contact equivalence.*

Consider the effect of such deformations on the space curve

$$a(u,\lambda)=0, \qquad b(x,u,\lambda)=0,$$

as far as its projection on the (x,λ)-plane is concerned. For simplicity take

$$a(u,\lambda)=u^2-\lambda, \qquad b(x,u,\lambda)=x.$$

Then the bifurcation curve in (x,u,λ)-space is a parabola (Fig. 6a) whose projection to the (x,λ)-plane is degenerate (Fig. 7a). If we consider the *equivalent* problems

$$u^2-\lambda=0, \qquad x\pm u=0,$$

then this parabola becomes *tilted* and the (x,λ)-diagrams are also parabolic; however, the projections of the two branches of the space curve are arranged differently depending on the sign; see Figs. 6b and 7b, 6c and 7c. The degeneracy in Fig. 7a is simply an artifact of the projection.

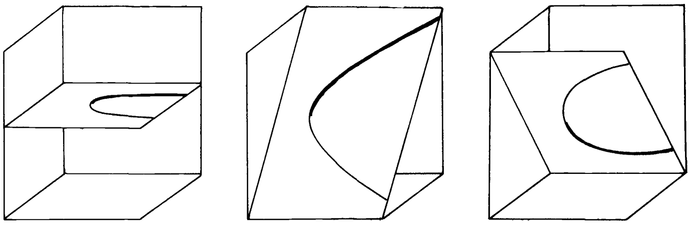


FIG. 6. *Effect of an equivalence for a simple example, drawn in (x,u,λ) space.*

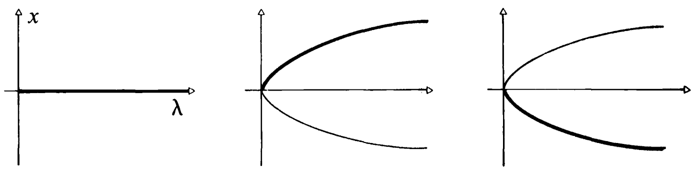


FIG. 7. *Projections of Fig. 6 into (x,λ) space. Note that equivalent problems can have different projections.*

In our interpretation of this theory, however, it is assumed that the intermediate variable u is not directly known. In consequence, certain apparent features of the bifurcation diagrams, plotted in the (x,λ)-plane, are not preserved by equivalence. (This is not surprising: it expresses the "lost information" when u is neglected.) The features that *are* preserved are limit points, including double limit points, and the way these connect together, but (for example) the curves that connect them can be twisted, made to cross or uncross, and so on, without changing the equivalence class. For example, the various diagrams of Fig. 8 are all equivalent.

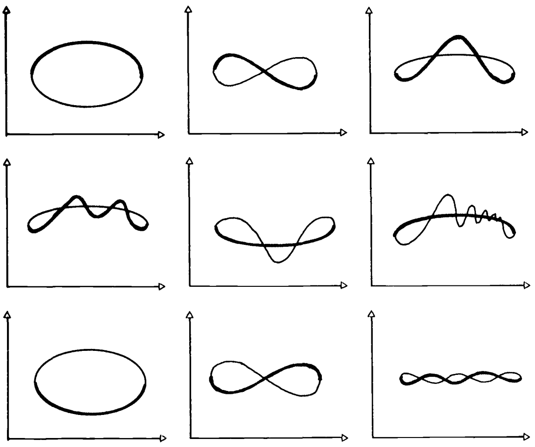


FIG. 8. *Samples of equivalent diagrams in (x,λ)-space.*

Transverse crossings of different branches (occurring for different u-values) are of course structurally stable to *small* deformations (equivalences close to the identity). However, not *all* rearrangements of branches are necessarily permitted. Consider for example the problem

$$\lambda^2+u^2-1=0, \qquad x-u^2+2=0,$$

whose solution is sketched in Fig. 9. Here there are two closed loops in the (x,λ)-bifurcation diagram, but each has a "front branch" (solid lines) and a "back branch" (dotted). Equivalent diagrams can make back branches intersect front branches at will, but *cannot* make two front branches cross, or two back branches. This is because equivalences preserve the vertical lines (u,λ) = constant that generate the cylinder $u^2+\lambda^2=1$. See Fig. 10.

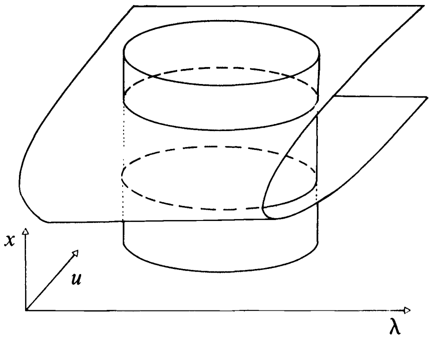


FIG. 9. *"Front" and "back" branches of bifurcation diagrams in (x,u,λ)-space.*

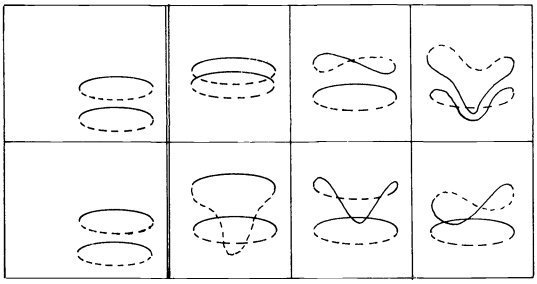


FIG. 10. *Possible and impossible projections into (x,λ)-space of diagrams equivalent to Fig.* 9.

Similar considerations are necessary in the interpretation of each type of diagram: in each case reference to the underlying three-dimensional geometry resolves the possibilities. Essentially the branches of the (x,λ)-diagram are partitioned into sets corresponding to "sheets" of the (u,λ)-surface, and the relative arrangement of branches *within each sheet* must be preserved.

5. Classification. In this section we present a classification of sequential bifurcation problems up to some reasonable degree of degeneracy.

In Table 1 we have listed six hierarchies of normal forms of SBP's, together with their unfoldings. The normal forms $(1)_{mn}$ through $(3)_{mn}$ are just sequences of standard one-dimensional bifurcation problems in the sense of Golubitsky and Schaeffer (G–S sense) [5]. Bifurcation diagrams associated with a universal unfolding of the normal form $(1)_{22}$ (which is the only codimension-1 SBP) have been already sketched in §2. Observe that in the case of normal forms $(6)_k$ the linear part $u+\varepsilon_1\lambda$ in b factors through a. This factoring forces the presence of the term $\varepsilon_2\lambda uk$ in order to keep the codimension finite.

TABLE 1

Hierarchies of sequential bifurcation problems. $\varepsilon_{1,2,3}=\pm 1$.

	a	b	Unfoldings	cod g
$(1)_{mn}$	$u^m+\varepsilon_1\lambda$	$x^n+\varepsilon_2 u$	$(u^j,0)$, $1\le j\le m-2$ $(0,x^i), 0\le i\le n-2$	$m+n-3$
$(2)_{mn}, n\ge 3$		$x^n+\varepsilon_2 ux$	$(0,x^i), 0\le i\le n-1$	$m+n-2$
$(3)_{mn}, n\ge 2$		$x^2+\varepsilon_2 u^n$	$(0,u^i), 0\le i\le n-1$	$m+n-2$
$(4)_m, m\ge 3$	$u^m+\varepsilon_1 u\lambda$	$x^2+\varepsilon_2 u+\varepsilon_3\lambda$	$(u^j,0), 0\le j\le m-1$	m
$(5)_k, k\ge 3$	$u^2+\varepsilon_1\lambda^k$	$x^2+\varepsilon_2 u+\varepsilon_3\lambda$	$(\lambda^j,0), 0\le j\le k-1$	k
$(6)_k, k\ge 2$	$u^2-\lambda^2$	$x^2+\varepsilon_1 u+\varepsilon_2\lambda+\varepsilon_3\lambda^k$	$(1,0), (0,\lambda^j), 0\le j\le k-1$	$k+1$

In Table 2 we have listed further normal forms with codimensions between three and five. For several of these normal forms we have added an asterisk to the codimension to denote that g has *modality* one, that is, one unfolding parameter is associated with a modal parameter of g. The *topological codimension* of these normal forms is cod $g-1$. (The meaning of "modality", "modal parameters", "topological codimension", etc., is the same as in imperfect bifurcation theory in the G–S sense, see [5], [6], [7].) The modal parameter for the normal forms (9b, c), (10c) and (12) is c, but modal

TABLE 2

Normal forms of sequential bifurcation problems with topological codimension less than 5. $\varepsilon_{1,2,3}=\pm 1$.

	a	b	Unfoldings	cod g
(7)	$u^2+\varepsilon_1\lambda$	$x^3+\varepsilon_2 u^2$	$(0,1), (0,x), (0,ux), (0,u)$	4
(8)	$u^3+\varepsilon_1\lambda^2$	$x^2+\varepsilon_2 u+\varepsilon_3\lambda$	$(0,1), (1,0), (u,0), (u\lambda,0)$	4
(9a)	$u^3+\varepsilon_1 u\lambda$	$x^2+\varepsilon_2 u+\varepsilon_3\lambda^2$	$\left\{\begin{matrix}(u^j,0),\\ 0\le j\le 2\end{matrix}\right\}$ $(\lambda,0)$	4
(9b)		$x^2+cu^2+\varepsilon_2\lambda,\ c\neq\varepsilon_1\varepsilon_2$	$(0,u); (0,u^2)$	5*
(9c)		$x^3+cux+\varepsilon_2 u+\varepsilon_3\lambda$	$(0,x); (0,ux)$	5*
(10a)	$u^2+\varepsilon_1\lambda^3$	$x^2+\varepsilon_2\lambda$	$\left\{\begin{matrix}(\lambda^j,0),\\ 0\le j\le 2\end{matrix}\right\}$ $(0,u)$	4
(10b)		$x^2+\varepsilon_2 u+\varepsilon_3\lambda^2$	$(0,1)$	4
(10c)		$x^3+cux+\varepsilon_2 u+\varepsilon_3\lambda$	$(0,x), (0,xu)$	5*
(11a)	$u^2+\varepsilon_1\lambda^2$	$x^2+L(u,\lambda)$	$\left\{\begin{matrix}(1,0),(0,1)\\ (0,\bar{L}(u,\lambda))\end{matrix}\right\}$	3*
(11b)		$x^3+L(u,\lambda)+x\bar{L}(u,\lambda)$	$(0,x)$	4*
(11c)		$x^3+L(u,\lambda)$	$(0,x), (0,x\bar{L}(u,\lambda))$	5*
(11d)		$x^2+uL(u,\lambda)$	$(1,0), (0,1), (0,u), (0,\lambda); (0,u\bar{L}(u,\lambda))$	5*
(12)	$u^2-\lambda^2$	$x^3+\varepsilon_1 u+\varepsilon_2\lambda+\varepsilon_3\lambda^2+cux,\ c\neq 0$	$(1,0), (0,1), (0,x), (0,\lambda); (0,ux)$	5*

boundaries are only present for (9b) and (12). In the normal forms (11a–d), L is a nonvanishing linear form in (u,λ) which must not factor through $u^2+\varepsilon_1\lambda^2$. L can be normalized to

$$(19)\qquad L(u,\lambda)=u\cos\phi-\lambda\sin\phi,\ \cos^2\phi+\varepsilon_1\sin^2\phi\neq 0.$$

The modal parameter is the angle ϕ (angle between the straight line $L=0$ and the λ-axis). $\bar{L}(u,\lambda)$ is a linear form linearly independent to L which can be chosen to be

$$(20)\qquad \bar{L}(u,\lambda)=u\sin\phi+\lambda\cos\phi.$$

Now we state the main result of this section:

CLASSIFICATION THEOREM. *Any sequential bifurcation problem with topological codimension less than five and modality less than two is equivalent to one of the germs of Tables* 1 *and* 2.

The proof of the classification theorem is very technical and will be omitted. The methods are similar to those used by Dangelmayr and Armbruster [4] in classifying Z_2-equivariant bifurcation problems with corank two in the context of imperfect bifurcations in the presence of symmetry [7]. The essential idea is to transform an arbitrary SBP to the form $g=(a,b)$ with $a\in\mathfrak{E}_{u\lambda}$ being a standard normal form of a one-dimensional bifurcation problem in the G–S sense and then looking for the possible b's which keep the topological codimension less than five and modality less than two.

Conditions which must be satisfied by an arbitrary SBP for being equivalent to one of the normal forms of Tables 1 and 2 with topological codimension less than four are summarized in the Appendix.

Note that we have classified SBP's, that is, $a_u(0,0)=b_x(0,0,0)=0$. If $a_u \neq 0$ we can solve $a=0$ for u and insert into b to obtain a one-dimensional bifurcation problem in (x,λ) in the standard G–S sense. This means that no information is lost due to the hidden variable u, and double-limit points in the (x,λ)-diagram are no longer structurally stable. On the other hand, if $b_x \neq 0$, g is equivalent to $(a(u,\lambda),x)$ and we also end up essentially with a one-dimensional bifurcation problem, but now in the hidden variable u. Observing only the (x,λ)-diagrams means now that a lot of information is lost because all essential bifurcation phenomena occur in the (u,λ)-plane.

6. The bifurcation diagrams. We now describe the bifurcation diagrams, plotted in (x,λ)-space as explained in §4, for sequential bifurcations of codimension≤two. The unfolding parameters are denoted by α and β where α corresponds to the unfolding $(0,1)$ for types $(1)_{22}$, $(1)_{23}$, $(3)_{22}$, to $(u,0)$ for type $(1)_{32}$ and to $(1,0)$ for type (11a). β is associated with $(0,x)$, $(0,u)$, $(u^2,0)$, $(0,1)$ for the types $(1)_{23}$, $(3)_{22}$, $(1)_{32}$, (11a), respectively.

Figures 12–18 show how the unfolding space is divided up into regions corresponding to structurally stable bifurcation diagrams. This is in exact analogy with Golubitsky and Schaeffer's decomposition by means of the bifurcation, hysteresis and double limit varieties; however, it should be noted that in the case of sequential

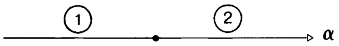


FIG. 12. *Decomposition of unfolding space for type* $(1)_{23}$.

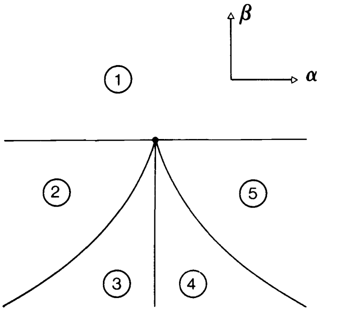


FIG. 13. *Decomposition of unfolding space for type* $(3)_{22}$, $\varepsilon_2=+1$.

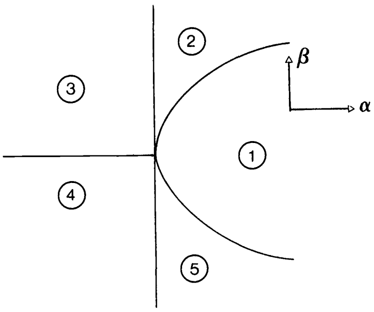


FIG. 14. *Decomposition of unfolding space for type* $(3)_{22}$, $\varepsilon_2=-1$.

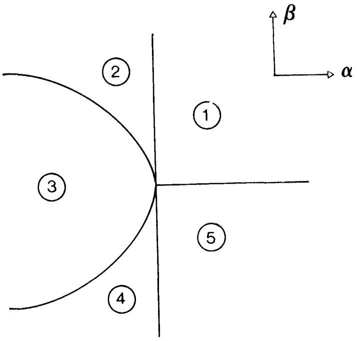


FIG. 15. *Decomposition of unfolding space for type* $(1)_{32}$.

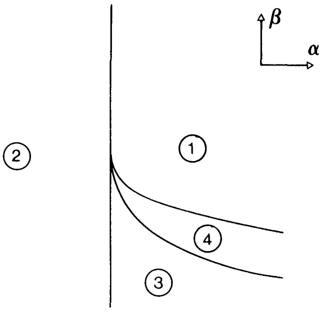


FIG. 16. *Decomposition of unfolding space for type* $(11a)$, $\varepsilon_1 = +1$.

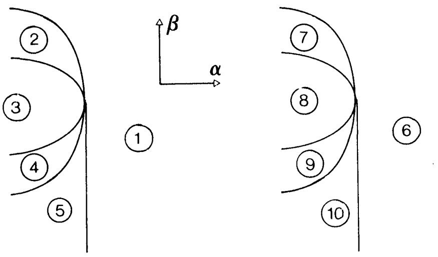


FIG. 17. *Decomposition of unfolding space for type* $(11a)$, $\varepsilon_1 = -1$.

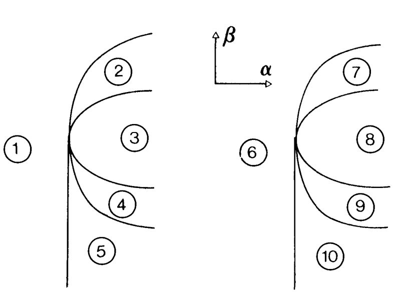


FIG. 18. *Structurally stable perturbations of bifurcation diagrams for type* $(1)_{22}$, *corresponding to regions marked in Fig.* 11. *For the assignment of stabilities, see text.*

bifurcations, double limit points fall into two types. Those occurring due to a fold in the (u, λ) bifurcation diagram are of codimension 0; but those involving limit points on two distinct "sheets" of the (u,λ) diagram are of codimension 1 and hence lead to changes in the topology of the (x, λ) bifurcation diagram.

For each of the regions marked in Figs. 12–18, the corresponding bifurcation diagram is shown in Figs. 19–25. The method used to find these diagrams can be illustrated on the problem $(1)_{23}$:

$$u^2 - \lambda = 0, \tag{21}$$

$$x^3 - u + \beta x + \alpha = 0. \tag{22}$$

(Whenever, as here, sign choices merely affect orientations, we have made them arbitrarily.) Note that (22) does not involve λ. (This feature is relatively common in low dimensions, at least if one makes careful choices of unfoldings, but becomes more unusual as the codimension increases—for example, it does not happen for type $(5)_3$. When it is absent, the geometry is a little harder to disentangle.) Equation (21) is a ruled parabolic surface; (22) is a ruled cubic surface; and their intersection can be

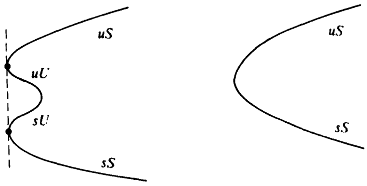


FIG. 19. *Diagrams corresponding to regions marked in Fig.* 12.

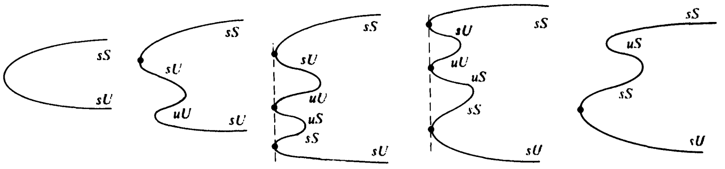


FIG. 20. *Diagrams corresponding to regions marked in Fig.* 13. *Region* 1 *corresponds to an empty diagram.*

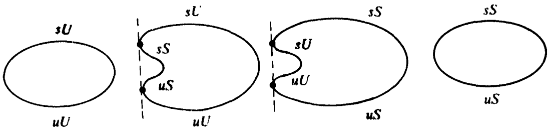


FIG. 21. *Diagrams corresponding to regions marked in Fig.* 14.

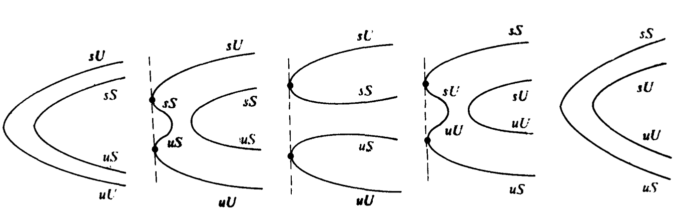


FIG. 22. *Diagrams corresponding to regions marked in Fig.* 15.

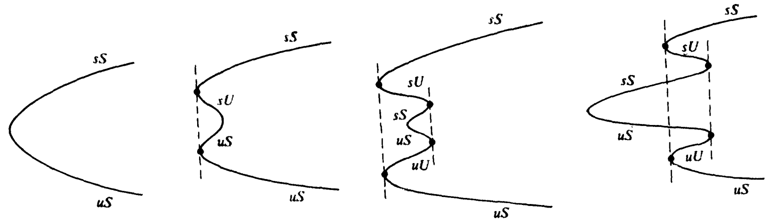


FIG. 23. *Diagrams corresponding to regions marked in Fig.* 16. *Regions* 1 *and* 6 *are empty. Regions* 7–10 *have diagrams like* 2–5, *except that the second stability coefficient is changed from S to U and vice versa.*

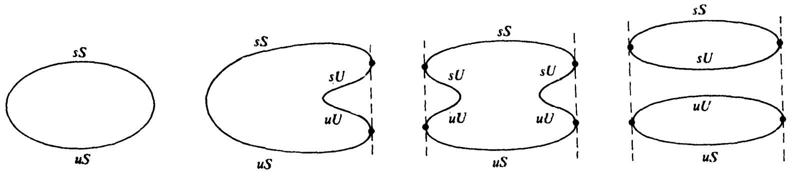


FIG. 24. *Diagrams corresponding to regions marked in Fig.* 17. *Regions* 6–10 *have diagrams like* 1–5, *but the second stability coefficient is changed from S to U and vice versa.*

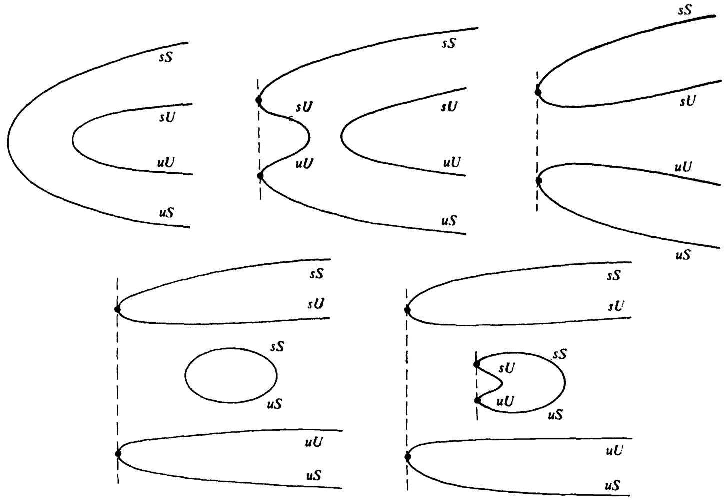


FIG. 25. *How the "coupling" of two bifurcation problems leads to a simple method for finding the diagram of the corresponding sequential bifurcation, by projection, in cases where the x-bifurcation is independent of* λ. *The example here is type* $(1)_{23}$.

visualized as the *projection* of the *curve* (22) in the (x,λ)-plane onto the paraboloid (Fig. 11). The topology of the result depends only on the *position* of the vertex of the paraboloid relative to the curve (2), and on the form of (2) itself. Note that if the vertex (here the origin) falls midway between two limit points of (2), then the projection has a double limit point involving two distinct sheets of the (u,λ)-equation. This is a general feature of the analysis, and means that "Maxwell set" considerations (and others of a similar nature) arise—see Poston and Stewart [11].

We have assigned stabilities to these bifurcation diagrams. Although stabilities are not invariants of contact equivalence, in general, they are for one-dimensional problems (up to an arbitrary choice on initial branches). That is, changes of stability happen in an invariant way. On a structurally stable diagram in one dimension, stability changes

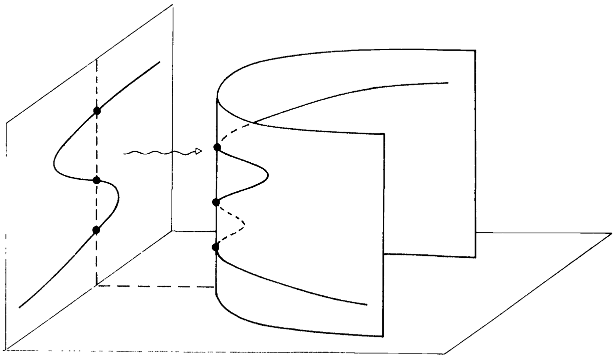


FIG. 11. *Decomposition of unfolding space into regions giving distinct types of structurally stable bifurcation diagrams, for type* $(1)_{22}$.

at a fold (limit point). Because of the "triangular" form of a sequential bifurcation, we can assign stability coefficients as follows. First consider only the (u,λ)-equation: assign stabilities here. They will change only at folds in the (u,λ)-diagram; that is, at the (structurally) *stable* double limit points of the (x,λ)-diagram. The other stability coefficient, on the other hand, will change only at limit points *not* coming from the (u,λ) folds.

In Figs. 19–25 we have assigned stabilities in an arbitrary way to initial branches, and derived all other assignments by the above rules. We write "s" and "u" in combination, in upper or lower case, to represent the two stability coefficients: the second (S and U) refers to the (u,λ)-stability; the first (s and u) to that involving x also. Thus the second coefficient changes at structurally stable double limit points, the first at all other single limit points. Alternative assignments, corresponding to different choices of stability on initial branches, are obtained by interchanging s and u throughout, or S and U, or both.

We turn now to individual cases. In Figs. 12, 19 we show type $(1)_{22}$, which may be interpreted as two limit-point bifurcations in series. The result superficially resembles the quartic limit point $x^4-\lambda=0$ under Z_2-symmetry, but note that the stabilities are different. Thus a sequential bifurcation can give rise to an apparent symmetry, but is here distinguished from a genuine symmetry by the stability assignments.

Figures 13, 20 refer to type $(1)_{23}$, limit point followed by hysteresis. Here *triple* limit points can stably occur. There are at most two stable solution branches.

Figures 14, 21 are type $(3)_{22}$ for $\varepsilon_2=+1$, a limit point followed by isola formation. There is at most one stable branch. Similarly Figs. 15, 22 are $(3)_{22}$ with $\varepsilon_2=-1$, limit-point/bifurcation.

Figures 16, 23 are $(1)_{32}$, hysteresis/fold. Note the differences from Fig. 9, where these are coupled in reverse order.

Figures 17, 24 show (11a) with $\varepsilon_1=1$, and a particular choice of modal parameter, $L(u,\lambda)=\gamma u-\lambda$, where we have confined to $|\gamma|<1$. This problem may be interpreted as isola formation followed by a limit point. Note that there are now double isolas, but only one has a stable branch. Figures 18, 25 show the corresponding results when $\varepsilon_1=-1$, bifurcation/limit point. Note particularly that *isolas* can form in this case for regions 4, 5, 9, 10.

7. An application to stirred tank reactors. In this section we shall sketch an application of sequential bifurcations to a problem involving chemical reactors coupled in sequence. In order to avoid undue length, we consider only the simplest model and omit many details of the calculations. A full account of a more realistic model, giving similar results, is in preparation.

Consider two stirred tank reactors coupled in sequence, producing a reaction $\mathcal{R}\to\mathcal{P}$. This system has been studied by Svoronos, Aris and Stephanopoulos [13], but we shall use a different bifurcation parameter here for simplicity.

A model for the dynamics of such a system is given by (Svoronos et al. [13])

$$V_1\frac{dc_1}{dt'}=F(c_0-c_1)-V_1k(T_1)c_1, \tag{23}$$

$$V_1\rho C_p\frac{dT_1}{dt'}=F\rho C_p(T_0-T_1)+V_1(-\Delta H)k(T_1)c_1-h_1S_1(T_1-T_{c1}), \tag{24}$$

$$V_2\frac{dc_2}{dt'}=F(c_1-c_2)-V_2k(T_2)c_2, \tag{25}$$

$$V_2\rho C_p\frac{dT_2}{dt'}=F\rho C_p(T_1-T_2)+V_2(-\Delta H)k(T_2)c_2-h_2S_2(T_2-T_{c2}). \tag{26}$$

Here the variables are as follows:

Tank i $(=1,2)$ has volume V_i and surface area S_i for heat transfer.
Reactant $\mathcal{R}$ has concentration c_0 and temperature T_0.
Reactant $\mathcal{R}$/product $\mathcal{P}$ has concentration c_1 and temperature T_1.
Product $\mathcal{P}$ has concentration c_2 and temperature T_2.
T_{ci} is ambient temperature for tank i.
h_i is a heat transfer coefficient.
t' is time (to be scaled later).
F is flow rate.
ρ is density and C_p specific heat of mixtures.
ΔH is the heat of reaction, and is negative (for an exothermic reaction).
$k(T)$ is a temperature-dependent reaction rate.

For the Arrhenius case we have

$$k(T)=\text{const.}\,e^{-ER/T}$$

where E is activation energy, R the Boltzmann constant.

The first step is to pass to dimensionless variables

$$x_i=\frac{c_0-c_i}{c_0},\quad y_i=\frac{T_i-T_0}{T_0},\quad t=\left(\frac{F}{V_i}\right)t'.$$

Now $-\infty<x_i<1$, $-1<y_i<\infty$, and the equations become

$$\frac{dx_1}{dt}=-x_1+D_1(1-x_1)\mathcal{A}'(y_1), \tag{27}$$

$$\frac{dy_1}{dt}=-y_1+\beta_1(\eta_1-y_1)+QD_1(1-x_1)\mathcal{A}'(y_1), \tag{28}$$

$$\alpha\frac{dx_2}{dt}=x_1-x_2+D_2(1-x_2)\mathcal{A}'(y_2), \tag{29}$$

$$\alpha\frac{dy_2}{dt}=y_1-y_2+\beta_2(\eta_2-y_2)+QD_2(1-x_2)\mathcal{A}'(y_2). \tag{30}$$

Here $D_i = V_i k(T_0)/F > 0$ is a Damköhler number, $\beta_i = h_i S_i / F\rho C_p > 0$, $Q = (-\Delta H)c_0/\rho C_p T_0 > 0$, $\alpha = V_2/V_1$, $\eta_i = (T_{ci} - T_0)/T_0$ and $-1 < \eta_i < \infty$, and $\mathcal{A}'(y) = k(T_0 + yT_0)/k(T_0)$. Following Svoronos et al. [13] we set

$$\alpha = 1, \quad D_1 = D_2 \equiv D, \quad \beta_1 = \beta_2 \equiv \beta.$$

This corresponds to (essentially) identical tanks.

We seek steady states by putting the time derivatives equal to zero. Using the x_i-equations we have

$$D\mathcal{A}'(y_1) = \frac{x_1}{1 - x_1}, \tag{31}$$

$$D\mathcal{A}'(y_2) = \frac{x_2 - x_1}{1 - x_2}. \tag{32}$$

Inserting these into the y_i-equations we get

$$(1+\beta)y_1 = \beta\eta_1 + Qx_1, \tag{33}$$

$$(1+\beta)y_2 = \beta\eta_2 + \beta\eta_1/(1+\beta) + Q(x_2 - \beta x_1/(1+\beta)). \tag{34}$$

Solving (31), (32) for x_i and substituting into (33), (34) we obtain a reduction to just two equations:

$$A(y_1, \delta) = -y_1(1+\beta) + \beta\eta_1 + \frac{Q}{1 + \delta\mathcal{A}(y_1)} = 0, \tag{35}$$

$$B(y_1, y_2, \delta) = -y_2(1+\beta) + y_1 + \beta\eta_2 + \frac{Q\delta\mathcal{A}(y_1)}{(1 + \delta\mathcal{A}(y_1))(1 + \delta\mathcal{A}(y_2))} = 0. \tag{36}$$

Here $\delta = 1/D$, $\mathcal{A}(y) = 1/\mathcal{A}'(y)$ and in the Arrhenius case $\mathcal{A}(y) = e^{-(\gamma y/(1+y))}$ where $\gamma = -E/RT_0 > 0$. Clearly (35), (36) define a sequential bifurcation in (y_1, y_2) with bifurcation parameter δ, for a given choice of the remaining parameters.

In a paper in preparation we shall analyse these equations for a class of functions A that includes the Arrhenius case. However, this greatly complicates the calculations, and for the purposes of this illustration we now make an approximation which, though standard in the literature, must be handled with some delicacy. It is known as the *positive experimental approximation.*

Assume Arrhenius kinetics and rescale further by setting

$$y_i \mapsto \gamma y_i, \quad \eta_i \mapsto \gamma\eta_i, \quad Q \mapsto \gamma Q$$

and then let $\gamma \to \infty$. The resulting equations may be written in the form

$$0 = A(y_1, \delta) = -y_1(1+\beta) + \beta\eta_1 + Qp_1/(1+p_1)\delta, \tag{37}$$

$$0 = B(y_1, y_2, \delta) = -y_2(1+\beta) + y_1 + \beta\eta_2 + Qp_2/(1+p_1)(1+p_2), \tag{38}$$

where

$$p_i = \delta e^{-y_i}. \tag{39}$$

Now the ranges of the variables are $-\infty < y_i, \eta_i < \infty$, and $Q, \delta, \beta > 0$.

Essentially, the approximation amounts to letting activation energy tend to infinity, so any results derived should be interpreted in this asymptotic sense.

We now analyse (37) and (38) as sequential bifurcations. The first step is to seek the codimension-1 "accidents" that delineate transitions between stable codimension-0 diagrams (analogous to the use of hysteresis, bifurcation and double limit varieties by Golubitsky and Schaeffer [6]). We have

$$(40) \qquad A=B=0, \qquad A_{y_1}=B_{y_2}=0$$

(subscripts denoting partial derivatives). The second pair of equation yields

$$(41) \qquad p_1=\frac{1+p_2+p_2^2}{p_2}>1,$$

$$(42) \qquad 1+\beta=\frac{Qp_1}{(1+p_1)^2}.$$

The other two equations let us express the η_i in terms of β, Q, p_i, and will be used later.

Now we look for phenomena of codimension>1. Note that $A_{y_1y_1}=0$ implies $p_1=1$ which contradicts (41). Hence at any sequential bifurcation point the local form of A must be $u^2-\lambda$, since $A\delta<0$. Thus any higher-codimension phenomena must be due to additional degeneracies in B.

The first case is when $B_{y_2y_2}=0$, which gives $p_2=1, p_1=3$ (from (41)) and $B_{y_1y_2y_2}<0$, leading to type $(1)_{23}$ sequential bifurcations. These degenerate to $(2)_{23}$ if in addition we have $B_{y_1}=0$, giving

$$(43) \qquad \beta=p_2, \qquad Q=\frac{(1+p_1)^2(1+p_2)}{p_1}.$$

Using this, we find a unique line of problems of type $(2)_{23}$, which can be expressed parametrically in terms of δ:

$$(44) \quad y_1=\ln\delta-1.10, \quad y_2=\ln\delta, \quad \beta=1, \quad Q=\tfrac{32}{3}, \quad \eta_1=2\ln\delta-4.86, \quad \eta_2=\ln\delta-2.90.$$

Further, since $B_{y_1y_2}<0$, the sign ε_2 in $(2)_{23}$ is $+1$.

In all other cases it follows that the only possible problems with codimension>1 and $B_{y_2y_2}\neq 0$ are types $(3)_{2n}$ where $n\geq 2$.

The degeneracy condition for $(3)_{22}$ is again $B_{y_1}=0$ (corresponding to tangential contact of the surfaces $A=0$ and $B=0$). This degenerates into $(3)_{23}$ if there is second order contact (equal curvature in some direction) which is the geometrical meaning of the degeneracy condition $D_{3,1}=0$. A straightforward computation yields

$$(45) \qquad (p_1^2-1)(p_2-1)+(p_2+1)=0$$

so, from (41),

$$p_2=\left(\frac{(\sqrt{5}-1)}{2}\right)^{1/2}=0.79, \qquad p_1=3.06.$$

Thus we have a unique $(3)_{23}$-line, expressible as

$$(46) \qquad \begin{aligned} &y_1=\ln\delta-1.12, \quad y_2=\ln\delta+0.24, \quad \beta=0.79, \quad Q=2.66, \\ &\eta_1=2.27\ln\delta-5.56, \quad \eta_2=\ln\delta+1.54. \end{aligned}$$

A further computation shows that $D_{3,3}>0$ and $B_{y_2y_2}<0$ in (46), so that the $(3)_{23}$ occurs with sign $\varepsilon_2=-1$.

These computations show that the most degenerate types occurring (in the approximated equations) are $(2)_{23}$ and $(3)_{23}$, hence that any other types are organized by these, according to the subordination diagram

$$\begin{tikzcd}
 & (u,x^3\pm x\lambda) \arrow[dr] \arrow[dddr] & & \\
 & & (u,x^3\pm\lambda) \arrow[dr] & \\
 & (1)_{23} \arrow[ur] \arrow[dddr] & & (u,x^2\pm\lambda) \\
(2)_{23} \arrow[uuur] \arrow[ur] \arrow[dr] & & (u,x^2\pm\lambda^2) \arrow[ur] & \\
 & (3)_{22} \arrow[ur] \arrow[dr] & & (u^2\pm\lambda,x) \\
(3)_{23} \arrow[ur] \arrow[dr] & & (1)_{22} \arrow[uuur] \arrow[ur] & \\
 & (u,x^2\pm\lambda^3) \arrow[uuur] & &
\end{tikzcd}$$

(where for simplicity we have ignored sign distinctions). Here an arrow indicates that the arrowhead occurs in an unfolding of the tail, and the singularities in brackets are ordinary bifurcations in normal form as in the Golubitsky–Schaeffer theory. The diagram above is not complete: a few (more singular) ordinary bifurcations also occur.

Since a much more detailed and general analysis is in preparation, we shall not describe here the geometry of types $(2)_{23}$ and $(3)_{23}$, but the calculations above exemplify the way that the singularity theory can lead to a fairly detailed understanding of the possible types of behaviour, and the way that the calculations are organized.

8. Conclusions. The theory of sequential bifurcations

$$(1) \qquad a(u,\lambda)=0, \qquad b(x,u,\lambda)=0$$

modelling two processes coupled in cascade, while analogous to that for ordinary bifurcation problems, has important differences from it. To preserve the coupled structure, it is necessary to modify the usual notions of structural stability, equivalence, codimension and unfoldings. (This is of course not surprising: there are already many variants of these ideas in the literature, selected to be appropriate to particular kinds of problems.)

In particular, certain phenomena can occur *typically* in (1) viewed as a sequential bifurcation, which do *not* occur if it is seen as simply a special case (a-equation independent of x) of a bifurcation in two variables (x,u). For example, double limit points are now a codimension-0 phenomenon, hence likely to be observed in an experiment modelled by such equations.

Further, (1) can have lower codimension, as a sequential bifurcation, than as a two-variable bifurcation.

This implies that the "hidden" parameter u is not just some kind of dummy behaviour of the "output" variable x. In principle it might be possible to decide whether an experiment is best described using such a hidden variable, by observing its bifurcations. To put it very loosely, *the presence of hidden variables can be detected via bifurcations.*

This is true only if one accepts the relevance of structural stability and unfolding theory, of course. However, it illustrates the way that these ideas lead to new predictions.

It also raises an intriguing point. In the analysis of systems of model equations, it is common practice to take advantage of any "accidental" simplification to (say) eliminate unwanted variables. If an equation is independent of a certain variable, it may be possible to solve it for another one, substitute this into one of the other equations, and so forth. As our example in §4 shows, while this is unobjectionable as a way of understanding the given system of equations, it may lead to unwarranted inferences about the likely *perturbations* of that system. In this sense, "accidental" simplifications are misleading and should not be taken advantage of! However, for each class of problems, there should be some natural structure which reflects the physical assumptions behind the model, and transformations or simplifications *which respect that structure* are of course permissible. There is much more to understanding a system than just solving "the equations"; in fact the *same* equations, under two different interpretations, can have quite distinct implications. This is because, as well as the equations, there is "hidden" mathematical structure: which transformations of the equations make sense.

The type of system chosen in this paper—two processes in series—is arguably the simplest case where such considerations appear. It is obviously the tip of an enormous iceberg: there are innumerable more or less complex ways to couple systems. Each has its own version of singularity/catastrophe/bifurcation theory, with its mathematics tailored to its structure. However, before embarking on extensive analyses of even more complicated versions of the theory, some additional experience of how these ideas work out in genuine applications appears advisable.

Appendix. We summarize conditions which must be satisfied by an arbitrary SBP g for being contact equivalent to one of the normal forms of Tables 1 and 2 with topological codimension less than four or to a member of the hierarchies $(1)_{mn}$, $(2)_{mn}$, $(4)_m$. The conditions are expressed in terms of the Taylor coefficients

$$a_{ij}=\frac{1}{i!j!}\frac{\partial^{i+j}a(0,0)}{\partial u^i\partial\lambda^j},\qquad b_{ijk}=\frac{1}{i!j!k!}\frac{\partial^{i+j+k}b(0,0,0)}{\partial x^i\partial u^j\partial\lambda^k}$$

of a and b and are summarized in Table 3. Recall that any SBP satisfies $a_{10}=b_{100}=0$. The first and second columns of Table 3 contain expressions which must be zero and nonzero, respectively (degeneracy and nondegeneracy conditions), if g is equivalent to a particular normal form. In the third and fourth columns we have summarized expressions whose signs determine ε_1 and ε_2. ε_3 is given by

$$\varepsilon_3=\begin{cases}\operatorname{sgn}(b_{200}b_{001}) & \text{for } (4)_m,\ (5)_3,\\ \operatorname{sgn}(b_{200}D_{6,2}) & \text{for } (6)_2\end{cases}$$

($D_{6,2}$ is defined in (A3) below). In the last two columns of Table 3 we have tabulated the ideals I and J which generate the space $P(g)$ (see §3, (17)). Here, $\mathfrak{m}$ stands for $\mathfrak{m}_{u\lambda}$ and $\mathfrak{m}_{xu\lambda}$ in the fifth and sixth column, respectively. The various expressions entering Table 3 are summarized in (A1)–(A4):

(A1)
$$\begin{aligned}D_{3,1}&=b_{020}-\tfrac{1}{4}b_{110}^2b_{200}-a_{20}b_{001}/a_{01},\\ D_{3,2}&=b_{200}b_{020}-\tfrac{1}{4}b_{110}^2,\\ D_{3,3}&=\sum_{i+j=3}b_{ij0}(-b_{110}/2b_{200})^i+a_{20}a_{11}b_{001}/a_{01}^2\\ &\qquad-(b_{001}a_{30}+b_{011}a_{20}-b_{110}b_{101}a_{20}/2b_{200})/a_{01};\end{aligned}$$

(A2)
$$D_{5,1}=a_{20}a_{02}-\tfrac{1}{4}a_{11}^2,\qquad D_{5,2}=d^3a(a_{11},-2a_{20}),$$

TABLE 3
Conditions for the main normal forms in Tables 1 *and* 2.

	Zero	Nonzero	$\varepsilon_1=\operatorname{sgn}(\cdot)$	$\varepsilon_2=\operatorname{sgn}(\cdot)$	J	I
$(1)_{mn}$	$a_{j0}(j<m)$ $b_{i00}(i<n)$	$a_{m0}, b_{n00}, a_{01}, b_{010}$	$a_{m0}a_{01}$	$b_{n00}b_{010}$	$\mathfrak{m}^{m+1}+\mathfrak{m}\langle\lambda\rangle$	$\langle x^{n+1}, xu, u^2, \lambda\rangle$
$(2)_{mn}$	$a_{j0}(j<m)$ $b_{i00}(i<n), b_{010}$	$a_{m0}, a_{01}, b_{n00}, b_{110}$	$a_{m0}a_{01}$	$b_{n00}b_{110}$	$\mathfrak{m}^{m+1}+\mathfrak{m}\langle\lambda\rangle$	$\langle x^{n+1}, x^2u, u^2, \lambda\rangle$
$(3)_{22}$	b_{010}	$a_{20}, a_{01}, b_{200}, D_{3,1}$	$a_{20}a_{01}$	$b_{200}D_{3,1}$	$\mathfrak{m}^3+\mathfrak{m}\langle\lambda\rangle$	$\mathfrak{m}^3+\mathfrak{m}\langle\lambda\rangle$
$(3)_{32}$	a_{20}, b_{010}	$a_{30}, a_{01}, b_{200}, D_{3,2}$	$a_{30}a_{01}$	$D_{3,2}$	$\mathfrak{m}^4+\mathfrak{m}\langle\lambda\rangle$	$\mathfrak{m}^3+\langle\lambda\rangle$
$(3)_{23}$	$b_{010}, D_{3,1}$	$a_{20}, a_{01}, b_{200}, D_{3,3}$	$a_{20}a_{01}$	$b_{200}D_{3,3}$	$\langle u^4, u^2\lambda, \lambda^2\rangle$	$\mathfrak{m}^4+\mathfrak{m}^2\langle\lambda\rangle+\langle\lambda^2\rangle$
$(4)_m$	$a_{j0}(j<m), a_{01}$	$a_{m0}, a_{11}, b_{200}, b_{010}, b_{001}$	$a_{m0}a_{11}$	$b_{200}b_{010}$	$\langle u^{m+1}, u^2\lambda, \lambda^2\rangle$	$\mathfrak{m}^3+\mathfrak{m}\langle u,\lambda\rangle$
$(5)_3$	$D_{5,1}$	$D_{5,2}, a_{20}, b_{200}, b_{010}, b_{001}$	$D_{5,2}$	$b_{200}b_{010}$	$\mathfrak{m}^4$	$\mathfrak{m}^3+\mathfrak{m}\langle u,\lambda\rangle$
$(6)_2$	$a_{01}, D_{6,1}$	$a_{20}, D_{5,1}<0, b_{200}, b_{010}, D_{6,2}$	$\varepsilon a_{20}b_{010}b_{200}$	$b_{200}b_{010}$	$\mathfrak{m}^4$	$\mathfrak{m}^3$
(11a)	a_{01}	$a_{20}, D_{5,1}, D_{6,1}$	$D_{5,1}$	–	$\mathfrak{m}^3$	$\mathfrak{m}^3+\mathfrak{m}\langle u,\lambda\rangle$
(11b)	a_{01}, b_{200}	$a_{20}, D_{6,1}, D_{5,1}, b_{300}, D_{11}$	$D_{5,1}$	–	$\mathfrak{m}^2$	$\langle x^2, u, \lambda\rangle^2$

(A3) $$D_{6,1}=d^2a(w,w), \qquad \varepsilon=\operatorname{sgn}\{2a_{20}b_{001}/b_{010}-a_{11}\},$$
$$D_{6,2}=d^2b(w,w)-\tfrac{1}{4}(db_x(w))^2-\tfrac{1}{2}\varepsilon|D_{5,1}|^{-1/2}d^3a(w,w,w),$$
$$w=(b_{001}, -b_{010}),$$

(A4) $$D_{11}=b_{001}b_{110}-b_{101}b_{010}.$$

In (A2) and (A3), d is the total derivative with respect to (u,λ) which has to be evaluated at $(x,u,\lambda)=(0,0,0)$. The angle ϕ entering the linear forms $L(u,\lambda)$ and $\bar{L}(u,\lambda)$ in the normal forms (11a, b) is given by

$$\tan\phi=-|a_{20}||D_{5,1}|^{-1/2}\left\{\frac{b_{001}}{b_{010}}-\frac{1}{2}\frac{a_{11}}{a_{20}}\right\}.$$

Acknowledgments. One of us (IS) wishes to thank the A. v. Humboldt Foundation for a visiting fellowship and Prof. W. Güttinger for the hospitality of his Institute.

We are grateful to Prof. Jim Damon for sending us a copy of [3] in preprint form and for answering our questions on the applicability of his work on unfoldings to problems mentioned in §3 above.

REFERENCES

[1] V. I. ARNOLD, *Classification of unimodal critical points of functions*, Functional Anal. Appl., 7 (1973), pp. 230–231.

[2] ______, *Classification of bimodal critical points of functions*, Functional Anal. Appl., 9 (1975), pp. 43–44.

[3] J. DAMON, *The unfolding and determinacy theorems for subgroups of A and K: A survey*, Proc. Arcata Conference on Singularities (1981), Symposia in Pure Math., American Mathematical Society, Providence, RI, to appear.

[4] G. DANGELMAYR AND D. ARMBRUSTER, *Classification of Z(2)-equivariant bifurcation problems with corank two*, Proc. London Math. Soc., (1982), 46 (1983), pp. 517-546.

[5] C. G. GIBSON, *Singular points of smooth mappings*, Research Notes in Mathematics, 25, Pitman, London and Boston, 1979.

[6] M. GOLUBITSKY AND D. SCHAEFFER, *A theory of imperfect bifurcation via singularity theory*, Comm. Pure Appl. Math., 32 (1979), pp. 21–98.

[7] ______, *Imperfect bifurcation in the presence of symmetry*, Comm. Pure Appl. Math., 67 (1979), pp. 205–232.

[8] ______, *A discussion of symmetry and symmetry breaking*, preprint, October 1981.

[9] J. MATHER, *Stability of C^∞-mappings* VI. *The nice dimensions*, in Proc. of the Liverpool Singularities Symposium, C. T. Wall, ed., Lecture Notes in Mathematics 192, Springer, New York, 1971, pp. 207–253.

[10] ______, *How to stratify mappings and jet spaces*, in Singularités d'applications différentiables, Plans-sur-Bex 1975, O. Burlet and F. Rouga, eds., Lecture Notes in Mathematics 535, Springer, New York, 1976, pp. 128–176.

[11] T. POSTON AND I. STEWART, *Catastrophe Theory and Its Applications*, Pitman, Boston and London, 1978.

[12] I. STEWART, *Applications of catastrophe theory in the physical sciences*, Physica, 2D (1981), pp. 245–305.

[13] S. SVORONOS, R. ARIS AND G. STEPHANOPOULOS, *On the behaviour of two stirred tanks in series*, Chem. Engng. Sci., 37 (1982), pp. 357–366.

[14] R. THOM, *Structural Stability and Morphogenesis*, Benjamin-Addison-Wesley, New York, 1975.

[15] E. C. ZEEMAN, *Classification of elementary catastrophes of codimension$\leq$5*, in Structural Stability, the Theory of Catastrophes and Applications in the Sciences, P. J. Hilton, ed., Lecture Notes in Mathematics 525, Springer, New York, 1977, pp. 263–327.

[16] ______, *Selected Papers* 1972–1977, Addison-Wesley, Reading, MA, 1977.

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics

002

BOUNDARY CONDITIONS AND MODE JUMPING IN THE VON KÁRMÁN EQUATIONS*

E. J. HOLDER[†] AND D. SCHAEFFER[‡]

Abstract. Mode jumping in the postbuckling response of a long rectangular plate is investigated with bifurcation theory methods near a double eigenvalue. It is found that mode jumping is predicted by the theory if clamped boundary conditions are imposed on the end faces of the plate, but not if simply supported boundary conditions are imposed.

Introduction. Consider a long rectangular plate in compression. At high loads such a plate buckles; the deformed pattern consists of a series of bulges (called buckles) along the length of the plate which alternate in sign. This pattern persists as the load Λ is increased beyond the buckling load Λ_c, up to a point. In certain experiments [10], when Λ is increased sufficiently far beyond Λ_c the plate jumps "suddenly and violently" to a new configuration, the number of buckles increasing by one. The phrase *mode jumping* has been coined to describe this phenomenon.

Several authors [1], [8a], [9] have suggested that perturbation of a multiple eigenvalue may provide an explanation for mode jumping. This work is based on the important paper of Bauer, Keller, and Reiss [1]. These authors showed that perturbing a bifurcation problem with a multiple eigenvalue often causes secondary bifurcation. The most relevant perturbed diagram which can occur in this way is sketched in Fig. 0.1. In this figure "s" and "u" refer to stable and unstable branches. We claim that this diagram exhibits mode jumping. To see this, consider gradually increasing the load Λ from zero. For $\Lambda<\Lambda_A$ the system will follow the trivial solution (i.e., no deflection), and for $\Lambda_A<\Lambda<\Lambda_B$ the system will follow the first bifurcating solution branch. However at Λ_B this branch loses stability through a subcritical secondary bifurcation, and there is no stable solution branch emanating from B for the system to follow when $\Lambda>\Lambda_B$. Thus, although strictly speaking it is impossible to determine the dynamic behavior of a system from a bifurcation diagram, presumably the system jumps to a new equilibrium along the other branch if Λ is increased beyond Λ_B.

These ideas apply to buckling plates as follows. For certain aspect ratios (i.e., ratio of length to width) the first bifurcation in the plate problem is from a double eigenvalue. Let l^* be such an aspect ratio. The plate problem for values of l close to l^* is then a perturbation of a bifurcation problem with a multiple eigenvalue, and the hope is that the perturbed bifurcation diagram may be that of Fig. 0.1.

However Fig. 0.1 does not invariably result when a double eigenvalue is perturbed. There are several other possibilities, one of which is shown in Fig. 0.2, and none of these other diagrams exhibit mode jumping. Exactly what does occur depends on the numerical values of the ratios of certain coefficients in the governing equations. Thus to determine whether a given theory predicts mode jumping, it is necessary to do specific calculations for that theory.

In this paper we study whether or not mode jumping is predicted by the above mechanism in the von Kármán equations with various boundary conditions. (The von

*Received by the editors December 13, 1982, and in revised form April 19, 1983.

[†] Dynetics, Inc., Huntsville, Alabama.

[‡] Department of Mathematics, Duke University, Durham, North Carolina 27706. The research of this author was supported by the National Science Foundation under grant MCS77-04148.

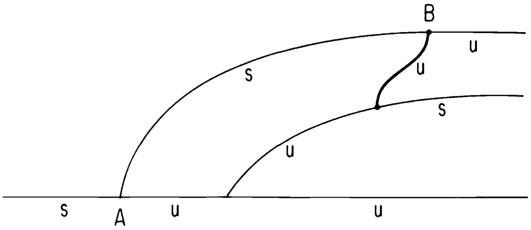


FIG. 0.1

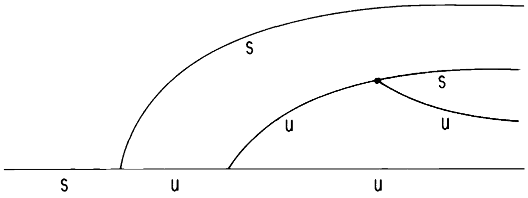


FIG. 0.2

Kármán equations for the displacement w and the Airy stress function ϕ are given in §2.) We consider two sets of boundary conditions for w,

I. simply supported on all four edges,

II. simply supported on (unloaded) sides, clamped on (loaded) ends;

and two sets of boundary conditions for ϕ,

A. Dirichlet ($\phi=\Delta\phi=0$),

B. Neumann ($\phi_N=(\Delta\phi)_N=0$).

We refer to the four cases here as IA, etc. Our conclusion is that mode jumping is predicted in cases IIA and IIB, not in IA or IB. The following remarks may be helpful for comparison with experiment. Simply supported boundary conditions I on w are a convenient mathematical abstraction but are notoriously difficult to achieve in experiments. Stein [10] considers II the most accurate description for w in his experiment. It is shown in [9] that conditions B result if one assumes that the compression of the plate is uniform along the loaded edges and that no tangential stresses are transmitted. The physical basis of conditions A is rather problematic, at best.

Case IA has been studied by several authors [4], [7], [8], [8a] despite its irrelevance for the experiment. Cases IB and IIB were studied in [9]. Thus case IIA is the only completely new case here, although in carrying out this research we found disagreement among the various results in the literature for case IA. This is discussed further in §4.

Our purpose in writing yet another paper on this much studied problem was two-fold: to fill a gap in the literature by analyzing the missing case IIA mentioned above and to emphasize the usefulness of singularity theory in applied problems. Let us elaborate on the second goal. The ultimate goal of the analysis in this and related problems is to understand how the solution depends on various parameters. The difficulties of achieving this goal stem from the fact that precisely at a bifurcation point the solution does *not* vary smoothly, and at a multiple eigenvalue the possible complications are legion. We believe that there is a significant conceptual gain, with no loss of computational power, in dividing this analysis into two distinct steps. We assume the reader is familiar with the Lyapunov–Schmidt technique to reduce a problem with bifurcation from an eigenvalue of multiplicity n to a system of n equations in n real unknowns. The first step in our proposed division of the analysis is to study how the coefficients in the reduced equation depend on the various parameters; the second step is to study how the solutions depend on these coefficients. The point is that all nonsmooth behavior is isolated in the second step. (Indeed, the mathematical discipline of singularity theory is concerned exactly with nonsmooth dependence of the solution of an equation on the coefficients.) The first step must be done separately for each new specific problem; this step poses no theoretical subtleties, just calculations that may be difficult. By contrast the theoretical analyses of the second step can often be applied to many different problems, with considerable savings in effort. Even in cases where the calculations of step one are too difficult to carry out, the theory in step two may provide useful information about the form of the solution.

In §1 of this paper we recall from [9] the theoretical framework for this problem provided by singularity theory. In §2, we give the properties of von Kármán equations relevant for this analysis, referring to [9] for proofs. We present the results of our calculations in §3, along with a brief discussion of the method of calculation. In §4, we list several independent checks on our calculations, the most noteworthy being an asymptotic analysis of a plate with a large aspect ratio.

It is rather unsatisfying that there is only the one reference [10] in this paper to experiment. However, we are not aware of other published experiments, although we have looked for them; nor are there any such references in the other analyses of the problem mentioned above.

1. The theoretical framework. Bauer, Keller and Reiss [1] observed that secondary bifurcation may occur in perturbing an idealized bifurcation problem with a double eigenvalue, and this is the idea behind the prediction of mode jumping with bifurcation theory. In the plate problem for certain aspect ratios the first bifurcation is from a double eigenvalue; for example this happens with boundary conditions II on w if the aspect ratio is $\sqrt{k(k+2)}$, k a positive integer. (See §2.) We consider changes of the aspect ratio as a perturbation of an idealized problem in which the eigenvalue is double. Thus for nearby aspect ratios one may expect mode jumping as a result of secondary bifurcation, *provided* the parameters which describe the bifurcation lie in the right ranges. The conditions on these parameters were formulated systematically in [6], [9], and we now review these conditions.

At a double eigenvalue the Lyapunov–Schmidt reduction of the von Kármán equations will lead to a system of equations of the form

$$(1.1) \qquad ax^3+bxy^2-p\lambda x=0, \qquad cx^2y+dy^3-q\lambda y=0,$$

where $\lambda=\Lambda-\Lambda_c$, the load minus the buckling load, and a,b,c,d,p,q are certain constants. Here x and y are *not* spatial coordinates, but unknown coefficients of the eigenfunctions in the Lyapunov–Schmidt reduction; more specifically in this reduction one seeks a solution $w=xw_1+yw_2+O(x^2+y^2)$ where w_1 and w_2 span the kernel of the linearized problem at the double eigenvalue. At first glance it appears that (1.1) is only valid in modulo higher order terms, but it was shown in [6] that apart from the exceptional cases (on the boundary between two regions) mentioned below the higher order terms *may be transformed away by an appropriate change of coordinates, at least locally.*

With appropriate scaling (1.1) may be reduced to the form

$$(1.2) \qquad x^3+\mu xy^2-\lambda x=0, \qquad \nu x^2y+y^3-\lambda y=0,$$

where

$$\mu = \frac{bq}{dp}, \qquad \nu = \frac{cp}{aq}. \tag{1.3}$$

Specifically scale x and y by the factors $(p/a)^{1/2}$ and $(q/d)^{1/2}$, respectively, and multiply the first and second equations by $(a/p^3)^{1/2}$ and $(d/q^3)^{1/2}$, respectively. The two parameters μ and ν are the only dimensionless parameters that may be associated to (1.1), and they provide a far more compact description of the problem than the five parameters used in [8]. Following the usage of singularity theory we shall refer to them as *modal parameters*.

Both equations in (1.2) may be factored giving rise to the following four sets of solutions:

(i) $x=y=0$, λ arbitrary,
(ii) $x=0, y=\pm\lambda^{1/2}$,
(iii) $y=0, x=\pm\lambda^{1/2}$,

$$\text{(iv)} \qquad x=\pm\left(\frac{\mu-1}{\mu\nu-1}\lambda\right)^{1/2}, \qquad y=\pm\left(\frac{\nu-1}{\mu\nu-1}\lambda\right)^{1/2}.$$

Depending on μ and ν, there may or not be a range of λ for which the radicals in (iv) are both real. In Fig. 1.1, we have identified five regions in the μ,ν plane (and two mirror images), and in Table 1.1 we have tabulated the properties of solutions (ii), (iii), and (iv). This tabulation is based on the normalization that the trivial solution (i) is stable for $\lambda<0$, unstable for $\lambda>0$. See [9] for proofs. In our calculations for the plate problem, the modal parameters always lie in regions 1 or 2.

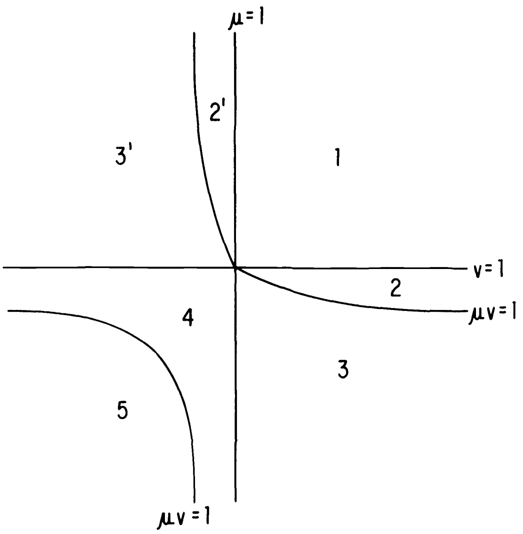


FIG. 1.1.

TABLE 1.1

Region number	Stability of (ii)	Stability of (iii)	Existence of (iv)	Stability of (iv)
1	s	s	$\lambda>0$	u
2	s	—	Never real	not applicable
3	s	—	Never real	not applicable
4	u	u	$\lambda>0$	s
5	u	u	$\lambda<0$	u

It was shown in [9] that a change in the aspect ratio away from the value giving a double eigenvalue modifies (1.2) as follows:

$$(1.4) \qquad x^3+\mu xy^2-\lambda x=0, \qquad \nu x^2y+y^3-(\lambda+\sigma)y=0.$$

(More accurately, (1.4) emerges after an appropriate change of coordinates.) The parameter σ splits the double eigenvalue into two simple bifurcations, at $\lambda=0$ and $\lambda=-\sigma$. In Fig. 1.2 and 1.3, we have drawn the bifurcation diagram of (1.4) (i.e., the solution set) when μ,ν belongs to region 1 or 2. We have labeled the solution branches s, $-$, u to indicate that the two eigenvalues of the differential of (1.4) are stable-stable, stable-unstable, or unstable-unstable respectively. All bifurcations in these figures are symmetric, but we have drawn the bifurcations which occur in the x,λ plane slightly asymmetric in an attempt at perspective. An inspection of these diagrams shows that mode jumping as a result of secondary bifurcation occurs if μ,ν lie in region 2 and $\sigma<0$, and not otherwise. Basically the difference between regions 1 and 2 consists in the fact that in region 1 both the x and y modes are stable for large λ, while in region 2 only the y mode is stable for large λ. Now consider region 2 when $\sigma<0$; the first bifurcation is into the x mode, which (by exchange of stability) is stable near the bifurcation point; mode jumping occurs at the point where the x mode loses stability as required in order to match properly with large λ. In contrast, in region 1 the secondary

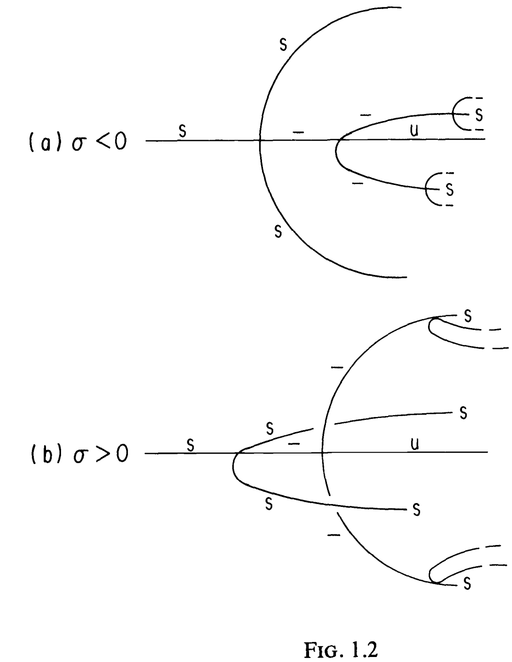


FIG. 1.2

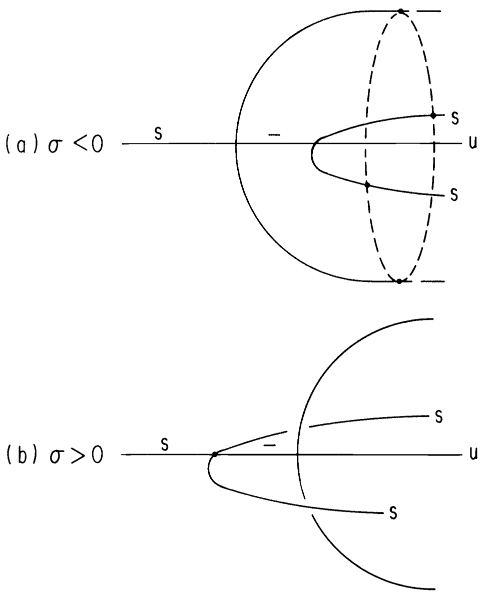


FIG. 1.3

bifurcation merely restabilizes the branch emerging from the second primary bifurcation.

Let us anticipate the results of the next section and indicate how these ideas apply to the plate problem. Here we consider only boundary conditions II on w. If the aspect ratio l satisfies

$$\sqrt{(k-1)(k+1)}<l<\sqrt{k(k+2)}\,, \tag{1.5}$$

the first bifurcation is into a mode with k buckles. At the upper limit in (1.5), two distinct modes with k and with $k+1$ buckles bifurcate simultaneously; (1.2) describes the bifurcation in this case, with x the amplitude of the mode with k buckles, and y, $k+1$ buckles. We show that for both cases IIA and IIB the modal parameters lie in region 2. When $\sigma<0$, which corresponds to a decrease in the aspect ratio from the value $\sqrt{k(k+2)}$, the mode with k buckles bifurcates first but loses stability slightly above the bifurcation point as indicated in Fig. 1.4a, leading to a mode jump. This analysis is rigorously applicable only locally (i.e., for aspect ratios close to $\sqrt{k(k+2)}$), but it provides a strong plausibility argument for mode jumping throughout the entire interval (1.5) and indeed for still smaller l, since the secondary solution branches which cause the mode jumping will tend to persist as l is decreased. We recall that Stein [10] observed repeated mode jumping, in each case the number of buckles increasing by 1. This fact supports the existence of secondary solution branches far from the region where rigorous analysis guarantees their existence.

In summary, the modal parameters are decisive as to whether mode jumping will occur, and we now turn to their calculation in the various cases.

2. Formulas for the modal parameters. Let us normalize the dimensions of the plate so that when undeformed it occupies the planar region

$$\Omega=\{(z_1,z_2)\colon 0<z_1<l\pi,\ 0<z_2<\pi\}.$$

The von Kármán equations are

$$(2.1)\qquad \Delta^2 w=[\phi,w]-\lambda\frac{\partial^2 w}{\partial z_1^2},\qquad \Delta^2\phi=-\frac{1}{2}[w,w],$$

where Δ^2 is the biharmonic operator in the plane and

$$(2.2)\qquad [u,v]=\frac{\partial^2 u}{\partial z_1^2}\frac{\partial^2 v}{\partial z_2^2}-2\frac{\partial^2 u}{\partial z_1\partial z_2}\frac{\partial^2 v}{\partial z_1\partial z_2}+\frac{\partial^2 u}{\partial z_2^2}\frac{\partial^2 v}{\partial z_1^2}.$$

We consider two sets of boundary conditions for w, namely

I. $w=\Delta w=0$ on $\partial\Omega$

and

II. $w=\Delta w=0$ for $0<x<l\pi, y=0,\pi$,
$w=w_N=0$ for $x=0,\ l\pi,\ 0<y<\pi$;

and two sets of boundary conditions for ϕ, namely

A. $\phi=\Delta\phi=0$ on $\partial\Omega$ (Dirichlet),

and

B. $\phi_N=(\Delta\phi)_N=0$ on $\partial\Omega$ (Neumann).

The subscript N indicates differentiation in the normal direction.

Equations (2.1) with any choice of boundary conditions admit only the trivial solution $w=\phi=0$ for small λ, but as λ is increased nontrivial solutions bifurcate. As explained in §1, we may extract the information about mode jumping by considering bifurcation from a double eigenvalue. If boundary conditions I on w are imposed, the first bifurcation is double if the aspect ratio equals $\sqrt{k(k+1)}$ for some positive integer k, in which case the associated eigenfunctions are

$$(2.3)\qquad w_1(z)=\sin\left(\frac{kz_1}{l}\right)\sin z_2,\qquad w_2(z)=\sin\left(\frac{(k+1)z_1}{l}\right)\sin z_2.$$

If boundary conditions II are imposed, the first bifurcation is double if $l=\sqrt{k(k+2)}$, and the eigenfunctions are

$$(2.4)\qquad \begin{aligned} w_1(z)&=\left\{\frac{k+2}{k}\sin\frac{kz_1}{l}-\sin\frac{(k+2)z_1}{l}\right\}\sin z_2,\\ w_2(z)&=\left\{\cos\frac{kz_1}{l}-\cos\frac{(k+2)z_1}{l}\right\}\sin z_2.\end{aligned}$$

Both these cases are derived in [9, §4] although case I is well known in the literature.

In performing the Lyapunov–Schmidt reduction of (2.1) a variational formulation of this equation is convenient. It was shown in [9, §3] that solutions of (2.1) may be characterized as extrema of the functional

$$(2.5)\qquad V(w)=\frac{1}{2}\|\Delta w\|^2-\frac{\lambda}{2}\left\|\frac{\partial w}{\partial z_1}\right\|^2+\frac{1}{8}\|\Delta^{-1}[w,w]\|^2$$

where $\|\cdot\|$ indicates the norm in $L^2(\Omega)$. The domain of V consists of functions in $\mathcal{H}_2(\Omega)$ satisfying

(i) $w=0$ on $\partial\Omega$,

or

(ii) $w=0$ on sides, $w=w_N=0$ on ends,

according as boundary conditions I or II are desired. The inverse Laplacian in (2.5) is computed with Dirichlet or Neumann boundary conditions in cases A and B above, respectively. Strictly speaking only case B was considered in [9], but the extension to case A is immediate.

The coefficients a,b,c,d,p,q of the Lyapunov–Schmidt reduction (1.1) are obtained by computing various derivatives of the full equations. As shown in [9, §5], it suffices to compute derivatives of the functional (2.5). This leads to the following formulas for the coefficients:

$$\begin{aligned}(2.6)\qquad & a=\frac{1}{2}\|\Delta^{-1}[w_1,w_2]\|^2,\\ & b=c=\frac{1}{2}\left\{2\|\Delta^{-1}[w_1,w_2]\|^2+(\Delta^{-1}[w_1,w_1],\Delta^{-1}[w_2,w_2])\right\},\\ & d=\frac{1}{2}\|\Delta^{-1}[w_2,w_2]\|^2,\\ & p=\left\|\frac{\partial w_1}{\partial z_1}\right\|^2,\\ & q=\left\|\frac{\partial w_2}{\partial z_1}\right\|^2.\end{aligned}$$

Here w_1 and w_2 are defined by (2.3) or (2.4) to achieve boundary conditions I or II on w, respectively, and the inverse Laplacian is computed with Dirichlet or Neumann boundary conditions to achieve boundary conditions A or B on ϕ, respectively. These formulas may be substituted into (1.3) to yield formulas for the modal parameters.

3. Method of calculation and results. The computer was required to compute the coefficients a,b,c,d in (2.6); the remaining coefficients p,q could be evaluated analytically. The computation began by expanding $[w_i,w_j]$ in terms of the orthonormal eigenfunctions of Δ^{-1}—a double sine series or a double cosine series according as boundary conditions A or B are desired for ϕ, respectively. We then multiplied each coefficient by the corresponding eigenvalue of Δ^{-1} to obtain the coefficients in the expansion of $\Delta^{-1}[w_i,w_j]$. Finally we evaluated the inner products in (2.6) by multiplying Fourier coefficients and adding.

Actually it is possible to perform these calculations analytically when ϕ satisfies boundary conditions B—indeed this was done in [9]. However we repeated these calculations numerically here as a check on our program.

We present our results in Tables 3.1–3.4 below. For each of the four cases IA, IB, IIA, IIB we have evaluated the modal parameters μ and ν for the ten smallest aspect ratios which lead to bifurcation from a double eigenvalue. (See §2.) We have also listed μ *times* ν as part of the output to assist the reader in identifying the region. As it happens, the region in which the modal parameters lie depends only on the choice of boundary conditions and not on the aspect ratio.

The calculations were performed on an IBM 5100. Because of the limited storage space on this machine, we only calculated the modal parameters to four significant figures. We summed up to 1000 terms of the double series in the worst case, which corresponds to $N=45$ in the truncation

$$\sum_{m,n}\{a_{mn}: m,n\geq 0, m+n\leq N\}.$$

Table 3.1
Case IA

k	μ	ν	$\mu\nu$
1	3.084	1.269	3.915
2	2.281	1.562	3.564
3	2.132	1.699	3.622
4	2.080	1.774	3.689
5	2.055	1.820	3.740
6	2.041	1.851	3.778
7	2.033	1.873	3.808
8	2.027	1.890	3.830
9	2.023	1.902	3.848
10	2.020	1.913	3.863

Table 3.2
Case IB

k	μ	ν	$\mu\nu$
1	4.802	1.201	5.767
2	4.154	1.846	7.670
3	3.840	2.160	8.295
4	3.659	2.341	8.566
5	3.541	2.459	8.707
6	3.459	2.541	8.790
7	3.398	2.602	8.841
8	3.352	2.649	8.876
9	3.315	2.685	8.901
10	3.285	2.715	8.919

Table 3.3
Case IIA

k	μ	ν	$\mu\nu$
1	1.982	.825	1.634
2	1.197	.959	1.148
3	1.106	.979	1.083
4	1.067	.987	1.053
5	1.046	.991	1.037
6	1.034	.993	1.027
7	1.026	.995	1.021
8	1.020	.996	1.016
9	1.016	.997	1.013
10	1.014	.997	1.011

Table 3.4
Case IIB

k	μ	ν	$\mu\nu$
1	3.378	.663	2.239
2	1.371	.877	1.202
3	1.199	.934	1.120
4	1.125	.958	1.078
5	1.086	.971	1.055
6	1.062	.979	1.040
7	1.048	.984	1.031
8	1.038	.988	1.025
9	1.030	.990	1.020
10	1.025	.992	1.016

4. Independent checks on the program. As stated in the introduction we found disagreement in the literature concerning case IA. We believe the results in [8] are in error. Because of this disagreement we checked our program very carefully as follows.

1. Our program was designed to handle four cases, with a minimum of program changes required for different cases. There are analytical results [9] for two of the four cases, and our results agree with these.

2. In case IA with $k=1$, there are two other calculations published [4], [7], and we agree with these to within numerical accuracy.

3. Analysis of the von Kármán equations in case IA shows that as $k\to\infty$ the modal parameters μ and ν tend to the value 2. This trend is apparent in the data of Table 3.1. By contrast, from the data of [8] one would conjecture that for large k, μ and ν are approximately equal to $k+1$.

The main content of the section is the asymptotic analysis referred to above. Before starting this analysis, we list formulas to help the reader in comparing our results with [8]. Specifically the modal parameters are given by

$$\mu=R_2\left(1-\frac{1}{R_1}\right),\qquad \nu=R_1\left(1-\frac{1}{R_2}\right), \tag{4.1}$$

where

$$R_1=\frac{\lambda^{(2)}_{M,M+1}}{A^2_M\lambda^{(2)}_M},\qquad R_2=\frac{\lambda^{(2)}_{M,M+1}}{A^2_{M+1}\lambda^{(2)}_{M+1}}. \tag{4.2}$$

In (4.2) we have used the notation of [8]. Their subscript M coincides with our k. The five parameters in (4.2) are related to properties of the solutions of (1.2), rather than of the equation, as are the modal parameters. Thus to derive (4.1) one must refer to the four solution branches (i)–(iv) of (1.2) listed in §1.

We give the asymptotic analysis only for case IA. Thus the appropriate eigenfunctions are given by (2.3), and the inverse Laplacians in (2.6) are to be taken with Dirichlet boundary conditions.

It is readily calculated in (2.6) that $p/q=(k/(k+1))^2$, so these two parameters make no contribution to (1.3) in the limit $k\to\infty$. Our principal task is to compute a,b,c,d in (2.6). The first step here is to show that

$$(4.3)\qquad \begin{aligned}[][w_1,w_1]&=-\cos 2z_2-\cos\frac{2kz_1}{l}+O(1/k),\\ [w_1,w_2]&=-\cos\frac{z_1}{l}\cos 2z_2-\cos\frac{(2k+1)z_1}{l}+O(1/k^2),\\ [w_2,w_2]&=-\cos 2z_2-\cos\frac{(2k+2)z_1}{l}+O(1/k),\end{aligned}$$

a calculation left to the reader. The error terms admit the indicated bounds in the sup norm.

In evaluating an expression $\Delta^{-1}f$ in the limit $k\to\infty$ it is convenient to introduce a scaled variable $\bar{z}_1=z_1/l$ where $l=\sqrt{k(k+1)}$. Then all boundary problems are formulated on the fixed domain $\Omega=(0,\pi)\times(0,\pi)$. We use the inner product

$$(4.4)\qquad (u,v)=\frac{1}{\pi^2}\int_0^\pi dz_1\int_0^\pi dz_2 u(z_1,z_2)v(z_1,z_2)$$

on this domain. Here and below we omit the bar over the scaled z_1 variable. To evaluate $\Delta^{-1}f$ we must solve a boundary problem

$$(4.5)\qquad \varepsilon^2\frac{\partial^2 u}{\partial z_1^2}+\frac{\partial^2 u}{\partial z_2^2}=f\quad\text{in }\Omega,\qquad u=0\quad\text{on }\partial\Omega,$$

where $\varepsilon=1/l$. We shall use the notation L_ε for the linear operator in (4.5). In one sense the limit of L_ε as $\varepsilon\to 0$ is a singular one. Indeed, the asymptotic solution of (4.5) involves boundary layers along the faces of $\partial\Omega$ where $z_1=0,\pi$. However, in the Hilbert space sense L_ε^{-1} remains bounded as $\varepsilon\to 0$. We claim in fact that

$$(4.6)\qquad \|L_\varepsilon^{-1}\|\le 1.$$

This follows from the fact that L_ε^{-1} is a self-adjoint operator with eigenfunctions $\sin mz_1\sin nz_2$, $m,n=1,2,\cdots$, and all the eigenvalues are less than 1.

Since Ω has finite measure, the error terms in (4.3) are small in $L^2(\Omega)$ as well as in $C(\Omega)$. Because of (4.6) these error terms will not contribute to formulas (2.6) in the limit $k\to\infty$.

We want to solve (4.5) asymptotically in the limit $\varepsilon\to 0$ with right-hand sides f given by (4.3). It is conceptually simpler to consider first a slightly more general problem,

$$(4.7)\qquad \begin{aligned}&\varepsilon^2\frac{\partial^2 u_\varepsilon}{\partial z_1^2}+\frac{\partial^2 u_\varepsilon}{\partial z_2^2}=f(\eta(\varepsilon)z_1,z_1,z_2)\quad\text{in }\Omega,\\ &u_\varepsilon=0\quad\text{on }\partial\Omega,\end{aligned}$$

where $\eta(\varepsilon)=1/\varepsilon+O(1)$ as $\varepsilon\to 0$ and f is 2π periodic in its first argument. In our choice of f in (4.7) we have anticipated the fact that (4.3) depends on the fast z_1 scale as well as the slow one. The leading term in the asymptotic solution of (4.7) comes from the reduced problem for $U(\xi,z_1,z_2)$, namely

$$\begin{aligned}&\frac{\partial^2 U}{\partial\xi^2}+\frac{\partial^2 U}{\partial z_2^2}=f(\xi,z_1,z_2)\quad\text{in }\mathbb{R}\times\Omega,\\&U\text{ is }2\pi\text{ periodic in }\xi,\\&U(\xi,z_1,0)=U(\xi,z_1,\pi)=0.\end{aligned}\tag{4.8}$$

No boundary conditions are imposed in (4.8) along the faces of $\mathbb{R}\times\Omega$ where $z_1=0,\pi$.

PROPOSITION. *If $f\in C^2(\mathbb{R}\times\bar{\Omega})$, then the norm in $L^2(\Omega)$ of the difference*

$$u_\varepsilon(z_1,z_2)-U(\eta(\varepsilon)z_1,z_1,z_2)$$

is $O(\sqrt{\varepsilon})$ as $\varepsilon\to 0$.

This proposition is proved in the appendix. We remark that the $O(\sqrt{\varepsilon})$ error comes primarily from neglect of the boundary layers along the faces with $z_1=0,\pi$. However we do not carry the expansion beyond the leading term; we have retained the only term which will contribute to (2.6) in the limit.

The appropriate data to substitute into (4.8) after scaled variables are introduced in (4.3) are

$$\begin{aligned}&f_1(\xi,z_1,z_2)=-\cos 2z_2-\cos 2\xi, &&\eta_1(\varepsilon)=k,\\&f_2(\xi,z_1,z_2)=-\cos z_1\cos 2z_2-\cos 2\xi, &&\eta_2(\varepsilon)=k+\tfrac{1}{2},\\&f_3(\xi,z_1,z_2)=-\cos 2z_2-\cos 2\xi, &&\eta_3(\varepsilon)=k+1.\end{aligned}\tag{4.9}$$

Our notation here may be confusing; essentially $k,l,$ and ε are all the same parameter, being related by

$$\frac{1}{\varepsilon}=l=\sqrt{k(k+1)}\,.\tag{4.10}$$

We use the asymptotic solution of (4.7) only for the discrete series $\varepsilon_k\to 0$ of values obtained by taking k an integer in (4.10).

We now solve (4.8) with the inhomogeneities given by (4.9). Define functions $\phi(z_2)$, $\psi(z_2)$, by the two-point boundary problems

$$\begin{aligned}&\phi''(z_2)=-\cos 2z_2, &&\phi(0)=\phi(\pi)=0,\\&\psi''(z_2)-4\psi(z_2)=-1, &&\psi(0)=\psi(\pi)=0.\end{aligned}$$

The solution of (4.8) with right-hand side f_i is

$$\begin{aligned}&U_1(\xi,z_1,z_2)=\phi(z_2)+\cos 2\xi\psi(z_2),\\&U_2(\xi,z_1,z_2)=\cos z_1\phi(z_2)+\cos 2\xi\psi(z_2),\\&U_3(\xi,z_1,z_2)=\phi(z_2)+\cos 2\xi\psi(z_2).\end{aligned}$$

On application of the proposition we obtain the following solutions of (4.7):

$$\begin{aligned}&u_1(z_1,z_2)=\phi(z_2)+\cos[2kz_1]\psi(z_2),\\&u_2(z_1,z_2)=\cos z_1\phi(z_2)+\cos[(2k+1)z_1]\psi(z_2),\\&u_3(z_1,z_2)=\phi(z_2)+\cos[(2k+2)z_1]\psi(z_2).\end{aligned}\tag{4.11}$$

Now we see from (2.6) that asymptotically

$$(4.12)\qquad a=\frac{1}{2}\|u_1\|^2,\qquad d=\frac{1}{2}\|u_3\|^2,$$
$$b=c=\frac{1}{2}\left\{2\|u_2\|^2+(u_1,u_3)\right\}.$$

But it follows from (4.11) that

$$\|u_1\|^2=\|u_3\|^2=\|\phi\|^2+\frac{1}{2}\|\psi\|^2,$$
$$\|u_2\|^2=\frac{1}{2}\|\phi\|^2+\frac{1}{2}\|\psi\|^2,$$
$$(u_1,u_3)=\|\phi\|^2,$$

where $\|\phi\|$ and $\|\psi\|$ are taken in $L^2(0,\pi)$. On substituting these into (4.12) and referring to (1.3) we see that the modal parameters μ and ν assume the asymptotic value 2, as claimed.

Appendix. Our proof of the proposition of §4 follows well-known principles—see for example [2]. We will need the following consequence of the maximum principle in the proof. Let $\Omega=(0,\pi)\times(0,\pi)$, and let w solve the boundary problem

$$(\mathrm{A}.1)\qquad \varepsilon^2\frac{\partial^2 w}{\partial z_1^2}+\frac{\partial^2 w}{\partial z_2^2}=g\quad\text{in }\Omega,\qquad w=\phi\quad\text{on }\partial\Omega.$$

LEMMA. *The solution of* (A.1) *satisfies the estimate*

$$(\mathrm{A}.2)\qquad \|w\|\le\|\phi\|+c\|g\|$$

in the sup norm, where C is some constant independent of ε.

A proof of this lemma is given on p. 153 of [3]. The reader may be concerned that the ellipticity constant in (A.1) tends to zero, as $\varepsilon\to 0$, although we assert that the constant C does not blow up. An examination of [3] shows that their proof works provided that the coefficient of one pure second derivative (here $\partial^2 w/\partial z_2^2$) remains bounded away from zero. (One takes a comparison function depending only on one variable.)

In proving the proposition it is helpful to compute the next term in the asymptotic expansion of the solution of (4.7), namely the boundary layers along the faces $z_1=0,\pi$. If U is defined by (4.8), let $V_i(z_1,z_2)$, $i=1,2$, be the bounded harmonic function on the half strip $(0,\infty)\times(0,\pi)$ with boundary values as follows:

$$V_i(z_1,0)=V_i(z_1,\pi)=0,\qquad i=1,2,$$
$$V_1(0,z_2)=U(0,0,z_2),$$
$$V_2(0,z_2)=U(\eta(\varepsilon)\pi,\pi,z_2).$$

Since the only boundary values of V_i are along the edge with $z_1=0$, V_i will be exponentially decaying in z_1 as $z_1\to\infty$.

Let

$$(\mathrm{A}.3)\qquad v_\varepsilon(z_1,z_2)=U(\eta(\varepsilon)z_1,z_1,z_2)-V_1(z_1/\varepsilon,z_2)-V_2((\pi-z_1)/\varepsilon,z_2).$$

We claim that the solution u_ε of (4.7) satisfies

$$u_\varepsilon - v_\varepsilon = O(\varepsilon), \tag{A.4}$$

this estimate holding in the sup norm. To prove this we show that $u_\varepsilon - v_\varepsilon$ satisfies a boundary problem of the form (A.1) where the inhomogeneity and boundary data satisfy an $O(\varepsilon)$ estimate, and then refer to the lemma to derive (A.4). Indeed since $L_\varepsilon V_1(z_1/\varepsilon, z_2) = 0$ and similarly for V_2, we have

$$L_\varepsilon(u_\varepsilon - v_\varepsilon) = f - L_\varepsilon U. \tag{A.5}$$

However,

$$\varepsilon^2 \frac{\partial^2}{\partial z_1^2} U(\eta(\varepsilon)z_1, z_1, z_2) = \varepsilon^2 \eta^2(\varepsilon) \frac{\partial^2 U}{\partial \xi^2} + \varepsilon^2 \eta(\varepsilon) \frac{\partial^2 U}{\partial \xi \partial z_1} + \varepsilon^2 \frac{\partial^2 U}{\partial z_1^2}.$$

Since $\varepsilon^2 \eta^2(\varepsilon) = 1 + O(\varepsilon)$ by hypothesis, we have

$$L_\varepsilon U = \frac{\partial^2 U}{\partial \xi^2} + \frac{\partial^2 U}{\partial z_1^2} + O(\varepsilon).$$

Recalling the definition (4.8) of U, we see that the RHS of (A.5) is $O(\varepsilon)$ as claimed. On the edges of $\partial\Omega$ with $z_2 = 0, \pi$, the boundary data of $u_\varepsilon - v_\varepsilon$ vanishes identically. Consider the edge with $z_1 = 0$; u_ε vanishes here, U and V_1 are both nonzero but cancel one another, and V_2 is exponentially small in ε. Similarly for $z_1 = \pi$. Thus the boundary data of $u_\varepsilon - v_\varepsilon$ certainly satisfies the required $O(\varepsilon)$ estimate.

We have therefore verified the sup norm estimate (A.4). Of course the same estimate holds in the L^2 norm. Moreover, the L^2 norm of the boundary layer terms V_i is $O(\sqrt{\varepsilon})$ because of the shrinking length scale. Thus in $L^2(\Omega)$, $v_\varepsilon = U + O(\sqrt{\varepsilon})$, which completes the proof of the proposition.

REFERENCES

[1] L. Bauer, H. Keller and E. Reiss, *Multiple eigenvalues lead to secondary bifurcation*, SIAM J. Appl. Math., 17 (1975), pp. 101–122.

[2] A. Bensoussan, J. L. Lions and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[3] L. Bers, F. John and M. Schechter, *Partial Differential Equations*, Johr Wiley, New York, 1964.

[4] S. N. Chow, J. Hale and J. Mallet-Paret, *Applications of generic bifurcations* I, II, Arch. Rat. Mech. Anal., 59 (1975), pp. 159–188; 62 (1976), pp. 209–235.

[5] M. Golubitsky and D. Schaeffer, *A theory for imperfect bifurcation via singularity theory*, Comm. Pure Appl. Math., 32 (1979), pp. 21–78.

[6] ———, *Imperfect bifurcation in the presence of symmetry*, Comm. Math. Phys., 67 (1979), pp. 205–232.

[7] R. Magnus and T. Poston, *On the full unfolding of the von Kármán equations at a double eigenvalue*, Battelle Math. Report 109, Geneva, 1977.

[8] B. Matkowsky and L. Putnick, *Multiple buckled states of rectangular plates*, Internat. J. Nonlin. Mech., 9 (1973), pp. 89–103.

[8a] B. Matkowsky, L. Putnick and E. Reiss, *Secondary states of rectangular plates*, SIAM J. Appl. Math., 38 (1980), pp. 38–51.

[9] D. Schaeffer and M. Golubitsky, *Boundary conditions and mode jumping in the buckling of a rectangular plate*, Comm. Math. Phys., 69 (1979), pp. 209–236.

[10] M. Stein, *Loads and deformations of buckled rectangular plates*, NASA Tech. Rep. R-40, 1959.

[11] G. Stroebel and W. Warner, *Stability and secondary bifurcation for von Kármán plates*, J. Elasticity, 3 (1973), pp. 185–202.

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics
003

A STUDY OF ONE-DIMENSIONAL SCHRÖDINGER EQUATION WITH NONLOCAL POTENTIAL*

A. H. NASR†

Abstract. For a linear bounded operator Q, it is proved that Cauchy problem

$$L\psi=-\frac{d^2\psi}{dx^2}+(Q\psi)(x)=s^2\psi(x), \qquad \psi(0)=0, \quad \psi'(0)=1$$

has a unique solution for $|s|$ greater than a fixed value s_0 which is proved to be best possible. When Q has the form $q(x)+V$, where V is a general linear operator with sufficiently small norm, it is proved that this Cauchy problem has a unique solution for all s. In this case, Green's function for the boundary value problem is calculated.

Key words. Cauchy problem, perturbation, Green's function

Introduction. In quantum mechanics, the state of a particle moving in the interval $[a,b]$ and interacting with a field of a given potential Q is described by a real valued function ψ satisfying the Schrödinger type equation

$$-\frac{d^2\psi}{dx^2}+(Q\psi)(x)=s^2\psi(x), \tag{1}$$

subject to the boundary conditions

$$\psi(a)=0, \qquad \psi(b)=0.$$

In most of the literature [1], [2], the potential Q is assumed to take the very special form

$$(Q\psi)(x)=q(x)\psi(x),$$

where $q(x)$ is a given real valued function.

In their book [3], De Alfaro and Regge point out that when the potential depends not only on the local value at x, but also on the values of the potential at all other points y, then a natural expression for Q is the following:

$$(Q\psi)(x)=q(x)\psi(x)+\int_a^b v(x,y)\psi(y)\,dy, \tag{2}$$

where v is a given symmetric kernel. De Alfaro and Regge studied (1) when Q has the form (2).

The aim of this paper is to investigate this problem, when it is assumed only that Q is an arbitrary linear bounded operator on a suitable functional space.

Under this general hypothesis, we shall study the Cauchy problem and the two-point boundary value problem for (1), proving the asymptotic behaviour of the eigenvalues and eigenfunctions and the existence of Green's function, as well as other results.

In the sequel, the interval $[a,b]$ is replaced by the interval $[0,\pi]$ for convenience.

*Received by the editors December 23, 1981, and in revised form September 29, 1982.

†Department of Mathematics, Faculty of Science, King Abdulaziz University, Geddah, P.O.B. 9028, Kingdom of Saudi Arabia. On leave from Ain Shams University, University College for Women, Heliopolis, Cairo, Egypt.

1. Cauchy problem for general bounded potential. Assuming that Q is a linear bounded operator of $C[0,\pi]$ into itself, we prove that the Cauchy problem associated with (1) has a unique solution for sufficiently large values of $|s|$.

THEOREM 1. *Let*

$$c_0 = \max_{0<\alpha<1} \frac{1-\cos\alpha\pi}{\alpha}.$$

Then for real s satisfying $|s|>c_0\|Q\|$, the Cauchy problem

$$-\frac{d^2\psi}{dx^2}+(Q\psi)(x)=s^2\psi(x), \qquad \psi(0)=0, \quad \psi'(0)=1 \tag{3}$$

has a unique solution in $C[0,\pi]$. Moreover, the above lower bound of $|s|$ cannot be improved, in the sense that there exist equations having no solutions when $s=c_0\|Q\|$.

Proof. Writing (3) in the form

$$\frac{d^2\psi}{dx^2}+s^2\psi(x)=(Q\psi)(x), \qquad \psi(0)=0, \quad \psi'(0)=0,$$

and using Lagrange's method of variation of constants [4], problem (3) may be replaced by the following equivalent integral equation:

$$\psi(x)=\frac{\sin sx}{s}+\frac{1}{s}\int_0^x \sin s(x-t)(Q\psi)(t)\,dt. \tag{4}$$

Consider the operator A: $C[0,\pi]\to C[0,\pi]$ acting by the formula

$$(Af)(x)=\frac{\sin sx}{s}+\frac{1}{s}\int_0^x \sin s(x-t)(Qf)(t)\,dt.$$

For this operator,

$$|Af-Ag|\le\frac{1}{|s|}\int_0^x|\sin s(x-t)||Q(f-g)|\,dt$$
$$\le\frac{\|Q\|}{|s|}\|f-g\|\int_0^x|\sin s(x-t)|\,dt.$$

In the integral

$$I(x)=\int_0^x|\sin s(x-t)|\,dt=\int_0^x|\sin|s|(x-t)|\,dt$$

the substitution $|s|(x-t)=y$, $-|s|\,dt=dy$ gives

$$I(x)=\frac{1}{|s|}\int_0^{|s|x}|\sin y|\,dy.$$

Consequently,

$$\max_{0\le x\le\pi} I(x)=\frac{1}{|s|}\int_0^{|s|\pi}|\sin y|\,dy.$$

Let $|s|=n+\alpha$, where n is an integer and $0\le\alpha<1$. Then

$$\frac{1}{|s|}\int_0^{|s|\pi}|\sin y|\,dy=\frac{1}{n+\alpha}\left[\int_0^{n\pi}|\sin y|\,dy+\int_{n\pi}^{(n+\alpha)\pi}|\sin y|\,dy\right]$$

$$=\frac{1}{n+\alpha}\left[n\int_0^{\pi}\sin y\,dy+\int_0^{\alpha\pi}\sin y\,dy\right]=\frac{2n+(1-\cos\alpha\pi)}{n+\alpha}.$$

Consider the function

$$f(n,\alpha)=\frac{2n+(1-\cos\alpha\pi)}{n+\alpha}.$$

From the definition of c_0, we have

$$1-\cos\alpha\pi\le\alpha c_0\quad\forall\alpha\in[0,1).$$

Then $f(n,\alpha)\le(2n+\alpha c_0)/(n+\alpha)$. Since $c_0>2$ (see Fig. 1), the function $(2n+\alpha c_0)/(n+\alpha)$ is easily seen to be decreasing. Hence

$$\max_n\frac{2n+\alpha c_0}{n+\alpha}=c_0.$$

Collecting these results, we conclude that

$$\max_{0\le x\le\pi}I(x)\le c_0.$$

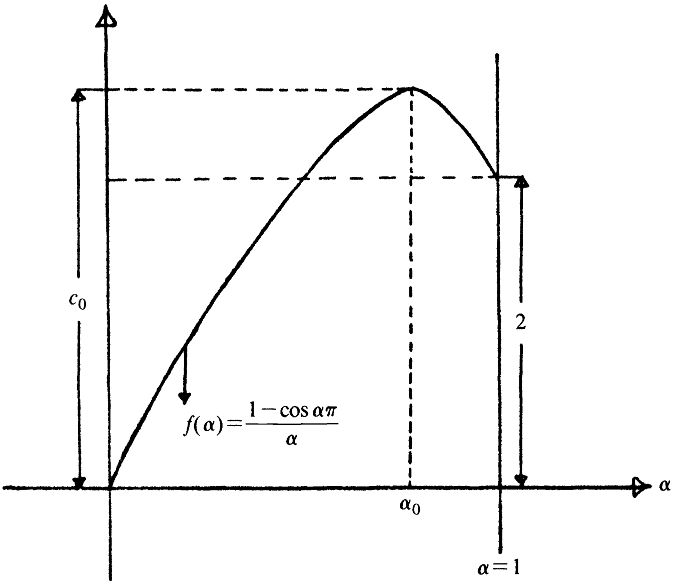


FIG. 1

Therefore,

$$\|Af-Ag\|\le\frac{c_0\|Q\|}{|s|}\|f-g\|.$$

If $c_0\|Q\|/|s|<1$ (i.e., $|s|>c_0\|Q\|$), the operator A would be a contraction and according to Banach's theorem of contraction mappings, it has a unique fixed point ψ which is the solution of (4) and, consequently, of (3).

To prove the second part of the theorem, consider the operator

$$(Q\psi)(x)=\frac{\alpha_0}{c_0}\psi(\pi),$$

where α_0 is the point at which the maximum of the function $(1-\cos\alpha\pi)/\alpha$ is attained (it is easy to see that $0<\alpha_0<1$), i.e.,

$$\frac{1-\cos\alpha_0\pi}{\alpha_0}=c_0. \tag{5}$$

It is quite clear that $\|Q\|=\alpha_0/c_0$ and hence the critical value of s equals α_0. Now consider the problem

$$\psi''+\alpha_0^2\psi=\frac{\alpha_0}{c_0}\psi(\pi), \qquad \psi(0)=0, \quad \psi'(0)=1. \tag{6}$$

Putting $\alpha_0\psi(\pi)/c_0=\theta$ and substituting into (4), we have

$$\psi(x)=\frac{\sin\alpha_0 x}{\alpha_0}+\frac{\theta}{\alpha_0}\int_0^x \sin\alpha_0(x-t)\,dt=\frac{\sin\alpha_0 x}{\alpha_0}+\frac{\theta}{\alpha_0^2}(1-\cos\alpha_0 x).$$

Applying the operator Q on both sides gives

$$\theta=\frac{\alpha_0}{c_0}\,\frac{\sin\alpha_0\pi}{\alpha_0}+\frac{\theta}{\alpha_0^2}\,\frac{\alpha_0}{c_0}(1-\cos\alpha_0\pi)$$
$$=\frac{\sin\alpha_0\pi}{c_0}+\frac{\theta}{\alpha_0 c_0}(1-\cos\alpha_0\pi).$$

Using equality (5) gives

$$\theta=\frac{\sin\alpha_0\pi}{c_0}+\theta, \quad \text{i.e.,} \quad \frac{\sin\alpha_0\pi}{c_0}=0,$$

which is a contradiction.

This proves that there does not exist any solution of problem (6).

THEOREM 2. *There exists a number* $M=1/(c_0\|Q\|(\alpha-1))$, *such that, for all real s in the region*

$$G=\{s\colon |s|\geq c_0\|Q\|\alpha=s_0,\ \alpha>1\}, \qquad \|\psi(x;s)\|\leq M.$$

Proof. For real s, it follows from (4) that

$$\|\psi\|\leq\frac{1}{|s|}+\frac{c_0\|Q\|}{|s|}\|\psi\|$$

or

$$\|\psi\|\leq\frac{1}{|s|-c_0\|Q\|}\leq\frac{1}{c_0\|Q\|(\alpha-1)}.$$

2. Cauchy problem in the case of small nonlocal perturbation. In this section we show that if Q has the special form

$$(Q\psi)(x)=q(x)\psi+(V\psi)(x),$$

where q is a continuous function and V a linear operator of $C[0,\pi]$ into itself with a sufficiently small norm, then the Cauchy problem has a unique solution for all values of s (this includes the case of a pure local potential when $V=0$).

THEOREM 3. *For* $\|V\|<1/\pi^2\cdot e^{\pi^2\|q\|}$, *the Cauchy problem*

$$-\frac{d^2\psi}{dx^2}+q(x)\psi(x)+(V\psi)(x)=s^2\psi(x), \qquad \psi(0)=0,\ \psi'(0)=1 \tag{7}$$

has a unique solution in $C[0,\pi]$ *for all values of the real parameter* s.

Proof. Write the equation in the form

$$\frac{d^2\psi}{dx^2}+s^2\psi(x)=q(x)\psi(x)+(V\psi)(x).$$

Using Lagrange's method of variation of constants [4], problem (8) is reduced to the equivalent integral equation

$$\begin{aligned}\psi(x)&=\frac{\sin sx}{s}+\int_0^x\frac{\sin s(x-t)}{s}q(t)\psi(t)\,dt+\int_0^x\frac{\sin s(x-t)}{s}(V\psi)(t)\,dt\\&=\frac{\sin sx}{s}+(A\psi)(x)+(B\psi)(x),\end{aligned} \tag{8}$$

where

$$(A\psi)(x)=\int_0^x\frac{\sin s(x-t)}{s}q(t)\psi(t)\,dt,$$

$$(B\psi)(x)=\int_0^x\frac{\sin s(x-t)}{s}(V\psi)(t)\,dt.$$

(for $s=0$, put $(\sin s(x-t))/s=(x-t)$, $(\sin sx)/s=x$).

For all s, the following estimates hold:

$$\left|\frac{\sin s(x-t)}{s}\right|\le\frac{|s||x-t|}{|s|}\le\pi. \tag{9}$$

Consequently,

$$\|B\psi\|\le\int_0^\pi\left|\frac{\sin s(x-t)}{s}\right|dt\cdot\|V\|\|\psi\|\le\pi^2\|V\|\|\psi\|,$$

i.e. $\|B\|\le\pi^2\|V\|$.

Now it will be proved that

$$\|A^r\|\le\frac{(\pi^2\|q\|)^r}{r!}, \qquad r=0,1,2,\cdots, \tag{10}$$

in fact (using (9)) we have

$$|(A\psi)(x)|\leq\pi\|q\|\|\psi\|\int_0^x dt=\pi\|q\|\|\psi\|x.$$

By mathematical induction we prove that

$$|(A^r\psi)(x)|\leq\pi^r\|q\|^r\|\psi\|\frac{x^r}{r!}. \tag{11}$$

Let (11) be true for some integer r; then

$$\begin{aligned}|A^{r+1}\psi|=|AA^r\psi|&=\int_0^x\frac{\sin s(x-t)}{s}q(t)(A^r\psi)(t)\,dt\\&\leq\int_0^x\left|\frac{\sin s(x-t)}{s}\right||q(t)||A^r\psi|\,dt\\&\leq\frac{\pi\|q\|\pi^r\|q\|^r\|\psi\|}{r!}\int_0^x t^r\,dt=(\pi\|q\|)^{r+1}\|\psi\|\frac{x^{r+1}}{(r+1)!}.\end{aligned}$$

Hence (11) is true for all r.

From (11) it follows that

$$\|A^r\psi\|\leq\frac{(\pi^2\|q\|)^r}{r!}\|\psi\|,$$

from which (10) follows.

Returning to (8) and writing it in the form

$$(E-A)\psi=\frac{\sin sx}{s}+B\psi,$$

we see that it is equivalent to the equation

$$\psi=(E-A)^{-1}\frac{\sin sx}{s}+(E-A)^{-1}B\psi=\varphi(x)+\sum_{r=0}^{\infty}A^rB\psi=\varphi(x)+L\psi, \tag{12}$$

where

$$\varphi(x)=(E-A)^{-1}\frac{\sin sx}{s}\quad\text{and}\quad L=\sum_{r=0}^{\infty}A^rB$$

(the series in (12) converges due to the estimate (10)).

Again, writing (12) in the form

$$(E-L)\psi=\varphi(x),$$

and remarking (using (11)) that

$$\|L\|\leq\sum_{r=0}^{\infty}\|A^r\|\|B\|\leq\|B\|\sum_{r=0}^{\infty}\frac{(\pi^2\|q\|)^r}{r!}=\|B\|e^{\pi^2\|q\|}\leq\pi^2\|V\|e^{\pi^2\|q\|},$$

from the conditions of the theorem it follows that $\|L\|<1$. Hence the inverse $(E-L)^{-1}$ exists and

$$\psi=(E-L)^{-1}\varphi=\sum_{r=0}^{\infty} L^r\varphi.$$

3. Spectrum of the boundary value problem. Here we study the spectrum of the operator $L=-d^2/dx^2+Q$ in the case where Q is a linear bounded self-adjoint operator in $L_2(0,\pi)$. The domain of definition consists of all twice differentiable functions $\psi\in L_2(0,\pi)$ satisfying the boundary conditions $\psi(0)=\psi(\pi)=0$. It is easy to see that L is a symmetric operator and thus its spectrum consists of real values only.

THEOREM 4. *The operator L has no continuous spectrum.*

Proof. Let $A=-d^2/dx^2$ with the same domain of definition as L. The operator A is positive definite and has an inverse A^{-1} acting by the formula

$$(A^{-1}f)(x)=\int_0^{\pi} g(x,t)f(t)\,dt,$$

where

$$g(x,t)=\frac{1}{\pi}\begin{cases} x(t-\pi), & x\leq t,\\ t(x-\pi), & x<t.\end{cases}$$

It is clear that A^{-1} is a complete continuous operator. Hence $A^{-1}Q$ is compact, i.e., Q is relatively compact with respect to A (see [5]). From the generalization of Weyl's theorem of relative completely continuous perturbations [5], we see that the continuous spectrum of $L=A+Q$ coincides with that of A, which is empty.

Now, consider the eigenvalue problem

$$-\frac{d^2\psi}{dx^2}+Q\psi=s^2\psi,\qquad \psi(0)=\psi(\pi)=0. \tag{13}$$

THEOREM 5. *The eigenvalues s^2 and the corresponding eigenfunctions $\psi(x;s)$ of problem* (13) *have the following asymptotes for real s*:

$$s=n+O\left(\frac{1}{n}\right),\qquad \psi(x;s)=\sin nx+O\left(\frac{1}{n}\right),\qquad n\to\infty,$$

where n is an integer.

Proof. Let $\psi(x;s)$ be the solution of problem (3). From §1, $\psi(x;s)$ satisfies the integral equation (4) and the first boundary condition of (13). To obtain the eigenvalues of problem (13), we assume that $\psi(x;s)$ satisfies the second boundary condition in (13), i.e.

$$\sin s\pi+\int_0^{\pi}\sin s(\pi-t)(Q\psi)(t)\,dt=\sin s\pi-R(s)=0. \tag{14}$$

The roots of this equation are the eigenvalues of problem (13). Now, we obtain the asymptotic behaviour of the roots of (14) as $s\to\infty$. Using Theorem 2, we have

$$\|Q\psi\|\leq\|Q\|\,\|\psi\|\leq\frac{\|Q\|}{|s|-2c_0\|Q\|}.$$

Hence,

$$|R(s)|\leq\frac{\|Q\|}{|s|-2c_0\|Q\|}\int_0^{\pi}|\sin s(\pi-t)|\,dt\leq\frac{c_0\|Q\|}{|s|-2c_0\|Q\|}=O\left(\frac{1}{s}\right).$$

Therefore, for sufficiently large values of $|s|$, the quantity $R(s)$ can be made enough small such that the curve $y=R(s)$ intersects the curve $y=\sin s\pi$ at points lying in the neighbourhood of the roots of the equation $\sin s\pi=0$, i.e. in the neighbourhood of the natural numbers. In fact, let $s_n=n+\delta_n$ be a root in the neighbourhood of the natural number n, then

$$\sin\delta_n\pi=O\left(\frac{1}{n+\delta_n}\right)=O\left(\frac{1}{n}\right),\quad \text{i.e.,}\quad \delta_n=O\left(\frac{1}{n}\right).$$

This means that the roots s_n of equation (14) have the following asymptotes.

$$s_n=n+O\left(\frac{1}{n}\right),\quad \text{for large } n.$$

Substituting these in (4), we get the asymptotes of the eigenfunctions (up to a multiplicative constant);

$$\psi(x;s_n)=\sin nx+O\left(\frac{1}{n}\right)\quad \text{as } s_n\to\infty.$$

Finally, we explicitly calculate the Green's function associated with problem (13), when Q has the special form

$$(Q\psi)(x)=q(x)\psi(x)+(V\psi)(x)$$

and V is a linear operator of $L_2(0,\pi)$ into itself with a sufficiently small norm.

THEOREM 6. *If zero does not belong to the spectrum of* $L_0=-d^2/dx^2+q(x)$, *then for sufficiently small (in norm) V, the problem*

$$(15)\qquad L_0\psi+V\psi=f,\quad \psi(0)=0,\quad \psi(\pi)=0,\quad f\in L_2(0,\pi)$$

has a unique solution in $L_2(0,\pi)$ *which can be expressed in the form* $\psi(x)=\int_0^\pi k(x,t)f(t)\,dt$ *(*$k(x,t)$ *is the Green's function for the problem).*

Proof. Let $g(x,t)$ be the Green's function for the operator L_0 provided with boundary conditions $\psi(0)=\psi(\pi)=0$ [4] (it exists due to the assumption that zero does not belong to the spectrum). Consequently, problem (15) is equivalent to the integral equation

$$(16)\qquad \psi(x)=\int_0^{\pi}g(x,t)f(t)\,dt+\int_0^{\pi}g(x,t)(V\psi)(t)\,dt=Gf+(GV)\psi$$

where

$$Gf=\int_0^{\pi}g(x,t)f(t)\,dt,\qquad (GV)\psi=\int_0^{\pi}g(x,t)(V\psi)(t)\,dt.$$

Equation (16) may be written in the form

$$(17)\qquad (E-GV)\psi=Gf.$$

If $\|GV\|<1$ (this can be satisfied if $\|V\|<1/\|G\|$), then the inverse $(E-GV)^{-1}$ exists (see [3]) and the solution of equation (17) is given by

$$\psi=(E-GV)^{-1}Gf=\sum_{r=0}^{\infty}(GV)^{r}Gf=\sum_{r=0}^{\infty}\hat{K}_{r}f \tag{18}$$

where

$$\hat{K}_r=(GV)^rG=\underbrace{(GV)(GV)\cdots(GV)}_{r\text{-times}}G.$$

Since G and V are symmetric operators, $\hat{K}_r$ is also symmetric, and acts by the formula

$$(\hat{K}_rf)(x)=\int_0^{\pi}g(x,x_r)V_{x_r}\int_0^{\pi}g(x_r,x_{r-1})\cdots\int_0^{\pi}g(x_2,x_1)V_{x_1}\int_0^{\pi}g(x_1,t)\,dt\,dx_1\cdots dx_r$$

($V_{x_i}\equiv V$ acting on the argument x_i).

From the linearity and continuity of V and using Fubini's theorem, the last equality may be written in the form

$$(\hat{K}_rf)(x)=\int_0^{\pi}\left[\underbrace{\int_0^{\pi}\cdots\int_0^{\pi}}_{r\text{-times}}g(x,x_r)V_{x_r}g(x_r,x_{r-1})\cdots g(x_2,x_1)V_{x_1}g(x_1,t)\,dx_1\cdots dx_r\right]f(t)\,dt,$$

i.e., the function

$$K_r(x,t)=\int_0^{\pi}\cdots\int_0^{\pi}g(x,x_r)V_{x_r}g(x_r,x_{r-1})\cdots g(x_1,x)V_{x_1}g(x_1,t)\,dx_1\cdots dx_r$$

is the symmetric kernel of the operator $\hat{K}_r$.

Put

$$K(x,t)=\sum_{r=0}^{\infty}K_r(x,t).$$

From the condition $\|GV\|<1$, it follows that the last series converges uniformly. Hence (18) takes the form

$$\psi(x)=\int_0^{\pi}K(x,t)f(t)\,dt.$$

Consequently, the function $K(x,t)$ is the Green's function for the operator L.

REFERENCES

[1] E. C. Titchmarsh, *Eigenfunction Expansions*, Vol. I, Clarendon Press, Oxford, 1946.

[2] B. M. Levitan and I. S. Sargsian, *Introduction to Spectral Theory*, Nauka, Moscow, 1970. (In Russian.)

[3] V. de Alfaro and T. Regge, *Potential Scattering*, North-Holland, Amsterdam, 1965.

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

[5] I. M. Glazman, *Direct methods of qualitative spectral analysis of singular differential operators*, Israel program of scientific translations, Jerusalem, 1965.

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics

004

A JUSTIFICATION OF THE KdV APPROXIMATION TO FIRST ORDER IN THE CASE OF N-SOLITON WATER WAVES IN A CANAL*

ROBERT L. SACHS[†]

Abstract. We consider the Euler equations for a perfect fluid in a flat-bottomed canal in the time-dependent case. A formal expansion procedure for small amplitude, long waves analogous to that of Friedrichs and Hyers for solitary waves is developed and leads to the Korteweg–de Vries equation (KdV for short) for the lowest order term. The higher order terms in the expansion satisfy the inhomogeneous version of the linearized KdV equation.

Of particular interest to us are those solutions of the KdV equation called N-solitons, which asymptotically separate into N travelling waves with distinct speeds. Using certain facts about the linearized KdV equation and some properties of the N-solitons, we prove that the next term in this expansion can be uniquely specified by certain asymptotic conditions and a symmetry requirement. This solution behaves like an N-soliton; asymptotically, it separates into N travelling waves with the same speeds and phases as those of the leading term.

AMS-MOS subject classification (1980). Primary 76B15; 35Q20; 35C05; 35G30

Key words. Euler equations, Korteweg–de Vries equation, N-solitons

1. Introduction. The Korteweg–de Vries equation (KdV for short) was originally derived in 1895 as an approximation for fluid flow in a flat-bottomed canal [14]. This nonlinear evolution equation for a function of one space variable has the rather remarkable property, discovered by Gardner, Greene, Kruskal and Miura [10], that it may be solved more or less exactly. In fact, a Hamiltonian structure can be introduced and the KdV equation may be regarded as a completely integrable Hamiltonian system. One very interesting class of solutions is the set of so-called N-solitons. These solutions behave, for large positive and negative times, like N exponentially decreasing "bumps" moving at distinct speeds. A natural question to ask is whether such "N-tuple waves" exist for the full set of Euler equations governing the fluid flow in a canal.

For $N=1$, such wave solutions, known as solitary waves, do in fact exist [3], [4], [9]. In [9], Friedrichs and Hyers gave a formal expansion procedure for the Euler equations in which a time-independent form of the KdV equation arose as the equation satisfied by the lowest order term. The higher ordrr terms of their expansion satisfied the inhomogeneous form of the linearization of the nonlinear ordinary differential equation for the leading term. With a symmetry condition added to the requirement of exponential decay, this equation could be solved uniquely. After reformulation of the problem, the convergence of this formal solution was shown by the implicit function theorem. Later Beale [4] simplified the argument by using a generalized implicit function theorem due to Zehnder [26]. In both of these approaches, the time-independent nature of the problem is relied upon from the beginning.

If we attempt to generalize these results to N-solitons for $N\geq 2$, the problem becomes unavoidably time-dependent. An essentially trivial step in both approaches to the solitary wave problem, namely inverting the linearized KdV operator, now becomes

*Received by the editors September 28, 1981, and in revised form January 13, 1983. This research was sponsored in part by the U.S. Army under contract DAAG29-80C-0041, and the National Science Foundation under grant MCS-77-01986, and in part by an American Mathematical Society Postdoctoral Research Fellowship.

[†] Mathematics Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53706.

a serious difficulty. Constructing a formal solution which behaves like an N-soliton requires solving the inhomogeneous linearized KdV equation with prescribed asymptotic behavior. We do this for the first order correction to an N-soliton by using the explicit form of the inhomogeneous term. For higher order corrections, the existence of *some* solution is guaranteed by Duhamel's principle and the solvability of the Cauchy problem for the linearized KdV equation [20]. However, in such an approach, initial values (say at $t=0$) "parameterize" the set of all solutions, and we cannot as yet single out those solutions with the desired asymptotic behavior. Indeed, such a procedure might conceivably be doomed to fail. The nonlinear interaction of the different wave modes might lead to the annihilation or creation of "soliton modes", so that an N-soliton solution for large negative times might end up as an $N \pm N'$ soliton for large positive times, with different soliton speeds and amplitudes, and a dispersive wave train might be created in the process. The result presented here shows that this possibility *can* be made to occur as at best a second order effect, if at all. Recent numerical studies by Fenton and Rienecker [7b] and Mirie and Su [17a] seem to indicate that the interaction of two solitary waves does not generate further solitary waves nor an exceedingly large oscillatory wave train. They do find a change in the wave speeds, however. In any case, analysis of the full expansion and a proof of the convergence are not known to date.

In this paper we present the following results:

(i) The time-dependent analogue of the formal expansion of Friedrichs–Hyers [9] is developed. For perturbations of a steady horizontal flow with Froude number near 1 which are of small amplitude, long wavelength and slow time variation, we consider a formal power series solution of the Euler equations. The small parameter ε is related to the Froude number. As in [9], the leading term satisfies the KdV equation and the higher order terms satisfy the inhomogeneous linearized KdV equation. However, in this case, both of these equations are time-dependent.

(ii) Using results on the solvability of the Cauchy problem for the linearized KdV equation [20] and certain facts about N-solitons, we analyze the first order term completely. In particular, we show that this term is uniquely determined by the following conditions:

(a) (symmetry) $u(x,t)=u(-x,-t)$;

(b) (asymptotic decay in moving frames) $u(ct+\xi,t)\to 0$ exponentially fast as $t\to +\infty$ for ξ fixed unless $c=c_j, j=1,\cdots,N$ where $\{c_j\}$ are the N-soliton speeds;

(c) (asymptotic shape) $\lim_{t\to\infty} u(c_jt+\xi,t)$ is an exponentially decreasing function of ξ.

This is the sense in which we use the term "justification" in the title of this paper. The first order correction to the KdV N-soliton, as chosen above, does not alter any of the essential features of the solution. After a long time, the water wave decomposes into N travelling waves with distinct speeds, each of which is exponentially decreasing in space when viewed from the appropriate moving frame of reference. The speeds are unchanged but the wave shapes are altered slightly.

Our result is most directly analogous to the expansion of Friedrichs and Hyers [9] for the solitary wave. Subsequent expansions of solitary waves have used the wave amplitude at the crest as small parameter rather than some measure of the supercriticality of the flow at ∞ (i.e. the Froude number). These papers, such as [7a] and the references therein, obtain uniform asymptotic expansions at the expense of expanding the Froude number in powers of the small parameter. In light of the successful proof of convergence of the Friedrichs–Hyers expansion [4], [9], the nonuniformity of this expansion is not as undesirable as it might seem. The reason for this may be seen by contrasting "soliton perturbations" with the classical Poincaré–Lindstedt perturbation theory for periodic solutions. Translation of a single soliton leads to a term of the form $\xi\ \mathrm{sech}^2\beta\xi\ \tanh\beta\xi$, which might be termed "secular". But this nomenclature may be misleading since in such a term the linear growth is overwhelmed by exponential decay, unlike the classical secular terms $t\cos\omega t$ of Poincaré–Lindstedt type. As mentioned in §4 below, terms resulting from variations in the soliton speeds are linearly growing modes in a particular moving frame and may be regarded as truly "secular".

The paper is organized as follows. Section 2 contains the time-dependent analogue of the formal expansion of Friedrichs and Hyers [9]. The basic facts concerning the Cauchy problem for the linearized KdV equation are presented in §3. Explicit solvability for this problem is related to the so-called inverse scattering method for solving the KdV equation [10], [20]. Using certain facts about N-solitons, which we present in the Appendix, and the particular terms arising in the expansion of §2, the first order correction to the N-soliton is analyzed in §4.

2. The Euler equations, the KdV limit and a formal expansion. In dimensionless variables, the Euler equations for a perfect fluid in a two-dimensional, flat-bottomed domain D, with a free boundary $y=\Gamma(t,x)$ as upper surface, subject only to gravitational acceleration g are (cf. Stoker [22]):

$$(2.1\ )\quad \begin{aligned} &\text{(i)} && \phi_{xx}+\phi_{yy}=0 \text{ in } D\equiv\{(x,y):0<y<\Gamma(t,x)\},\\ &\text{(ii)} && \phi_y=0 \text{ along } y=0,\\ &\text{(iii)} && \phi_t+\tfrac{1}{2}\left(\phi_x^2+\phi_y^2\right)+\gamma y=\text{constant along } y=\Gamma(t,x),\\ &\text{(iv)} && \Gamma_t+\phi_x\cdot\Gamma_x-\phi_y=0 \text{ along } y=\Gamma(t,x)\end{aligned}$$

where $\phi=\phi(x,y,t)$ is the velocity potential and $\gamma\equiv gh/U^2$ where h is the length scaling and U is the velocity scaling. $\gamma^{-1/2}$ is called the Froude number or reduced depth and is a parameter of the problem. The linear theory of water waves [22] predicts $\gamma=1$, while the existence of solitary waves occurs for $\gamma<1$ but sufficiently close to 1. From now on, we assume

$$(2.2)\qquad 0<1-\gamma\ll\gamma<1 \text{ and in fact, we define a small parameter } \varepsilon \text{ by } \gamma=e^{-3\varepsilon}.$$

In this section, we will consider flows which are very nearly the trivial flow of constant horizontal speed 1 given by the solution $\phi=x$, $\Gamma=1$, $\gamma=1$ of (2.1). Introducing auxiliary variables ξ', η' which vary over a fixed horizontal strip $0<\eta'<1$, we may eliminate the unknown free surface at the expense of defining x, y as functions of ξ', η', t. In steady flow problems, t does not appear and $\xi'+i\eta'$ is usually the complex potential function, but for time-dependent problems, we express both the potential function ϕ and the physical coordinates x, y in terms of ξ', η' and t. Provided the mapping $(\xi',\eta')\mapsto(x,y)$ is invertible for every t, solving the problem in the ξ', η' plane is equivalent to solving the original system in the x, y plane. In the neighborhood of the trivial horizontal flow, this mapping is roughly the identity map; hence it will be invertible.

After expressing the problem in these new independent variables, we will introduce a new dependent complex variable, $\lambda'-i\theta'$, defined as the logarithm of the complex velocity W ($W\equiv\phi_x-i\phi_y$). By differentiating with respect to ξ' along $\eta'=1$, ϕ is eliminated and a new system of equations for x, y, λ', θ' is obtained. Defining a small parameter $a\equiv\varepsilon^{1/2}$, we rescale the independent variables ξ', η', t and the small dependent variables λ', θ', $x'\equiv x-\xi'$, $y'\equiv y-\eta'$. The system (2.1) in the rescaled variables, ξ,

η, τ and $\hat{x}$, $\hat{y}$, λ, θ respectively, becomes:

$$\text{(2.3) (i)}\qquad \left(\varepsilon^{1/2}\frac{\partial}{\partial\xi}+i\frac{\partial}{\partial\eta}\right)(\hat{x}+i\varepsilon^{1/2}\hat{y})=\left(\varepsilon^{1/2}\frac{\partial}{\partial\xi}+i\frac{\partial}{\partial\eta}\right)(\lambda-i\varepsilon^{1/2}\theta)$$

$$=0 \quad \text{in } 0<\eta<1$$

(scaled Cauchy–Riemann equation),

$$\text{(ii)}\qquad \theta=0, \qquad \hat{y}=0 \quad \text{along } \eta=0,$$

$$\text{(iii)}\qquad e^{\varepsilon\lambda}\left\{\cos(\varepsilon^{3/2}\theta)\left[\varepsilon\lambda_\tau+\varepsilon^2(\lambda_\tau\hat{x}_\xi-\lambda_\xi\hat{x}_\tau)+\varepsilon^3(\theta_\tau\hat{y}_\xi-\theta_\xi\hat{y}_\tau)\right]\right.$$
$$\left.+\frac{\sin(\varepsilon^{3/2}\theta)}{\varepsilon^{3/2}}\left[-\varepsilon^3\theta_\tau+\varepsilon^4(\lambda_\tau\hat{y}_\xi-\lambda_\xi\hat{y}_\tau+\theta_\xi\hat{x}_\tau-\theta_\tau\hat{x}_\xi)\right]\right\}$$
$$+e^{2\varepsilon\lambda}\lambda_\xi+e^{-3\varepsilon}\hat{y}=0 \quad \text{along } \eta=1 \qquad \text{(Bernoulli's law)},$$

$$\text{(iv)}\qquad e^{\varepsilon\lambda}\hat{y}_\xi\cos(\varepsilon^{3/2}\theta)-e^{\varepsilon\lambda}(1+\varepsilon\hat{x}_\xi)\cdot\frac{\sin(\varepsilon^{3/2}\theta)}{\varepsilon^{3/2}}$$
$$+\varepsilon\hat{y}_\tau+\varepsilon^2(\hat{x}_\xi\hat{y}_\tau-\hat{x}_\tau\hat{y}_\xi)=0 \quad \text{along } \eta=1$$

(free boundary streamline condition).

We proceed to derive system (2.3) below and then discuss a formal expansion procedure using power series in ε.

2.1. Reformulation via a conformal mapping. Introduce complex variables into (2.1) as follows: $z\equiv x+iy$, $F(z,t)\equiv\phi(z,t)+i\psi(z,t)$. The complex velocity $W(z,t)\equiv F_z(z,t)=\phi_x-i\phi_y$, so $\operatorname{Re}W=\phi_x$, the horizontal velocity, and $-\operatorname{Im}W=\phi_y$, the vertical velocity. Since we are considering flows near the trivial one, for which the free surface is $\Gamma(t,x)\equiv 1$, we assume that there exists a complex variable $\zeta'=\xi'+i\eta'$ defined on the fixed strip $\{(\xi',\eta')\,|\,0<\eta'<1\}$ and a conformal mapping $z=z(\zeta',t)$ such that the boundaries of the flow domain, $y=0$ and $y=\Gamma(t,x)$ (where $y=\operatorname{Im}z$), correspond to the boundaries $\eta'=0$, $\eta'=1$ respectively.

Given the existence of such a mapping, we define new dependent variables implicitly:

$$\text{(2.4)}\qquad \begin{aligned} &f(\zeta',t)\equiv F(z(\zeta',t),t),\\ &w(\zeta',t)\equiv W(z(\zeta',t),t) \quad \text{so that}\\ &w(\zeta',t)=\frac{f_{\zeta'}}{z_{\zeta'}}. \end{aligned}$$

Thus ζ' derivatives of f are expressible in terms of w and z. Substitution in (2.1) and differentiation with respect to ξ' along $\eta'=1$ yields a system with w, z as dependent variables, namely:

(2.5) (i) $w(\zeta',t)$ and $z(\zeta',t)$ are holomorphic functions of ζ' in $0<\operatorname{Im}\zeta'<1$,

(ii) $\operatorname{Im}w=0$, $\operatorname{Im}z=0$ along $\eta'=0$,

(iii) $\operatorname{Re}(w_t z_{\xi'}-w_{\xi'}z_t)+\frac{1}{2}(|w|^2)_{\xi'}+\gamma\operatorname{Im}(z_{\xi'})=0$ along $\eta'=1$,

(iv) $\operatorname{Im}(z_t/z_{\xi'})+\operatorname{Im}(w/\bar{z}_{\xi'})=0$ along $\eta'=1$.

(This last condition comes from the relations $\Gamma_x=y_{\xi'}/x_{\xi'}$, $\Gamma_t=y_t-x_t y_{\xi'}/x_{\xi'}$ on $\eta=1$.)

It is convenient to replace w by $\lambda' - i\theta'$, defined by the relation

$$w = e^{\lambda' - i\theta'}. \tag{2.6}$$

This substitution was introduced by Levi-Civita [16] in the periodic case of infinite depth; it has the virtues of simplifying the $|w|$ differentiation, ensuring $w \neq 0$ for any solution and making $\lambda' - i\theta' \equiv 0$ the trivial flow. Upon substitution, we obtain

(2.7) (i) $z(\zeta', t)$, $(\lambda' - i\theta')(\zeta', t)$ are holomorphic in ζ' for $0 < \operatorname{Im} \zeta' < 1$,

(ii) $\theta' = 0, y = 0$ along $\eta' = 0$,

(iii) $\operatorname{Re}\left(e^{\lambda' - i\theta'}\left[(\lambda'_t - i\theta'_t) z_{\xi'} - (\lambda'_{\xi'} - i\theta'_{\xi'}) z_t\right]\right) + e^{2\lambda'} \lambda'_{\xi'} + \gamma \operatorname{Im} z_{\xi'} = 0$ along $\eta' = 1$,

(iv) $\operatorname{Im}(z_t / z_{\xi'}) + \operatorname{Im}\left(e^{\lambda' - i\theta'} / \bar{z}_{\xi'}\right) = 0$ along $\eta' = 1$.

We note that this is a system of equations for two holomorphic functions on a strip which are real for ζ' real (the bottom) and satisfy a pair of coupled nonlinear time-dependent boundary conditions along the top of the strip. Kano and Nishida [11] used essentially the system (2.5), along with some basic facts about harmonically conjugate functions on a strip, to obtain a nonlinear expression for the t-derivatives of x and ϕ along $\eta' = 1$, for which a solution will exist to the Cauchy problem for small times (see also [19]).

We will now consider a particular limiting case of system (2.7) corresponding to long wavelength, small amplitude waves of slow time variation and will obtain the Korteweg–de Vries equation in the limit. We assume that, as $|\zeta'| \to \infty$, $\lambda' - i\theta' \to 0$ and $z_{\xi'} \to 1$, so the perturbations from the steady flow vanish asymptotically. The limiting case is given by the following rescaling:

Define new independent variables

$$\xi = a\xi', \quad \eta = \eta', \quad \tau = a^3 t, \quad \text{where } a^2 = \varepsilon, \tag{2.8}$$

and new dependent variables $\hat{x}, \hat{y}, \theta, \lambda$ by

$$\begin{aligned} a\hat{x}(\xi, \eta, \tau, \varepsilon) &= \operatorname{Re}(z(\zeta', t)) - \xi', \\ a^2 \hat{y}(\xi, \eta, \tau, \varepsilon) &= \operatorname{Im}(z(\zeta', t)) - \eta', \\ a^2 \lambda(\xi, \eta, \tau, \varepsilon) &= \lambda'(\zeta', t), \\ a^3 \theta(\xi, \eta, \tau, \varepsilon) &= \theta'(\zeta', t). \end{aligned} \tag{2.9}$$

Substituting these variables into the system (2.7) gives system (2.3) above, which we have therefore derived.

2.2. A formal solution procedure. If we consider the system (2.3) and assume expansions for $\lambda, \theta, \hat{x}, \hat{y}$ of the form

$$\begin{aligned} \lambda(\xi, \eta, \tau, \varepsilon) &= \sum_{j=0}^{\infty} \lambda^{(j)}(\xi, \eta, \tau) \varepsilon^j, \\ \theta(\xi, \eta, \tau, \varepsilon) &= \sum_{j=0}^{\infty} \theta^{(j)}(\xi, \eta, \tau) \varepsilon^j, \\ \hat{x}(\xi, \eta, \tau, \varepsilon) &= \sum_{j=0}^{\infty} \hat{x}^{(j)}(\xi, \eta, \tau) \varepsilon^j, \\ \hat{y}(\xi, \eta, \tau, \varepsilon) &= \sum_{j=0}^{\infty} \hat{y}^{(j)}(\xi, \eta, \tau) \varepsilon^j, \end{aligned} \tag{2.10}$$

then (2.3)(i) implies the system

$$\lambda_{\xi}^{(j)}+\theta_{\eta}^{(j)}=0, \qquad -\lambda_{\eta}^{(j)}+\theta_{\xi}^{(j-1)}=0,$$
$$\hat{x}_{\xi}^{(j)}-\hat{y}_{\eta}^{(j)}=0, \qquad \hat{x}_{\eta}^{(j)}+\hat{y}_{\xi}^{(j-1)}=0. \tag{2.11}$$

In particular, $\lambda_{\eta}^{(0)}=0$; $\hat{x}_{\eta}^{(0)}=0$. From the boundary condition (2.3)(ii) on θ, $\hat{y}$ at $\eta=0$, this implies

$$\lambda^{(0)}=\lambda^{(0)}(\xi,\tau), \qquad \hat{x}^{(0)}=\hat{x}^{(0)}(\xi,\tau),$$
$$\theta^{(0)}=-\lambda_{\xi}^{(0)}(\xi,\tau)\eta, \qquad \hat{y}^{(0)}=\hat{x}_{\xi}^{(0)}(\xi,\tau)\eta, \tag{2.12}$$

so

$$\lambda^{(1)}(\xi,\eta,\tau)=-\tfrac{1}{2}\lambda_{\xi\xi}^{(0)}(\xi,\tau)\eta^2+Q^{(1)}(\xi,\tau),$$
$$\hat{x}^{(1)}(\xi,\eta,\tau)=-\tfrac{1}{2}\hat{x}_{\xi\xi}^{(0)}(\xi,\tau)\eta^2+P^{(1)}(\xi,\tau),$$
$$\theta^{(1)}(\xi,\eta,\tau)=\tfrac{1}{6}\lambda_{\xi\xi\xi}^{(0)}(\xi,\tau)\eta^3-Q_{\xi}^{(1)}(\xi,\tau)\eta,$$
$$\hat{y}^{(1)}(\xi,\eta,\tau)=-\tfrac{1}{6}\hat{x}_{\xi\xi\xi}^{(0)}(\xi,\tau)\eta^3+P_{\xi}^{(1)}(\xi,\tau)\eta.$$

Proceeding inductively, with $Q^{(0)}\equiv\lambda^{(0)}$, $P^{(0)}\equiv\hat{x}^{(0)}$, we have

$$\lambda^{(k)}(\xi,\eta,\tau)=\sum_{j=0}^{k}\frac{(-1)^j\eta^{2j}}{(2j)!}\left(\frac{\partial}{\partial\xi}\right)^{2j}Q^{(k-j)}(\xi,\tau),$$
$$\hat{x}^{(k)}(\xi,\eta,\tau)=\sum_{j=0}^{k}\frac{(-1)^j\eta^{2j}}{(2j)!}\left(\frac{\partial}{\partial\xi}\right)^{2j}P^{(k-j)}(\xi,\tau), \tag{2.13}$$

with similar expressions for $\theta^{(j)}$, $\hat{y}^{(j)}$ involving odd powers of η and $\partial/\partial\xi$.

Thus if $\lambda^{(j)}$, $j<k$, are known, and similarly for $\hat{x}^{(j)}$, there are two unknown functions $P^{(k)}(\xi,\tau)$, $Q^{(k)}(\xi,\tau)$ which arise in terms of order ε^k and higher. Substituting these series (2.13) into the two boundary conditions at $\eta=1$, namely

$$\lambda_{\xi}e^{2\varepsilon\lambda}+e^{-3\varepsilon}\hat{y}_{\xi}+e^{\varepsilon\lambda}\Big\{\cos(\varepsilon^{3/2}\theta)\left[\varepsilon\lambda_{\tau}+\varepsilon^2(\lambda_{\tau}\hat{x}_{\xi}-\lambda_{\xi}\hat{x}_{\tau})+\varepsilon^3(\theta_{\tau}\hat{y}_{\xi}-\theta_{\xi}\hat{y}_{\tau})\right]$$
$$+\frac{\sin(\varepsilon^{3/2}\theta)}{\varepsilon^{3/2}}\left[-\varepsilon^3\theta_{\tau}+\varepsilon^4(\lambda_{\tau}\hat{y}_{\xi}-\lambda_{\xi}\hat{y}_{\tau}-\theta_{\tau}\hat{x}_{\xi}+\theta_{\xi}\hat{x}_{\tau})\right]\Big\}=0$$
$$\text{along } \eta=1$$

and

$$\hat{y}_{\xi}\cos(\varepsilon^{3/2}\theta)-(1+\varepsilon\hat{x}_{\xi})\frac{\sin(\varepsilon^{3/2}\theta)}{\varepsilon^{3/2}}+e^{-\varepsilon\lambda}\{\varepsilon\hat{y}_{\tau}+\varepsilon^2(\hat{x}_{\xi}\hat{y}_{\tau}-\hat{x}_{\tau}\hat{y}_{\xi})\}=0 \quad \text{along } \eta=1,$$

we obtain, setting $\varepsilon=0$,

$$Q_{\xi}^{(0)}+P_{\xi\xi}^{(0)}=0 \tag{2.14}$$

from each equation.

Terms of order ε in the boundary conditions set $\eta=1$ are

$$\lambda_{\xi}^{(1)}+2\lambda^{(0)}\lambda_{\xi}^{(0)}+\hat{y}_{\xi}^{(1)}-3\hat{y}_{\xi}^{(0)}+\lambda_{\tau}^{(0)}=0,$$
$$\hat{y}_{\xi}^{(1)}-\theta^{(1)}-\hat{x}_{\xi}^{(0)}\theta^{(0)}+\hat{y}_{\tau}^{(0)}=0, \tag{2.15}$$

which imply

$$-\tfrac{1}{2}Q^{(0)}_{\xi\xi\xi}+Q^{(1)}_{\xi}+2Q^{(0)}Q^{(0)}_{\xi}-\tfrac{1}{6}P^{(0)}_{\xi\xi\xi\xi}+P^{(1)}_{\xi\xi}-3P^{(0)}_{\xi\xi}+Q^{(0)}_{\tau}=0,$$
$$P^{(1)}_{\xi\xi}-\tfrac{1}{6}P^{(0)}_{\xi\xi\xi\xi}-\tfrac{1}{6}Q^{(0)}_{\xi\xi\xi}+Q^{(1)}_{\xi}+P^{(0)}_{\xi}Q^{(0)}_{\xi}+P^{(0)}_{\xi\tau}=0$$

so that $P^{(1)}_{\xi\xi}+Q^{(1)}_{\xi}$ is known in terms of $Q^{(0)}$, $P^{(0)}$ and drops out upon subtracting these two equations. If we integrate (2.14), we have $Q^{(0)}+P^{(0)}_{\xi}=0$ (by boundary conditions $Q^{(0)}$, $P^{(0)}_{\xi}\to 0$ as $|\xi|\to\infty$), so that the two boundary conditions for order ε imply:

$$2Q^{(0)}_{\tau}+3Q^{(0)}Q^{(0)}_{\xi}+3Q^{(0)}_{\xi}-\tfrac{1}{3}Q^{(0)}_{\xi\xi\xi}=0, \tag{2.16}$$

which is a form of the KdV equation.

Remark. The formal expansion of Friedrichs and Hyers [9] for the solitary wave has the time-independent form of (2.16) as the equation for the leading term.

If we pick any $Q^{(0)}$ satisfying (2.16), we obtain $P^{(0)}$ by integration, since $P^{(0)}_{\xi}=-Q^{(0)}$, if we add the normalization $P^{(0)}(0)=0$.

The order ε^k terms in the boundary conditions yield two equations of the form:

$$\begin{aligned}&\lambda^{(k)}_{\xi}+2\lambda^{(0)}\lambda^{(k-1)}_{\xi}+2\lambda^{(k-1)}\lambda^{(0)}_{\xi}+\hat{y}^{(k)}_{\xi}-3\hat{y}^{(k-1)}_{\xi}+\lambda^{(k-1)}_{\tau}=R_{k-2},\\ &\hat{y}^{(k)}_{\xi}-\theta^{(k)}-\hat{x}^{(0)}_{\xi}\theta^{(k-1)}-\hat{x}^{(k-1)}_{\xi}\theta^{(0)}+\hat{y}^{(k-1)}_{\tau}=S_{k-2}\end{aligned} \tag{2.17}$$

where R_l, S_l (and later $\tilde{R}_l$, $\tilde{S}_l$) depend only on $P^{(j)}$, $Q^{(j)}$ for $j\le l$.

Thus $P^{(k)}_{\xi}+Q^{(k)}_{\xi\xi}=\tilde{R}_{k-1}$ and again, by subtraction,

$$2Q^{(k-1)}_{\tau}+3\left(Q^{(0)}Q^{(k-1)}\right)_{\xi}+3Q^{(k-1)}_{\xi}-\tfrac{1}{3}Q^{(k-1)}_{\xi\xi\xi}=\tilde{S}_{k-2} \tag{2.18}$$

where we used $P^{(k-1)}_{\xi}+Q^{(k-1)}_{\xi\xi}=\tilde{R}_{k-2}$ to eliminate $\hat{x}^{(k-1)}$ in (2.18). Inductively, we find a formal solution using the power series (2.13) by solving (2.18) for $Q^{(k-1)}$ and then obtaining $P^{(k-1)}$ by the relation $P^{(k-1)}_{\xi}+Q^{(k-1)}_{\xi\xi}=\tilde{R}_{k-2}$.

In particular, the terms of order ε^2 in the boundary conditions are, in the notation of (2.10) above,

$$\begin{aligned}\text{(i)}\quad &\lambda^{(2)}_{\xi}+y^{(2)}_{\xi}+2\lambda^{(0)}\lambda^{(1)}_{\xi}+2\lambda^{(0)}_{\xi}\lambda^{(1)}-3y^{(1)}_{\xi}+\lambda^{(1)}_{\tau}\\ &\quad+\tfrac{9}{2}y^{(0)}_{\xi}+\lambda^{(0)}\lambda^{(0)}_{\tau}+2(\lambda^{(0)})^2\lambda^{(0)}_{\xi}+\lambda^{(0)}_{\tau}x^{(0)}_{\xi}-\lambda^{(0)}_{\xi}x^{(0)}_{\tau}=0,\\ \text{(ii)}\quad &y^{(2)}_{\xi}-\theta^{(2)}-\theta^{(1)}\left(\lambda^{(0)}+x^{(0)}_{\xi}\right)+\lambda^{(1)}y^{(0)}_{\xi}+\lambda^{(0)}y^{(1)}_{\xi}\\ &\quad+\tfrac{1}{2}(\lambda^{(0)})^2y^{(0)}_{\xi}-\theta^{(0)}\left(\lambda^{(1)}+\left(\frac{\lambda^{(0)^2}}{2}+x^{(1)}_{\xi}+x^{(0)}_{\xi}\lambda^{(0)}\right)\right)\\ &\quad\quad+y^{(1)}_{\tau}+x^{(0)}_{\xi}y^{(0)}_{\tau}-x^{(0)}_{\tau}y^{(0)}_{\xi}=0.\end{aligned} \tag{2.19}$$

Upon substituting (2.13) and using the equations derived for the lower order terms, subtraction and some algebra yield the linearized KdV equation for $Q^{(1)}$, namely,

(2.20)

$$\begin{aligned}&-2Q^{(1)}_{\tau}+\frac{1}{3}Q^{(1)}_{\xi\xi\xi}-3Q^{(0)}_{\xi}Q^{(1)}-3Q^{(1)}_{\xi}-3Q^{(0)}Q^{(1)}_{\xi}\\ &\quad=-\frac{19}{180}Q^{(0)}_{\xi\xi\xi\xi\xi}+\frac{5}{2}Q^{(0)}_{\xi\xi\xi}-\frac{9}{4}Q^{(0)}_{\xi}+\frac{19}{6}Q^{(0)}_{\xi}Q^{(0)}_{\xi\xi}+\frac{1}{12}Q^{(0)}Q^{(0)}_{\xi\xi\xi}+\frac{5}{2}\left(Q^{(0)}\right)^2Q^{(0)}_{\xi}.\end{aligned}$$

The nontrivial step is solving (2.18), the linearized KdV equation with inhomogeneous terms. For water wave solutions of the system (2.3) which behave like N-tuple solitary waves, we would choose for $\lambda^{(0)}$ an N-soliton solution of the KdV equation and then solve (2.18) with this $\lambda^{(0)}$, seeking solutions with the appropriate asymptotic behavior.

For the remainder of this paper, we consider the linearized KdV equation. As we have seen, it arises in the study of small amplitude, long wavelength, slow time variations of a steady flow of a perfect fluid over a flat bottom with Froude number near 1. If we seek solutions describing a "nonlinear superposition" of N solitary waves of distinct speeds, the first approximant will be an N-soliton solution of the KdV equation, and the higher order corrections will satisfy the inhomogeneous form of the linearized KdV equation (linearized about the N-soliton).

We shall consider the Cauchy problem for the linearized KdV equation. By Duhamel's principle, this amounts to solving the inhomogeneous equation. By the change of variables,

$$X-9T=\xi, \quad -6T=\tau, \quad q(X,T)\equiv\tfrac{3}{2}\lambda(\xi,\tau) \tag{2.21}$$

we obtain the usual form of the KdV equation:

$$q_T+q_{XXX}-6qq_X=0. \tag{2.22}$$

We note that the τ-independent solution of the KdV equation is a function of $X-9T$; this gives the one soliton with speed 9, which explicitly is $\lambda^{(0)}=-3\operatorname{sech}^2(\frac{3}{2}\xi)$, the first order term in the expansion of Friedrichs and Hyers.

In the remaining sections, we shall use the letters x, y, t, u, v etc. for meanings other than those of the above section. Since these different meanings occur in separate places, this should cause no confusion for the reader.

3. Some results on the Cauchy problem for the linearized KdV equation. In this section, we summarize the results of [20] regarding the Cauchy problem:

$$u_t+u_{xxx}-6(qu)_x=0, \qquad u(x,0)=\phi(x), \tag{3.1}$$

where $q(x,t)$ satisfies the KdV equation

$$q_t+q_{xxx}-6qq_x=0. \tag{3.2}$$

By Duhamel's principle, the inhomogeneous form of (3.1) is solvable if the Cauchy problem is.

In [20], an explicit formula for the solution of problem (3.1) is given, using certain functions arising from the Schrödinger equation,

$$-f''(x,k,t)+q(x,t)f(x,k,t)=k^2f(x,k,t), \tag{3.3}$$

where the potential $q(x,t)$ satisfies:

$$\int_{-\infty}^{\infty}(1+x^2)|q(x,t)|dx<\infty \quad \text{for every } t \text{ fixed}. \tag{3.4}$$

The fundamental discovery of Gardner, Greene, Kruskal and Miura [10], later formulated abstractly by Lax [15], is that if $q(x,t)$ evolves according to the KdV equation (3.2), the spectrum of the Schrödinger equation (3.3) is fixed and the associated scattering data evolves in a simple way. We shall use this information below, but first introduce some notation and basic facts about the scattering theory for (3.3). This information (and much more) may be found in [7].

Let $f_{\pm}(x,k,t)$ denote the Jost solutions of (3.3), i.e.

$$f_+(x,k,t)\sim e^{ikx+4ik^3t} \quad \text{as } x\to+\infty,\ t \text{ fixed},$$
$$f_-(x,k,t)\sim e^{-ikx-4ik^3t} \quad \text{as } x\to-\infty,\ t \text{ fixed},$$

and both satisfy (3.3). We define the transmission coefficient, $T(k,t)$, in terms of the Wronskian of f_+, f_- as follows:

$$\text{(3.5)}\qquad \begin{aligned}\frac{1}{T(k,t)}&=\frac{1}{2ik}[f_+(x,k,t),f_-(x,k,t)]\\ &=\frac{f_+'(x,k,t)f_-(x,k,t)-f_-'(x,k,t)f_+(x,k,t)}{2ik}.\end{aligned}$$

(We shall always use the notation: $'\equiv\partial/\partial x$, $^{\cdot}\equiv\partial/\partial k$.) It is not hard to show that $T(k,t)=T(k)$ is independent of t and that under the normalization of f_+, f_-, $T(k)$ is meromorphic in the upper half-plane $\operatorname{Im} k>0$ with poles at $k=i\beta_j$, $j=1,\cdots,N$, where each energy $-\beta_j^2$ is a bound state energy for (3.3). N is finite by a classical estimate involving $\int_{-\infty}^{\infty}(1+|x|)|q(x)|\,dx<\infty$. $T(k)$ is also continuous and nonzero for real $k\neq0$. For notational ease, we introduce for $j=1,\cdots,N$ the following pair of functions:

$$\text{(3.6)}\qquad F_j(x,t)=f_+^2(x,i\beta_j,t),\qquad G_j(x,t)=c_jf_+(x,i\beta_j,t)^{\cdot}g_j(x,t)$$

where

$$g_j(x,t)=\frac{1}{i}\frac{d}{dk}\left[f_-(x,k,t)-\frac{f_-(x,i\beta_j,t)}{f_+(x,i\beta_j,t)}f_+(x,k,t)\right]\Bigg|_{k=i\beta_j}$$

and c_j is chosen so that $\int_{-\infty}^{\infty}F_j'(x,0)G_j(x,0)\,dx=1$ for $j=1,\cdots,N$.

The principal result of [20] is the following:

THEOREM 3.1. *Suppose $q(x,t)$ satisfies* (3.4). *If $\phi(x)$ is continuous and integrable, the solution of* (3.1) (*in the sense of distributions*) *is given by*:

$$\text{(3.7)}\quad \begin{aligned}u(x,t)=\int_{-\infty}^{\infty}\frac{dk}{4\pi ik}T^2(k)\bigg\{\int_{-\infty}^{\infty}\frac{\partial}{\partial x}&[f_+^2(x,k,t)f_-^2(y,k,0)\\ &-f_-^2(x,k,t)f_+^2(y,k,0)]\phi(y)\,dy\bigg\}\\ +\sum_{j=1}^{N}\int_{-\infty}^{\infty}&[F_j'(x,t)G_j(y,0)-G_j'(x,t)F_j(y,0)]\phi(y)\,dy.\end{aligned}$$

For a proof of this theorem, see [20].

Remark. It is known ([10, Thm. 3.6] or [20]) that the functions $(f_{\pm}^2)'(x,k,t)$, $F_j'(x,t)$, $G_j'(x,t)$ all satisfy the linearized KdV equation (3.1). The formula for $u(x,t)$ resembles the Fourier decomposition of $\phi(x)$, where the derivatives of the squared eigenfunctions replace the usual exponentials and the presence of a nonzero potential $q(x,t)$ can lead to the discrete terms $F_j'(x,t)$, $G_j'(x,t)$. In fact, when $q(x,t)\equiv0$, (3.7) reduces to the usual Fourier transform solution of the Cauchy problem

$$\text{(3.8)}\qquad v_t+v_{xxx}=0,\qquad v(x,0)=\phi(x),$$

namely

$$v(x,t)\equiv\frac{1}{\pi}\int_{-\infty}^{\infty}dk\cdot e^{2ikx+8ik^3t}\left\{\int_{-\infty}^{\infty}e^{-2iky}\phi(y)\,dy\right\} \tag{3.9}$$

since for $q\equiv 0$, $T(k)\equiv 1$ and $f_\pm(x,k,t)\equiv e^{\pm(ikx+4ik^3t)}$.

Noting that the solution $u(x,t)$ given by (3.7) consists of two pieces—a discrete sum and an integral—we analyze them separately. The sum corresponds to variations in the soliton part of the function $q(x,t)$ and decomposes into travelling waves with positive velocities as t becomes large. For the water wave problem of §2, these terms are of considerable interest. The k-space integral part of (3.7) forms a dispersive wave train and will be seen to behave like the solution $v(x,t)$ of the Airy equation (3.8). In particular, for initial data which is somewhat smoother and more rapidly decaying than was assumed in Theorem 3.1 above, we show that this part of the solution $u(x,t)$ is smoother for $t>0$ but, as $x\to-\infty$, it decays less rapidly. We present these results for linearizations about N-soliton solutions of the KdV equation. Similar analysis applies for a more general class of KdV solutions satisfying (3.4); we omit such a discussion for the sake of brevity and restrict our attention to the N-soliton case.

Slower decay as $x\to-\infty$ for $t>0$ occurs because of the dispersive nature of the oscillating solutions $(d/dx)(f_\pm^2(x,k,t))$ of the linearized KdV equation (see [25] for a general discussion of dispersive waves). In particular, the asymptotic behavior of $f_\pm^2(x,k,t)$, as $x\to\pm\infty$ respectively, is given by the exponentials $e^{\pm i\theta(x,k,t)}$, where we define

$$\theta(x,k,t)=2kx+8k^3t. \tag{3.10}$$

Indeed, we write $f_+(x,k,t)=e^{ikx+4ik^3t}m_+(x,k,t)$, $f_-(x,k,t)=e^{-ikx-4ik^3t}m_-(x,k,t)$ with $\lim_{x\to+\infty}m_+=\lim_{x\to-\infty}m_-=1$. These waves propagate with a negative velocity $-4k^2$ so that waves with large wave numbers contribute to the solution near $x=-\infty$ almost instantaneously.

The same exponentials, $e^{\pm i\theta}$, form the solution of the linearized equation for $q=0$ (see (3.9) above), namely

$$v_t+v_{xxx}=0,$$

as is seen by Fourier transform, and arise in the asymptotics of $f_\pm^2(x,k,t)$, which, by the trace formula of Deift–Trubowitz [7], lead to a solution of the full KdV equation

$$q_t+q_{xxx}-6qq_x=0. \tag{3.11}$$

(In [7], $q(x,0)$ is written as an integral over the real line in k:

$$q(x,0)=\int_{-\infty}^{\infty}\frac{2i}{\pi}kR(k,0)f_+^2(x,k,0)\,dk+\sum_{j=1}^{N}a_jf_j^2(x,i\beta_j,0). \tag{3.12}$$

An approach to the KdV equation itself using (3.12) will appear in a subsequent paper by the present author.) The smoothness and decay properties of the solution of the Cauchy problem for the KdV equation were analyzed by Tanaka [24] and later Cohen Murray [5] using Faddeev–Marchenko inverse scattering theory rather than the then-unknown trace formula (3.12); asymptotic analysis of the KdV equation also appeared in [1], [2], where the more delicate regions $x/t=O(1)$ as $t\to+\infty$ were also discussed in the absence of solitons.

Our analysis for the linearized KdV equation proceeds in direct analogy with (3.8); the chief difference is the presence of extra factors multiplying the exponentials and their derivatives, which must be considered in all arguments. The techniques used will be primarily integration by parts and stationary phase analysis. In the limits we consider, the stationary phase points may tend to $\pm\infty$, which complicates matters slightly. As in [5], we will work in a shrinking neighborhood of the stationary phase points, whose size is proportional to a small negative power of $|x|$. This variation of the usual stationary phase argument [8] is used to control the error terms arising at the stationary phase points. The smoothness argument relies on the observation [5] that for $t>0$, we may rewrite the x-derivative of θ in terms of the k-derivatives of θ as follows:

$$(3.13)\qquad (\theta_x)^2=4k^2=\frac{\theta_k}{6t}-\frac{x}{3t}.$$

We will use this to re-express u_{xx} as a function which is smoother than u_{xx} might otherwise appear to be.

Our results are summarized in the following theorem:

THEOREM 3.2. *Assume $\phi(x)$, the initial data for the linearized* KdV *equation* (3.1), *has four continuous derivatives and that, for some fixed $l\geq 4$,*

$$(3.14)\qquad \left(1+|x|^l\right)\left(\frac{d}{dx}\right)^\alpha \phi(x)\in L^1 \quad \text{for } 0\leq\alpha\leq 4.$$

Then, in the case when $q(x,t)$ is an N-soliton solution of the KdV *equation, $u(x,t)$, defined in* (3.7) *above, has the following properties*:

(3.15) (i) *$u(x,t)$ is a classical solution of* (3.1) *for $t>0$ with $u(x,0)=\phi(x)$*;

(ii) *$\partial_t^r\partial_x^s u(x,t)$ is continuous for $t>0$ for all nonnegative integers r, s satisfying $3r+s\leq 2l+2$*;

(iii) $\lim_{x\to+\infty}|u(x,t)\cdot x^l|=0$, *$t>0$ fixed*;

(iv) *$|x|^{9/4}|u(x,t)|$ is bounded as a function of x for $t>0$ fixed (even as $x\to-\infty$)*;

(v) *$|u(ct+\delta,t)|t^{1/2}$ is bounded for $c<0$ as $t\to+\infty$, δ fixed.*

The proof of Theorem 3.2 is given in the three lemmas below, in which the smoothness and the limiting behavior are discussed separately. Proofs of some of these facts are deferred to the Appendix. First, we present some facts concerning the Jost functions $f_\pm(x,k,t)$ in the N-soliton case, where we choose the phases of the waves so that $q(-x,-t)=q(x,t)$. (Recalling the scaling done in §2, we see that this normalization is reasonable.)

The explicit form of the N-soliton leads to an algebraic expression for the Jost functions (see also [6]). In the proof of Theorem 3.2, we shall exploit certain properties of these functions, which we state here and prove in the Appendix. Define, for $j=1,\cdots,N$,

$$(3.16)\qquad \xi_j\equiv x-4\beta_j^2 t,\qquad \psi_j=\begin{cases}\cosh(\beta_j\xi_j), & j \text{ odd},\\ \sinh(\beta_j\xi_j), & j \text{ even},\end{cases}$$

and consider the $N\times N$ Wronskian determinant (in x):

$$(3.17)\qquad w(x,t)\equiv\det\begin{pmatrix}\psi_1 & \psi_2 & \cdots & \psi_N\\ \psi_1^{(1)} & \psi_2^{(1)} & \cdots & \psi_N^{(1)}\\ \vdots & & & \\ \psi_1^{(N-1)} & \psi_2^{(N-1)} & \cdots & \psi_N^{(N-1)}\end{pmatrix}$$
$$\equiv W_N(\psi_1,\cdots,\psi_N).$$

In the Appendix, we show $w(x,t)>0$. The N-soliton solution of the KdV equation is given by

$$(3.18)\qquad q(x,t)\equiv -2\frac{d^2}{dx^2}\log w(x,t),$$

while the eigenfunctions $f_\pm(x,k,t)$ are given by the ratios

$$(3.19)\qquad f_\pm(x,k,t)\equiv\frac{W_{N+1}\left(\psi_1,\psi_2,\cdots,\psi_N,e^{\pm i(kx+4k^3t)}\right)}{w(x,t)\Pi_{j=1}^N\pm(ik-\beta_j)}$$

where W_{N+1} is the $(N+1)\times(N+1)$ Wronskian determinant. Writing $f_\pm(x,k,t)=m_\pm(x,k,t)e^{\pm i\theta(x,k,t)}$, we deduce the following properties of the factors $m_\pm(x,k,t)$ from (3.19):

(3.20) (i) $m_\pm(x,k,t)$ are rational functions of k. Their denominations and numerators are polynomials of degree N in k; both denominators are in fact precisely $\Pi_{j=1}^N(k+i\beta_j)$ while the numerators are polynomials $k^N+A_1^\mp(x,t)k^{N-1}+\cdots+A_{N-1}^\pm(x,t)k+A_N^\mp(x,t)$ where each coefficient $A_l^\mp(x,t)$ is a rational function of $\{e^{\beta_j\xi_j}\}$ which is bounded. The denominator of each $A_l^\mp(x,t)$ is $w(x,t)$, which we show in the Appendix is a sum of terms $\exp(\Sigma_{j=1}^N\varepsilon_j\beta_j\xi_j)$ over all possible choices $\varepsilon_j=\pm 1$ with positive coefficients for each term.

(ii) $(d/dk)m_\pm(x,k,t)$ is a rational function of k which decays like $|k|^{-2}$ as $|k|\to\infty$.

(iii) $$\left(\frac{d}{dx}\right)^j m_\pm(x,k,t)=\frac{\Sigma_{l=1}^N\left(\frac{d}{dx}\right)^j\left(A_l\pm(x,t)\right)\cdot k^{N-l}}{\Pi_{j=1}^N(k+i\beta_j)}$$

so all x-derivatives of $m_\pm$ decay like $|k|^{-1}$ as $|k|\to\infty$.

(iv) We also have: $T(k)\equiv\prod_{j=1}^N\frac{k+i\beta_j}{k-i\beta_j}$.

(3.21) $F_j'(x,t)$ and $G_j'(x,t)$ are real analytic in x, t and for fixed t, they decay· exponentially fast as $|x|\to\infty$ (see Appendix).

By (3.21), all the smoothness and decay properties of Theorem 3.2 are satisfied by $F_j'(x,t)$ and $G_j'(x,t)$. Therefore, we consider the function $\tilde{u}(x,t)$ given by:

(3.22)
$$\tilde{u}(x,t)\equiv\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{T^2(k)}{4\pi ik}\frac{d}{dx}\{f_+^2(x,k,t)f_-^2(y,k,0)-f_-^2(x,k,t)f_+^2(y,k,0)\}\phi(y)\,dy\,dk.$$

Note that the integrand is continuous, even at $k=0$ (since $f_+(x,0,t)$, $f_-(x,0,t)$ are linearly dependent). Formula (3.22) suggests the following definitions:

$$\tilde{\phi}_\pm(k)\equiv\int_{-\infty}^{\infty} f_\pm^2(y,k,0)\phi(y)\,dy=\int_{-\infty}^{\infty} m_\pm^2(y,k,0)e^{\pm 2iky}\phi(y)\,dy. \tag{3.23}$$

We will analyze $\tilde{\phi}_\pm(k)$ just as in the usual Fourier transform case, using (3.20) to control the extra terms. Thus we shall see that $\tilde{u}(x,t)$, given by (3.22), and $v(x,t)$, the solution to the linearized problem for $q\equiv0$ given by (3.9), behave quite similarly.

The first part of Theorem 3.2 is contained in the following lemma.

LEMMA 3.3. *If* $(1+|x|^l)(d/dx)^\alpha\phi(x)\in L^1$ *for all* $0\le\alpha\le4$ *where* $l\ge4$ *is fixed, then the functions* $\partial_t^r\partial_x^s\tilde{u}(x,t)$, *where* $\tilde{u}(x,t)$ *is given by* (3.17), *are continuous for all nonnegative integers* r, s *satisfying* $3r+s\le 2l+2$.

Proof. The idea of the proof is as follows: We show that $\tilde{\phi}_\pm(k)$ decay rapidly enough as $|k|\to\infty$ that we can differentiate (4.9) twice with respect to x and still have a convergent integral. Then, using (3.13) to eliminate the $-4k^2$ factor arising from the exponentials and the estimate (3.20)(iii) to control derivatives of $m_\pm(x,k,t)$, we show that the integral for $\tilde{u}_{xx}$ can be differentiated twice. Repeating this argument, we obtain the desired result. A full proof appears in the Appendix below.

The decay as $x\to+\infty$, $t>0$ fixed and finite, is given by:

LEMMA 3.4. *For* $t>0$ *finite, fixed,* $|u(x,t)x^l|\to0$ *as* $x\to+\infty$.

Proof. Once again, we need only consider $\tilde{u}(x,t)$ since $F_j'(x,t)$, $G_j'(x,t)$ decay exponentially. For $x>0$, $t>0$ we note that $\theta_k=2x+24k^2t>0$ and in fact:

$$\left|\frac{\theta_{kk}}{\theta_k}\right|=\left|\frac{48kt}{2x+24k^2t}\right|\le\frac{|24kt|}{2\cdot x^{1/2}(12k^2t)^{1/2}}=\left(\frac{x}{12t}\right)^{-1/2} \tag{3.24}$$

since $1/(a^2+b^2)\le 1/2ab$ for $a,b>0$. Note also $\theta_{kkk}=48t$, which is bounded. Write

$$\tilde{u}(x,t)=\int_{-\infty}^{\infty}\rho(x,k,t)e^{i\theta}\,dk \tag{3.25}$$

by defining

$$\begin{aligned}\rho(x,k,t)e^{i\theta}\equiv{}&\frac{T^2(k)}{4\pi ik}\frac{d}{dx}\left[m_+^2(x,k,t)e^{i\theta}\tilde{\phi}_-(k)\right]\\&+\frac{T^2(-k)}{4\pi ik}\frac{d}{dx}\left[m_-^2(x,-k,t)e^{i\theta}\tilde{\phi}_+(-k)\right].\end{aligned} \tag{3.26}$$

We note $\rho(x,k,t)$ is continuous in k and decays like $|k|^{-4}$ as $|k|\to\infty$, as does $(\partial/\partial k)^\gamma\rho(x,k,t)$ for $0\le\gamma\le l$.

Integrating (3.25) by parts l times in k, we have

$$\tilde{u}(x,t)=(i)^l\int_{-\infty}^{\infty}\left[\left(\frac{\partial}{\partial k}\frac{1}{\theta_k}\right)^l\rho(x,k,t)\right]e^{i\theta}\,dk. \tag{3.27}$$

Using (3.24) and the obvious bound $(1/\theta_k)\le 1/2x$ for $x>0$ we obtain an estimate, for $t>0$ fixed,

$$|\tilde{u}(x,t)|\le(2x)^{-l}C(t)\quad\text{for } x\ge M>0 \tag{3.28}$$

where $C(t)$ is polynomial in $t^{1/2}$ of degree at most $l-1$.

Moreover, since the integrand in (3.27) is integrable, by a simple modification of the usual Riemann–Lebesgue lemma (namely, pick κ with $\int_{-\infty}^{-\kappa}+\int_{\kappa}^{\infty}<\varepsilon$ then approximate by a smooth function and integrate by parts), we can show $x^l\tilde{u}(x,t)\to 0$ as $x\to+\infty$ for $t>0$ fixed, which proves Lemma 3.4.

Finally we discuss decay as $x\to-\infty$ for $t>0$ fixed in the Appendix, where we prove the following result.

LEMMA 3.5. *As $x\to-\infty$ for $t>0$ fixed, $\tilde{u}(x,t)\cdot|x|^{9/4}$ remains bounded.*

To finish the proof of Theorem 3.2, we remark that if $x=ct+\delta$, $c<0$ then the roots of $\theta_k=0$ remain bounded as $t\to+\infty$, and the usual method of stationary phase applies [8]. This gives a decay rate of $t^{-1/2}$ and finishes the proof of Theorem 3.2.

4. Global behavior and uniqueness for the first order term in the N-soliton water wave problem. In §2 above, a formal expansion procedure was given for the Euler equations for a fluid in a flat-bottomed canal which was near the constant horizontal flow of Froude number 1. We now show that the choice of an N-soliton solution of the KdV equation as leading term in this expansion results in an equation for the first order term which has a unique "N-tuple wave" solution if we add a symmetry requirement. As noted previously, this term satisfies the inhomogeneous form of the linearized KdV equation:

$$Lu\equiv u_t+u_{xxx}-6(qu)_x=h(x,t) \tag{4.1}$$

where $q(x,t)$ is an N-soliton and $h(x,t)$ is a term which depends only on $q(x,t)$. A simple calculation shows that in fact $h(x,t)$ is a linear combination of the functions:

$$q_x,\quad qq_x,\quad q_{xxx},\quad q^2q_x,\quad qq_{xxx},\quad q_xq_{xx},\quad q_{xxxxx}.$$

More precisely, in the untransformed coordinates of §2, we have:

$$\begin{aligned}(4.2)\quad & -2Q_\tau^{(1)}+\tfrac{1}{3}Q_{\xi\xi\xi}^{(1)}-3Q_\xi^{(0)}Q^{(1)}-3Q^{(0)}Q_\xi^{(1)}-3Q_\xi^{(1)}\\ &=-\tfrac{19}{180}Q_{\xi\xi\xi\xi\xi}^{(0)}+\tfrac{5}{2}Q_{\xi\xi\xi}^{(0)}-\tfrac{9}{4}Q_\xi^{(0)}+\tfrac{19}{6}Q_\xi^{(0)}Q_{\xi\xi}^{(0)}+\tfrac{1}{12}Q^{(0)}Q_{\xi\xi\xi}^{(0)}+\tfrac{5}{2}\left(Q^{(0)}\right)^2Q_\xi^{(0)}.\end{aligned}$$

Using the transformation of variables of (2.32) above, namely $X-9T=\xi$, $-6T=\tau$, $q(X,T)=\frac{3}{2}\lambda(\xi,\tau)$, we get a similar expression in x, t variables.

We remark that $h(x,t)$ contains terms of the form $F_j'(x,t)$, which satisfy the linearized equation. It is rather surprising that these "secular terms" [13] do not give rise to resonant solutions. The usual choice for the solution to $Lu=F_j'$ would be tF_j', which grows linearly in t in the moving frame in which $\xi_j\equiv x-4\beta_j^2t$ remains constant as $t\to\infty$. However, the function $G_j'(x,t)$ is a solution of the homogeneous equation of the form (see the discussion in the Appendix)

$$G_j'(x,t)=c_j\left[\left(x-12\beta_j^2t\right)F_j'(x,t)+F_j(x,t)+H_j'(x,t)\right] \tag{4.3}$$

where $H_j'(x,t)$ is a rational function of the exponentials $\{e^{\beta_l\xi_l}\}$ which decays as $|x|\to\infty$ for $t>0$ fixed. Thus

$$\frac{G_j'(x,t)}{c_j}-\left(x-4\beta_j^2t\right)F_j'(x,t)-F_j(x,t)-H_j'(x,t)\equiv-8\beta_j^2tF_j'(x,t) \tag{4.4}$$

and since $L(G_j')=0$, we have:

$$L\left[\left(x-4\beta_j^2t\right)F_j'(x,t)+F_j(x,t)+H_j'(x,t)\right]=L\left[8\beta_j^2tF_j'(x,t)\right]=8\beta_j^2F_j'(x,t). \tag{4.5}$$

The function $(x-4\beta_j^2t)F_j'(x,t)+F_j(x,t)+H_j'(x,t)$ has the property that it is bounded as $t\to\infty$ in any moving frame, even $\xi_j=$constant, so we have found a "nonresonant" solution for the "secular" term $F_j'(x,t)$. For the "secular" forcing terms $G_j'(x,t)$, the growth in the obvious solution is quadratic in t as $t\to\infty$ with ξ_j fixed and, to the best of our knowledge, no nonresonant solutions of (4.1) exist. By the absence of these "secular" terms, the perturbation we consider is rather special. It is however, quite likely that symmetry considerations will rule out the presence of such terms to any order in the expansion of §2 above.

In order to study (4.1) when $h(x,t)$ is a linear combination of the functions listed above, we use the following representation of the N-soliton solutions of the KdV equation (see Gardner, Greene, Kruskal and Miura [10], Tanaka [23], and Deift and Trubowitz [7]):

$$q(x,t)=\sum_{j=1}^{N} a_jF_j(x,t) \tag{4.6}$$

where $F_j(x,t)$ is as usual the squared eigenfunction $f_+^2(x,i\beta_j,t)$. Then, using the third order equation satisfied by the squared eigenfunctions [20], we have:

$$q'''=\sum_{j=1}^{N} a_j\left[4\left(q(x,t)+\beta_j^2\right)F_j'+2q'F_j\right]=6qq'+\sum_{j=1}^{N}4a_j\beta_j^2F_j'; \tag{4.7}$$

therefore

$$qq'''=6q^2q'+\sum_{j=1}^{N}4a_j\beta_j^2qF_j',$$

$$q^{(5)}=6qq'''+18q'q''+\sum_{j=1}^{N}\left[16a_j\beta_j^2\left(q+\beta_j^2\right)F_j'+8a_j\beta_j^2q'F_j\right].$$

Thus our particular forcing term $h(x,t)$ is in the span of the functions:

$$\left\{F_j'(x,t),\,q(x,t)F_j'(x,t),\,q'(x,t)F_j(x,t),\,q^2q'(x,t),\,q'q''\right\},\qquad j=1,\cdots,N. \tag{4.8}$$

We prove:

LEMMA 4.1. *Suppose $h(x,t)$ is a linear combination of the functions in* (4.8). *Then there exists a solution to the linearized* KdV *equation* (4.1), *$Lu=h$, which is an N-tuple solitary wave in the following sense*:

(i) *$u(x,t)\to 0$ exponentially fast as $|x|\to\infty$ for t fixed*;

(ii) *$u(ct+\delta,t)\to 0$ exponentially fast as $t\to+\infty$ if $c\neq 4\beta_j^2, j=1,\cdots,N$*;

(iii) *$\lim_{t\to\infty}u(4\beta_j^2t+\xi,t)$ exists and is an exponentially decreasing function of ξ*;

(iv) *In fact, $u(x,t)$ is a sum of terms which are either*

(a) *rational functions in the N-exponentials $\{\exp\{\beta_j(x-4\beta_j^2t)\}\}$ with the same denominator $[w(x,t)]^2$ as $q(x,t)$, or*

(b) *of the form $(x-4\beta_j^2t)F_j'(x,t)$.*

Proof of Lemma 4.1. As above, we write

$$L(u)\equiv u_t+u_{xxx}-6(qu)_x. \tag{4.9}$$

From [10, Thm. 3.6], $F_j'(x,t)$ satisfy $L(u)=0$ and $F_j(x,t)$ satisfy the adjoint equation

$$v_t+v_{xxx}-6qv_x=0. \tag{4.10}$$

Thus $L(f_j)=-6F_jq'$; $L(tF_j')=F_j'$. Also, $L((x-12\beta_j^2t)F_j'+F_j)=6qF_j'$. From the KdV equation, $L(q^2)=6q'q''-6q^2q'$ and $L(q'')=12q'q''$. A basis for the solution is therefore given by

$$\left\{F_j, tF_j', \left(x-12\beta_j^2t\right)F_j'+F_j, q^2, q''\right\}. \tag{4.11}$$

Using $L(G_j')=0$ and (4.2), an equivalent basis is:

$$\left\{F_j, \left(x-4\beta_j^2t\right)F_j', H_j', q^2, q''\right\}. \tag{4.12}$$

The functions F_j, q^2, q'' have properties (i)–(iv) of Lemma 4.1 since $F_j(x,t)$ is exponentially decreasing. In the Appendix, we show that $(x-4\beta_j^2t)F_j'(x,t)$ is bounded and satisfies (i)–(iv), and prove that $H_j'(x,t)$ is a rational function of the exponentials which has properties (i)–(iv) as well. Assuming these results, the lemma is proved.

Remark (i) *Uniqueness.* If we choose the phases of the N-solution so that $q(-x,-t)=q(x,t)$ then the soliton in Lemma 4.1 is unique provided we require:

(a) $u(-x,-t)=u(x,t)$,

(b) $u(ct+\delta,t)$ is bounded for all c as $t\to\infty$,

(c) $u(ct+\delta,t)\to 0$ exponentially fast if $c\neq 4\beta_j^2$.

Proof. From the results of §3, the kernel of L is spanned by $F_j'(x,t)$, $G_j'(x,t)$ and $\{(f_\pm^2)'(x,k,t)|-\infty<k<\infty\}$. The functions F_j' violate (a); $G_j'(x,t)$ violates (b) for $c=4\beta_j^2$; by stationary phase analysis, $(f_\pm^2)'(x,k,t)$ violates (c) for $c=-12k^2$ (the decay is algebraic, not exponential since dispersion prevents cancellation). Thus $u(x,t)$ as given in Lemma 4.1 is unique, since in this case, the functions given in (4.11) satisfy all these conditions.

Remark (ii). *Higher order terms.* Even in the time-independent case, explicit expressions for the higher order terms of the formal expansion involve more complicated, transcendental functions. For even the second order term, functions like

$$\log(1-\tanh\beta x)\cdot\operatorname{sech}^2\beta x$$

occur. Thus algebraic methods will not readily yield solvability results like Lemma 4.1 for the higher order terms.

Appendix. In this Appendix, we collect certain facts about N-solitons of the KdV equation and the associated eigenfunctions $f_+(x,k,t)$ in this case. These properties are well known ([6], [10], [13], [21], [23]), so we sketch the proofs for the most part. The functions $G_j(x,t)$ do not appear in these papers, so results regarding these eigenfunctions are presented and proved in full.

With the choice of phases so that $q(x,t)=q(-x,-t)$, the N-soliton $q(x,t)$ with bound states $-\beta_N^2<-\beta_{N-1}^2<\cdots<-\beta_1^2<0$ where $\beta_j>0$, it is given explicitly by the following formulae:

Let $\xi_j\equiv x-4\beta_j^2t$ and define

$$\xi_j(x,t)=\begin{cases}\cosh(\beta_j\xi_j) & \text{if } j \text{ is odd},\\ \sinh(\beta_j\xi_j) & \text{if } j \text{ is even}.\end{cases} \tag{A.1}$$

Let

$$w(x,t)\equiv W_N(\psi_1,\psi_2,\cdots,\psi_N)=\det\begin{vmatrix} \psi_1 & \psi_2 & \cdots & \psi_N \\ \psi_1^{(1)} & \psi_2^{(1)} & \cdots & \psi_N^{(1)} \\ \vdots & \vdots & & \\ \psi_1^{(N-1)} & \psi_2^{(N-1)} & \cdots & \psi_N^{(N-1)} \end{vmatrix}$$

the Wronskian determinant in x of $\psi_1,\cdots,\psi_N$. Then

$$q(x,t)\equiv -2\frac{d^2}{dx^2}\log w(x,t).$$

This definition is sensible because $w(x,t)>0$, which we show below.

LEMMA A.1. $w(x,t)>0$. *In fact,* $w(x,t)$ *is a sum of exponentials with positive coefficients.*

Proof.

$$w(x,t)=W_N\left(\cosh\beta_1\xi_1,\,\sinh\beta_2\xi_2,\cdots,\frac{\exp\{\beta_N\xi_N\}+(-1)^{N-1}\exp\{-\beta_N\xi_N\}}{2}\right).$$

By the multilinearity of the determinant, this is the sum over all possible choices $\varepsilon_j=\pm1$ of $2^{-N}W_N(\exp\{\varepsilon_1\beta_1\xi_1\},\ \varepsilon_2\exp\{\varepsilon_2\beta_2\xi_2\},\ \exp\{\varepsilon_3\beta_3\xi_3\},\ \varepsilon_4\exp\{\varepsilon_4\beta_4\xi_4\}\cdots)$ i.e. we have upon evaluating the Vandermonde determinants,

$$\text{(A.2)}\quad w(x,t)=2^{-N}\sum_{\substack{\text{all choices}\\ \varepsilon_j=\pm1}}\exp\left\{\left(\sum_{l=1}^{N}\varepsilon_l\beta_l\varepsilon_l\right)\right\}\varepsilon_2\varepsilon_4\cdots\varepsilon_{2[n/2]}\cdot\prod_{j<k}(\varepsilon_k\beta_k-\varepsilon_j\beta_j).$$

Since $0<\beta_1<\beta_2<\cdots<\beta_N$, the number of negative factors in the product can be found explicitly. Namely, if $\varepsilon_{i_1},\cdots,\varepsilon_{i_r}$ are the negative indices for a given choice of the ε's, we obtain $(i_1-1)+\cdots+(i_r-1)$ negative factors in $\Pi_{j<k}(\varepsilon_k\beta_k-\varepsilon_j\beta_j)$. The extra factor $\varepsilon_2\varepsilon_4\cdots\varepsilon_{2[n/2]}$ adds an additional (-1) factor for each i_j which is even. For any choice of r and $i_1,\cdots,i_r$, this means that there are an even number of (-1) factors; thus every term in the sum (A.2) has a positive coefficient. We remark that all exponents $\Sigma_{j=1}^{N}\varepsilon_j\beta_j\xi_j$ occur in $w(x,t)$ and that $w(x,t)=w(-x,-t)$ since changing $\{\varepsilon_j\}\to\{-\varepsilon_j\}$ does not alter the coefficients in (A.2). This proves Lemma A.1.

The eigenfunctions $f_+(x,k,t)$ are given explicitly by:

$$\text{(A.3)}\qquad f_+(x,k,t)\equiv\frac{W_{N+1}(\psi_1,\psi_2,\cdots,\psi_N,\exp\{ikx+4ik^3t\})}{w(x,t)\Pi_{l=1}^{N}(ik-\beta_l)}.$$

The normalization $f_+\sim\exp\{ikx+4ik^3t\}$ as $x\to+\infty$ for t fixed is satisfied, as is seen by looking at the leading term, which has exponent $\Sigma_{j=1}^{N}\beta_j\xi_j$ (i.e. pick the term with all ε_k's$=+1$ in (A.2) and the corresponding expansion of the numerator). From the fact ([6], [7]) that $T(k)\equiv\Pi_{j=1}^{N}(k+i\beta_j)/(k-i\beta_j)$, we obtain a similar expression for $f_-(x,k,t)$ using $T(k)f_-(x,k,t)\equiv f_+(x,k,t)$. The proof that $f_\pm(x,k,t)$ satisfy the Schrödinger equation with potential $q(x,t)$ defined as above is given in Deift [6], the basic idea is to use Jacobi's identity for the Wronskians and induction on N.

From (A.3) and the expression for $f_-(x,k,t)$, it is easy to see that

$$f_+(x,i\beta_j,t)+(-1)^{j+1}f_-(x,i\beta_j,t)\equiv0.$$

Also, from (A.2) and (A.3), it is clear that $f_+(x,i\beta_j,t)$ decays like $\exp\{-\beta_j|\xi_j|\}$ as $|\xi_j|\to\infty$ since the exponentials $\exp\{\pm\beta_j\xi_j\}$ in the numerator will cancel each other, while remaining in the denominator $w(x,t)$.

The factor is $g_j(x,t)$ defined in (3.6) as

$$g_j(x,t)\equiv\frac{1}{i}\frac{d}{dk}\left(f_-(x,k,t)+(-1)^{j+1}f_+(x,k,t)\right)\Big|_{k=i\beta_j}. \tag{A.4}$$

Differentiating the exponential $\exp\{ikx+4ik^3t\}$ gives a term $2(x-12\beta_j^2t)f_+(x,i\beta_j,t)$, while differentiating the factors $(\pm ik)^{l-1}$ which occur in the $(l,N+1)$ entry in the Wronskian gives terms having the form $(c_\varepsilon\exp\{\lambda(x,t)\})/(w(x,t))$ where $\lambda(x,t)\equiv\pm\beta_j\xi_j+\sum_{l=1}^N\varepsilon_l\beta_l\xi_l$ whose sum we denote by $h_j(x,t)$. Thus $g_j(x,t)$ grows like $\exp\{\beta_j|\xi_j|\}$ as $|\xi_j|\to\infty$ and is of the form $2(x-12\beta_j^2t)f_+(x,i\beta_j,t)+h_j(x,t)$ where $h_j(x,t)$ is a rational function of the N exponentials $\exp\{\beta_l\xi_l\}$ with denominator $w(x,t)$, growing like $\exp\{\beta_j|\xi_j|\}$ as $|\xi_j|\to\infty$. Since $G_j(x,t)\equiv c_jf_+(x,i\beta_j,t)g_j(x,t)$, multiplying by $f_+(x,i\beta_j,t)$ we have

$$\begin{aligned}\frac{G_j(x,t)}{c_j}&\equiv2\left(x-12\beta_j^2t\right)f_+^2(x,i\beta_j,t)+f_+(x,i\beta_j,t)h_j(x,t)\\&\equiv2\left(x-12\beta_j^2t\right)F_j(x,t)+H_j(x,t).\end{aligned} \tag{A.5}$$

Since $f_+(x,i\beta_j,t)$ is rational with denominator $w(x,t)$ and decays like $\exp\{-\beta_j|\xi_j|\}$ as $|\xi_j|\to\infty$, $H_j(x,t)$ is rational with denominator $(w(x,t))^2$ and is bounded as $|\xi_j|\to\infty$. Since all the other exponentials occur in the numerator with growth at most $\exp\{\sum_{l\neq j}2\beta_l|\xi_l|\}$ and these terms are balanced by those in the denominator, $H_j(x,t)$ is bounded for all x, t real. It then follows that $H_j'(x,t)$ is a sum of terms which decrease exponentially fast as $t\to\infty$ except in the frames $x-4\beta_j^2t=$constant, where their limit is an exponentially decreasing function of the variable $\xi_j\equiv x-4\beta_j^2t$. Note, however, that for $t\to\infty$, $H_j'(x,t)$ "decouples" into N exponentially decreasing bumps moving at the speeds $4\beta_j^2$ with the same phases as $q(x,t)$ as $t\to\infty$; unlike $F_j'(x,t)$, these terms give rise to contributions in *all* N moving frames. These are the basic properties used in the discussion in §§3 and 4 above.

Proof of Lemma 3.3. *Step* 1. We show that $(d/dk)^\gamma\bar{\phi}_\pm(k)$ exists and is $O(|k|^{-4})$ as $k\to\infty$ for $0\leq\gamma\leq l$. If we integrate (3.23) by parts, we have:

$$\tilde{\phi}_\pm(k)=\frac{1}{\pm2ik}\int_\infty^\infty\frac{d}{dy}\left(m_\pm^2(y,k,0)\phi(y)\right)e^{\pm2iky}\,dy. \tag{A.6}$$

The integral is absolutely convergent by our assumptions on ϕ and the properties of $m_\pm(y,k,0)$ listed above. In fact, we may integrate by parts four times with respect to y, obtaining

$$\tilde{\phi}_\pm(k)=\frac{1}{(2ik)^4}\int_{-\infty}^\infty\left(\frac{d}{dy}\right)^4\left(m_\pm^2(y,k,0)\phi(y)\right)e^{\pm2iky}\,dy, \tag{A.7}$$

and the integral is still absolutely convergent. Thus $\tilde{\phi}_\pm(k)$ is $O(|k|^{-4})$ as $|k|\to\infty$.

Since $(1+|x|^l)\phi(x)\in L^1$, the k-derivatives of $\tilde{\phi}_\pm(k)$ of order less than or equal to l all exist. Integrating these expressions by parts four times, we find

$$\left(\frac{d}{dk}\right)^\gamma\tilde{\phi}_\pm(k)=O\left(|k|^{-4}\right)\quad\text{as }|k|\to\infty\quad\text{for }0\leq\gamma\leq l, \tag{A.8}$$

which completes Step 1.

Step 2 (smoothness for $t>0$). Writing

(A.9)

$$\tilde{u}(x,t)\equiv\int_{-\infty}^{\infty}\frac{dk}{4\pi ik}T^2(k)\frac{d}{dx}\left[\left(m_+^2(x,k,t)e^{i\theta}\tilde{\phi}_-(k)\right)-\left(m_-^2(x,k,t)e^{-i\theta}\tilde{\phi}_+(k)\right)\right],$$

we have a continuous integrand which decays like $|k|^{-4}$ as $|k|\to\infty$. Therefore we may differentiate twice with respect to x and still have a convergent integral. This yields:

(A.10)

$$\tilde{u}_{xx}(x,t)=\int_{-\infty}^{\infty}\frac{dk}{4\pi ik}T^2(k)\left(\frac{d}{dx}\right)^3\left\{m_+^2(x,k,t)e^{i\theta}\tilde{\phi}_-(k)-m_-^2(x,k,t)e^{-i\theta}\tilde{\phi}_+(k)\right\}.$$

Since $\theta_x=2k$ and $m'_\pm(x,k,t)$ decays like $|k|^{-1}$, the terms on the integrand in which $m'_\pm(x,k,t)$ or a higher derivative appears all have decay like $|k|^{-4}$ or faster; these terms can therefore be differentiated twice more with respect to x. The remaining terms, in which the exponential is differentiated three times, are:

$$\text{(A.11)}\quad \int_{-\infty}^{\infty}\frac{T^2(k)}{2\pi}(-4k^2)\left\{m_+^2(x,k,t)e^{i\theta}\tilde{\phi}_-(k)+m_-^2(x,k,t)e^{-i\theta}\tilde{\phi}_+(k)\,dk\right\}$$

$$=\int_{-\infty}^{\infty}\frac{T^2(k)}{2\pi}\left(\frac{x}{3t}-\frac{\theta_k}{6t}\right)\left\{m_+^2(x,k,t)e^{i\theta}\tilde{\phi}_-(k)+m_-^2(x,k,t)e^{-i\theta}\tilde{\phi}_+(k)\right\}dk$$

$$=\int_{-\infty}^{\infty}\frac{T^2(k)}{2\pi}\frac{x}{3t}\left\{m_+^2(x,k,t)e^{i\theta}\tilde{\phi}_-(k)+m_-^2(x,k,t)e^{-i\theta}\tilde{\phi}_+(k)\right\}dk$$

$$+\frac{1}{12\pi it}\int_{-\infty}^{\infty}\left\{\left[\frac{d}{dk}\left(T^2(k)m_+^2(x,k,t)\tilde{\phi}_-(k)\right)\right]e^{i\theta}\right.$$

$$\left.-\left[\frac{d}{dk}\left(T^2(k)m_-^2(x,k,t)\tilde{\phi}_+(k)\right)\right]e^{-i\theta}\right\}dk$$

where we integrated by parts in the second term. Each of these terms has continuous k-integrands which decay like $|k|^{-4}$ or better; hence they are also twice differentiable with respect to x. Therefore $\tilde{u}(x,t)$ has in fact four continuous x-derivatives for $t>0$. Repeating this argument iteratively, we obtain $u(x,t)$ has $2l+2$ continuous x-derivatives (since we can only bound $(d/dk)^\gamma\tilde{\phi}_\pm(k)$ for $0\leq\gamma\leq l$, we may repeat the argument l times).

To handle t-derivatives, we note that differentiating directly in t brings down a factor $\theta_t\equiv 8k^3$, which does not a priori lead to a convergent integral. However, multiplication by $8k^3$ may be expressed in the sense of distributions as $(\partial/\partial x)^3$ plus convergent integrals. Since $(\partial/\partial x)^3\tilde{u}(x,t)$ is continuous, so is $\tilde{u}_t$. The equation (3.1) then gives higher regularity and the desired result. This proves Lemma 3.3.

Proof of Lemma 3.5. Define $\alpha\equiv(-x/12t)^{1/2}$. For $x<0$, $t>0$, $\theta_k=2x+24k^2t=0$ for $k=\pm\alpha$. As $x\to-\infty$, (for fixed $t>0$), $\alpha\to+\infty$. Let $k=\alpha\kappa$. Then

$$\text{(A.12)}\qquad \tilde{u}(x,t)=\int_{-\infty}^{\infty}\rho(x,k,t)e^{i\theta}\,dk=\alpha\int_{-\infty}^{\infty}\rho(x,\alpha\kappa,t)e^{i\lambda\tilde{\theta}}\,d\kappa$$

where $\lambda \equiv 2|x|^{3/2}/(12t)^{1/2}$ and $\tilde{\theta}(\kappa) \equiv -\kappa+\kappa^3/3$, so that $\tilde{\theta}(\kappa)=0$ for $\kappa=\pm 1$. *In the usual stationary phase method*, the chief contribution to the integral comes from the terms

$$\rho(x,\alpha,t)\int_1^{1\pm\varepsilon} e^{i\lambda[-\kappa+\kappa^3/3]}\,d\kappa \quad \text{and} \quad \int_1^{-1\pm\varepsilon}\rho(x,\alpha,t)e^{i\lambda\tilde{\theta}}\,d\kappa$$

where the κ value is frozen at $\kappa=\pm 1$ in the function ρ. The extra term $\int[\rho(x,\alpha\kappa,t)-\rho(x,\alpha,t)]e^{i\lambda\tilde{\theta}}\,d\kappa$ is of lower order for large λ by bounds on the derivative. In the case considered here, $\alpha\to\infty$ so this error term may become large. To counteract this, as in [5], we consider a very small interval about the stationary phase points $\kappa=\pm 1$, of order $|x|^{-\upsilon}$ for instance. We estimate as follows:

$$\text{(A.13)} \qquad \alpha\int_1^{1+\varepsilon}\rho(x,\alpha\kappa,t)e^{i\lambda\tilde{\theta}}\,d\kappa$$

$$=\alpha\left[\int_1^{1+\varepsilon}\rho(x,\alpha,t)e^{i\lambda\tilde{\theta}}\,d\kappa+\int_1^{1+\varepsilon}(\rho(x,\alpha\kappa,t)-\rho(x,\alpha,t))e^{i\lambda\tilde{\theta}}\,d\kappa\right]$$

$$=\alpha\rho(x,\alpha,t)\int_1^{1+\varepsilon}e^{i\lambda\tilde{\theta}}\,d\kappa+\alpha^2\cdot\frac{d}{dk}\rho(x,\alpha\bar{\kappa},t)\int_1^{1+\varepsilon}(\kappa-1)e^{i\lambda\tilde{\theta}}\,dk$$

$$\text{for some } \bar{\kappa}\in[1,1+\varepsilon].$$

The first term is estimated as in the usual method of stationary phase; the second term leads after an integration by parts to an estimate of the form $C\alpha^{-2}\lambda^{-1}\varepsilon$ since $d\rho/dk$ decays like $|k|^{-4}$. With $\varepsilon=|x|^{-\upsilon}$, $0<\upsilon<\frac{1}{2}$, this term is of order $|x|^{-5/2-\upsilon}$ as $|x|\to\infty$. The first term decays like $\alpha^{-3}\cdot\lambda^{-1/2}=O(|x|^{-9/4})$ as $|x|\to\infty$ and is the leading term.

Similar estimates hold on the intervals $[1-\varepsilon, 1]$, $[-1-\varepsilon, -1]$, $[-1, -1+\varepsilon]$.

On the interval $[1+\varepsilon, \infty)$, we estimate as follows:

$$\text{(A.14)} \qquad \int_{1+\varepsilon}^{\infty}\rho(x,\alpha\kappa,t)e^{i\lambda\tilde{\theta}}\,d\kappa=\frac{-\alpha e^{i\lambda[-2/3+\varepsilon^2+\varepsilon^3/3]}}{i\lambda\varepsilon(2+\varepsilon)}\rho(x,\alpha(1+\varepsilon),t)$$

$$-\frac{\alpha}{i\lambda}\int_{1+\varepsilon}^{\infty}\left[\frac{d}{dk}\left(\frac{\rho(x,\alpha\kappa,t)}{\kappa^2-1}\right)\right]e^{i\lambda\tilde{\theta}}\,d\kappa,$$

which leads to a bound of the form

(A.15)

$$\left|\alpha\int_{1+\varepsilon}^{\infty}\rho(x,\alpha\kappa,t)e^{i\lambda\tilde{\theta}}\,d\kappa\right|\le(\lambda^{-1}\alpha^{-3}\varepsilon^{-1})\cdot C_1+\lambda^{-1}\alpha^{-2}\varepsilon^{-1}\cdot C_2+\lambda^{-1}\alpha^{-3}\varepsilon^{-2}\cdot C_3$$

$$=O\left(|x|^{-3+2\upsilon}\right) \quad \text{since } 0<\upsilon<\tfrac{1}{2}.$$

A similar estimate holds on the interval $(-\infty, -1-\varepsilon]$.

Finally, on the interval $[-1+\varepsilon, 1-\varepsilon]$, we have:

$$\text{(A.16)} \qquad \alpha\int_{-1+\varepsilon}^{1-\varepsilon}\rho(x,\alpha\kappa,t)e^{i\lambda\tilde{\theta}}\,d\kappa=\left.\frac{\alpha\cdot\rho(x,\alpha\kappa,t)e^{i\lambda\tilde{\theta}}}{i\lambda(\kappa^2-1)}\right|_{k=-1+\varepsilon}^{\kappa-1-\varepsilon}$$

$$-\frac{\alpha}{i\lambda}\int_{-1+\varepsilon}^{1-\varepsilon}\left[\frac{d}{d\kappa}\left(\frac{\rho(x,\alpha\kappa,t)}{\kappa^{2-1}}\right)\right]e^{i\lambda\tilde{\theta}}\,d\kappa.$$

Integrating by parts three more times, we obtain:

(A.17)

$$\alpha\int_{-1+\varepsilon}^{1-\varepsilon}\rho(x,\alpha\kappa,t)e^{i\lambda\tilde{\theta}}\,d\kappa=\alpha\left(\sum_{l=0}^{3}\left[\left(\frac{-1}{i\lambda(\kappa^2-1)}\frac{d}{d\kappa}\right)^{l}\frac{(x,\alpha\kappa,t)}{i\lambda(\kappa^2-1)}\right]\right)e^{i\lambda\tilde{\theta}}\Bigg|_{\kappa=-1+\varepsilon}^{\kappa=1-\varepsilon}$$
$$+\alpha\int_{-1+\varepsilon}^{1-\varepsilon}\left[\left(\frac{d}{d\kappa}\frac{1}{i\lambda(\kappa^2-1)}\right)^{4}\rho(x,\alpha\kappa,t)\right]e^{i\lambda\tilde{\theta}}\,d\kappa.$$

The boundary terms lead to estimates, for $l=0$, 1, 2, 3, of the form

$$\lambda^{-1}\alpha^{-3}\varepsilon^{-1},\qquad \lambda^{-2}\alpha^{-3}\varepsilon^{-3}(1+\alpha\varepsilon),$$
$$\lambda^{-3}\alpha^{-3}\varepsilon^{-5}(1+\alpha\varepsilon+\alpha^2\varepsilon^2),\qquad \lambda^{-4}\alpha^{-3}\varepsilon^{-7}(1+\alpha\varepsilon+\alpha^2\varepsilon^2+\alpha^3\varepsilon^3),$$

respectively.

Since α is $O(|x|^{1/2})$, λ is $O(|x|^{3/2})$ and $\varepsilon=|x|^{-v}$, the inequality $v<\frac{1}{2}$ implies that the leading term is the first, i.e. $\lambda^{-1}\alpha^{-3}\varepsilon^{-1}=O(|x|^{-3+v})$ as $|x|\to\infty$.

The remaining integral

$$\int_{-1+\varepsilon}^{1-\varepsilon}\left[\left(\frac{d}{d\kappa}\frac{1}{i\lambda(\kappa^2-1)}\right)^{4}\rho(x,\alpha\kappa,t)\right]e^{i\lambda\tilde{\theta}}\,d\kappa$$

has a bound of the form

$$C\lambda^{-4}(\alpha^5\varepsilon^{-4}+\alpha^4\varepsilon^{-5}+\alpha^3\varepsilon^{-6}+\alpha^2\varepsilon^{-7}+\alpha\varepsilon^{-8})$$

which has leading term

$$\lambda^{-4}\alpha^5\varepsilon^{-4}=O\left[|x|^{-7/2+4v}\right]\quad\text{as } |x|\to\infty.$$

This can be chosen to be of lower order than the contributions from the stationary phase points, by making

$$-\tfrac{7}{2}+4v<-\tfrac{9}{4},\quad\text{which is true for } v<\tfrac{5}{16}.$$

This completes the proof of Lemma 3.5.

Acknowledgments. Some of these results appear in the author's doctoral dissertation (New York University, October 1980). It is a pleasure to thank Percy Deift and Tom Beale for helpful discussions, and, most especially, Jürgen Moser for his encouragement and advice.

REFERENCES

[1] M. J. Ablowitz and A. C. Newell, *The decay of the continuous spectrum for solutions of the Korteweg–de Vries equation*, J. Math. Phys., 14 (1973), pp. 1277–1284.

[2] M. J. Ablowitz and H. Segur, *Asymptotic solutions of the Korteweg–de Vries equation*, Stud. Appl. Math., 57 (1977), pp. 13–44.

[3] C. J. Amick and J. F. Toland, *Finite amplitude solitary waves*, Arch. Rational Mech. Anal., 76 (1981), pp. 9–95.

[4] J. T. Beale, *The existence of solitary water waves*, Comm. Pure Appl. Math., 30 (1977), pp. 373–389.

[5] A. COHEN, *Existence and regularity for the solutions of the Korteweg–de Vries equation*, Arch. Rational Mech. Anal., 71 (1979), pp. 143–175.

[6] P. DEIFT, *Applications of a commutation formula*, Duke Math. J., 45 (1978), pp. 267–310.

[7] P. DEIFT AND E. TRUBOWITZ, *Inverse scattering on the line*, Comm. Pure Appl. Math., 32 (1979), pp. 121–251.

[7a] J. D. FENTON, *A ninth-order solution for the solitary wave*, J. Fluid Mech., 53 (1972), pp. 257–271.

[7b] J. D. FENTON AND M. M. RIENECKER, *A Fourier method for nonlinear waves: Solitary wave interactions*, J. Fluid Mech., 118 (1982), pp. 411–443.

[8] B. FRIEDMAN, *Lectures on Applications-Oriented Mathematics*, Holden-Day, San Francisco, 1969.

[9] K. O. FRIEDRICHS AND D. H. HYERS, *The existence of solitary waves*, Comm. Pure Appl. Math., 7 (1954), pp. 517–550.

[10] C. GARDNER, J. GREENE, M. KRUSKAL AND R. MIURA, *Korteweg–de-Vries equation and generalizations*, VI. *Methods for exact solution*, Comm. Pure Appl. Math., 27 (1974), pp. 97–133.

[11] T. KANO AND T. NISHIDA, *Sur les ondes de surface de l'eau avec une justification mathématique des équations des ondes en eau peu profonde*, J. Math. Kyoto Univ., 19 (1979), pp. 335–370.

[12] D. KAUP, *Closure of the squared Zakharov–Shabat eigenstates*, J. Math. Anal. Appl., 54 (1976), pp. 849–864.

[13] J. P. KEENER AND D. W. MCLAUGHLIN, *Solitons under perturbations*, Phys. Rev. A, 16 (1977), pp. 777–790.

[14] D. J. KORTEWEG AND G. DE VRIES, *On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary wave*, Philos. Magazine, 39 (1895), pp. 422–433.

[15] P. D. LAX, *Integrals of nonlinear equations of evolution and solitary waves*, Comm. Pure Appl. Math., 21 (1968), pp. 467–490.

[16] T. LEVI-CIVITA, *Détermination rigoureuse des ondes permanentes d'ampleur finie*, Math. Ann., 93 (1925), pp. 265–314.

[17] H. P. MCKEAN AND E. TRUBOWITZ, *Hill's operator and hyperelliptic function theory in the presence of infinitely many branch points*, Comm. Pure Appl. Math., 29 (1976), pp. 143–226.

[17a] R. M. MIRIE AND C. H. SU, *Collisions between two solitary waves, Part* 2. *A numerical study*, J. Fluid Mech., 115 (1982), pp. 475–492.

[18] A. C. NEWELL, *The inverse scattering transform*, in Solitons, R. K. Bullough and P. J. Caudrey, eds., Springer-Verlag, New York, 1980.

[19] T. NISHIDA, *Nonlinear Hyperbolic Equations and Related Topics in Fluid Dynamics*, 78.02, Publications Mathématiques D'Orsay, Orsay, France, 1978.

[20] R. L. SACHS, *Completeness of derivatives of squared Schrödinger eigenfunctions and explicit solutions of the linearized* KdV *equation*, this Journal, 14 (1983), pp. 674–683.

[21] A. C. SCOTT, F. Y. F. CHU AND D. W. MCLAUGHLIN, *The soliton: A new concept in applied science*, Proc. IEEE, 61 (1973), pp. 1443–1483.

[22] J. J. STOKER, *Water Waves*, Interscience, New York, 1957.

[23] S. TANAKA, *On the N-tuple wave solution of the Korteweg–de-Vries equation*, Publ. Res. Inst. Math. Sci., 8 (1972-3), pp. 419–427.

[24] ———, *Korteweg–de-Vries equation: Construction of solutions in terms of scattering data*, Osaka J. Math., 11 (1974), pp. 49–59.

[25] G. B. WHITHAM, *Linear and Nonlinear Waves*, John Wiley, New York, 1974.

[26] E. ZEHNDER, *Generalized implicit function theorems with applications to some small divisor problems*, Comm. Pure Appl. Math., 28 (1975), pp. 91–140.

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics
005

POSITIVE SOLUTIONS FOR TEMPERATURE-DEPENDENT TWO-GROUP NEUTRON FLUX EQUATIONS: EQUILIBRIA AND STABILITIES*

ANTHONY W. LEUNG† AND GEN-SHUN CHEN‡

Abstract. A two-group neutron flux diffusion-reaction system for the nuclear fission reactor is studied. The reaction rates are essentially assumed to depend on temperature, introducing a third equation. Conditions for existence or nonexistence of positive steady state are found for a system of three elliptic equations under Dirichlet boundary conditions.

The case of a system of two equations simulating prompt-feedback condition is also studied. Here, a simple sufficient condition for positive steady state is found. Stability for the trivial solution is also considered. The method of upper and lower solutions is used in the analysis.

1. Introduction. This article is concerned with the study of the nonlinear two-group diffusion equation:

$$(1.1)\qquad \begin{aligned}&\Delta u+H_1(T)u+H_2(T)v=0,\\ &\Delta v+Q_1(T)u-Q_2(T)v=0, \qquad \text{in } \mathfrak{D},\\ &\Delta T-cT+G_1(T)u+G_2(T)v=0,\end{aligned}$$

of the bare, homogeneous nuclear fission reactor, where u and v are respectively the fast and thermal neutron flux, T is the core temperature above averaging coolant temperature, and $\mathfrak{D}$ is a bounded domain (representing the core) in R^n, $n\geq 2$. In (1.1), $\Delta\equiv\sum_{i=1}^{n}\partial^2/\partial x_i^2$; $H_2(T)$, $Q_i(T)$ and $G_i(T)$, $i=1,2$ are positive functions of temperature; c is a positive constant. In more conventional notation of nuclear engineering:

$$\begin{aligned}&H_1(T)\equiv\sigma_1^{-1}\left[\nu_1\Sigma_{f_1}(T)-\Sigma_R(T)\right], \qquad H_2(T)\equiv\sigma_1^{-1}\nu_2\Sigma_{f_2}(T),\\ &Q_1(T)\equiv\sigma_2^{-1}\Sigma_{S_{12}}(T), \qquad Q_2(T)\equiv\sigma_2^{-1}\Sigma_{a_2}(T),\\ &G_1(T)\equiv\theta_1\Sigma_{f_1}(T), \qquad G_2(T)\equiv\theta_2\Sigma_{f_2}(T).\end{aligned}$$

We classify the fast and thermal neutron into groups 1 and 2 respectively. For $i=1,2$, the parameter σ_i is the diffusion coefficient of group i; Σ_{f_i} is the fission "macroscopic cross section" in group i (i.e. Σ_{f_i} is the probability per unit path length traveled that a neutron in energy group i will undergo fission); ν_i is the average number of fast neutrons released during fission induced by a neutron in group i. Σ_R is the removal "macroscopic cross section" characterizing the probability that a neutron will be removed from group 1 (i.e. Σ_R is the probability per unit length traveled that a neutron in the fast group will undergo a collision causing its own removal through absorption or slowing down to the thermal group). $\Sigma_{S_{12}}$ is the macroscopic group-transfer cross section (probability of collision causing transfer from fast to thermal group); Σ_{a_2} is the absorption macroscopic cross section for the thermal group. Finally, θ_i is the effective energy released in each fission for group i, and c is the cooling constant. The detail definitions can be found in [3, p. 288].

*Received by the editors July 9, 1982, and in revised form May 3, 1983.

†Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio 45221. The research of this author was partially supported by the National Science Foundation under grant MCS 80-01851.

‡Department of Chemical and Nuclear Engineering, University of Cincinnati, Cincinnati, Ohio 45221.

System (1.1) essentially assumes that the cross sections, Σ_{f_i}, Σ_R, etc., depend on temperature, whose changes and effects can briefly be explained as follows. Power levels of nuclear reactors are adjusted by moving the control rods. This will cause changes in temperature and neutron flux in the components of the reactor core. However, the atomic concentrations of materials in the core depend sensitively on temperature. As temperature changes, they may contract, expand or change phase. This in turn will cause a change in the macroscopic cross section. Furthermore, temperature changes may directly affect the microscopic cross section (e.g. via the Doppler effect). The readers are referred to [6, p. 264], or [3] for further explanations. Such a two group temperature-dependent model has been studied by e.g. [7], with results not as readily applicable as those obtained below. For more general multigroup models, see e.g. [3, p. 301].

We now clarify our notation, conventions and assumptions in this article. $\mathcal{D}$ is a bounded domain in R^n, $n\geq 2$, whose boundary $\delta\mathcal{D}$ is C^2 smooth (i.e. can be locally represented as $x_i=\phi(x)$ for some i,ϕ with continuous second derivatives and independent of x_i); $\bar{\mathcal{D}}$ denotes $\mathcal{D}$ closure. The functions H_i, Q_i, G_i, $i=1,2$ are Lipschitz continuous for $T\geq 0$. Let $\tilde{h}_i\equiv\inf\{H_i(T), T\geq 0\}$, $\bar{h}_i\equiv\sup\{H_i(T), T\geq 0\}$, $i=1,2$; and similarly define $\tilde{q}_i$, $\bar{q}_i$, $\tilde{g}_i$, $\bar{g}_i$ to be the corresponding inf and sup for Q_i and G_i. We assume that:

$$(1.2)\qquad \begin{aligned} &-\infty<\tilde{h}_1\leq H_1(T)\leq\bar{h}_1<\infty, \qquad 0<\tilde{h}_2\leq H_2(T)\leq\bar{h}_2<\infty,\\ &0<\tilde{q}_i\leq Q_i(T)\leq\bar{q}_i<\infty, \qquad 0<\tilde{g}_i\leq G_i(T)\leq\bar{g}_i<\infty, \qquad i=1,2. \end{aligned}$$

Let $\lambda_1>0$ denote the first eigenvalue of the eigenvalue problem: $\Delta w+\lambda w=0$ in $\mathcal{D}$, $w=0$ in $\delta\mathcal{D}$, where $\omega(x)$ is the corresponding normalized eigenfunction with $\max\{\omega(x)| x\in\bar{\mathcal{D}}\}=1$. For positive integers n, $C^n(\mathcal{D})$ and $C^n(\bar{\mathcal{D}})$ denote n times continuously differentiable functions in $\mathcal{D}$ and $\bar{\mathcal{D}}$ respectively.

In §2, we consider equations (1.1) in $\mathcal{D}$ with nonnegative or zero Dirichlet boundary conditions on $\delta\mathcal{D}$. Theorem 2.1 finds conditions when "nontrivial" nonnegative solutions cannot exist (such conditions for the corresponding time-dependent parabolic problem would imply solutions, with nonnegative initial conditions, "blow up" as $t\to+\infty$). Theorem 2.3 finds other conditions when "nontrivial" nonnegative solutions cannot exist (such conditions for the corresponding time-dependent parabolic problem would imply solutions tending to zero as $t\to+\infty$). Theorem 2.4 gives necessary conditions for a nontrivial nonnegative equilibrium solution to exist.

Due to the difficulty explained at the end of §2, we consider a simpler model for two group neutron-flux in §3, when the temperature feedback is prompt. Theorem 3.1 finds sufficient conditions for the existence of nontrivial nonnegative equilibrium solution. Theorem 3.2 gives other conditions when the corresponding time-dependent parabolic problem would have nonnegative solutions tending to zero as $t\to+\infty$.

2. Temperature dependent equations, conditions for positive equilibrium. We first find conditions which imply that "nontrivial" nonnegative solutions cannot be finite, and hence cannot exist. We can interpret these as "blow-up" conditions.

THEOREM 2.1. *Suppose that*

$$(2.1)\qquad \tilde{h}_1\equiv\inf\{H_1(s)|s\geq 0\}<\lambda_1, \quad \textit{and}$$

$$(2.2)\qquad \tilde{q}_1\tilde{h}_2>(\lambda_1+\bar{q}_2)(\lambda_1-\tilde{h}_1).$$

Then (1.1) *with boundary conditions*

$$u(x)=u^0(x)\geq 0, \quad v(x)=v^0(x)\geq 0, \quad T(x)=T^0(x)\geq 0$$

for $x\in\delta\mathcal{D}$ (*here*, $u^0(x), v^0(x), T^0(x)$ *are given continuous functions on* $\delta\mathcal{D}$) *has no solution* $(\hat{u}(x), \hat{v}(x), \check{T}(x))$ *with the properties that*:

(i) *each component is in* $C^2(\mathcal{D})\cap C^1(\bar{\mathcal{D}})$,

(ii) $\hat{u}(x)\geq 0$, $\hat{v}(x)\geq 0$, $\hat{T}(x)\geq 0$ *in* $\bar{\mathcal{D}}$,

(iii) $\hat{u}(x)\not\equiv 0$ *in* $\mathcal{D}$.

(*Consequently, if further* $v^0(x)\equiv 0$ *on* $\delta\mathcal{D}$, *then the only solution with properties* (i) *and* (ii) *is* $(0,0,\check{T}(x))$, *where* $\Delta\check{T}(x)-c\check{T}(x)=0$ *in* $\mathcal{D}$, $\check{T}(x)=T^0(x)\geq 0$ *on* $\delta\mathcal{D}$.)

Proof. Assume that $(\hat{u}(x),\hat{v}(x),\hat{T}(x))$ exist with properties (i), (ii) and (iii). We will construct a family of lower bounds for the solution, parametrized by $\delta>0$. As $\delta\to\infty$, the lower bounds will tend to ∞. First, let $\tilde{\tilde{h}}_1<\tilde{h}_1$ so that (2.2) is still satisfied with $\tilde{h}_1$ replaced by $\tilde{\tilde{h}}_1$. For each $\delta>0$, define $u_\delta(x)=\delta\omega(x)$, $v_\delta(x)=\delta\tilde{h}_2^{-1}(\lambda_1-\tilde{\tilde{h}}_1)\omega(x)$, $T_\delta(x)=\delta\tilde{g}_1(\lambda_1+c)^{-1}\omega(x)$ for $x\in\bar{\mathcal{D}}$. For all $v(x)\geq v_\delta(x)$, $T(x)\geq T_\delta(x)$, we have

$$\begin{aligned}(2.3)\quad &\Delta u_\delta(x)+H_1(T(x))u_\delta(x)+H_2(T(x))v(x)\\ &\qquad\geq -\lambda_1\delta\omega(x)+\tilde{h}_1\delta\omega(x)+\tilde{h}_2\delta\tilde{h}_2^{-1}\big(\lambda_1-\tilde{\tilde{h}}_1\big)\omega(x)>0\end{aligned}$$

in $\mathcal{D}$. For all $u(x)\geq u_\delta(x)$, $T(x)\geq T_\delta(x)$, we have

$$\begin{aligned}(2.4)\quad &\Delta v_\delta(x)+Q_1(T(x))u(x)-Q_2(T(x))v_\delta(x)\\ &\qquad\geq -\lambda_1\delta\tilde{h}_2^{-1}\big(\lambda_1-\tilde{\tilde{h}}_1\big)\omega(x)+\tilde{q}_1\delta\omega(x)-\bar{q}_2\delta\tilde{h}_2^{-1}\big(\lambda_1-\tilde{\tilde{h}}_1\big)\omega(x)\\ &\qquad=\delta\omega(x)\big[\tilde{q}_1-\tilde{h}_2^{-1}(\lambda_1+\bar{q}_2)\big(\lambda_1-\tilde{\tilde{h}}_1\big)\big]>0\end{aligned}$$

in $\mathcal{D}$. (By (2.2) and the choice of $\tilde{\tilde{h}}_1$). For all $u(x)\geq u_\delta(x)$, $v(x)\geq v_\delta(x)$, we have

$$\begin{aligned}(2.5)\quad &\Delta T_\delta(x)-cT_\delta(x)+G_1(T_\delta(x))u(x)+G_2(T_\delta(x))v(x)\\ &\qquad\geq -\delta\tilde{g}_1\omega(x)+\tilde{g}_1\delta\omega(x)+\tilde{g}_2v_\delta(x)=\tilde{g}_2v_\delta(x)>0\end{aligned}$$

in $\mathcal{D}$.

We now show that properties (i) to (iii) imply that $\hat{u}(x)>0$, $\hat{v}(x)>0$ and $\hat{T}(x)>0$ for $x\in\mathcal{D}$. Let $C>|\tilde{h}_1|$, we have $\Delta\hat{u}-C\hat{u}=-H_2(\hat{T})\hat{v}-(H_1(\hat{T})+C)\hat{u}\leq 0$ in $\mathcal{D}$, $\hat{u}\geq 0$ in $\bar{\mathcal{D}}$. The maximum principle implies that $\hat{u}(x)>0$ in $\mathcal{D}$, otherwise (iii) is violated. Considering the equation for $\hat{v}$, we can similarly find that either $\hat{v}(x)>0$ in $\mathcal{D}$ or $\hat{v}(x)\equiv 0$ in $\bar{\mathcal{D}}$. However (iii) implies that the trivial function is not a solution for the second equation in (1.1); hence $\hat{v}(x)>0$ in $\mathcal{D}$. Similarly we find $\hat{T}(x)>0$ in $\mathcal{D}$. Moreover, applying the maximum principle at the boundary (see e.g. [11, p. 67]), we find that outward normal derivatives $\partial/\partial\eta$ of $\hat{u},\hat{v}$ or $\hat{T}$ must be negative at those boundary points where the corresponding function is 0.

From the above paragraph, we see that for $\delta>0$ sufficiently small, $0<u_\delta(x)<\hat{u}(x)$, $0<v_\delta(x)<\hat{v}(x)$, $0<T_\delta(x)<\hat{T}(x)$ for $x\in\mathcal{D}$. Let $\mathcal{S}=\{s>0|\hat{u}(x)>u_t(x),\hat{v}(x)>v_t(x),\hat{T}(x)>T_t(x)$ for all $0\leq t\leq s, x\in\mathcal{D}\}$. Suppose $\mathcal{S}$ has an upper bound; let its lub be $\bar{\delta}$. There must be a point in $\mathcal{D}$ where $u_{\bar{\delta}}=\hat{u}$ or $v_{\bar{\delta}}=\hat{v}$, or $T_{\bar{\delta}}=\hat{T}$. Otherwise, we consider for $C>|\tilde{h}_1|$ that

(2.6)

$$\begin{aligned}\Delta(\hat{u}-u_{\bar{\delta}})-C(\hat{u}-u_{\bar{\delta}})&=\{\Delta\hat{u}+H_1(\hat{T})\hat{u}+H_2(\hat{T})\hat{v}\}-\{\Delta u_{\bar{\delta}}+H_1(\hat{T})u_{\bar{\delta}}+H_2(\hat{T})\hat{v}\}\\ &\quad-\{H_1(\hat{T})+C\}(\hat{u}-u_{\bar{\delta}})<0\end{aligned}$$

in $\mathfrak{D}$. (The last inequality is true because $\hat{T}(x) \geq T_{\bar{\delta}}(x)$, $\hat{v}(x) \geq v_{\bar{\delta}}(x)$ and inequality (2.3) can be applied with $\delta = \bar{\delta}$). Together with $\hat{u} - u_{\bar{\delta}} \geq 0$ in $\bar{\mathfrak{D}}$, this implies that $\partial \hat{u}/\partial \eta < \partial u_{\bar{\delta}}/\partial \eta$ at those points at the boundary where $\hat{u} = u_{\bar{\delta}}$. Consequently, for sufficiently small $\varepsilon > 0$, $u_{\bar{\delta}+\varepsilon}(x) < \hat{u}(x)$ for all $x \in \mathfrak{D}$. Similarly we deduce from inequalities (2.4) and (2.5) that

$$\Delta(\hat{v} - v_{\bar{\delta}}) - Q_2(\hat{T})(\hat{v} - v_{\bar{\delta}}) < 0, \tag{2.7}$$

$$\Delta(\hat{T} - T_{\bar{\delta}}) - (c + D)(\hat{T} - T_{\bar{\delta}}) < 0 \tag{2.8}$$

in $\mathfrak{D}$ (where $D \geq [\max_{x \in \bar{\mathfrak{D}}} \hat{u}(x)] \cdot K_1 + [\max_{x \in \bar{\mathfrak{D}}} \hat{v}(x)] \cdot K_2$, and K_i are respectively Lipschitz constants for $G_i(T)$ for $T \geq 0, i = 1, 2$). As before, we have $v_{\bar{\delta}+\varepsilon}(x) < \hat{v}(x)$ and $T_{\bar{\delta}+\varepsilon}(x) < \hat{T}(x)$ for all $x \in \mathfrak{D}$, if $\varepsilon > 0$ is small enough. This violates the definition of $\bar{\delta}$, and we conclude that there must be a point in $\mathfrak{D}$ where $u_{\bar{\delta}} = \hat{u}$ or $v_{\bar{\delta}} = \hat{v}$ or $T_{\bar{\delta}} = \hat{T}$.

Suppose $u_{\bar{\delta}}(\bar{x}) = \hat{u}(\bar{x})$, $\bar{x} \in \mathfrak{D}$. Inequality (2.6) and maximum principle again implies that $u_{\bar{\delta}}(x) \equiv \hat{u}(x)$ in $\bar{\mathfrak{D}}$. This implies that $0 = \Delta \hat{u} + H_1(\hat{T})\hat{u} + H_2(\hat{T})\hat{v} = \Delta u_{\bar{\delta}} + H_1(\hat{T}) u_{\bar{\delta}} + H_2(\hat{T})\hat{v} > 0$. The last inequality is true by letting $\delta = \bar{\delta}$ and $v = \hat{v}$, $T = \hat{T}$ in (2.3). Suppose $v_{\bar{\delta}}(\bar{x}) = \hat{v}(\bar{x})$ or $T_{\bar{\delta}}(\bar{x}) = \hat{T}(\bar{x})$, $\bar{x} \in \mathfrak{D}$, we use inequalities (2.7) or (2.8) to conclude $v_{\bar{\delta}}(x) \equiv \hat{v}(x)$ or $T_{\bar{\delta}}(x) \equiv \hat{T}(x)$ in $\bar{\mathfrak{D}}$ respectively. In the first case, $0 = \Delta \hat{v} + Q_1(\hat{T})\hat{u} - Q_2(\hat{T})\hat{v} = \Delta v_{\bar{\delta}} + Q_1(\hat{T})\hat{u} + H_2(\hat{T})\hat{v}_{\bar{\delta}} > 0$ by (2.4); and in the second case, $0 = \Delta \hat{T} - c\hat{T} + G_1(\hat{T})\hat{u} + G_2(\hat{T})\hat{v} = \Delta T_{\bar{\delta}} - cT_{\bar{\delta}} + G_1(T_{\bar{\delta}})\hat{u} + G_2(T_{\bar{\delta}})\hat{v} > 0$. These contradictions show that the set $\mathfrak{S}$ is unbounded. However, as $\delta \to +\infty$, $u_\delta(x)$, $v_\delta(x)$ and $T_\delta(x)$ tends to $+\infty$. This proves the nonexistence of $(\hat{u}(x), \hat{v}(x), \hat{T}(x))$.

COROLLARY 2.2. *Suppose that the assumptions* (2.1) *and* (2.2) *are replaced by the single condition*

$$\tilde{h}_1 \equiv \inf\{H_1(s) | s \geq 0\} \geq \lambda_1. \tag{2.9}$$

Then the boundary value problem described in Theorem 2.1 *has no solution with properties as described there.*

Proof. Choose a positive constant $k, 0 < k < \lambda_1$ so that

$$\tilde{q}_1 \tilde{h}_2 > (\lambda_1 + \bar{q}_2)(\lambda_1 - k).$$

Define $u_\delta(x)$ and $T_\delta(x)$ as in the proof of Theorem 21. Redefine

$$v_\delta(x) = \delta \tilde{h}_2^{-1}(\lambda_1 - k)\omega(x) \quad \text{in } \bar{\mathfrak{D}}.$$

We have, for all $v(x) \geq v_\delta(x)$, $T(x) \geq T_\delta(x)$

$$\Delta u_\delta(x) + H_1(T(x)) u_\delta(x) + H_2(T(x)) v(x) \geq -\lambda_1 \delta \omega(x) + \tilde{h}_1 \delta \omega(x) + \tilde{h}_2 v_\delta(x) > 0$$

in $\mathfrak{D}$, $\delta > 0$ (because $\tilde{h}_1 \geq \lambda_1$). Inequality (2.4) remains true with $\tilde{\tilde{h}}_1$ everywhere replaced by k. The rest of the proof will be exactly the same as Theorem 2.1.

We next find conditions which imply that nonnegative solutions have to be identically zero. We can interpret these as "decay" conditions.

THEOREM 2.3. *Suppose that*

$$\bar{h}_1 = \sup\{H_1(s) | s \geq 0\} < \lambda_1, \quad \textit{and} \tag{2.10}$$

$$\bar{q}_1 \bar{h}_2 < (\tilde{q}_2 + \lambda_1)(\lambda_1 - \bar{h}_1). \tag{2.11}$$

Then (1.1) *with boundary conditions*

$$u(x) = v(x) = T(x) = 0, \quad \textit{for } x \in \delta \mathfrak{D},$$

has the solution $(0,0,0)$ *as the only solution with the properties that each component is in* $C^2(\mathcal{D})\cap C^1(\bar{\mathcal{D}})$ *and nonnegative in* $\bar{\mathcal{D}}$.

Proof. We will consider a parabolic system related to (1.1):

$$\begin{aligned}&\frac{\partial\tilde{u}}{\partial t}=\Delta\tilde{u}+H_1(\tilde{T})\tilde{u}+H_2(\tilde{T})\tilde{v},\\ &\frac{\partial\tilde{v}}{\partial t}=\Delta\tilde{v}+Q_1(\tilde{T})\tilde{u}-Q_2(\tilde{T})\tilde{v}, \qquad \text{in } \mathcal{D}\times(0,\infty),\\ &\frac{\partial\tilde{T}}{\partial t}=\Delta\tilde{T}-c\tilde{T}+G_1(\tilde{T})\tilde{u}+G_2(\tilde{T})\tilde{v},\end{aligned}\tag{2.12}$$

$$\tilde{u}(x,t)=\tilde{v}(x,t)=\tilde{T}(x,t)=0, \qquad (x,t)\in\delta\mathcal{D}\times[0,\infty).\tag{2.13}$$

Here, $\tilde{u},\tilde{v},\tilde{T}$ are function in $\bar{\mathcal{D}}\times[0,\infty)$. We will prove that all solutions of (2.12), (2.13) with components in $C^2(\mathcal{D}\times(0,\infty))\cap C^1(\bar{\mathcal{D}}\times[0,\infty))$, and initial conditions which are nonnegative for all $x\in\bar{\mathcal{D}}$, $t=0$, will tend to zero as $t\to+\infty$. Consequently, the equilibrium solution as stated in the theorem can only be the trivial one, $(0,0,0)$.

We proceed to utilize some differential inequalities. Hypothesis (2.11) implies that there is some $\sigma, 0<\sigma<c$, so that $(\tilde{q}_2+\lambda_1-\sigma)\bar{q}_1^{-1}>\bar{h}_2(\lambda_1-\bar{h}_1-\sigma)^{-1}$. Let C_1, C_2 be positive constants so that $(\tilde{q}_2+\lambda_1-\sigma)\bar{q}_1^{-1}>C_1C_2^{-1}>\bar{h}_2(\lambda_1-\bar{h}_1-\sigma)^{-1}$. Let $k>0$ be a constant such that $kC_1\omega(x)\geq\tilde{u}(x,0)$ and $kC_2\omega(x)\geq\tilde{v}(x,0)$, for each $x\in\bar{\mathcal{D}}$. Choose $C_3>0$ so that $C_3>\max\{(\bar{g}_1kC_1+\bar{g}_2kC_2)(c-\sigma)^{-1},\max_{x\in\bar{\mathcal{D}}}\tilde{T}(x,0)\}$. Finally, define $\underline{w}_1\equiv\underline{w}_2\equiv\underline{w}_3\equiv0$, $\bar{w}_1=kC_1\omega(x)e^{-\sigma t}$, $\bar{w}_2=kC_2\omega(x)e^{-\sigma t}$ and $\bar{w}_3=C_3e^{-\sigma t}$, for all $(x,t)\in\bar{\mathcal{D}}\times[0,\infty)$. Let

$$\begin{aligned}J\equiv\{(x,t,\check{u},\check{v},\check{T})|(x,t)\in\mathcal{D}\times(0,\infty),\ &\underline{w}_1(x,t)\leq\check{u}\leq\bar{w}_1(x,t),\\ &\underline{w}_2(x,t)\leq\check{v}\leq\bar{w}_2(x,t),\ \underline{w}_3(x,t)\leq\check{T}\leq\bar{w}_3(x,t)\}.\end{aligned}$$

Clearly, we have

$$\Delta\underline{w}_1+H_1(\check{T})\underline{w}_1+H_2(\check{T})\check{v}-\frac{\partial}{\partial t}(\underline{w}_1)=H_2(\check{T})\check{v}\geq0,\tag{2.14}$$

$$\Delta\underline{w}_2+Q_1(\check{T})\check{u}-Q_2(\check{T})\underline{w}_2-\frac{\partial}{\partial t}(\underline{w}_2)=Q_1(\check{T})\check{u}\geq0,\tag{2.15}$$

$$\Delta\underline{w}_3-c\underline{w}_3+G_1(\underline{w}_3)\check{u}+G_2(\underline{w}_3)\check{v}-\frac{\partial}{\partial t}(\underline{w}_3)=G_1(\underline{w}_3)\check{u}+G_2(\underline{w}_3)\check{v}\geq0,\tag{2.16}$$

for all $(x,t,\check{u},\check{v},\check{T})\in J$. On the other hand,

$$\begin{aligned}&\Delta\bar{w}_1+H_1(\check{T})\bar{w}_1+H_2(\check{T})\check{v}-\frac{\partial}{\partial t}(\bar{w}_1)\\ &\qquad\leq k\omega(x)e^{-\sigma t}\{[-\lambda_1+\bar{h}_1+\sigma]C_1+\bar{h}_2C_2\}<0,\end{aligned}\tag{2.17}$$

$$\begin{aligned}&\Delta\bar{w}_2+Q_1(\check{T})\check{u}-Q_2(\check{T})\bar{w}_2-\frac{\partial}{\partial t}(\bar{w}_2)\\ &\qquad\leq k\omega(x)e^{-\sigma t}\{[-\lambda_1-\tilde{q}_2+\sigma]C_2+\bar{q}_1C_1\}<0,\end{aligned}\tag{2.18}$$

$$\begin{aligned}&\Delta\bar{w}_3-c\bar{w}_3+G_1(\bar{w}_3)\check{u}+G_3(\bar{w}_3)\check{v}-\frac{\partial}{\partial t}(\bar{w}_3)\\ &\qquad\leq e^{-\sigma t}\Big\{(-c+\sigma)C_3+\bar{g}_1\max_{x\in\bar{\mathcal{D}}}\omega(x)kC_1+\bar{g}_2\max_{x\in\bar{\mathcal{D}}}\omega(x)kC_2\Big\}<0,\end{aligned}\tag{2.19}$$

for all $(x,t,\check{u},\check{v},\check{T})\in J$, because of the choice of C_1, C_3, C_3. Moreover, by assumption and the choice of k and C_3, we have

$$\underline{w}_1(x,0)\le\tilde{u}(x,0)\le\bar{w}_1(x,0),\quad \underline{w}_2(x,0)\le\tilde{v}(x,0)\le\bar{w}_2(x,0),$$
$$\underline{w}_3(x,0)\le\tilde{T}(x,0)\le\bar{w}_3(x,0) \tag{2.20}$$

for $x\in\bar{\mathcal{D}}$, and

$$\underline{w}_1(x,t)=\tilde{u}(x,t)=\bar{w}_1(x,t),\quad \underline{w}_2(x,t)=\tilde{v}(x,t)=\bar{w}_2(x,t),$$
$$\underline{w}_3(x,t)=\tilde{T}(x,t)<\bar{w}_3(x,t) \tag{2.21}$$

for $(x,t)\in\delta\mathcal{D}\times[0,\infty)$. Therefore if such a solution $(\tilde{u}(x,t),\tilde{v}(x,t),\tilde{T}(x,t))$ exists in $\bar{\mathcal{D}}\times[0,\infty)$, it will satisfy

$$\underline{w}_1(x,t)\le\tilde{u}(x,t)\le\bar{w}_1(x,t),\quad \underline{w}_2(x,t)\le\tilde{v}(x,t)\le\bar{w}_2(x,t),$$
$$\underline{w}_3(x,t)\le\tilde{T}(x,t)\le\bar{w}_3(x,t) \tag{2.22}$$

for all $(x,t)\in\bar{\mathcal{D}}\times[0,\infty)$, by inequalities (2.14) to (2.21). (See e.g. [4, Cor. 1] or [10, Lemma 2.1] for a variant of the comparison principles used here.)

A solution $(\hat{u}(x),\hat{v}(x),\hat{T}(x))$ of the boundary value problem described in the statement of the theorem, with properties as stated, will be a solution of (2.12), (2.13) with the appropriate smoothness and nonnegativity condition at $t=0$. Let $(\tilde{u}(x,t),\tilde{v}(x,t),\tilde{T}(x,t))=(\hat{u}(x),\hat{v}(x),\hat{T}(x))$, $(x,t)\in\bar{\mathcal{D}}\times[0,\infty)$. Inequalities (2.22) imply that

$$0\le\hat{u}(x)\le kC_1\omega(x)e^{-\sigma t},\quad 0\le\hat{v}(x)\le kC_2\omega(x)e^{-\sigma t},\quad 0\le\hat{T}(x)\le C_3e^{-\sigma t}$$

for $(x,t)\in\bar{\mathcal{D}}\times[0,\infty)$. Consequently, $(\hat{u}(x),\hat{v}(x),\hat{T}(x))\equiv(0,0,0)$.

Remarks. In the proof of Theorem 2.3, we show that a solution of (2.12), (2.13) with nonnegative initial conditions and appropriate smoothness properties will tend to zero uniformly as $t\to+\infty$ (under assumptions (2.10), (2.11)). Similarly, adapting this proof and that of Theorem 2.1, we should be able to show that such a solution of (2.12), (2.13), (under assumptions (2.1) and (2.2)) will tend to $+\infty$ as $t\to+\infty$.

Summarizing Theorem 2.1, Corollary 2.2 and Theorem 2.3, and reversing our point of view, we obtain the following theorem.

THEOREM 2.4. *Suppose that* (1.1) *with boundary conditions*

$$u(x)=v(x)=T(x)=0 \quad \textit{for } x\in\delta\mathcal{D} \tag{2.23}$$

has a solution $(\underline{\hat{u}}(x),\hat{v}(x),\hat{T}(x))\not\equiv(0,0,0)$ *with each component in* $C^2(\mathcal{D})\cap C^1(\bar{\mathcal{D}})$ *and nonnegative in* $\mathcal{D}$. *Then all the following conditions must be satisfied*:

(i) $\tilde{h}_1<\lambda_1$,

(ii) $\tilde{q}_1\tilde{h}_2\le(\lambda_1+\bar{q}_2)(\lambda_1-\underline{h}_1)$,

(iii) $(\lambda_1+\tilde{q}_2)(\lambda_1-\bar{h}_1)\le\bar{q}_1\bar{h}_2$ *if* $\bar{h}_1<\lambda_1$.

In view of the similarity of the two inequalities in (ii) and (iii) above (one simply interchanges max and min of H_i,Q_i and reverses the order relation), it is interesting to consider the situation when all rates are constants.

COROLLARY 2.5. *Suppose that* $H_i(T)$, $Q_i(T)$, $i=1,2$ *are all constant functions, where* $H_i(T)=h_i$, $Q_i(T)=q_i$, $i=1,2$. *Then* (1.1) *with boundary conditions* (2.23) *has a solution* $(\hat{u},\hat{v},\hat{T})\not\equiv(0,0,0)$, *with each component* $C^2(\mathcal{D})\cap C^1(\bar{\mathcal{D}})$ *and nonnegative in* $\bar{\mathcal{D}}$ *if and only if*

$$q_1h_2=(\lambda_1+q_2)(\lambda_1-h_1) \quad \textit{and} \quad h_1<\lambda_1. \tag{2.24}$$

Under this situation, $\hat{u}(x)=k(\lambda_1+q_2)\omega(x)$, $\hat{v}(x)=kq_1\omega(x)$ for any constant $k>0$. (Note: Hypotheses (1.2) still apply.)

Proof. Suppose $h_1>\lambda_1$; then (i) in Theorem 2.4 implies that no such solution $(\hat{u},\hat{v},\hat{T})$ exists. If $h_1<\lambda_1$ and $q_1h_2\neq(\lambda_1+q_2)(\lambda_1-h_1)$, then (ii) and (iii) in Theorem 2.4 also imply that such a solution does not exist. Consequently condition (2.24) is necessary. Conversely, supposing (2.24) is satisfied, one easily sees that $\hat{u}(x)=k(\lambda_1+q_2)\omega(x)$, $\hat{v}(x)=kq_1\omega(x)$ are solutions of the first two equations of (1.1). For the third equation, we have

$$\Delta T-cT+G_1(T)k(\lambda_1+q_2)\omega(x)+G_2(T)kq_1\omega(x)=0 \quad \text{in } \mathfrak{D},$$

$T=0$ on $\delta\mathfrak{D}$, $(k>0)$. From the properties of G_i, $i=1,2$, one sees that a large constant function is an upper solution, and 0 is a lower solution. Hence there exists a solution $T=\hat{T}(x)\geq 0$, for the corresponding k.

Remarks. In view of the remarks following Theorem 2.3, and Theorem 2.4 it becomes apparent that nontrivial nonnegative solutions of (1.1) under homogeneous Dirichlet conditions are not likely to exist, unless $H_1(T)<\lambda_1$ for large values of T and $H_1(T)>\lambda_1$ for small values of T. One might also postulate that the other rates, e.g. $Q_2(T)$, vary significantly with T. However, sufficient conditions which prevent blow-up (like Theorem 2.1) and decay (like Theorem 2.2) are found to be very difficult to be simultaneously satisfied, so that nontrivial nonnegative equilibrium might exist. In the next section, a simpler set of equations is studied. The rate H_1 will be influenced directly by u without first changing T. This is known as "prompt" feedback. In fact, the third equation for temperature change will be eliminated, and reaction rates are assumed to be promptly affected by u and v. Under this situation, conditions for nontrivial nonnegative equilibrium are more readily found. The results in the next section should be helpful for suggesting simple further assumptions on (1.1) to produce desired results.

3. Prompt-feedback equations, equilibrium and stability. In this section, we consider a simpler model where the reaction rates (i.e. cross sections) are functions of u,v. In other words, the feedback is prompt, and does not have to be regulated through change in temperature indirectly. More precisely, we have

$$\text{(3.1)} \qquad \begin{aligned} &\Delta u+H_1(u,v)u+H_2(u,v)v=0,\\ &\Delta v+Q_1(u,v)u-Q_2(u,v)v=0, \end{aligned} \quad \text{in } \mathfrak{D},$$

$$\text{(3.2)} \qquad u(x)=v(x)=0, \qquad x\in\delta\mathfrak{D}.$$

The functions $uH_1(u,v)$, $vH_2(u,v)$, $uQ_1(u,v)$, $vQ_2(u,v)$ are assumed to belong to the class C^α in the first closed quadrant, i.e. they are locally Hölder continuous in (u,v) with Hölder exponent α, $0<\alpha<1$. Let $C^{2+\alpha}(\overline{\mathfrak{D}})$ denote the Banach space of real-valued functions continuous in $\overline{\mathfrak{D}}$, with first and second derivatives also continuous in $\overline{\mathfrak{D}}$, with finite value for the usual norm $|u|_{\mathfrak{D}}^{(2+\alpha)}$. We assume the boundary $\delta\mathfrak{D}$ belongs to class $C^{2+\alpha}$. (See e.g. [5] for details of these symbols.) The following three conditions will be selectively assumed in the following theorems.

(3.3) There are positive constants, h_2', h_2'', q_1', q_1'', q_2' and q_2'', such that for all (u,v) in the first closed quadrant:

$$0<h_2'\leq H_2(u,v)\leq h_2'', \quad 0<q_1'\leq Q_1(u,v)\leq q_1'', \quad 0<q_2'\leq Q_2(u,v)\leq q_2''.$$

(3.4) $H_1(0,0)>\lambda_1$; in the first closed quadrant $H_1(u,v)$ is continuously differentiable with respect to v, and $|\partial/\partial v[H_1(u,v)]|<K$ for all such (u,v), where K is some positive constant.

(3.5) There exist positive constants p and U^* such that $H_1(u,v)\leq -p$ for all $u\geq U^*$, $v\geq 0$.

THEOREM 3.1. *Assume* (3.3), (3.4), *and let there exist positive numbers U^* and p as described in* (3.5) *with p further satisfying*:

$$q_1''h_2''<q_2'p. \tag{3.6}$$

Then the boundary value problem (3.1), (3.2) *has a solution $(\hat{u}(x),\hat{v}(x))$ with components in $C^{2+\alpha}(\bar{\mathcal{D}})$ and $\hat{u}(x)>0$, $\hat{v}(x)>0$ in $\mathcal{D}$.*

Proof. We will construct upper and lower solutions for (3.1), (3.2) and apply a theorem in [13] to conclude the existence of solution. By (3.4), there is a small constant $k>0$ so that $H_1(u,0)>\lambda_1$ for $0<u<k$. Choose $0<\varepsilon<\min\{k,K^{-1}h_2'\}$, and $0<\delta<\varepsilon q_1'[\lambda_1+q_2'']^{-1}$. Define lower solutions as

$$u_1(x)=\varepsilon\omega(x), \qquad v_1(x)=\delta\omega(x), \tag{3.7}$$

for $x\in\bar{\mathcal{D}}$. Define upper solutions as

$$u_2(x)=U^*, \qquad v_2(x)=pU^*/(h_2'') \tag{3.8}$$

for $x\in\bar{\mathcal{D}}$. We now check the appropriate inequalities for u_i, v_i, $i=1,2$.

$$\Delta u_1+H_1(u_1,v)u_1+H_2(u_1,v)v\leq[-\lambda_1+H_1(\varepsilon\omega(x),v)]\varepsilon\omega(x)+h_2'v \tag{3.9}$$

for $v\geq 0$, $x\in\mathcal{D}$. However, $[-\lambda_1+H_1(\varepsilon\omega(x),0)]\varepsilon\omega(x)>0$ in $\mathcal{D}$ and the expression on the right side of inequality (3.9) is an increasing function of v, for $v\geq 0$, each $x\in\mathcal{D}$ (by the choice of ε). Consequently, we have

$$\Delta u_1(x)+H_1(u_1(x),v(x))u_1(x)+H_2(u_1(x),v(x))v(x)\geq 0 \tag{3.10}$$

for all $v(x)\geq v_1(x)>0$, $x\in\mathcal{D}$. For $v_1(x)$, we have the inequality

$$\begin{aligned}\Delta v_1(x)+Q_1(u(x),v_1(x))u(x)-Q_2(u(x),v_1(x))v_1(x)&\\ \geq[-\lambda_1-q_2'']\delta\omega(x)+q_1'u(x)&>0\end{aligned} \tag{3.11}$$

in $\mathcal{D}$, for all $u(x)>u_1(x)=\varepsilon\omega(x)$ (by the choice of δ). For $u_2(x)$, $v_2(x)$, we have

$$\begin{aligned}\Delta u_2(x)+H_1(u_2(x),v(x))u_2(x)+H_2(u_2(x),v(x))v(x)&\\ \leq -pU^*+h_2''[pU^*/h_2'']&=0\end{aligned} \tag{3.12}$$

in $\mathcal{D}$, for all $v_1(x)\leq v(x)\leq v_2(x)$; and

$$\begin{aligned}\Delta v_2(x)+Q_1(u(x),v_2(x))u(x)-Q_2(u(x),v_2(x))v_2(x)&\\ \leq q_1''U^*-q_2'[pU^*/h_2'']=U^*[q_1''-(q_2'p)/h_2'']&<0\end{aligned} \tag{3.13}$$

in $\mathcal{D}$, for all $u_1(x)\leq u(x)\leq u_2(x)$ (by hypothesis (3.6)). By [13], (3.10) to (3.13) imply that there is a solution $(\hat{u}(x),\hat{v}(x))$ as described in the statement of the theorem; moreover $\varepsilon\omega(x)\leq\hat{u}(x)\leq U^*$, $\delta\omega(x)\leq\hat{v}(x)\leq pU^*/(h_2'')$, in $\bar{\mathcal{D}}$.

Remarks. If we consider the initial boundary value problem:

$$(3.14)\qquad \left.\begin{aligned}\frac{\partial u}{\partial t}&=\Delta u+H_1(u,v)u+H_2(u,v)v,\\ \frac{\partial v}{\partial t}&=\Delta v+Q_1(u,v)u-Q_2(u,v)v,\end{aligned}\right\}\quad \text{in } \mathcal{D}\times(0,\infty).$$

$$(3.15)\qquad u(x,t)=v(x,t)=0, \qquad (x,t)\in\delta\mathcal{D}\times[0,\infty),$$

$$(3.16)\qquad u(x,0)=u_0(x),\quad v(x,0)=v_0(x),\quad x\in\overline{\mathcal{D}}.$$

Then inequalities (3.10) to (3.13) imply that any solution $(u(x,t),v(x,t))$ with components in $C^2(\mathcal{D}\times(0,\infty))\cap C^1(\overline{\mathcal{D}}\times[0,\infty))$ and

$$(3.17)\qquad u_1(x)\le u_0(x)\le u_2(x),\quad v_1(x)\le v_0(x)\le v_2(x),\quad x\in\overline{\mathcal{D}}$$

must satisfy

$$(3.18)\qquad u_1(x)\le u(x,t)\le u_2(x),\qquad v_1(x)\le v(x,t)\le v_2(x),$$

for all $(x,t)\in\overline{\mathcal{D}}\times[0,\infty)$. (See e.g. [4, Cor. 1]).

In case hypothesis (3.4) is violated, a theorem analogous to Theorem 2.3 can be readily proved, as follows.

THEOREM 3.2. *Assume* (3.3), *and there exists a constant* h_1'' *such that* $H_1(u,v)\le h_1''$ *for all* $u\ge 0$, $v\ge 0$, *with* h_1'' *satisfying*:

$$(3.19)\qquad h_1''<\lambda_1,\quad \textit{and}$$

$$(3.20)\qquad q_1''h_2''<(q_2'+\lambda_1)(\lambda_1-h_1'').$$

Then all solutions of the initial boundary value problem (3.14) *to* (3.16), *with components in* $C^2(\mathcal{D}\times(0,\infty))\cap C^1(\overline{\mathcal{D}}\times[0,\infty))$, *and* $u_0(x)\ge 0$, $v_0(x)\ge 0$ *in* $\mathcal{D}$, *must satisfy*:

$$(3.21)\qquad (u(x,t),v(x,t))\to(0,0)\quad \textit{as } t\to+\infty$$

uniformly for $x\in\overline{\mathcal{D}}$

Proof. Let $\sigma>0$ be sufficiently small so that $\lambda_1-h_1''-\sigma>0$, $q_2'+\lambda_1-\sigma>0$ and $q_1''(q_2'+\lambda_1-\sigma)^{-1}<(\lambda_1-h_1''-\sigma)/(h_2'')$. Let C_1, C_2 be positive constants, so that $q_1''(q_2'+\lambda_1-\sigma)^{-1}<C_2C_1^{-1}<(\lambda_1-h_1''-\sigma)/(h_2'')$. Finally, let $k>0$ be large enough so that $kC_1\omega(x)>u_0(x)$ and $kC_2\omega(x)>v_0(x)$ for all $x\in\overline{\mathcal{D}}$. Define $\underline{u}(x,t)\equiv\underline{v}(x,t)\equiv 0$ and $\bar{u}(x,t)=kC_1\omega(x)e^{-\sigma t}$, $\bar{v}(x,t)=kC_2\omega(x)e^{-\sigma t}$ for $(x,t)\in\overline{\mathcal{D}}\times[0,\infty)$. One readily verifies that

$$\Delta\underline{u}+H_1(\underline{u},v(x))\underline{u}+H_2(\underline{u},v(x))v(x)-\frac{\partial}{\partial t}(\underline{u})\ge 0$$

for $x\in\mathcal{D}$, all $\underline{v}(x)\le v(x)\le\bar{v}(x)$; and

$$\Delta\bar{u}(x)+H_1(\bar{u}(x),v(x))\bar{u}(x)+H_2(\bar{u}(x),v(x))v(x)-\frac{\partial}{\partial t}(\bar{u})$$
$$\le k\omega(x)e^{-\sigma t}[(-\lambda_1+h_1''+\sigma)C_1+h_2''C_2]<0$$

for $x\in\mathcal{D}$, all $\underline{v}(x)\le v(x)\le\bar{v}(x)$.

Analogous inequalities are satisfied by $\underline{v}$ and $\bar{v}$, and the remainder of the proof is similar to that of Theorem 2.3. Details are omitted.

Remarks. Under additional assumptions, it can be proved that the positive equilibrium found in Theorem 3.1 is stable, as a solution of (3.14) to (3.16). However, these assumptions are quite inelegant, and more natural conditions can probably be obtained through investigating the temperature feedback equations in §2.

REFERENCES

[1] C. Y. CHAN, *Existence, uniqueness, upper and lower bounds of solutions of nonlinear space-time nuclear reactor kinetics*, SIAM J. Appl. Math., 27 (1974), pp. 72–82.

[2] D. S. COHEN, *Positive solution of nonlinear eigenvalue problem: application to nonlinear reactor dynamics*, Arch. Rat. Mech. Anal., 26 (1967), pp. 305–315.

[3] J. J. DUDERSTADT AND L. J. HAMILTON, *Nuclear Reactor Analysis*, John Wiley, New York, 1976.

[4] P. C. FIFE AND M. M. TANG, *Comparison principles for reaction-diffusion systems: irregular comparison functions and applications to questions of stability and speed of propagation of disturbances*, J. Differential Equations, 40 (1981), pp. 168–185.

[5] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[6] S. GLASSTONE AND A. SESONSKE, *Nuclear Reactor Engineering*, Van Nostrand, New York, 1981.

[7] W. E. KASTENBERG, *A stability criterion for space-dependent nuclear-reactor systems with variable temperature feedback*, Nuc. Sci. Eng., 37 (1969), pp. 19–29.

[8] ———, *Comparison theorems for nonlinear multicomponent diffusion systems*, J. Math. Anal. Appl., 29 (1970), pp. 299–304.

[9] W. E. KASTENBERG AND P. L. CHAMBRÉ, *On the stability of nonlinear space-dependent reactor kinetics*, Nucl. Sci. Eng., 31 (1968), pp. 67–79.

[10] A. LEUNG, *Equilibrium and stabilities for competing-species reaction-diffusion equations with Dirichlet boundary data*, J. Math. Anal. Appl., 73 (1980), pp. 204–218.

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

[12] I. STAKGOLD AND L. E. PAYNE, *Nonlinear problems in nuclear reactor analysis*, Proc. Conference on Nonlinear Problems in Physical Sciences and Biology, Lecture Notes in Mathematics, 322, Springer, New York, 1972, pp. 298–307.

[13] L. Y. TSAI, *Nonlinear boundary value problems for systems of second order elliptic differential equations*, Bull. Inst. Math. Acad. Sinica, 5 (1977), pp. 157–165.

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics
006

INTERACTIONS OF FAST AND SLOW WAVES IN PROBLEMS WITH TWO TIME SCALES*

JOHN W. BARKER†

Abstract. We consider certain symmetric hyperbolic systems of nonlinear partial differential equations whose solutions vary on two time scales, a slow scale t and a fast scale t/ε. We show that if the initial data are not prepared correctly for the suppression of the fast scale motion, but contain errors of amplitude $O(\varepsilon^{\mu})$, so that only μ time derivatives of the solution are bounded at $t=0$, then fast waves of amplitude $O(\varepsilon^{\mu})$ will be present in the solution, but the error introduced in the slow scale motion by nonlinear interactions of these waves will be of amplitude only $O(\varepsilon^{2\mu})+O(\varepsilon^{\mu+1})$. This holds for any $\mu>0$, and therefore extends the earlier results of Kreiss. In consequence, the effects of the fast waves can be controlled more easily than was previously thought, and their neglect in some physical situations is partially justified.

1. Introduction. This paper is concerned with symmetric hyperbolic systems of partial differential equations which have solutions varying on two distinct time scales, a "slow" scale t, and a "fast" scale t/ε, where ε is some small parameter. Such systems arise in the description of several physical situations.

For example, the shallow water equations with Coriolis force, used in meteorology as a simple model for describing the large scale motion of the atmosphere in midlatitudes, are, in dimensionless form:

$$\begin{aligned} &u_t+uu_x+vu_y+\mathrm{Ro}^{-1}(\phi_x-fv)=0,\\ &v_t+uv_x+vv_y+\mathrm{Ro}^{-1}(\phi_y+fu)=0,\\ &\phi_t+(u\phi)_x+(v\phi)_y+\frac{\mathrm{Ro}}{\varepsilon^2}(u_x+v_y)=0, \end{aligned} \tag{1.1}$$

where (u,v) is the velocity field, ϕ the geopotential height, f the (nondimensionalized) Coriolis parameter, Ro the Rossby number, and ε the ratio of a typical fluid speed to $(gh_0)^{1/2}$, h_0 being the mean fluid depth.

This is a hyperbolic system, and can easily be put in symmetric form. Under the assumption that both Ro and ε are small, and that $\varepsilon=\mathrm{Ro}^{3/2}$, the linearized system has normal modes which fall into two classes, Rossby modes, which vary on the "slow" time scale of a day or so, and inertia-gravity modes, which oscillate on the "fast" scale of a few hours, t/ε. For a more detailed discussion of this system, the reader is referred to [4].

Other, better models for atmospheric motions in common use, such as the full equations for a compressible, isentropic fluid with Coriolis force, and the primitive equations, also describe motions on two time scales when the Rossby number is small. However, it should be noted that the primitive equations are not hyperbolic and therefore will not be covered by the results of this paper.

Now, it is widely believed that the inertia-gravity modes are of little importance in determining the weather, at least on time scales of a couple of days. They are certainly present in the atmosphere, but their amplitude is smaller than that of the Rossby modes by a factor of ε or ε^2. Thus, in numerical models of the system, it is considered sufficient to compute accurately only the Rossby motion.

*Received by the editors March 15, 1982, and in revised form March 9, 1983. This research was supported by the Office of Naval Research under contract N00014-80-C-0076.

†Department of Applied Mathematics, California Institute of Technology, Pasadena, California 91125.

Unfortunately, it has been observed that unless care is taken in choice of the initial data for the numerical model, large amplitude fast scale motion is excited early in the calculation, obscuring and possibly destroying the underlying slow scale motion. In particular, data obtained from measurement or observation of a real physical system will excite fast waves not present physically, because of inevitable errors in the collection process.

Various "initialization schemes" for eliminating these spurious fast scale waves have been proposed by meteorologists. These include the early schemes of Charney [6] and Phillips [17], based on quasi-geostrophic theory, the normal mode initialization scheme [7], [19] and its later nonlinear improvements based on the "two-timing" method [1], [2], [14], and the dynamic initialization scheme [15], [16]. A discussion of these schemes and some of their merits may be found in [4]. We shall be most concerned here with another scheme, proposed by Kreiss [4], [5], [12], [13] and based on his "bounded derivative principle". Most of the above-mentioned schemes can be regarded as special cases of this scheme, which is applicable in a wider range of problems than any other, including limited area forecasting with open boundaries, is based on rigorous mathematical theory, and is relatively simple to apply.

In brief, the bounded derivative method is based on the observation that time derivatives of the slow scale motion are $O(1)$, whereas those of the fast scale motion $\sim\varepsilon^{-1}$. Thus, solutions in which the fast scale motion is of amplitude $O(\varepsilon^p)$ must have p time derivatives bounded independently of ε at all times, in particular at $t=0$. Kreiss has shown, using fairly standard energy estimates, that if the initial data are adjusted to ensure that the solution and a number p of its time derivatives are $O(1)$ at $t=0$, then they will remain so on some finite time interval $[0,T]$, where T is independent of ε, i.e. the fast scale motion present in the solution will have amplitude $O(\varepsilon^p)$ on $[0,T]$. This result holds for any symmetric hyperbolic system under appropriate assumptions concerning the smoothness of the coefficients and the definition of the two time scales, which enable the equations to be written in a "normal form" whose structure is used in the proof. The reader is referred to [13] for details.

In nonlinear problems, the need to estimate a certain number of time derivatives of the solution simultaneously, in order to obtain a closed system, requires that $p\geq p_c=[n/2]+2$, where n is the number of space dimensions and $[r]$ is the largest integer less than or equal to r. In one space dimension then, $p_c=2$, in two or three, $p_c=3$. However, numerical experiments in a two-dimensional nonlinear case have found taking $p=2$ to be sufficient to control the fast scale motion [4].

Thus, the first object of this paper to see whether less restrictive conditions for controlling the fast scale motion can be deduced.

A second related question that we shall consider concerns the amplitude of the interactions between the fast and slow scale motions. The importance of this can be seen, for example, in the meteorological problem referred to above. Even if the weather at a particular time were determined solely by the Rossby wave field at that time, the developement of that field could be influenced by interactions with any inertia-gravity waves present. Again, numerical experiments with the shallow water equations [7] have indicated that such interactions may be small.

The third question that we shall be able to investigate concerns the behavior of the solution in the limit $\varepsilon\to 0$, in particular the conditions on the initial data which will ensure that solutions of the symmetric hyperbolic system will converge to solutions of the limiting ($\varepsilon=0$) system, which is usually not hyperbolic. This is of interest in a wide variety of physical and mathematical contexts, several of which were described in [11]. One example is the incompressible limit of a compressible fluid, where the small parameter is the Mach number, the fast scale motion is composed of acoustic waves, and the slow scale motion is essentially incompressible flow. Another is the limit of small Alfven number in magnetohydrodynamics, where the fast motion consists of Alfven waves. In both these cases, the governing equations form a symmetric hyperbolic system.

A particularly simple example of the magnetohydrodynamic case is discussed by Gustafsson [9]. This concerns a plasma, surrounded by vacuum and confined between two infinitely long cylindrical walls. The problem is assumed to be longitudinally uniform and radially symmetric, so reduces to a one-dimensional problem with governing equations

$$\begin{aligned} &\rho_t+v\rho_x+\rho v_x=0,\\ &v_t+\frac{a^2}{p}\rho_x+vv_x+\frac{B}{4\pi\mu\rho}B_x=0,\\ &B_t+Bv_x+vB_x=0, \end{aligned} \tag{1.2}$$

where ρ is the density, v the velocity, B the magnetic induction, a the sound speed, μ the permeability, and p the pressure, a given function of ρ. These equations are hyperbolic, and can easily be symmetrized. When the Alfven number $(v_0/B_0)(4\pi\mu\rho_0)^{1/2}$ is small, the system contains two time scales but, as in the meteorological case, only motion on the slow scale is of interest.

In the next section, we state our result. In §3, some lemmas necessary for its proof are stated and proved. The result itself is proved under some extra assumptions in §4, and without these in §5.

2. Statement of present result. For our study, we consider the following system:

$$\begin{aligned} &W_t=\frac{1}{\varepsilon}P_0\left(\frac{\partial}{\partial \mathbf{x}}\right)W+P_1\left(W,\mathbf{x},t,\varepsilon,\frac{\partial}{\partial \mathbf{x}}\right)+F(\mathbf{x},t,\varepsilon),\\ &W(\mathbf{x},0)=W_s(\mathbf{x},0)+\varepsilon^{\mu}f(\mathbf{x}),\\ &W(\mathbf{x}+2\pi\mathbf{e}_j,t)\equiv W(\mathbf{x},t),\qquad 1\le j\le n, \end{aligned} \tag{2.1a}$$

where $\mathbf{x}=(x_1,\cdots,x_n)$, $\mathbf{e}_j$ is the unit vector in the jth direction, and

$$\begin{aligned} &P_0\left(\frac{\partial}{\partial \mathbf{x}}\right)=\sum_{j=1}^{n}A_j\frac{\partial}{\partial x_j}+C,\\ &P_1\left(W,\mathbf{x},t,\varepsilon,\frac{\partial}{\partial \mathbf{x}}\right)=\sum_{j=1}^{n}\frac{\partial}{\partial x_j}\left[\Phi_j(W,\mathbf{x},t,\varepsilon)\right]+\Phi_0(W,\mathbf{x},t,\varepsilon). \end{aligned} \tag{2.1b}$$

We make the following assumptions:

(A1) $A_j=A_j^*$ and $\Phi_{jw}=\Phi_{jw}^*$, $1\le j\le n$; this ensures the system is symmetric and hyperbolic.

(A2) $C=-C^*$, and F and each Φ_j, $0\le j\le n$, is C^∞ in all its arguments, 2π-periodic in $\mathbf{x}$, and, together with its $\mathbf{x}$ and t derivatives, is bounded independently of ε and uniformly in W (at least in some neighbourhood of the solution); this ensures the solution does not grow on a time scale t/ε.

(A3) $W(\mathbf{x},t)$ is a solution of the first and third equations all of whose time derivatives are bounded independently of ε, $\mu>0$, and $f(\mathbf{x})$ and all its derivatives are 2π-periodic in $\mathbf{x}$ and bounded independently of ε; this ensures exactly μ time derivatives of the solution are bounded at $t=0$.

Here the large part of the spatial operator, P_0, consisting of all terms inversely proportional to ε, has been written separately from the rest, P_1, which is thus bounded independently of ε. We require one more, important, assumption:

(A4) each eigenvalue κ of the symbol $P_0(i\omega)$ is either zero for all ω or satisfies:

$$|\kappa(\omega)|\geq\lambda \tag{2.1c}$$

for some positive constant λ, independent of ε, for all ω, except possibly at a finite number of values where some of these eigenvalues may also pass through zero.

Some limitations of the system (2.1) should perhaps be noted here. Firstly, no coefficient of the time derivative of W in (2.1a) has been allowed, even though in many physical examples, including all those mentioned above except (1.2), a nonlinear coefficient arises because of the symmetry requirement. We shall discuss systems with such a "symmetrizer" in a later paper, and show that a similar, though slightly weaker, result holds. Secondly, P_0 has been assumed to have constant coefficients. Our results can probably be extended to the case where its coefficients depend on x and t without any severe difficulty (the normal forms developed by Kreiss [13] and Tadmor [18] would be needed), but this has not been attempted here. Thirdly, periodic boundary conditions have been assumed.

We can now state our main result.

THEOREM 1. *There exist constants* ε_0, K_0, K_1, δ *and* T, *each independent of* ε *and strictly positive*, *such that*, *under the assumptions* (A1)–(A4), *the solution of* (2.1) *satisfies*:

$$\begin{aligned}\|(W-W_s)(\cdot,t)\|&\leq\varepsilon^{\mu}K_0,\\ \|\overline{(W-W_s)}(\cdot,t)\|&\leq(\varepsilon^{2\mu}+\varepsilon^{\mu+1})K_1,\end{aligned} \tag{2.2}$$

for all $t\in[0,T]$ *and all* $\varepsilon<\varepsilon_0$, *where*

$$\overline{(W-W_s)}(\mathrm{x},t)=\int_{t-\delta}^{t+\delta}(W-W_s)(\mathrm{x},\tau)\,d\tau$$

and $\|\cdot\|$ *is the usual* L_2*-norm.*

If we now note that a function that "varies only on the slow time scale" is one whose time derivatives are all bounded independently of ε, whereas a function that "oscillates on the fast time scale" can be characterized by the existence of some constant δ, independent of ε but with $\varepsilon\ll\delta\ll1$, such that for any (x,t):

$$\int_{t-\delta}^{t+\delta}w(x,t)\,dt=O(\varepsilon w(x,t)), \tag{2.3}$$

then we may restate the theorem, perhaps more illuminatingly, by saying that the solution of (2.1) has the form

$$W(\mathrm{x},t)=W_s(\mathrm{x},t)+\varepsilon^{\mu}\tilde{w}(\mathrm{x},t)+O(\varepsilon^{2\mu})+O(\varepsilon^{\mu+1}), \tag{2.4}$$

where $\tilde{w}$ is $O(1)$ and oscillates on the fast time scale.

Finally, the theorem may be expressed in words as follows: if the initial data for (2.1) are not chosen correctly for the suppression of the fast scale motion, but contain errors of amplitude $O(\varepsilon^{\mu})$ (so that only μ time derivatives are bounded initially), for any $\mu>0$, then fast scale motion of amplitude $O(\varepsilon^{\mu})$ will be present in the solution, but the resulting change in the slow scale motion will be of amplitude only $O(\varepsilon^{2\mu})+O(\varepsilon^{\mu+1})$, on some $O(1)$ time interval $[0,T]$.

This result is not trivial since the system (2.1) is nonlinear, and it might be thought that "nonlinear resonance" could occur, which would lead to an $O(\varepsilon^{\mu})$ change in the slow scale motion. For example, in the O.D.E. system

$$u_t=\frac{i}{\varepsilon}\lambda u,\quad v_t=\frac{i}{\varepsilon}\mu v,\quad w_t=(uv)^{1/2},\qquad u(0)=\varepsilon u_0,\quad v(0)=\varepsilon v_0,\quad w(0)=0 \tag{2.5}$$

the solution has

$$w(t)=\varepsilon(u_0v_0)^{1/2}\int_0^t e^{(\lambda+\mu)\tau/2\varepsilon}\,d\tau, \tag{2.6}$$

which is $O(\varepsilon^2)$, unless $\lambda+\mu=0$, when resonance occurs and w both is $O(\varepsilon)$ and varies on the slow time scale. Our result is that for the system (2.1) this does not happen.

In regard to the three questions discussed in the introduction, the consequences of our result are thus:

(i) The slow scale motion can be computed with error only $O(\varepsilon^{2\mu})+O(\varepsilon^{\mu+1})$ by choosing initial data so that only μ time derivatives are bounded at $t=0$, for any $\mu>0$. This would require carrying the fast scale motion along in the computation of course, and filtering it out at the end, so this may not be useful.

(ii) If fast scale motions of amplitude $O(\varepsilon^{\mu})$ are in fact present in the physical system being modelled, they can be omitted from the computed solution without introducing an error greater than $O(\varepsilon^{2\mu})+O(\varepsilon^{\mu+1})$ in the slow scale motion.

(iii) If the initial data are within $O(\varepsilon^{\mu})$ of some values that completely suppress the fast motion, then, as $\varepsilon\to 0$, the solution of the hyperbolic system will converge to a solution of the limiting ($\varepsilon=0$) equations with the limiting initial data, and at any ε, the difference between the solution and the limiting solution will be $O(\varepsilon^{\mu})$.

The proof of Theorem 1, given in the following sections, is really quite simple, relying only on a straightforward asymptotic expansion of the solution and standard energy estimates.

3. Some required lemmas.

The first and most important lemma we need states that the x-derivatives of the solution are bounded independently of ε, even if the t-derivatives are not.

LEMMA 3.1. *Suppose $W(\mathrm{x},t)$ satisfies the symmetric hyperbolic system* (2.1) *under assumptions* (A1)–(A3), *except that $\mu=0$ is also allowed. Then there exist constants T,ε_0 and K_{rs}, independent of ε, such that*

$$\left\|\frac{\partial^{r+s}W}{\partial x_1^{r_1}\cdots\partial x_n^{r_n}\partial t^s}(\cdot,t)\right\|\le\varepsilon^{-s}K_{rs} \tag{3.1}$$

for all $t\in[0,T]$, all $\varepsilon\le\varepsilon_0$, and all nonnegative $r_1,\cdots,r_n$, $r,s,r_1+\cdots+r_n=r$.

Proof. This is a standard result, and can be proved using simple energy estimates [5].

LEMMA 3.2. *Suppose $w(x,t)$ satisfies*

$$w_t=\frac{1}{\varepsilon}Aw_x+[B(x,t)w]_x,\qquad w(x,0)=f(x),\qquad w(x+2\pi,t)\equiv w(x,t) \tag{3.2}$$

where A is a constant, symmetric, nonsingular matrix, B is symmetric, both B and f are 2π-periodic in x and C^∞ in their arguments, f satisifes

$$\int_0^{2\pi} f(x)\,dx=0,$$

and there exist constants P_{rs}, R_s, and ε_1, independent of ε, such that

$$\left|\frac{\partial^{r+s}B}{\partial x^r\partial t^s}(\cdot,t)\right|\le P_{rs}, \qquad \left\|\frac{\partial^s f}{\partial x^s}\right\|\le R_s$$

for all nonnegative r,s and all $\varepsilon\le\varepsilon_1$.

Then there exist constants K_1, ε_0 and δ, independent of ε, such that

$$\|\bar{w}(\cdot,t)\|\le\varepsilon K_1 \tag{3.3}$$

for all $\varepsilon\le\varepsilon_0$, where

$$\bar{w}(x,t)=\int_{t-\delta}^{t+\delta} w(x,\tau)\,d\tau.$$

That is, w is oscillates on the fast time scale.

Proof. By the previous lemma and Sobolev's inequality, $|w|_\infty$ is $O(1)$. Let

$$P_0 w\equiv Aw_x+\varepsilon[B(x,t)w]_x.$$

Now, P_0^{-1} exists and is bounded independently of ε on the space S_p^0 of 2π-periodic, once-differentiable functions with zero mean value over x. Since (3.2) has conservation form, and f has zero mean value, w belongs to this space for all t. Also, since A is constant, P_0^{-1} commutes with $\partial/\partial t$ at leading order. Thus

$$w_t=\frac{1}{\varepsilon}P_0 w \Rightarrow w=\varepsilon P_0^{-1}w_t=\varepsilon\frac{\partial}{\partial t}P_0^{-1}w+O(\varepsilon^2 w).$$

Thus, for any δ with $\varepsilon\ll\delta\ll 1$

$$\int_{t-\delta}^{t+\delta} w(x,\tau)\,d\tau=\varepsilon\left[P_0^{-1}w\right]_{t-\delta}^{t+\delta}+O(\varepsilon^2 w)\le 2\varepsilon|P_0^{-1}||w|_\infty+O(\varepsilon^2 w)=O(\varepsilon),$$

whence the result follows.

LEMMA 3.3. *Suppose $w(x,t)$ satisfies*

$$w_t=\frac{1}{\varepsilon}Aw_x+[B(x,t)w]_x+F(x,t), \tag{3.4}$$
$$w(x,0)=0, \qquad w(x+2\pi,t)\equiv w(x,t),$$

where A and B are as in Lemma 3.2, F is C^∞ in x and t and 2π-periodic in x with

$$\int_0^{2\pi} F(x,t)\,dx=0, \tag{3.5a}$$

and there exist constants Q_{s0}, independent of ε, such that

$$\left\|\frac{\partial^s F}{\partial x^s}(\cdot,t)\right\|\le Q_{s0}, \qquad s=0,1 \tag{3.5b}$$

for all t. Then:

(i) *If F varies only on the slow time scale at leading order, i.e. if there exist constants* Q_{s1}, *independent of* ε, *such that*

$$\left\| \frac{\partial^{s+1}F}{\partial x^s \partial t}(\cdot,t) \right\| \le Q_{s1}, \qquad s=0,1 \tag{3.5c}$$

for all t, then for any fixed T, independent of ε, *there exist constants K and* ε_0, *independent of* ε, *such that*

$$\|w(\cdot,t)\| \le \varepsilon K \tag{3.6}$$

for all $t \in [0,T]$ *and all* $\varepsilon < \varepsilon_0$.

(ii) *If* $\bar{F}$ *oscillates on the fast time scale at leading order, i.e. if* $\|\bar{F}(\cdot,t)\| \le \varepsilon \bar{Q}$ *for some constant* $\bar{Q}$, *all t, then for any fixed T, independent of* ε, *there exist constants* K_0, K_1 *and* ε_0, *independent of* ε, *such that*

$$\|w(\cdot,t)\| \le K_0, \qquad \|\bar{w}(\cdot,t)\| \le \varepsilon K_1 \tag{3.7}$$

for all $t \in [0,T]$ *and all* $\varepsilon < \varepsilon_0$. *Thus, w also oscillates on the fast time scale at leading order.*

Proof. (i) Since A is constant and symmetric, B is symmetric, its derivatives are bounded, and the boundary conditions are periodic:

$$\frac{1}{2}\frac{d}{dt}\|w\|^2 = (w,(Bw)_x) + (w,F) \le \text{const}\left\{\|w\|^2 + \|F\|^2\right\},$$

$$\frac{1}{2}\frac{d}{dt}\|w_t\|^2 = (w_t,(Bw_t)_x) + (w_t,(B_t w)_x) + (w_t,F_t)$$
$$\le \text{const}\left\{\|w_t\|^2 + \|w\|^2 + \|F_t\|^2\right\}.$$

Since $\|F\|$, $\|F_t\|$, $w(x,0)$ and $w_t(x,0)$ are bounded independently of ε, it follows that $\|w_t\|$ is also so bounded on any $O(1)$ time interval $[0,T]$. Since $\|F_x\|$ is bounded independently of ε, it can be shown in similar manner that $\|w_{xt}(\cdot,t)\|$ is also $O(1)$ on $[0,T]$. Thus with P_0 as in the previous proof:

$$w_t = \frac{1}{\varepsilon}P_0 w + F \Rightarrow w = \varepsilon P_0^{-1}(w_t - F)$$
$$\Rightarrow |w|_\infty \le \varepsilon\,\text{const.}\left(|w_t|_\infty + |F|_\infty\right)$$
$$\le \varepsilon\,\text{const.}\left(\|w_t\| + \|w_{xt}\| + \|F\| + \|F_x\|\right) = O(\varepsilon)$$

using Sobolev's inequality. The result follows.

(ii) Here

$$w = \varepsilon P_0^{-1}(w_t - F)$$
$$\Rightarrow \int_{t-\delta}^{t+\delta} w(x,\tau)\,d\tau = \left[\varepsilon P_0^{-1} w\right]_{t-\delta}^{t+\delta} + O(\varepsilon^2 w)$$
$$- P_0^{-1}\int_{t-\delta}^{t+\delta} F(x,\tau)\,d\tau + O(\varepsilon F)$$
$$\Rightarrow \|\bar{w}\| = O(\varepsilon\|w\|) + O(\|\bar{F}\|) + O(\varepsilon\|F\|) = O(\varepsilon)$$

as required.

LEMMA 3.4. *Suppose $w(x,t)$ satisfies*

$$w_t=[B(x,t)w]_x+[B_1(x,t).F_t(x,t)]_x,\qquad (3.8)$$
$$w(x,0)=0,\qquad w(x+2\pi,t)\equiv w(x,t),$$

where B and F are as in the previous lemma (either case), except that (3.7b) *must hold for $s=0,1,2,3$, i.e. F has three $O(1)$ space derivatives, and B_1 is C^∞ in x and t, 2π-periodic in x, and, together with all its derivatives, is bounded independently of ε. Then there exist constants M, T and ε_0, independent of ε, such that for $\varepsilon<\varepsilon_0$ and $t\in[0,T]$*

$$|w(\cdot,t)|_\infty\leq M.\qquad (3.9)$$

Note this is true whatever the magnitude of the time derivatives of F, which may be $O(\varepsilon^{-1})$.

Proof. Let $w^{(1)}$ satisfy

$$w_t^{(1)}=[B_1(x,t).F_t(x,t)]_x,$$
$$w^{(1)}(x,0)=0,\qquad w^{(1)}(x+2\pi,t)\equiv w^{(1)}(x,t).$$

Then

$$w^{(1)}(x,t)=\left\{\int_0^t B_1(x,t)F_t(x,t)\,dt\right\}_x$$
$$=\left\{[B_1(x,t)F(x,t)]_0^t-\int_0^t B_{1t}(x,t)F(x,t)\,dt\right\}_x .$$

Thus, $\|w^{(1)}\|$ can be bounded in terms of $\|F\|$ and $\|F_x\|$ (and norms of B_1 and its derivatives of course). Differentiating with respect to x, it follows that $\|w_x^{(1)}\|$ can be bounded in terms of $\|F\|$, $\|F_x\|$ and $\|F_{xx}\|$, while $\|w_{xx}^{(1)}\|$ can be bounded in terms of these and $\|F_{xxx}\|$. By assumption, all these norms are $O(1)$, so $\|w^{(1)}\|$ certainly satisfies a bound of the form (3.9).

Let $w^{(1)}=w-w^{(1)}$:

$$w_t^{(2)}=[B(x,t)w^{(2)}]_x+[B(x,t)w^{(1)}]_x,$$
$$w^{(2)}(x,0)=0,\qquad w^{(2)}(x+2\pi,t)\equiv w^{(2)}(x,t).$$

By Duhamel's principle, and the bounds on B and its derivatives,

$$\|w^{(2)}(\cdot,t)\|\leq\text{const}\sup_{0\leq\tau\leq t}\left\{\|w^{(1)}(\cdot,\tau)\|+\|w_x^{(1)}(\cdot,\tau)\|\right\}$$

and:

$$\|w_x^{(2)}(\cdot,t)\|\leq\text{const}\sup_{0\leq\tau\leq t}\left\{\|w^{(1)}(\cdot,\tau)\|+\|w_x^{(1)}(\cdot,\tau)\|+\|w_{xx}^{(1)}(\cdot,\tau)\|\right\}$$

on any $O(1)$ time interval $[0,T]$. The result follows.

4. Proof of restricted form of Theorem 1. In this section, we prove Theorem 1 under the additional assumptions that:

(A5) $F=0$, i.e. the system is unforced for $t>0$.

(A6) $C=0$ and $\Phi_0=0$, so no undifferentiated terms appear and the system has conservation form.

(A7) $f \in S_p^0$, the space of 2π-periodic functions with zero mean value over x; coupled with (A6), this means that the mean value of the perturbation will be zero for all time.

(A8) It is possible to write $W=(U,V)^T$, where U consists of the fast scale variables and V of the slow scale variables.

(A9) The problem is restricted to one space dimension.

In consequence of these assumptions, (2.1) may be written as

$$\text{(4.1a)} \qquad \begin{aligned} &U_t = \frac{1}{\varepsilon} A U_x + [\Phi(U,V,x,t)]_x, \qquad V_t = [\Psi(U,V,x,t)]_x \\ &U(x+2\pi,t) \equiv U(x,t), \qquad V(x+2\pi,t) \equiv V(x,t) \end{aligned}$$

where (A1) and (A4) now allow us to assume A is nonsingular, as well as real symmetric, and imply that the matrix

$$\begin{pmatrix} \frac{1}{\varepsilon} A + \Phi_U & \Phi_V \\ \Psi_U & \Psi_V \end{pmatrix}$$

is real symmetric, while (A2) implies that Φ and Ψ are C^∞ functions of all their arguments, 2π-periodic in x, and, together with their x and t derivatives, are bounded independently of ε and uniformly in (U, V), at least in a neighborhood of the solution.

The requirement that p time derivatives of the solution of (4.1a) are bounded at $t=0$ places no restriction on $V(x,0)$, but determines $U(x,0)$ to within $O(\varepsilon^p)$. In fact, $U(x,t)=O(\varepsilon)$ in any solution with one or more bounded time derivatives. Thus, we assume $W_s=(\varepsilon U_s, V_s)^T$. Also, there is nothing to be gained from considering a perturbation in the smooth solution, so we take initial conditions:

$$\text{(4.1b)} \qquad U(x,0) = \varepsilon U_s(x,0) + \varepsilon^\mu f(x), \qquad V(x,0) = v_s(x,0)$$

where f satisfies (A3) and in view of (A7):

$$\text{(4.1c)} \qquad \int_0^{2\pi} f(x)\,dx = 0.$$

Theorem 1 now states that there exist constants ε_0, K_0, K_1, δ and T, each independent of ε and strictly positive, such that the solution of (4.1) satisfies

$$\text{(4.2)} \qquad \begin{aligned} &\|U(\cdot,t) - \varepsilon U_s(\cdot,t)\| \le \varepsilon^\mu K_0, \\ &\|\overline{U(\cdot,t) - \varepsilon U_s(\cdot,t)}\| + \|V(\cdot,t) - V_s(\cdot,t)\| \le (\varepsilon^{2\mu} + \varepsilon^{\mu+1}) K_1 \end{aligned}$$

for all $t \in [0,T]$ and all $\varepsilon < \varepsilon_0$, or alternatively that the solution of (4.1) has the form

$$\text{(4.3)} \qquad \begin{aligned} &U(x,t) = U_s(x,t) + \varepsilon^\mu \tilde{u}(x,t) + O(\varepsilon^{2\mu}) + O(\varepsilon^{\mu+1}), \\ &V(x,t) = V_s(x,t) + O(\varepsilon^{2\mu}) + O(\varepsilon^{\mu+1}) \end{aligned}$$

on some $O(1)$ time interval, where $\tilde{u}$ is $O(1)$ and oscillates on the fast time scale.

We shall now prove this. Let u,v be the perturbation in the solution:

$$\text{(4.4)} \qquad u(x,t) = U(x,t) - \varepsilon U_s(x,t), \qquad v(x,t) = V(x,t) - V_s(x,t).$$

The equations satisfied by u and v are

$$\begin{aligned}(4.5)\qquad & u_t=\frac{1}{\varepsilon}Au_x+[B_{11}(x,t)u+B_{12}(x,t)v]_x+[\varphi(u,v,x,t)]_x,\\ & v_t=[B_{21}(x,t)u+B_{22}(x,t)v]_x+[\psi(u,v,x,t)]_x,\\ & u(x,0)=\varepsilon^{\mu}f(x), \qquad v(x,0)=0,\\ & u(x+2\pi,t)\equiv u(x,t), \qquad v(x+2\pi,t)\equiv v(x,t),\end{aligned}$$

where $(B_{11}u+B_{12}v)$ is the linear part and ϕ the quadratic and higher part of

$$[\Phi(\varepsilon u_s+u,v_s+v,x,t)-\Phi(\varepsilon u_s,v_s,x,t)],$$

and $(B_{21}u+B_{22}v)$ is the linear part and ψ the quadratic and higher part of

$$[\Psi(\varepsilon u_s+u,v_s v,x,t)-\Psi(\varepsilon u_s,v_s,x,t)].$$

Each of B_{ij}, $i,j=1,2$, is C^{∞} in x and t, 2π-periodic in x, and, together with all its x and t derivatives, bounded independently of ε, for all sufficiently small ε. The same may be assumed of ϕ and ψ, with bounds uniform in u and v, since such bounds are needed only in a neighborhood of the solution, and ϕ and ψ may be altered elsewhere without affecting the solution.

In the system (4.5), we are hoping to show that v is an order in ε smaller than u. Therefore, as a first approximation we neglect v. Neglecting also nonlinear terms, since u is expected to be $O(\varepsilon^{\mu})$, let u_0 satisfy

$$\begin{aligned}(4.6)\qquad & u_{0t}=\frac{1}{\varepsilon}Au_{0x}+[B_{11}(x,t)u_0]_x,\\ & u_0(x,0)=\varepsilon^{\mu}f(x), \qquad u_0(x+2\pi,t)\equiv u_0(x,t).\end{aligned}$$

By Lemma 3.2,

$$(4.7)\qquad u_0(x,t)=\varepsilon^{\mu}\tilde{u}_0(x,t)$$

where $\tilde{u}_0$ oscillates on the fast time scale, but, together with all its x-derivatives, is $O(1)$. Also, (4.6) may be written as

$$\varepsilon u_{0t}=P_0u_0,$$

where $P_0=A\partial/\partial x+O(\varepsilon)$ is nonsingular on S_p^0, to which u_0 belongs. Therefore

$$(4.8)\qquad u_0=\varepsilon P_0^{-1}u_{0t}=\varepsilon\frac{\partial}{\partial t}(P_0^{-1}u_0)+\varepsilon P_0^{-2}P_{0_t}u_0=\varepsilon\frac{\partial}{\partial t}(P_0^{-1}u_0)+O(\varepsilon^2u_0),$$

since $P_0^{-1}=O(1)$ and $P_{0t}=\varepsilon[B_{11t}\partial/\partial x+B_{11xt}]=O(\varepsilon)$. Essentially, we have used the fact that P_0^{-1} and $\partial/\partial t$ commute at leading order.

Next let a first approximation to v be v_1, satisfying

$$\begin{aligned}(4.9)\qquad & v_{1t}=[B_{22}(x,t)v_1]_x+[B_{21}(x,t)u_0]_x+[\psi(u_0,0,x,t)]_x,\\ & v_1(x,0)=0, \qquad v_1(x+2\pi,t)\equiv v_1(x,t).\end{aligned}$$

Using (4.8), it can be seen that the linear forcing term in this equation is

$$(4.10)\qquad [B_{21}(x,t)u_0]_x=\varepsilon[B_{21}(x,t)\cdot(P_0^{-1}u_0)_t]_x+O(\varepsilon^2u_0).$$

Thus, if we note that P_0^{-1} is bounded, and from Lemma 3.1 that the x-derivative does not alter the order in ε, Lemma 3.4 implies that this term makes a contribution to v_1 of amplitude $O(\varepsilon u_0)=O(\varepsilon^{\mu+1})$. Also, the nonlinear forcing term $[\psi(u_0,0,x,t)]$ is of amplitude $O(u_0^2)=O(\varepsilon^{2\mu})$; by Duhamel's principle, it makes a contribution to v_1 of amplitude $O(\varepsilon^{2\mu})$. Thus:

$$v_1(x,t)=\varepsilon^{\mu+1}v_1^{(1)}(x,t)+\varepsilon^{2\mu}v_1^{(2)}(x,t), \tag{4.11}$$

where $v_1^{(1)}$ and $v_1^{(2)}$ are both bounded independently of ε, but may vary on both the fast and slow time scales at leading order.

Now return to the u equation, and let u_1 satisfy

$$\begin{gathered} u_{1t}=\frac{1}{\varepsilon}Au_{1x}+[B_{11}(x,t)u_1]_x+[B_{12}(x,t)v_1]_x+[\varphi(u_0,0,x,t)]_x, \\ u_1(x,0)=0, \qquad u_1(x+2\pi,t)\equiv u_1(x,t). \end{gathered} \tag{4.12}$$

By Lemma 3.1, the forcing terms are of the same order in ε as if they were not differentiated by x, and therefore they have amplitude $O(\varepsilon^{\mu+1})+O(\varepsilon^{2\mu})$. They may vary on both fast and slow time scales at these orders in ε, but, by Lemma 3.3, terms that vary only on the slow time scale make a contribution to the solution that is smaller by a factor of ε than the terms themselves. Thus it is sufficient to solve

$$\begin{gathered} u_{1t}=\frac{1}{\varepsilon}Au_{1x}+[B_{11}(x,t)u_1]_x+(I-S)[B_{12}(x,t)v_1+\varphi(u_0,0,x,t)]_x, \\ u_1(x,0)=0, \qquad u_1(x+2\pi,t)\equiv u_1(x,t) \end{gathered} \tag{4.13}$$

where S is the time-averaging operator, given by

$$Sw(t)=\int_{t-\delta}^{t+\delta}w(\tau)\,d\tau$$

for some δ with $\varepsilon\ll\delta\ll 1$. This has a solution

$$u_1(x,t)=\varepsilon^{\mu+1}\tilde{u}_1^{(1)}(x,t)+\varepsilon^{2\mu}\tilde{u}_1^{(2)}(x,t)+\text{lower order terms}, \tag{4.14}$$

where both $\tilde{u}_1^{(1)}$ and $\tilde{u}_1^{(2)}$ are $O(1)$ and oscillate on the fast time scale.

Also from (4.13), arguing as from (4.6):

$$\begin{gathered} \varepsilon u_{1t}=P_0u_1+\varepsilon G \quad \text{(say)} \\ \Rightarrow u_1=\varepsilon P_0^{-1}u_{1t}-\varepsilon P_0^{-1}G=\varepsilon\frac{\partial}{\partial t}(P_0^{-1}u_0)-\varepsilon P_0^{-1}G+O(\varepsilon^2u_1). \end{gathered} \tag{4.15}$$

Here, G is $O(\varepsilon^{2\mu})+O(\varepsilon^{\mu+1})$. Thus, the next approximation to the v equation,

$$\begin{gathered} v_{2t}=[B_{22}(x,t)v_2]_x+[B_{21}(x,t)u_1]_x+[\psi(u_0+u_1,v_1,x,t)-\psi(u_0;0,x,t)]_x, \\ v_2(x,0)=0, \qquad v_2(x+2\pi,t)\equiv v_2(x,t), \end{gathered} \tag{4.16}$$

has a solution which, using Lemmas 3.1 and 3.4, and Duhamel's principle, is of the form:

$$\begin{aligned} v_2(x,t)&=O(\varepsilon u_1)+O(\varepsilon G)+O(u_0(u_1+v_1)) \\ &=O(\varepsilon^{\mu+2})+O(\varepsilon^{2\mu+1})+O(\varepsilon^{3\mu}). \end{aligned} \tag{4.17}$$

This iteration between the two equations can be continued to obtain an asymptotic expansion of the solution to (4.5) to any desired order. All remaining terms will be of the same order in ε as v_2 or smaller. This is so, because the remainder terms $(u-u_0-u_1)$ and $(v-v_1-v_2)$ satisfy a symmetric system, which is well-posed with a growth constant independent of ε. Thus, by Duhamel's principle, this system will have a solution of the same order in ε as the forcing terms, and, using Lemma 3.1, these are no larger than v_2.

All terms of amplitude ε^{μ}, $\varepsilon^{2\mu}$ or $\varepsilon^{\mu+1}$ in the solution of (4.5) are thus given by the linear systems (4.6), (4.9) and (4.12).

5. Completion of proof of Theorem 1. The proof given in the previous section relied essentially only on the bounds on the x-derivatives of the solution of (4.1) given by Lemma 3.1, which ensured that the forcing terms at successive stages in the iteration did indeed become smaller, and on the nonsingularity of the large part of the spatial operator, P_0, on S_p^0, which enabled the time derivative of the fast part of the solution to be expressed in terms of the fast part itself. Lemma 3.1 covers the system (2.1) as well. Thus, in relaxing the assumptions (A5)–(A9), we need only be concerned with writing the large part of the spatial operator in an appropriate, nonsingular form.

(a) *Undifferentiated terms.* Suppose assumption (A6) is relaxed, and we consider a system of the form

$$(5.1)\qquad \begin{gathered} U_t=\frac{1}{\varepsilon}AU_x+[\Phi(U,V,x,t)]_x+\Gamma(U,V,x,t),\\ V_t=[\Psi(U,V,x,t)]_x+\Omega(U,V,x,t),\\ U(x+2\pi,t)\equiv U(x,t),\qquad V(x+2\pi,t)\equiv V(x,t), \end{gathered}$$

where Γ and Ω are C^∞ functions of all their arguments, 2π-periodic in x, bounded, together with their derivatives, independently of ε, and all other symbols are as before.

Subtracting out a smooth solution, we obtain, analogously to (4.5),

$$(5.2)\qquad \begin{aligned} u_t&=\frac{1}{\varepsilon}Au_x+[B_{11}u+B_{12}v+\varphi(u,v)]_x+C_{11}u+C_{12}v+\gamma(u,v),\\ v_t&=[B_{22}v+B_{21}u+\psi(u,v)]_x+C_{22}v+C_{21}u+\omega(u,v) \end{aligned}$$

say, with initial and boundary conditions as for (4.5). The spatial operator is, to leading order, unchanged, but the mean value of u,

$$\langle u\rangle=\frac{1}{2\pi}\int_0^{2\pi}u(x,t)\,dx,$$

is no longer zero for all time, so there is apparently no unique inverse. This problem arises because the mean value of u is really a slow scale variable, i.e. it has at least one time derivative bounded independently of ε, so it should really be grouped with v rather than with the rest of u.

A separate equation can be formed for $\langle u\rangle$ by integrating the first equation of (5.2). Using the periodic boundary conditions, one obtains

$$(5.3)\qquad \begin{aligned} \langle u\rangle_t&=\langle C_{11}u\rangle+\langle C_{12}v\rangle+\langle\gamma(u,v)\rangle\\ &=\langle C_{11}\rangle\langle u\rangle+F(u,\langle u\rangle,v,\langle v\rangle), \end{aligned}$$

say. Subtracting this from the unaveraged equation, and writing u' for $u-\langle u\rangle$, gives

(5.4)

$$\begin{aligned}u'_t=&\frac{1}{\varepsilon}Au'_x+[B_{11}u'+B_{12}v+\varphi(u',v)]_x\\&+C_{11}u'+C_{12}v+\gamma(u'+\langle u\rangle,v)+C_{11}\langle u\rangle\\&\quad-\{\langle C_{11}u'\rangle+\langle C_{12}v\rangle+\langle\gamma(u'+\langle u\rangle,v)\rangle+\langle C_{11}\rangle\langle u\rangle\}.\end{aligned}$$

This replaces the first equation in (5.2), while the second equation in (5.2) is augmented by (5.3).

Now the mean value of u' is zero for all time, and the proof can proceed as before. The first approximation to u', u'_0, is, as before, $O(\varepsilon^{\mu})$, oscillates on the fast time scale, and satisfies (4.6) with slightly modified P_0. (4.8) can now be used in the second equation of (5.2) as before, and also in (5.3), to show that both v and $\langle u\rangle$ are at leading order $O(\varepsilon^{2\mu})+O(\varepsilon^{\mu+1})$. The rest of the iteration proceeds as before.

(b) *Large undifferentiated terms.* Suppose an undifferentiated term is added to the large part of the spatial operator in (4.1):

$$(5.5)\qquad \begin{aligned}U_t&=\frac{1}{\varepsilon}(AU_x+CU)+[\Phi(U,V,x,t)]_x,\\V_t&=[\Psi(U,V,x,t)]_x,\end{aligned}$$

where C is constant and antisymmetric to ensure that the x-derivatives remain bounded. The large part of the spatial operator, $A\partial/\partial x+C$, may have eigenfunctions with zero eigenvalue, corresponding to the constant function in case (a). Separate equations for those parts of u parallel to such eigenfunctions must be formed, just as a separate equation for the mean value of u was in (a). The spatial operator will then be nonsingular if restricted to the remainder of S_p.

This allows (A6) and (A7) to be completely relaxed.

(c) *More space dimensions and nonseparation of scales.* There is no difficulty in relaxing (A9) and going to more space dimensions, except that it may no longer be possible to separate equations for the fast and slow variables as was assumed in (4.1). However, such a separation can be carried out in Fourier space by means of a projection. Exactly how this is done is described by Kreiss [5]; the projection is defined by the construction of unitary transformations at each point in Fourier space; the assumption (A4) is required for this to be possible. The large part of the spatial operator is then nonsingular on the range of the projection, which is also the space spanned by the fast scale variables, and so our proof goes through as before.

This takes care of assumption (A8) also.

(d) *Forcing terms.* Finally, note that use of Duhamel's principle allows estimates to be obtained for forced systems, provided the forcing is on the slow time scale, in accordance with (A2). This allows the last extra assumption of §4, (A5), to be relaxed.

Thus, the proof of §4 can indeed be extended to prove Theorem 1 in full generality.

Acknowledgement. The author is deeply indebted to Prof. H.-O. Kreiss for his suggestion of, and help and encouragement with, this work.

REFERENCES

[1] F. Baer, *Adjustment of initial conditions required to suppress gravity oscillations in nonlinear flows*, Beitr. Phys. Atmos., 50 (1977), pp. 350–366.

[2] F. Baer and J. J. Tribbia, *On complete filtering of gravity modes through initialization*, Monthly Weather Rev., 105 (1977), pp. 1536–1539.

[3] G. Browning, *A new system of equations for numerical weather forecasting*, Ph. D. thesis, New York Univ., New York, 1979.

[4] 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.

[5] G. Browning and H.-O. Kreiss, *Problems with different time scales for nonlinear partial differential equations*, SIAM J. Appl. Math., 42 (1982), pp. 704–718.

[6] J. Charney, *The use of the primitive equations of motion in numerical prediction*, Tellus, 7 (1955), pp. 22–26.

[7] R. E. Dickinson and D. L. Williamson, *Free oscillations of a discrete stratified fluid with application to numerical weather prediction*, J. Atmos. Sci., 29 (1972), pp. 623–640.

[8] B. Gustafsson, *Numerical solution of hyperbolic systems with different time scales using asymptotic expansions*, J. Comp. Phys., 36 (1980), pp. 209–235.

[9] ———, *Asymptotic expansions for hyperbolic problems with different time scales*, SIAM J. Numer. Anal., 17 (1980), pp. 623–634.

[10] B. Gustafsson and H.-O. Kreiss, *Difference approximations of hyperbolic problems with different time scales* I: *The reduced problem*, Dept. Computer Science, Uppsala Univ., Report 86, 1980.

[11] S. Klainerman and A. Majda, *Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids*, Comm. Pure Appl. Math., 34 (1981), pp. 481–524.

[12] H.-O. Kreiss, *Problems with different time scales for ordinary differential equations*, SIAM J. Numer. Anal., 16 (1979), pp. 980–998.

[13] ———, *Problems with different time scales for partial differential equations*, Comm. Pure Appl. Math., 33 (1980), pp. 399–439.

[14] B. Machenhauer, *On the dynamics of gravity oscillations in a shallow water model, with applications to normal mode initialization*, Beitr. Phys. Atmos., 50 (1977), pp. 253–271.

[15] K. Miyakoda and R. W. Moyer, *A method of initialization for dynamical weather forecasting*, Tellus, 20 (1968), pp. 115–128.

[16] Ta. Nitta and J. B. Hovermale, *A technique of objective analysis and initialization for the primitive forecast equations*, Monthly Weather Rev., 97 (1969), pp. 652–658.

[17] N. A. Phillips, *On the problem of initial data for the primitive equations*, Tellus, 12 (1960), pp. 121–126.

[18] E. Tadmor, *Hyperbolic systems with different time scales*, to appear (1983).

[19] D. L. Williamson, *Normal mode initialization procedure applied to forecasts with the global shallow water equations*, Monthly Weather Rev., 104 (1976), pp. 195–206.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, May 1984

© 1984 Society for Industrial and Applied Mathematics
007

INVARIANT MANIFOLDS AND THE ONSET OF REVERSAL IN THE RIKITAKE TWO-DISK DYNAMO*

MARCY BARGE†

Abstract. In this paper we prove a stable manifold theorem for noncompact solution curves of an ordinary differential equation and apply the theorem to establish the existence of two-dimensional stable and unstable manifolds for a particular solution curve of the Rikitake equations. Some information on the configuration of these surfaces in phase space is obtained yielding a rough description of the oscillatory nature of typical solutions.

Introduction. The nondimensional Rikitake equations

$$\dot{x}=-\mu x+yz, \quad \dot{y}=-\mu y+(z-A)x, \quad \dot{z}=1-xy, \tag{0.1}$$

model the currents (x and y) and angular velocities (z and $z-A$) in a pair of identical homoplanar coupled disk dynamos driven by constant torque [3], [6]. This system was conceived by Rikitake as a model of the Earth's magnetohydrodynamic dynamo.

An interesting facet of the model is that, for some values of the parameters μ and A and most initial conditions, numerical solutions to (0.1) have been observed by a number of people to exhibit a complicated pattern of changes in sign in the x and y components that seems to persist for all time. This empirical data together with the constant negative divergence (for $\mu>0$) of the vector field defined by (0.1), so that Lebesgue volume is compressed exponentially along solutions, has led to the question of the possible existence of a strange attractor—a nontrivial invariant set that attracts nearby solutions—in the phase space of (0.1).

This paper is the first in a series devoted to the study of solutions to (0.1). Here we will establish the existence of two invariant surfaces in the phase space of (0.1). One surface is composed of solutions that, eventually, have the z coordinate increasing monotonically without bound and x and y coordinates decaying exponentially to zero. The other surface consists of solutions that in reversed time have, eventually, the z coordinate decreasing monotonically without bound and x and y coordinates decaying exponentially to zero. These surfaces can be thought of as stable and unstable manifolds of the z-axis or as stable and unstable manifolds of points at infinity on the z-axis. The latter point of view, entailing a compactification of phase space, will be adopted in order to gain information about the configuration of these surfaces in space. We will describe the development of the complicated dynamics of (0.1), in particular the onset of reversal behavior, in terms of a change in configuration of the above surfaces as A is increased from 0.

In §1 a qualitative description of the relatively simple dynamics of (0.1) when $A=0$ will be given. In this case the 2-dimensional stable manifold of the z-axis and the 2-dimensional unstable manifold of the z-axis coincide, preventing dynamo reversal (switch in the signs of both x and y).

In §2 it is proven that these stable and unstable manifolds persist for all $A>0$ but no longer coincide. The proof is based on Theorem 2.1, a stable manifold theorem for the noncompact case. Specifically, the theorem provides a means of establishing existence and uniqueness of stable and unstable manifolds for an unbounded solution

* Received by the editors June 3, 1982, and in revised form May 19, 1983.

† Department of Mathematics, University of Wyoming, Laramie, Wyoming 82071.

curve. As the proof of Theorem 2.1 is highly technical, we postpone it until §3. Roughly speaking, uniformity assumptions are used in lieu of compactness.

Section 2 concludes with a brief description of a typical solution to (0.1) with $A>0$ in terms of the configuration of the various invariant surfaces.

The existence of these invariant surfaces rules out the bounded limit surface for (0.1) proposed by Cook and Roberts [3]. The second paper in this series [2] will provide more detailed information on these invariant surfaces and use this knowledge to describe the topology of a strange unbounded attractor for the Rikitake equations.

1. The dynamics at $A=0$. We consider (0.1) with $A=0$, $\mu>0$:

$$\dot{x}=-\mu x+yz,\quad \dot{y}=-\mu y+xz,\quad \dot{z}=1-xy. \tag{1.1}$$

Clearly, the planes $y=x$ and $y=-x$ are invariant. On $y=-x$ (the Bullard antidynamo) integral curves are given by

$$y=-x,\qquad x^2+z^2+2\mu z+\ln x^2=c \tag{1.2}$$

or by $x=y=0$.

The curves (1.2) are doubly asymptotic to the z-axis. Solutions on the plane $y=-x$ approach the z-axis asymptotically, in the Euclidean metric, as $t\to\pm\infty$.

On the plane $y=x$ (the Bullard dynamo) integral curves are given by

$$y=x,\qquad x^2+z^2-2\mu z-\ln x^2=c,\quad c\geq 1-\mu^2 \tag{1.3}$$

or by $x=y=0$.

The curves (1.3) include two equilibria $(1,1,\mu)$ and $(-1,-1,\mu)$, at $c=1-\mu^2$, and are periodic solutions for $c>1-\mu^2$.

To determine the behavior of solutions near the above periodic solutions, suppose $x_0\notin\{0,1,-1\}$, let (x_0,x_0,μ) be initial conditions on $y=x$, and let $\phi_t(x_0,x_0,\mu)=(x(t),y(t),z(t))$ be the solution to (1.1) with these initial conditions. The variational equation along this solution is

$$\dot{w}=A(t)w,\qquad A(t)=\begin{bmatrix}-\mu & z(t) & x(t)\\ z(t) & -\mu & x(t)\\ -x(t) & -x(t) & 0\end{bmatrix}. \tag{1.4}$$

A solution to (1.4) is

$$w(t)=\begin{pmatrix}1\\-1\\0\end{pmatrix}\exp\left(-\int_0^t(\mu+z(t))\,dt\right).$$

Equations (1.1) have the following symmetry on $y=x$:

$$(\dot{x}(a,a,\mu+b),\dot{z}(a,a,\mu+b))=(-\dot{x}(a,a,\mu-b),\dot{z}(a,a,\mu-b)).$$

From this symmetry we can conclude that

$$\int_0^\tau(\mu+z(t))\,dt=2\mu\tau$$

where $\tau=\tau(x_0)$ is the period of the periodic solution $\phi_t(x_0,x_0,\mu)$.

Thus, integration of (1.4) with initial conditions

$$w_0 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$$

over a period gives

$$w(\tau) = w_0 \exp(-2\mu\tau).$$

It follows that for each (x,y,z) on the stable manifold of the periodic solution through (x_0,x_0,μ) there are a phase ρ and constants $k,\alpha>0$ such that

$$|\phi_t(x,y,z) - \phi_{t+\rho}(x_0,x_0,\mu)| \leq ke^{-\alpha t}.$$

Furthermore, these 2-dimensional local stable manifolds foliate a neighborhood of $\{(x,y,z)|x=y, x\neq 0,1,-1\}$. See [1].

To complete the description of solutions near the plane $y=x$, we observe that the eigenvalues of the linear part of the vector field (1.1) at the equilibria are -2μ and $\pm 2i$. Thus, for $\mu>0$ each equilibrium has a 1-dimensional stable manifold. It follows that there is a neighborhood of $\{(x,y,z)|x=y, x\neq 0\}$ such that every solution entering this neighborhood approaches a periodic solution on $y=x$ with exponential rate and asymptotic phase, or approaches an equilibrium with exponential rate.

To obtain global information, consider

$$u = x^2 - y^2.$$

The total derivative of u along solution curves of (1.1) is

$$\frac{du}{dt} = 2x\dot{x} - 2y\dot{y} = -2\mu u.$$

Thus, for every solution $(x(t),y(t),z(t))$ of (1.1) we have

$$(x(t))^2 - (y(t))^2 = e^{-2\mu t}\left((x(0))^2 - (y(0))^2\right),$$

and each solution approaches the union of the $y=x$ and $y=-x$ planes asymptotically (in the Euclidean metric) as $t\to\infty$.

The result of the above is that as $t\to\infty$ solutions to (1.1) are either unbounded or approach a periodic solution or equilibrium with exponential rate. The only solutions bounded as $t\to-\infty$ are the periodic solutions and equilibria on the plane $y=x$.

The solutions to the Rikitake equations of most interest are those that display change in x and y from both positive to both negative or vice versa. Such solutions correspond to reversal in the polarity of the magnetic field of the coupled dynamos. At $A=0$ the invariance of the plane $y=-x$ prevents the existence of such solutions, but for $A>0$ such orbits are observed numerically. In order to see how such solutions arise for $A>0$ we will view the planes $y=x$ and $y=-x$ (at $A=0$) as center manifolds of new equilibria in an extension of the Rikitake equations for which phase space is compact.

In cylindrical coordinates

$$x = r\cos\theta, \quad y = r\sin\theta, \quad z = z$$

(0.1) becomes

(1.5)

$$\dot{r} = [-\mu + (2z-A)\cos\theta\sin\theta]r, \quad \dot{\theta} = (z-A)\cos^2\theta - z\sin^2\theta, \quad \dot{z} = 1 - r^2\cos\theta\sin\theta.$$

Divide the right-hand side of (1.5) by

$$\rho(z)=\sqrt{1+z^2} \quad \text{and} \quad \eta(r)=(1+r^2)$$

to get

(1.6)

$$\dot r=\frac{[-\mu+(2z-A)\cos\theta\sin\theta]\,r}{\rho(z)\eta(r)}, \quad \dot\theta=\frac{(z-A)\cos^2\theta-z\sin^2\theta}{\rho(z)\eta(r)}, \quad \dot z=\frac{1-r^2\cos\theta\sin\theta}{\rho(z)\eta(r)}.$$

Oriented solution curves of (1.6) are the same as those of (1.5), and (1.6) can be continuously extended to the solid cylinder $\{(r,\theta,z)\,|\,|r|\le\infty,|z|\le\infty\}$ by setting

(1.7)

$$\begin{aligned}
&\dot r=[2r\cos\theta\sin\theta]/\eta(r),\\
&\dot\theta=[\cos^2\theta-\sin^2\theta]/\eta(r), && \text{at } z=\infty,\\
&\dot z=0,\\
&\dot r=[-2r\cos\theta\sin\theta]/\eta(r),\\
&\dot\theta=[\sin^2\theta-\cos^2\theta]/\eta(r), && \text{at } z=-\infty,\\
&\dot z=0,\\
&\dot r=0,\\
&\dot\theta=0, && \text{at } r=\infty.\\
&\dot z=[-\cos\theta\sin\theta]/\rho(z),
\end{aligned}$$

Equations (1.6) and (1.7) are continuous on the solid cylinder (with the obvious topology) and everywhere tangent to the boundary. Furthermore, the flow on the boundary of the cylinder is independent of the parameters μ and A.

It is seen that the point $(0,0,\infty)$ is a hyperbolic equilibrium of (1.7) with stable manifold $\{(r,-\pi/4,\infty)\,|\,|r|<\infty\}$ and unstable manifold $\{(r,+\pi/4,\infty)\,|\,|r|<\infty\}$. Similarly, $(0,0,-\infty)$ is a hyperbolic equilibrium of (1.7) with stable manifold $\{(r,\pi/4,-\infty)\,|\,|r|<\infty\}$ and unstable manifold $\{(r,-\pi/4,\infty)\,|\,|r|<\infty\}$.

At $A=0$ the planes $y=x$ and $y=-x$ have the following characterization in this system. The half-plane $\{(r,\pi/4,z)\,|\,0<r<\infty,|z|<\infty\}$ is the unique center manifold of the equilibrium $(\sqrt2,\pi/4,\mu)$, the half-plane $\{(r,\pi/4,z)\,|\,-\infty<r<0,|z|<\infty\}$ is the unique center manifold of the equilibrium $(-\sqrt2,\pi/4,\mu)$, and the closed plane $\{(r,-\pi/4,z)\,|\,|r|\le\infty,|z|\le\infty\}$ is the closure of both the unique center stable manifold of $(0,0,\infty)$ and of the unique center unstable manifold of $(0,0,-\infty)$.

The coincidence of the 2-dimensional center stable manifold of $(0,0,\infty)$ and the 2-dimensional center unstable manifold of $(0,0,-\infty)$ and the neutral stability of the center manifolds of $(\pm\sqrt2,\pi/4,\mu)$ is highly unusual. This situation is forced by the symmetries in (1.6) at $A=0$. Some of the symmetry is lost for $A>0$, and we will see in the next section that, although the various center manifolds persist, the center stable manifold of $(0,0,\infty)$ and the center unstable manifold of $(0,0,-\infty)$ split apart and the finite equilibria become unstable.

2. Stable and unstable manifolds for $A>0$. In this section the persistence of the various center manifolds observed above at $A=0$ will be established for all $A>0$, and information on their configuration in phase space, allowing for the existence of magnetic field reversals in the coupled dynamo system, will be obtained. A subsequent paper will give more complete information on the configuration of these manifolds in phase space and will use this geometry to investigate a noncompact strange attractor containing both periodic and nonperiodic reversal behavior.

The eigenvalues of (0.1) linearized at the equilibria $(\pm K, \pm 1/K, \mu K^2)$ are -2μ and $\pm(K^2+1/K^2)^{1/2}i$ (recall that $A=\mu(k^2-1/K^2)$). Thus, for $\mu>0$, each equilibrium has a 1-dimensional stable manifold and a 2-dimensional center manifold. In [3] Cook and Roberts showed that the 2-dimensional center manifolds are center unstable for $\mu>0$ and $A>0$.

The following terminology will be used in Theorem 2.1. We will say that

$$\dot{x}=A(t)x, \qquad x\in\mathbb{R}^n$$

has an exponential dichotomy of type $(n,,k,\alpha)$ on $\mathbb{R}^+$ if there exist positive constants α, K, M and projection P of rank k such that, for fundamental matrix $X(t)$ and all $V\in\mathbb{R}^n$,

$$|X(t)PV|\leq K(\exp(-\alpha(t-s)))|X(s)PV| \quad \text{for } t\geq s\geq 0,$$
$$|X(t)(I-P)V|\geq K(\exp(\alpha(t-s)))|X(s)(I-P)V| \quad \text{for } t\geq s\geq 0$$

and

$$\|X(t)PX^{-1}(t)\|\leq M \quad \text{for } t\geq 0,$$

where $|\cdot|$ is the Euclidean norm and $\|\ \|$ the induced matrix norm.

In Theorem 2.1 we consider

$$\dot{z}=f(z,t), \qquad z\in\mathbb{R}^n. \tag{2.1}$$

Let $\phi_t(z,s)$ denote a solution to (2.1) with $\phi_s(z,s)=z$.

THEOREM 2.1. *Assume that*

i) $f(0,t)=0$ *for* $t\geq 0$;

ii) $f(z,\ )\in C^1(\mathbb{R}^+,\mathbb{R}^n)$, *and there exist constants* $c>0$, $\varepsilon>0$ *such that* $f(\cdot,t)\in C^2(\mathbb{R}^n,\mathbb{R}^n)$ *for* $t\geq 0$, $|z|\leq\varepsilon$ *and*

$$\left|\frac{\partial^2 f^i}{\partial z_j \partial z_k}(z,t)\right|\leq c \quad \textit{for } |z|\leq\varepsilon,\ t\geq 0, \textit{ and all } i,j,k\in\{1,\cdots,n\};$$

and

iii) $\dot{x}=A(t)x, A(t)=D_z f(0,t)$, *has an exponential dichotomy of type* (n,k,α) *on* $\mathbb{R}^+$ *and* $\|A(t)\|$ *is bounded for* $t\geq 0$.

Then there is a continuous 1*-parameter family of k-dimensional continuous manifolds*

$$W_s\subseteq\mathbb{R}^n, \quad s\in\mathbb{R}^+, \quad 0\in W_s$$

such that

$$\phi_{s+t}(W_s,s)\subseteq W_{s+t} \quad \textit{for } t\geq 0, \qquad \lim_{t\to\infty}\phi_t(z,s)=0 \quad \textit{for } z\in W_s$$

and, if $\lim_{t\to\infty}\phi_t(z,s)=0$, *then* $\phi_t(z,s)\in W_t$ *for large enough t.*

The proof of Theorem 2.1 is given in §3. The theorem and its proof are along the lines of Fenichel [5], but with compactness replaced by the uniformity assumptions in ii) and iii).

We define sets S and S' and their saturations Σ and Σ' below. Theorem 2.1 and Proposition 2.2 will be used in Proposition 2.3 to determine that Σ and Σ' are the unique stable and unstable manifolds of the z-axis. Proposition 2.4 begins a description of the configuration of these sets in phase space.

Let $\phi_t(x_0,y_0,z_0)=(x(t),y(t),z(t))$ denote the solution to (0.1) with initial value $\phi_0(x_0,y_0,z_0)=(x_0,y_0,z_0)$. Define

$$S=\{(x_0,y_0,z_0)|A\leq z_0, x(t)y(t)\leq 0 \text{ for all } t\geq 0\},$$
$$S'=\{(x_0,y_0,z_0)|z_0\leq 0, x(t)y(t)\leq 0 \text{ for all } t\leq 0\},$$
$$\Sigma=\bigcup_{t\in\mathbb{R}}\phi_t(S) \quad\text{and}\quad \Sigma'=\bigcup_{t\in\mathbb{R}}\phi_t(S').$$

PROPOSITION 2.2. *For* $\mu>0$ *and* $A\geq 0$ *we have*

i) $\lim_{t\to\infty}[(x(t))^2+(y(t))^2]=0$ *if and only if* $(x_0,y_0,z_0)\in\Sigma$,

and

ii) $\lim_{t\to-\infty}[(x(t))^2+(y(t))^2]=0$ *if and only if* $(x_0,y_0,z_0)\in\Sigma'$.

Proof. i) Suppose $(x_0,y_0,z_0)\in\Sigma$. Then, for some $T\in\mathbb{R}$ we have

$$\phi_t(x_0,y_0,z_0)=(x(t),y(t),z(t))\in S \quad \text{for all } t\geq T.$$

Moreover, for $(x(t),y(t),z(t))\in S$ we have

$$\frac{d}{dt}\left[(x(t))^2+(y(t))^2\right]=-2\mu\left[(x(t))^2+(y(t))^2\right]+2[2(z(t))-A]x(t)y(t)$$
$$\leq -2\mu\left[(x(t))^2+(y(t))^2\right],$$

so that $\lim_{t\to\infty}[(x(t))^2+(y(t))^2]=0$.

Conversely, suppose $\lim_{t\to\infty}[(x(t))^2+(y(t))^2]=0$. Then there exists T_1 such that

$$(x(t))^2+(y(t))^2\leq 1 \quad \text{for all } t\geq T_1.$$

Then

$$(x(t))(y(t))\leq\tfrac{1}{2} \quad \text{for all } t\geq T_1,$$

so that

$$\dot{z}(t)=1-x(t)y(t)\geq\tfrac{1}{2} \quad \text{for all } t\geq T_1.$$

Thus, there is a $T_2\geq T_1$ so that

$$z(t)\geq\mu K^2$$

and $z(t)$ increases monotonically for $t\geq T_2$. Now for $t\geq T_2$ we have

$$\frac{d}{dt}[x(t)y(t)]=z(t)(y(t))^2-2\mu x(t)y(t)+(z(t)-A)(x(t))^2\geq 0.$$

That is, $x(t)y(t)$ is nondecreasing for $t\geq T_2$.

Suppose now that $(x_0,y_0,z_0)\notin\Sigma$. Then there are arbitrarily large values of t for which $x(t)y(t)>0$. Let $T_3\geq T_2$ be such that

$$x(T_3)y(T_3)>0.$$

Then for $t\geq T_3$

$$x(t)y(t)\geq x(T_3)y(T_3)>0,$$

and for $t \geq T_3$ we have

$$\frac{d}{dt}\left[(x(t))^2+(y(t))^2\right]=-2\mu\left[(x(t))^2+(y(t))^2\right]+2[2z(t)-A]x(t)y(t)$$
$$\geq -2\mu\left[(x(t))^2+(y(t))^2\right]+2\mu(k^2+1/k^2)x(T_3)y(T_3).$$

But, by assumption, $\lim_{t\to\infty}[(x(t))^2+(y(t))^2]=0$, so for all t sufficiently large, the above inequality implies that

$$\frac{d}{dt}\left[(x(t))^2+(y(t))^2\right]>0,$$

a contradiction. Therefore $(x_0,y_0,z_0)\in\Sigma$.

The proof of ii) is similar.

PROPOSITION 2.3. *For $\mu>0$ and $A\geq 0$, Σ and Σ' are continuous 2-dimensional manifolds.*

Proof. It suffices to prove that S and S' are continuous 2-dimensional manifolds.

Let $\phi_t(x_0,y_0,z_0)=(x(t),y(t),z(t))$ denote a solution to (0.1). We will establish the existence of a 2-manifold $M\supset\{(0,0,z)|z>\mu+A\}$ that has the following properties: If $(x_0,y_0,z_0)\in M$, then $(x(t))^2+(y(t))^2\to 0$ as $t\to\infty$ and if $(x(t))^2+(y(t))^2\to 0$, then $\phi_t(x_0,y_0,z_0)\in M$ for all sufficiently large t. By Proposition 2.2 the existence of such an M implies that S is itself a continuous 2-manifold.

By rescaling (0.1) near the z-axis and eliminating z, we will obtain a 2-dimensional system of nonautonomous equations. Theorem 2.1 will be applied to these equations to find a unique family of stable manifolds. M will then be constructed from this family.

On the set $\{(x,y,z)|x^2+y^2\leq 1, 0<z\}$ the oriented solution curves of (0.1) are the same as those for

$$\dot{x}=\frac{-\mu x+yz}{(1-xy)z},\quad \dot{y}=\frac{-\mu y+(z-A)x}{(1-xy)z},\quad \dot{z}=\frac{1}{z}. \tag{2.2}$$

Fix $z_0>\mu+A$. The solution to (2.2) with initial conditions (x_0,y_0,z_0) has the form

$$\tilde{\phi}_t(x_0,y_0,z_0)=\left(\tilde{\phi}_t^1(x_0,y_0,z_0),\tilde{\phi}_t^2(x_0,y_0,z_0),\left(2t+z_0^2\right)^{1/2}\right).$$

To investigate these solutions, consider

$$\dot{x}=\frac{-\mu x+y\left(2t+z_0^2\right)^{1/2}}{(1-xy)\left(2t+z_0^2\right)^{1/2}}=f^2(x,y,t), \tag{2.3}$$
$$\dot{y}=\frac{-\mu y+\left(\left(2t+z_0^2\right)^{1/2}-A\right)x}{(1-xy)\left(2t+z_0^2\right)^{1/2}}=f^2(x,y,t).$$

Then $f(x,y,t)=(f^1(x,y,t),f^2(x,y,t))$ satisfies the conditions of Theorem 3.1. Indeed, properties i) and ii) are obvious. To verify property iii) make the change of coordinates

$$\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}.$$

Then (2.3) becomes

$$\frac{d}{dt}\begin{pmatrix} u \\ v \end{pmatrix}=g(u,v,t)$$

with

$$D_{(u,v)}g(0,0,t)=\begin{pmatrix} 1-\left[\dfrac{\mu+A/2}{\left(2t+z_0^2\right)^{1/2}}\right] & \dfrac{A}{2\left(2t+z_0^2\right)^{1/2}} \\ \dfrac{-A}{2\left(2t+z_0^2\right)^{1/2}} & -1-\left[\dfrac{\mu+(A/2)}{\left(2t+z_0^2\right)^{1/2}}\right] \end{pmatrix}=(a_{ij}).$$

We have

$$|a_{11}|=1-\left[\frac{\mu+(A/2)}{\left(2t+z_0^2\right)^{1/2}}\right]\geq 1-\left[\frac{\mu+(A/2)}{(\mu+A)}\right]+\delta$$

$$\geq\left[\frac{A}{2\left(2t+z_0^2\right)^{1/2}}\right]+\delta=|a_{21}|+\delta$$

and

$$|a_{22}|\geq|a_{12}|+\delta$$

for $\delta>0$ small and all $t\geq 0$. Thus, according to Coppel [4, Prop. 6.3], $D_{(u,v)}g(0,0,t)$ has an exponential dichotomy of type $(2,1,\alpha)$ for some $\alpha>0$. Since $D_{(x,y)}f(0,0,t)$ is similar to $D_{(u,v)}g(0,0,t)$, f also satisfies property iii). Thus, we have verified the conditions of Theorem 2.1 for (2.3).

Now let W_s be the collection of stable manifolds for (2.3) at $(0,0)$ and define

$$M=\left\{(x,y,z)\middle|(x,y)\in W_\xi,\xi=\frac{z^2-z_0^2}{2},z\geq z_0>\mu+A\right\}.$$

M is a continuous 2-manifold with the desired properties. This completes the proof of Proposition 2.3.

Embed S and S' in the solid cylinder viewed as the phase space of (1.6) and (1.7). The following proposition describes the configuration of S and S'.

Let $\mathrm{cl}(S)$ denote the closure of S in the solid cylinder.

PROPOSITION 2.4.

i) *S separates $\{(r,\theta,z)|0<r<\infty,A\leq z<\infty,-\pi/4\leq\theta\leq 0\}$ into two connected components.*

ii) *S separates $\{(r,\theta,z)|0<r<\infty,A\leq z<\infty,3\pi/4\leq\theta\leq\pi\}$ into two connected components.*

iii) $\{(\pm\infty,-\pi/4,z)|A\leq z\leq\infty\}\cup\{(r,-\pi/4,\infty)|-\infty\leq r\leq\infty\}\subseteq\mathrm{cl}(S)$.

iv) *S' separates $\{(r,\theta,z)|-\pi/2\leq\theta\leq-\pi/4,0<r<\infty,-\infty<z\leq 0\}$ into two connected components.*

v) *S' separates $\{(r,\theta,z)|\pi/2\leq\theta\leq 3\pi/4,0<r<\infty,-\infty<z\leq 0\}$ into two connected components.*

vi) $\{(\pm\infty,-\pi/4,z)|-\infty\leq z\leq 0\}\cup\{(r,-\pi/4,-\infty)|-\infty\leq r\leq\infty\}\subseteq\mathrm{cl}(S')$.

Proof. i) Let $W=\{(r,\theta,z)|0<r<\infty,A\leq z<\infty,-\pi/4\leq\theta\leq 0\}$, let $U_1=\{(r,\theta,z)|(r,\theta,z)\in W$ and there exists a $T\geq 0$ with $\theta(T)=-\pi/4$, and $-\pi/4\leq\theta(t)\leq 0$ for $0\leq t\leq T\}$, and let $U_2=\{(r,\theta,z)|(r,\theta,z)\in W$ and there exists a $T\geq 0$ with $\theta(T)=0$, and $-\pi/4\leq\theta(T)\leq 0$ for $0\leq t\leq T\}$. Then U_1 and U_2 are each connected and relatively open in W. Furthermore, $S\cap W=W-(U_1\cup U_2)$.

Part ii) follows from i) and symmetry.

We prove part iii). For $z<0$ or $z \geq A$ in (1.6) we have

$$\dot{\theta}=0 \quad \text{if and only if} \quad \tan\theta = \pm\left(1-\left(\frac{A}{z}\right)\right)^{1/2}.$$

Let

$$\tilde{W}=\left\{(r,\theta,z)\,\middle|\,(r,\theta,z)\in W \text{ and } -1\leq \tan\theta \leq -\left(1-\left(\frac{A}{z}\right)\right)^{1/2}\right\}.$$

Then the flow of (1.6) is out of $\tilde{W}$ along $\tan\theta=-1$ and along $\tan\theta=-(1-(A/z))^{1/2}$. Thus $S\cap W\subseteq \tilde{W}$. Part i) implies that for each r satisfying $0<r<\infty$ there is a θ in the interval $-\pi/4\leq\theta\leq 0$ such that $(r,\theta,\infty)\in \mathrm{cl}(S)$. Since $S\cap W\subseteq\tilde{W}$, it follows that $\theta=-\pi/4$. Thus $\{(r,-\pi/4,\infty)\,|\,0\leq r\leq\infty\}\subseteq \mathrm{cl}(S)$. By symmetry, $\{(r,-\pi/4,\infty)\,|\,-\infty\leq r\leq 0\}\subseteq \mathrm{cl}(S)$.

We also know from i) that for each ξ in the interval $A<\xi<\infty$ there is a θ_ξ in the interval $-\pi/4\leq\theta_\xi\leq 0$ such that $(\infty,\theta_\xi,\xi)\in\mathrm{cl}(S)$. Since $S\cap W\subseteq\tilde{W}$ we conclude that $-1\leq\tan\theta_\xi\leq -(1-(A/z))^{1/2}$.

Now (1.7) and continuity of the flow $\phi_t(r,\theta,z)$ in (r,θ,z) imply that $(\infty,\theta_\xi,z)\in \mathrm{cl}(S)\cap\tilde{W}$ for all z in the interval $A\leq z\leq\infty$. But since $\lim_{z\to\infty} -(1-(A/z))^{1/2}=-1$ we must have $\theta_\xi=-\pi/4$.

Parts iv), v) and vi) are proved similarly.

From Proposition 2.4 we see that $\Sigma\cup\{(r,-\pi/4,\infty)\,|\,|r|<\infty\}$ is the unique global center stable manifold of $(0,0,\infty)$ and that $\Sigma'\cup\{(r,-\pi/4,-\infty)\,|\,|r|<\infty\}$ is the unique global center unstable manifold of $(0,0,-\infty)$.

In order to give a rough description of a typical solution to the Rikitake equations, we will define four disjoint regions of the $z=\mu K^2$ plane through which solutions may pass.

Let c and c' denote the curves of first intersection of Σ and Σ' with the disk $z=\mu K^2$ (see Fig. 1). Then we have

$$c=\{\phi_t(r,\theta,z)|\,(r,\theta,z)\in S, z(t)=\mu K^2, \text{ and } z(\tau)>\mu K^2 \text{ for } t>\tau\}$$

and

$$c'=\{\phi_t(r,\theta,z)|\,(r,\theta,z)\in S', z(t)=\mu K^2, \text{ and } z(\tau)<\mu K^2 \text{ for } \tau<t\}.$$

As a consequence of Proposition 2.4 and continuity in initial conditions, we see that for small $A>0$, $\mathrm{cl}(c)\cup\mathrm{cl}(c')$ separates the disk $\{(r,\theta,\mu K^2)\,|\,|r|\leq\infty\}$ into four connected components. Label these $E_\pm$ for the connected component containing the equilibrium $e_\pm=(\pm(K^2+1/K^2)^{1/2}, \tan^{-1}(1/K^2),\mu K^2)$ and $W_\pm$ for the connected component containing

$$\{(r,-\pi/4,\mu K^2)\,|\,0<|r|<\infty, \pm r>0\}.$$

Note that for $A=0$, c and c' coincide so that W_+ and W_- are empty. As a consequence, when $A=0$, a solution starting in E_+ (E_-) can never intersect E_- (E_+). However, for $A>0$, c and c' have split apart, creating the windows W_+ and W_- through which solutions may make the transition from oscillation about one equilibrium to oscillation about the other.

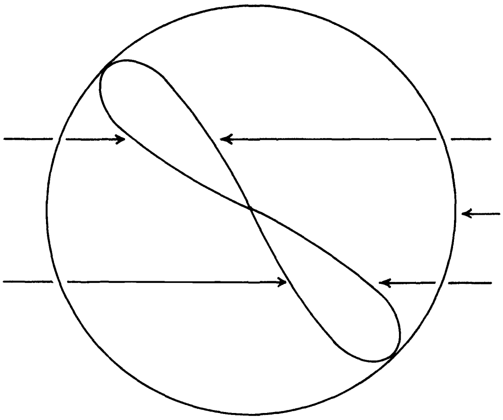


FIG. 1. *c and c' are the first intersections of Σ and Σ' with the disk $z=\mu K^2$.*

A crude description of a typical solution to (0.1) with $\mu>0$ and small $A>0$, displaying dynamo reversal, is as follows. A solution with initial conditions near the equilibrium e_+ is attracted towards the 2-dimensional center unstable manifold of e_+ and oscillates about e_+ with increasing amplitude until it enters the window W_+. The solution is then attracted towards the center unstable manifold of e_- (reversal has taken place); it oscillates about e_- with increasing amplitude until it enters the window W_-; the solution is then attracted towards the center unstable manifold of e_+ (another reversal); and so on.

A much more detailed description of solutions to the Rikitake equations in terms of sequences of intersections with the components $E_\pm$ and $W_\pm$ will be the subject of a forthcoming paper [2].

3. Proof of Theorem 2.1.

LEMMA 1 (Coppel). *Assume* iii), *of Theorem* 2.1. *Then there are* $S(t), B(t) \in C^1(\mathbb{R}^+, \mathbb{R}^{n^2})$ *with* $\|S(t)\|$, $\|S^{-1}(t)\|$, $\|S'(t)\|$ *bounded and* $B(t)=S^{-1}(t)A(t)S(t)-S^{-1}(t)S'(t)$ *such that* $B(t)$ *commutes with* P *for all* $t\geq 0$, *and* $\dot{y}=B(t)y$ *has an exponential dichotomy of type* (n,k,α) *on* $\mathbb{R}^+$.

This is [4, Lemma 5.3]. $\|S'(t)\|$ is bounded since $S'(t)=A(t)S(t)-S(t)B(t)$.

By a linear (constant) change of coordinates in $\mathbb{R}^N$ we may assume the projection P is $\left(\begin{smallmatrix} I_k & 0\\ 0 & 0\end{smallmatrix}\right)$. Now change coordinates in $\mathbb{R}^n\times\mathbb{R}^+$ by $(z,t)=(S(t)\xi,t)$. Then $\dot{z}=f(z,t)$ is transformed into

$$\dot{\xi}=S^{-1}(t)[f(S(t)\xi,t)+S'(t)\xi]=g(\xi,t)$$

where

$$D_\xi g(0,t)=S^{-1}(t)D_z f(0,t)S(t)+S^{-1}(t)S'(t)=B(t).$$

Since $B(t)$ commutes with P,

$$B(t)=\begin{pmatrix} B_1(t) & 0\\ 0 & B_2(t)\end{pmatrix}$$

where $B_1(t)$ is $k\times k$. Also, $g(\xi,t)$ satisfies property ii) of the theorem since $\|S(t)\|$, $\|S^{-1}(t)\|$ and $\|S'(t)\|$ are bounded for $t\geq 0$.

Thus, without loss of generality, we may assume in property iii) that

$$D_z f(0,t) = \begin{pmatrix} A_1(t) & 0 \\ 0 & A_2(t) \end{pmatrix}$$

where $A_1(t)$ is $k \times k$ and that

$$\begin{aligned} |v_1(t)| &\leq K e^{-\alpha(t-s)} |v_1(s)|, & t \geq s \geq 0, \\ |v_2(t)| &\geq K e^{\alpha(t-s)} |v_2(s)|, & t \geq s \geq 0, \end{aligned}$$

$v_i(t)$ being any solution to $\dot{v}_i = A_i(t) v_i$, $i=1,2$.

LEMMA 2. *Given $T>0$ there are a $K_1 = K_1(T) < \infty$ and an $\varepsilon_1 = \varepsilon_1(T) > 0$ such that*

$$\left| \frac{\partial \phi_t^i}{\partial z_j}(z,s) \right| \leq K_1 \quad \textit{and} \quad \left| \frac{\partial^2 \phi_t^i}{\partial z_j \partial z_k}(z,s) \right| \leq K_1$$

for $|z| \leq \varepsilon_1$, $|t-s| \leq T$, $t,s \geq 0$, and $i,j,k \in \{1,\cdots,n\}$.

Proof. Let $\varepsilon>0$ be as in assumption ii) of Theorem 2.1. Since $\|D_z f(0,t)\|$ and $|(\partial^2 f^i/\partial z_j \partial z_k)(z,t)|$ are bounded for $|z| \leq \varepsilon$, $t \geq 0$, so is $\|D_z f(z,t)\|$. This implies that $|f(z,t)| \leq M_1 |z|$ for $|z| \leq \varepsilon$, $t \geq 0$ and some $M_1 < \infty$. From this it follows that $\|\phi_t(z,s)\| \leq \varepsilon$ for $|z| \leq \varepsilon_1 = \varepsilon / e^{M_1 T}$, $|t-s| \leq T$ and $t,s \geq 0$. Now let K_1 be the finite supremum over $|z| \leq \varepsilon_1$, $|t-s| \leq T$, $t,s \geq 0$, of the norms of the solutions

$v = D_z \phi_t(z,s)$ to

$$\dot{v} = \left(D_z f|_{(\phi_t(z,s),t)}\right) v, \qquad v(s) = I$$

and

$u = (\partial^2 \phi_t^i(z,s)/\partial z_j \partial z_k)_{i=1}^n$ to

$$\dot{u} = \left(D_z f|_{(\phi_t(z,s),t)}\right) u + \left(\sum_{l,m} \frac{\partial^2 f^i}{\partial z_m \partial z_l}(\phi_t(z,s),t) \frac{\partial \phi_t^m}{\partial z_j}(z,s) \frac{\partial \phi_t^l}{\partial z_k}(z,s) \right)_{i=1}^n,$$

$$u(s) = 0.$$

Note that, for $0 \leq T \leq T' < \infty$, we may take $K_1(T') \geq K_1(T)$ and $\varepsilon(T') \leq \varepsilon(T)$. In particular, $\sup_{0 \leq T \leq T'} K_1(T) < \infty$ and $\inf_{0 \leq T \leq T'} \varepsilon(T) > 0$.

Now let $F_s(z) = F_s^T(z) = \phi_{s+T}(z,s)$, and for $z = (z,y) \in \mathbb{R}^k \times \mathbb{R}^{n-k}$ write $F_s^{-1}(z) = (g_s^T(x,y), h_s^T(x,y))$.

LEMMA 3. $D_y g_s^T(0,0) = 0$, $D_x h_s(0,0) = 0$, *and for any $\eta > 1$ there is a $T = T(\eta)$ so that* $\|D_x g_s^T(0,0)\|_{\min} \geq \eta > 1$ *and* $\|D_y h_s^T(0,0)\| \leq 1/\eta < 1$ *for all $s \geq 0$.*

Proof. $D_z \phi_t(0,s)$ is a fundamental matrix for

$$\dot{v} = D_z f(0,t), \qquad v(s) = I.$$

Since

$$D_z f(0,t) = \begin{pmatrix} A_1(t) & 0 \\ 0 & A_2(t) \end{pmatrix},$$

the first two inequalities in the lemma are obvious. Now write

$$D_z\phi_t(0,s)=\begin{pmatrix} D_z\phi_t^1(0,s) & 0 \\ 0 & D_y\phi_t^2(0,s) \end{pmatrix}.$$

Then

$$\left|D_x\phi_t^1(0,s)v_1\right|\le Ke^{-\alpha(t-s)}|v_1| \qquad \text{for } t\ge s\ge 0,$$
$$\left|D_y\phi_t^2(0,s)v_2\right|\ge Ke^{\alpha(t-s)}|v_2| \qquad \text{for } t\ge s\ge 0.$$

Since $D_xg_s^T(0,0)=(D_x\phi^1_{T+s}(0,s))^{-1}$ and $D_yh_s^T(0,0)=(D_y\phi^2_{T+s}(0,s))^{-1}$, we have $\|D_xg_s^T(0,0)\|_{\min}\ge e^{\alpha T}/K$ and $\|D_yh_s(0,0)\|\le e^{-\alpha T}/K$. Now let T be large enough.

Let λ and ε be positive and define $S_\lambda^\varepsilon=\{u\colon \mathbb{R}_\varepsilon^k\times\mathbb{R}^+\to\mathbb{R}^{n-k}|\ u$ is continuous and $u_s=u(\cdot,s)$ is Lipschitz with constant $\le\lambda$ for all $s\ge 0\}$. If we let

$$d(u,u')=\sup_{\substack{s\ge 0\\ 0<|x|\le\varepsilon}}\left\{\frac{|u_s(x)-u_s'(x)|}{|x|}\right\},$$

S_λ^ε becomes a complete metric space. The stable manifolds will be obtained as the graphs of a fixed point of a graph transform on S_λ^ε for sufficiently small ε.

From now on let $\eta>1$ be fixed, let T be large enough so that the conclusions of Lemma 3 hold, and let $g_s=g_s^T$, $h_s=h_s^T$.

Let $\varepsilon_1=\varepsilon_1(T)$ be as in Lemma 2.

LEMMA 4. *There exist $c_1=c_1(T)<\infty$ and $c_2=c_2(T)<\infty$ such that*

$$|g_s(x_2,y_2)-g_s(x_1,y_1)|\ge\eta|x_2-x_1|-c_1(\bar\varepsilon_1)\big[|y_2-y_1|+|x_2-x_1|\big],$$
$$|h_s(x_2,y_2)-h_s(x_1,y_1)|\le\frac{1}{\eta}|y_2-y_1|+c_2(\bar\varepsilon_1)\big[|y_2-y_1|+|x_2-x_1|\big]$$

for any $\bar\varepsilon_1\le\varepsilon_1$ provided $|(x_i,y_i)|\le\bar\varepsilon_1$ for $i=1,2$.

Proof. Let $k_s(x,y)=g_s(x,y)-D_xg_s(0,0)x$. Then

$$Dk_s(x,y)=Dg_s(x,y)-Dg_s(0,0)=\left(\sum_{l=1}^n\frac{\partial^2g_s^i}{\partial z_l\partial z_j}(\xi_{ij})z_l\right)_{k\times n}$$

where z_l is the lth component of $z=(x,y)$ and $|\xi_{ij}|\le|z|$. Then, for $|(x,y)|\le\varepsilon_1(T)$ as in Lemma 2, we have $\|Dk_s(x,y)\|\le C_1|(x,y)|$ for all $s\ge 0$. Here C_1 depends directly on the $K_1(T)$ of Lemma 2. We now have $|k_s(x_2,y_2)-k_s(x_1,y_1)|\le C_1(\bar\varepsilon_1)[|y_2-y_1|+|x_2-x_1|]$ provided $|(x_i,y_i)|\le\bar\varepsilon_1\le\varepsilon_1$, and so

$$\begin{aligned}|g_s(x_2,y_2)-g_s(x_1,y_1)|&\ge|D_xg_s(0,0)(x_2-x_1)|-|k_s(x_2,y_2)-k_s(x_1,y_1)|\\ &\ge\|D_xg_s(0,0)\|_{\min}|x_2-x_1|-c_1(\bar\varepsilon_1)\big[|y_2-y_1|+|x_2-x_1|\big]\\ &\ge\eta|x_2-x_1|-c_1(\bar\varepsilon_1)\big[|y_2-y_1|+|x_2-x_1|\big].\end{aligned}$$

For the inequality involving h_s, let $\bar k_s(x,y)=h_s(x,y)-D_yh_s(0,0)y$. The proof is similar.

Let $u\in S_\lambda^{\varepsilon_2}$ and define $l_s(x)=g_s(x,u_{s+T}(x))$.

LEMMA 5. *For $\varepsilon_2=\varepsilon_2(T,\lambda)$ small enough, l_s is one-to-one and $l_s(\mathbb{R}^k_{\varepsilon_2})\supseteq\mathbb{R}^k_{\varepsilon_2}$.*

Proof.

$$\begin{aligned}|l_s(x_2)-l_s(x_1)|&=|g_s(x_2,u_{s+T}(x_2))-g_s(x_1,u_{s+T}(x_1))|\\ &\geq\eta|x_2-x_1|-c_1(\bar{\varepsilon}_1)\big[|u_{s+T}(x_2)-u_{s+T}(x_1)|+|x_2-x_1|\big]\\ &\geq\eta|x_2-x_1|-c_1(\bar{\varepsilon}_1)(1+\lambda)|x_2-x_1|=(\eta-c_1(\bar{\varepsilon}_1)(1+\lambda))|x_2-x_1|\end{aligned}$$

provided $|(x_i,u_{s+T}(x_i))|\leq\bar{\varepsilon}_1\leq\varepsilon_1$, $i=1,2$. Now choose $\bar{\varepsilon}_1>0$ small enough so that $\eta-c_1(\bar{\varepsilon}_1)(1+\lambda)>1$ and let $\varepsilon_2=\bar{\varepsilon}_1/(1+\lambda^2)^{1/2}$. Then, if $|x_i|\leq\varepsilon_2$, we have $|l_s(x_2)-l_s(x_1)|\geq|x_2-x_1|$, so that l_s is one-to-one. Also, for $|x|\leq\varepsilon_2$, $|l_s(x)|>|x|$. This, together with the continuity of l_s and the fact that $l_s(0)=0$, implies that $l_s(\mathbb{R}^k_{\varepsilon_2})\supseteq\mathbb{R}^k_{\varepsilon_2}$.

Let $u\in S^{\varepsilon_2}_\lambda$, $0<\varepsilon_3<\varepsilon_2$, and define $(\Gamma u)_s(x)=h_s(l_s^{-1}(x),u_{s+T}(l_s^{-1}(x)))$ for $|x|\leq\varepsilon_3$.

LEMMA 6. *For $\varepsilon_3>0$ sufficiently small, Γ: $S^{\varepsilon_3}_\lambda\to S^{\varepsilon_3}_\lambda$.*

Proof. We must show that $\operatorname{Lip}((\Gamma u)_s)\leq\lambda$ for all $s\geq0$. So let $|x_i|\leq\varepsilon_3$, $i=1,2$. Then by Lemma 5 there exist x_i' with $|x_i'|\leq\varepsilon_3$ such that $l_s(x_i')=x_i$, $i=1,2$. Then

$$\begin{aligned}&\frac{|(\Gamma u_s)(x_2)-(\Gamma u)_s(x_1)|}{|x_2-x_2|}\\ &\quad=\frac{|h_s(x_2',u_{s+T}(x_2'))-h_s(x_1',u_{s+T}(x_1'))|}{|g_s(x_2',u_{s+T}(x_2'))-g_s(x_1',u_{s+T}(x_1'))|}\\ &\quad\leq\frac{1/\eta|u_{s+T}(x_2')-u_{s+T}(x_1')|+c_2(\varepsilon_3)(1+\lambda^2)^{1/2}\big(|u_{s+T}(x_2')-u_{s+T}(x_1')|+|x_2'-x_1'|\big)}{\eta|x_2'-x_1'|-c_1(\varepsilon_3)(1+\lambda^2)^{1/2}\big(|u_{s+T}(x_2')-u_{s+T}(x_1')|+|x_2'-x_1'|\big)}\\ &\quad\leq\frac{\lambda/\eta+c_2(\varepsilon_3)(1+\lambda^2)^{1/2}(1+\lambda)}{\eta-c_1(\varepsilon_3)(1+\lambda^2)^{1/2}(1+\lambda)}.\end{aligned}$$

Now let $\varepsilon_3>0$ be small enough so that $\lambda/\eta+c_2(\varepsilon_3)(1+\lambda^2)^{1/2}(1+\lambda)\leq\lambda$ and $\eta-c_1(\varepsilon_3)(1+\lambda^2)^{1/2}(1+\lambda)\geq1$.

LEMMA 7. *For $\varepsilon_4>0$ sufficiently small Γ: $S^{\varepsilon_4}_\lambda\to S^{\varepsilon_4}_\lambda$ is a contraction.*

Proof. Let $u,u'\in S^{\varepsilon_4}_\lambda$ and let $g_s(x',u_{s+T}(x'))=x\in\mathbb{R}^k_{\varepsilon^4}$. Let $\varepsilon_4\leq\varepsilon_3/(1+\lambda)$. Then

$$\begin{aligned}&|(\Gamma u)_s(x)-(\Gamma u')_s(x)|\\ &\quad\leq|(\Gamma u)_s(g_s(x',u_{s+T}(x')))-(\Gamma u')_s(g_s(x',u'_{s+T}(x')))|\\ &\qquad\qquad+|(\Gamma u')_s(g_s(x',u_{s+T}(x')))-(\Gamma u')_s(g_s(x',u'_{s+T}(x')))|\\ &\quad\leq|h_s(x',u_{s+T}(x'))-h_s(x',u'_{s+T}(x'))|+\lambda|g_s(x',u_{s+T}(x'))-g_s(x',u'_{s+T}(x'))|\\ &\quad\leq\left(\frac{1}{\eta}+C_2(\varepsilon_4)(1+\lambda)\right)|u_{s+T}(x')-u'_{x+T}(x')|+\lambda C_1(\varepsilon_4)(1+\lambda)|u_{s+T}(x')-u'_{s+T}(x')|\\ &\quad=\left[\left(\frac{1}{\eta}+C_2(\varepsilon_4)(1+\lambda)+\lambda C_1(\varepsilon_4)(1+\lambda)\right)\right]|u_{s+T}(x')-u'_{s+T}(x')|\end{aligned}$$

by Lemma 4.

Now let ε_4 be small enough so that $[(1/\eta + C_2(\varepsilon_4)(1+\lambda)) + \lambda C_1(\varepsilon_4(1+\lambda))] = \gamma < 1$. From Lemma 5 $|x| \geq |x'|$, so that

$$d(\Gamma u, \Gamma u') = \sup_{\substack{s\geq 0 \\ 0<|x|\leq\varepsilon_4}} \left\{ \frac{|(\Gamma u)_s(x) - (\Gamma u')_s(x)|}{|x|} \right\}$$

$$\leq \gamma \sup_{\substack{s\geq 0 \\ 0<|x'|\leq\varepsilon_4}} \left\{ \frac{|u_{s+T}(x') - u'_{s+T}(x')|}{|x'|} \right\}$$

$$< \sup_{\substack{s\geq 0 \\ 0<|x'|\leq\varepsilon_4}} \left\{ \frac{|u_s(x') - u'_s(x')|}{|x'|} \right\} = d(u,u').$$

Fix $\lambda > 0$ and let T be large enough so that the conclusions of Lemma 3 hold. We will use the notation: $\varepsilon(T) > 0$ is small enough so that $\Gamma_T\colon S_\lambda^{\varepsilon(T)} \to S_\lambda^{\varepsilon(T)}$ is a contraction with unique fixed point U^T. The manifolds W_s will be constructed from the graph of u^{2T}, the fixed point of Γ_{2T}.

First note that for $(x, u_s^T(x)) \in \text{graph}(u_s^T)$, we have

$$\phi_{T+s}((x,u_s(x)),s) = \phi_{T+s}\big(\big(g_s^T(x',u_{s+T}(x'),h_s(x',u_{s+T}(x'))),s\big)\big)$$

$$= (x', u_{s+T}(x')) \in \text{graph}\big(u_{s+T}^T\big).$$

Now let $T' \geq T$ so that $\varepsilon(T') \leq \varepsilon(T)$ (see the note following Lemma 3). Then $\Gamma_T\colon S_\lambda^{\varepsilon(T')} \to S_\lambda^{\varepsilon(T')}$ is a contraction with unique fixed point $u^T|_{\mathbb{R}^k_{\varepsilon(T')}\times\mathbb{R}^+}$.

Suppose $MT' = NT$ for some positive integers M and N. Then $\Gamma_{T'}^M = \Gamma_T^M$ on $S_\lambda^{\varepsilon(T')}$ because $(\Gamma_{T'}^M u)_s(x) = y$ if and only if there is an $x' \in \mathbb{R}^k_{\varepsilon(T')}$ such that $\phi_{s+MT'}((x,y),x) = (x', u_{s+MT'}(x'))$. Thus, in this case, $u^T|_{\mathbb{R}^k_{\varepsilon(T')}\times\mathbb{R}^+} = u^{T'}$ so that $\phi_{s+T'}((x,u_s^T(x)),s) \in \text{graph}(u_{s+T'}^T|_{\mathbb{R}^k_{\varepsilon(T')}})$ provided $|x| \leq \varepsilon(T')$.

For $T' \in [T, 2T]$ rationally related to T we have $\phi_{s+T'}(\text{graph}\, u_s^{2T}, s) \in \text{graph}(u_{s+T'}^{2T})$. Since ϕ and $\text{graph}(u^{2T})$ are continuous, we must have $\phi_{s+T'}(\text{graph}(u_s^{2T}), s) \subseteq \text{graph}(u_{s+T'}^{2T})$ for all $T' \in [T, 2T]$.

If $T' > 2T$, say $T' = NT + T''$ with $T \leq T'' < 2T$, then

$$\phi_{s+T'}\big((x,u_s^{2T}(x)),s\big) = \phi_{(s+T'')+NT}\big(\phi_{s+T''}\big((x,u_s^{2T}(x)),s\big),s+T''\big)$$

$$= \phi_{(s+T'')+NT}\big((x',u_{s+T''}^{2T}(x')),s+T''\big) \in \text{graph}\big(u_{s+T'}^{2T}\big).$$

Now define

$$W_s = \begin{cases} \text{graph}\big(u_s^{2T}\big), & s \geq T, \\ \phi_s\big(\text{graph}\big(u_T^{2T}\big)\big), & 0 \leq s \leq T. \end{cases}$$

Then $\phi_{s+T}(W_s, s) \subseteq W_{s+T}$ for all $t \geq 0$ and if $(x,y) \in W_s$, $\lim_{t\to\infty} \phi_{s+t}((x,y),s) = 0$ since $|l_s(x)| \geq \mu|x| > |x|$ (see the proof of Lemma 5).

The uniqueness of the family W_s follows from the next two lemmas. Let $\phi_{t+s}(z,s) = (\phi_{t+s}^1(z,s), \phi_{t+s}^2(z,s)) \in \mathbb{R}^k \times \mathbb{R}^{n-k}$.

LEMMA 8. *If* $\lim_{t\to\infty} \phi_{t+s}(z,s) = 0$ *and* $|\phi_{t+s}^2(z,s)|/|\phi_{t+s}^1(z,s)|$ *is bounded for* $t \geq T' \geq 0$ *then* $\phi_{t+s}(z,s) \in W_{t+s}$ *for all t larger than some* $T'' < \infty$.

Proof. Say $|\phi^2_{t+s}(z,s)|/|\phi^1_{t+s}(z,s)|\le\lambda<\infty$ for $t\ge T'$. Let T be as in the definition of the W_s and let $T''>T'$ be large enough so that $|\phi^1_{t+s}(z,s)|\le\varepsilon(2T)$ for $t\ge T''$. Define $u\in S_\lambda^{\varepsilon(2T)}$ by

$$u_{t+s}(x)=\begin{cases}\left(\dfrac{\phi^2_{t+s}(z,s)}{|\phi^1_{t+s}(z,s)|}\right)|x|, & t\ge T'',\\[2ex] \left(\dfrac{\phi^2_{T''+s}(z,s)}{|\phi^1_{T''+s}(z,s)|}\right)|x|, & t<T''.\end{cases}$$

Then $u\in S_\lambda^{\varepsilon(2T)}$ and $\lim_{n\to\infty}\Gamma^n_{2T}u=u^{2T}$, the fixed point of Γ_{2T}. But $(\Gamma^n u)_{t+s}(\phi^1_{t+s}(z,s))=\phi^2_{t+s}(z,s)$ for all $t\ge T''$. Thus $u^{2T}_{t+s}(\phi^1_{t+s}(z,s))=\phi^2_{t+s}(z,s)$ so $\phi_{t+s}(z,s)\in W_{t+s}$ for all $t\ge T''$.

LEMMA 9. *Let $\gamma, 0<\gamma<\eta-1$, be given. Then there is an $\varepsilon=\varepsilon(\gamma)>0$ so that if $|x|\le\gamma|y|$ and $|(x,y)|\le\varepsilon$ we have $|\phi^1_{T+s}((x,y),s)|\le\gamma|\phi^2_{T+s}((x,y),s)|$ and $|\phi_{T+s}((x,y),s)|\ge|(x,y)|$ for all $s\ge0$. T is as in the definition of W_s.*

Proof.

$$\begin{aligned}|\phi^1_{T+s}((x,y),s)|&=|D_x\phi^1_{T+s}((0,0),s)x+\psi^1_s(x,y)|\\&\le\frac{1}{\eta}|x|+\frac{1}{2}K_1|(x,y)|^2\le\left(\frac{1}{\eta}\gamma+\frac{1}{2}K_1(1+\gamma)^2|y|\right)|y|\end{aligned}$$

and

$$\begin{aligned}|\phi^1_{T+s}((x,y),s)|&=|D_y\phi^2_{T+s}((0,0),s)y+\psi^2_s(x,y)|\\&\ge\eta|y|-\frac{1}{2}K_1|(x,y)|^2\ge\left(\eta-\frac{1}{2}K_1(1+\gamma)^2|y|\right)|y|\end{aligned}$$

for all $s\ge0$ and ε sufficiently small (Lemmas 2 and 3). Now let $\varepsilon>0$ be small enough so that

$$\frac{|\phi^1_{T+s}((x,y),s)|}{|\phi^2_{T+s}((x,y),s)|}\le\frac{(\gamma/\eta+(1/2)K_1(1+\gamma)^2|y|)|y|}{(\eta-(1/2)K_1(1+\gamma)^2|y|)|y|}\le\gamma$$

and small enough so that

$$|\phi_{T+s}((x,y),s)|\ge|\phi^2_{T+s}((x,y),s)|\ge\left(\eta-\tfrac{1}{2}K_1(1+\gamma)^2|y|\right)|y|\ge(1+\gamma)|y|\ge|(x,y)|.$$

Now suppose $\phi_{t+s}(z,s)\notin W_{t+s}$ for any $t\ge0$ and that $\lim_{t\to\infty}\phi_{t+s}(z,s)=0$. Then $|\phi^1_{t+s}(z,s)|\le\gamma|\phi^2_{t+s}(t,s)|$ for arbitrarily large t and fixed γ, $0<\gamma<\eta-1$, by Lemma 8. Now let T_1 be large enough so that $|\phi_{t+s}(z,s)|\le\frac{1}{2}\varepsilon(\gamma)$ for all $t\ge T_1$ and let $t_1\ge T_1$ be such that $|\phi^1_{t_1+s}(z,s)|\le\gamma|\phi^2_{t_1+s}(z,s)|$. Then either

$$|\phi_{nT+t_1+s}(z,s)|=|\phi_{nT+(t_1+s)}(\phi_{t_1+s}(z,s)),(t_1+s)|\ge|\phi_{t_1+s}(z,s)|$$

for all integers $n\ge0$, in which case $\lim_{t\to\infty}\phi_{t+s}(z,s)\ne0$; or there exists a positive integer N such that $|\phi_{NT+t_1+s}(z,s)|\ge\varepsilon(\gamma)$, a contradiction to the choice of t_1. Thus, if $\phi_{t+s}(z,s)\notin W_{t+s}$ for any $t\ge0$, we have $\lim_{t\to\infty}\phi_{t+s}(z,s)\ne0$.

REFERENCES

[1] B. AULBACH, *Behavior of solutions near manifolds of periodic solutions*, J. Differential Equations, 39 (1981), pp. 345–377.

[2] M. BARGE, *A geometric model of the Rikitake attractor*, in preparation.

[3] A. E. COOK AND P. H. ROBERTS, *The Rikitake two-disc dynamo system*, Proc. Cambridge Philos. Soc., 68 (1970), pp. 547–569.

[4] W. A. COPPEL, *Dichotomies in Stability Theory*, Lecture Notes in Mathematics, 629, Springer-Verlag, New York/Berlin, 1978.

[5] N. FENICHEL, *Asymptotic stability with rate conditions*, Indiana J. Math., 23 (1974), pp. 1109–1137.

[6] T. RIKITAKE, *Oscillations of a system of disk dynamos*, Proc. Cambridge Philos. Soc., 54 (1958), pp. 89–105.

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics
008

COMPETITIVE AND COOPERATIVE TRIDIAGONAL SYSTEMS OF DIFFERENTIAL EQUATIONS*

JOHN SMILLIE†

Abstract. A vector field in $\mathbb{R}^n$ determines a competitive (or cooperative) system of differential equations provided all the off-diagonal terms of the Jacobian matrix are nonpositive (or nonnegative). We show that when in addition $\partial F_i/\partial X_j$ is zero for $|i-j|>1$, all solutions must converge to equilibria or diverge in a strong sense.

A solution of a general system of ordinary differential equations can display quite chaotic behavior. It is therefore useful to know easily verifiable conditions on the equations which guarantee that the qualitative behavior of the solutions is simple. We will describe a family of differential equations for which the qualitative behavior of the solutions is as simple as possible.

Let $\Omega \subset \mathbb{R}^n$ be an open set. A system of equations $\dot{X}=F(X)$ defined for X in Ω is said to be competitive if $\partial F_i/\partial X_j \leq 0$ for $i \neq j$. It is said to be cooperative if $\partial F_i/\partial X_j \geq 0$ for $i \neq j$. Borrowing terminology from [3], we say that the system is tridiagonal if $\partial F_i/\partial X_j = 0$ for $|i-j|>1$. We say that a competitive (or cooperative) system is strongly competitive (or strongly cooperative) if the terms $\partial F_i/\partial X_j$ for $|i-j|=1$ are nowhere zero.

Cooperative tridiagonal systems arise in the construction of approximations to certain nonlinear parabolic equations defined on the interval when one makes the space variable discrete (compare [3]). Competitive tridiagonal systems arise in the study of competing species when one assumes a one-dimensional niche space and further assumes that each species competes only with its nearest neighbors in this space.

In [5] Smale shows that general competitive systems can contain quite chaotic attractors. General cooperative systems can contain chaotic invariant sets, but according to Hirsch (see [2]) these cannot be attractors. The following theorem shows that the situation for tridiagonal systems is much more restrictive.

THEOREM. *Let F be a strongly competitive or strongly cooperative tridiagonal system of ordinary differential equations defined on $\Omega \subset \mathbb{R}^n$. Assume that the component functions F_i are $n-1$ times differentiable. Let ϕ be a solution of $\dot{\phi}=F(\phi)$ defined on a maximal interval of the form $[0,a)$ for $0<a\leq\infty$. Then either $\lim_{t\to a}\phi(t)$ exists and is an equilibrium point or as $t\to a$ $\phi(t)$ eventually leaves any compact set.*

The assumption that F be strongly competitive (or cooperative) is equivalent in the tridiagonal case to the assumption that F be competitive (or cooperative) and as in [2] that the matrix of first derivatives of F be indecomposable. The hypothesis on the differentiability of F can be weakened somewhat. In particular if n is 3, 4 or 5, it is sufficient that the F_i be once differentiable. Unlike [1] and [2] we make no convexity assumptions on Ω.

In the case $n=2$ this theorem has been proved many times. See [2] for references. A continuous analogue of this theorem for semilinear parabolic equations is proved in [4]. The proof of our theorem follows the statement and proof of a proposition. A corollary to the proposition may have some independent interest. (Compare [2, Thm. 2.7].)

*Received by the editors November 2, 1982, and in revised form May 18, 1983. This work was supported by a National Science Foundation Postdoctoral Fellowship.

†Graduate School of the City University of New York, New York, New York 10036

Remarks. In a tridiagonal system a change of variables $Y_i=(-1)^i X_i$ converts a competitive system to a cooperative system and vice versa. In fact any tridiagonal system for which $\partial F_i/\partial X_{i-1}\cdot\partial F_i/\partial X_{i+1}\geq 0$ for all i is equivalent to a cooperative system by a change of variables of the form $Y_i=\pm X_i$. Changing F to $-F$ changes cooperative systems to competitive systems and reverses the direction of orbits; thus our results can be stated for "negative semi-orbits" defined on intervals $(-a,0]$. In what follows we consider only cooperative systems.

Let v be a vector in $\mathbb{R}^n$, all of whose coordinates v_i, $i=1,\cdots,n$ are nonzero. Define $\sigma(v)$ to be the number of indices $i\in\{1,\cdots,n-1\}$ for which v_i and v_{i+1} have opposite signs. The function σ has an extension to a continuous function $\bar\sigma$ defined on a slightly larger set. Let $\Lambda\subset\mathbb{R}^n$ be the set of vectors v such that $v_1\neq 0$, $v_n\neq 0$ and if $v_i=0$ for $i\in\{2,\cdots,n-1\}$, then v_{i-1} and v_{i+1} are nonzero and have opposite signs. Let $\bar\sigma$ be the unique continuous function defined on Λ that agrees with σ where both are defined.

Let $\phi(t)$ be a nonconstant solution defined for $t\geq 0$. We write $\bar\sigma(t)$ for $\bar\sigma(\dot\phi(t))$. The main technical result of the paper is the following.

PROPOSITION. *The set of $t\geq 0$ for which $\bar\sigma(t)$ is not defined is discrete. Around values of t for which $\bar\sigma(t)$ is defined, $\bar\sigma$ is constant. At values where $\bar\sigma$ is not defined, $\bar\sigma$ jumps to a strictly smaller value.*

COROLLARY. *$\phi_1(t)$ and $\phi_n(t)$ are eventually monotone.*

Proof of Corollary. $\bar\sigma(\dot\phi(t))$ is a decreasing, nonnegative and integer valued function; thus it is eventually constant. According to the proposition, eventually $\dot\phi(t)$ remains in Λ, the domain of $\bar\sigma$. $\dot\phi(t)\in\Lambda$ implies that $\dot\phi_1(t)$ and $\dot\phi_n(t)$ are nonzero. Thus eventually $\dot\phi_1$ and $\dot\phi_n$ maintain constant sign and ϕ_1 and ϕ_n are monotone.

Proof of Proposition. The fact that $\bar\sigma$ is constant where it is defined follows from the fact that $\bar\sigma$ is continuous and integer valued in its domain of definition. Let t_0 be a positive real number at which $\bar\sigma$ is not defined. We must show that $\bar\sigma$ is defined in a neighborhood of t_0 and that $\bar\sigma$ decreases at t_0.

For the remainder of the proof we write $\dot\phi$ and $\dot\phi^{(j)}$ for $\dot\phi(t_0)$ and $\dot\phi^{(j)}(t_0)$.

DEFINITION. For $i\in\{1,\cdots,n\}$, set $k(i)=0$ if $\phi_i^{(1)}\neq 0$. For $j<n$, set $k(i)=j$ if $\phi_i^{(1)}=\phi_i^{(2)}=\cdots=\phi_i^{(j)}=0$ and $\phi_i^{(j+1)}\neq 0$. Set $k(i)=n$ if $\phi_i^{(1)}=\phi_i^{(2)}=\cdots=\phi_i^{(n)}=0$. If $k(i)<n$, define $\operatorname{sgn}(i)$ to be $+1$ if $\phi_i^{(k(i)+1)}$ is positive and -1 if it is negative. $k(i)$ is essentially the order of the zero of $\dot\phi_i(t)$ at $t=t_0$.

Since F_i is $n-1$ times differentiable, ϕ_i is n-times differentiable. Thus the derivatives in the above definition are defined.

DEFINITION. Let $S\subset\{2,\cdots,n-1\}$ consist of indices i for which $k(i-1)=k(i+1)<n$ and $\operatorname{sgn}(i-1)=\operatorname{sgn}(i+1)$.

The following lemma shows that for indices not in S we have some control over $k(i)$ and $\operatorname{sgn}(i)$.

LEMMA 1. *If $i\in\{1,\cdots,n\}$ and $k(i)\geq 1$, then $k(i)\geq\min\{k(i-1),k(i+1),n-1\}+1$. If $k(i)\geq 1$ and $i\notin S$, then $k(i)=\min\{k(i-1),k(i+1),n-1\}+1$ and* $\operatorname{sgn}(i)$*, if defined, is given by*

$$\operatorname{sgn}(i)=\begin{cases}\operatorname{sgn}(i-1) & \textit{if } k(i-1)\leq k(i+1),\\ \operatorname{sgn}(i+1) & \textit{if } k(i+1)\leq k(i-1).\end{cases}$$

Proof of Lemma. Let $m=\min\{k(i-1),k(i+1),n-1\}$. We will prove the formulas:

$$\text{(a)}\qquad \phi_i^{(1)}=\phi_i^{(2)}=\cdots=\phi_i^{(m+1)}=0,$$

$$\text{(b)}\qquad \phi_i^{(m+2)}=\phi_{i-1}^{(m+1)}\frac{\partial F_i}{\partial X_{i-1}}+\phi_{i+1}^{(m+1)}\frac{\partial F_i}{\partial X_{i+1}}.$$

First we observe that (a) and (b) imply the lemma. (a) implies that $k(i)\geq m+1$. If $\phi_i^{(m+1)}$ is not zero or if $m=n-1$, then $k(i)=m+1$. Assume $\phi_i^{(m+2)}$ is zero and $m<n-1$. Since $\partial F_i/\partial X_{i-1}$ and $\partial F_i/\partial X_{i+1}$ are positive, and at least one of $\phi_{i-1}^{(m+1)}$ and $\phi_{i+1}^{(m+1)}$ is nonzero, the only way that the right-hand side of (b) could be zero would be for both $\phi_{i-1}^{(m+1)}$ and $\phi_{i+1}^{(m+1)}$ to be nonzero and to have opposite signs. In this case $i\in S$. The assertion about $\operatorname{sgn}(i)$ is easily verified.

We prove (a) and (b) by induction. Our inductive hypothesis is that the following equations $A(j)$ and $B(j)$ are valid.

$$(A(j))\qquad \phi_i^{(1)}=\phi_i=\cdots=\phi_i^{(j+1)}=0,$$

$$(B(j))\qquad \phi_i^{(j+2)}=\phi_{i-1}^{(j+1)}\frac{\partial F_i}{\partial X_{i-1}}+\phi_i^{(j+1)}\frac{\partial F_i}{\partial X_i}+\phi_{i+1}^{(j+1)}\frac{\partial F_i}{\partial X_{i+1}}.$$

We begin our induction at $j=0$. The assertion $A(0)$ is $\phi_i^{(1)}=0$. This is equivalent to the hypothesis $k(i)\geq 1$ of the lemma. Since ϕ is a solution of $\dot\phi=F(\phi)$, we have $\dot\phi_i=F_i(\phi_{i-1},\phi_i,\phi_{i+1})$. Differentiating this equation gives $B(0)$.

Now we show that for $j<m$, $A(j)$ and $B(j)$ imply $A(j+1)$ and $B(j+1)$. To prove $A(j+1)$ we need to show that $\phi_i^{(j+2)}$ is zero. $\phi_i^{(j+2)}$ is the left-hand side of $B(j)$. We claim that the ϕ's on the right-hand side of $B(j)$ vanish. $\phi_{i-1}^{(j+1)}$ and $\phi_{i+1}^{(j+1)}$ vanish because $k<m$. $A(j)$ asserts that $\phi_i^{(j+1)}$ vanishes. Since the ϕ's vanish, $\phi_i^{(j+2)}$ is zero. To prove $B(j+1)$ we differentiate $B(j)$ and use the fact that the ϕ's that occur in $B(j)$ vanish at t_0.

Induction establishes $A(m)$ and $B(m)$. Note that $A(m)$ implies that the middle term on the right-hand side of $B(m)$ vanishes. Discarding this term gives (b).

LEMMA 2. *For $i\notin S$, $k(i)<n$.*

Proof. Define a new function $\bar k$ on $\{1,\cdots,n\}$ as follows. If $i\notin S$, set $\bar k(i)=k(i)$. If $i\in S$, set $\bar k(i)=k(i+1)$. We claim:

$$(*)\qquad |\bar k(i)-\bar k(i+1)|\leq 1.$$

If $i\in S$ then $\bar k(i)=k(i+1)=\bar k(i+1)$. If $i+1\in S$ then $\bar k(i+1)=k(i+2)=k(i)$. In either case $|\bar k(i)-k(i+1)|=0$. Assume that neither i nor $i+1$ are in S. According to Lemma 1, $k(i)=\min\{k(i-1), k(i+1), n-1\}+1$ so $k(i)\leq k(i+1)+1$. Similarly $k(i+1)\leq k(i)+1$. Thus $(*)$ holds.

If $\dot\phi(t_0)$ were 0, then by uniqueness of solutions ϕ would be an equilibrium solution. Thus for some j, $\dot\phi_j(t_0)\neq 0$. Equivalently $k(j)=0$. Since $k(j)=0$, j is not in S so $\bar k(j)=0$. Let m be the index for which $\bar k$ achieves its maximal value. By repeated use of $(*)$ we see that $\bar k(m)\leq|j-m|\leq n-1$. Thus $\bar k(i)<n$ for all i and $k(i)<n$ for all $i\notin S$.

Let $v=(v_1,v_2,\cdots,v_n)\in\mathbb{R}^n$. Let $0\leq i<j\leq n$ be given. Let $v^{i,j}$ be the vector $(v_i,v_{i+1},\cdots,v_j)$. Say that we are given $0=i_0<i_1\cdots<i_k=n$; then, if $\bar\sigma(v^{i_l,i_{l+1}})$ is defined for each l, $\bar\sigma(v)$ is defined and $\bar\sigma(v)=\sum_{l=0}^{k}\bar\sigma(v^{i_l,i_{l+1}})$. We adopt the notation $\bar\sigma_{i,j}(v)$ for $\bar\sigma(v^{i,j})$.

We will partition $\{1,\cdots,n\}$ into subintervals $\{i,\cdots,j\}$ so that on each subinterval there is some ε such that $\bar\sigma_{i,j}(\dot\phi(t))$ is defined for $|t-t_0|<\varepsilon$ and $t\neq t_0$, and such that $\bar\sigma_{i,j}$ does not increase. This will establish the discreteness of the set of points at which $\bar\sigma$ is not defined.

Take for the indices $0=i_0<i_1\cdots<i_k=n$ the collection of all indices not in S. We distinguish three types of intervals $\{i_l,i_{l+1}\}$.

An interval $\{i,j\}$ has type 1 if there is some index l, $i<l<j$ with $l\in S$. An interval has type 2 if it is of the form $\{i,i+1\}$ with $k(i)=k(i+1)$. An interval has type 3 if it is of the form $\{i,i+1\}$ with $k(i)\neq k(i+1)$.

We analyze each type of interval separately.

Type 1. If S contains 2 successive indices i and $i+1$, then by definition $i \in S$ implies $k(i+1)<n$ and $i+1 \in S$ implies $k(i)<n$. Applying Lemma 1 twice gives $k(i)>k(i+1)$ and $k(i+1)>k(i)$. This contradiction implies elements of S are "isolated" and that an interval of type 1 has the form $\{i-1, i+1\}$ with $i \in S$. Taylor's theorem implies that for $|t-t_0|$ small, and $k(j)<n$, the sign of $\dot{\phi}_j(t)$ is the same as that of $\operatorname{sgn}(j) \cdot (t-t_0)^{k(j)}$. Since $k(i-1)=k(i+1)<n$ and $\operatorname{sgn}(i-1)=\operatorname{sgn}(i+1)$ for $|t-t_0|$ small and $t \neq t_0$, $\dot{\phi}_{i-1}(t)$ and $\dot{\phi}_{i+1}(t)$ have opposite sign. Thus $\bar{\sigma}_{i-1, i+1}(t)$ is defined and equal to 1. Note that if $k(i-1)=k(i+1)=0$, then $\bar{\sigma}_{i-1,, i+1}(t)$ is also defined for $t=t_0$.

Type 2. For $|t-t_0|$ small the signs of $\dot{\phi}_1(t)$ and $\dot{\phi}_{i+1}(t)$ are the same as those of $\operatorname{sgn}(i)(t-t_0)^{k(i)}$ and $\operatorname{sgn}(i+1)(t-t_0)^{k(i)}$. Thus $\bar{\sigma}_{i, i+1}(t)$ has the same value for $t<t_0$ and $t>t_0$. Note that if $k(i)=k(i+1)=0$, $\bar{\sigma}_{i, i+1}(t)$ is actually defined at $t=t_0$.

Type 3. The argument in Lemma 2 shows that if $k(i) \neq k(i+1)$, then $|k(i)-k(i+1)|=1$. Without loss of generality we may assume that $k(i+1)=k(i)+1$. According to Lemma 2, $k(i+1)<n$; so by Lemma 1 $k(i+1)=\min\{k(i), k(i+2)\}+1$. This implies that $k(i)=\min\{k(i), k(i+2)\}$ so $k(i) \leq k(i+2)$. It follows from Lemma 1 that $\operatorname{sgn}(i+1)=\operatorname{sgn}(i)$. For $|t-t_0|$ small the signs of $\dot{\phi}_i(t)$ and $\dot{\phi}_{i+1}(t)$ are the same as those of $\operatorname{sgn}(i) \cdot (t-t_0)^{k(i)}$ and $\operatorname{sgn}(i) \cdot (t-t_0)^{k(i)+1}$ respectively. For $t<t_0$ these signs are opposite and for $t>t_0$ they are the same. Thus $\bar{\sigma}_{i, i+1}(t)$ decreases at t_0.

We have shown that $\bar{\sigma}$ is not increasing on any of the intervals in the partition. We conclude the proof of the proposition by showing that $\bar{\sigma}$ is actually decreasing on some interval in the partition. Assume not. Then all intervals are of type 1 and type 2. The function k of Lemma 2 is constant on intervals of type 1 and type 2. So k is constant on $\{1, \cdots, n\}$. As in Lemma 2, $\bar{k}(i)=0$ for some i, hence $\bar{k}(i)=0$ for all i. If $\{i,j\}$ is an interval of type 1 or 2 and $\bar{k}(i)=k(i)$ is zero, then $\bar{\sigma}_{i, j}(t)$ is defined at $t=t_0$. Since $\bar{\sigma}_{i, j}(t_0)$ is defined for all intervals $\{i,j\}$ in the partition, then $\bar{\sigma}(t_0)$ is defined. But this is contrary to our assumption that $\bar{\sigma}(t_0)$ was not defined. This completes the proof of the proposition.

Proof of theorem. Let $\phi(t)$ be a nonconstant solution of $\phi=F(\phi)$ defined on a maximal interval $[0, a)$ for $0<a\leq\infty$. Let L_ω consist of points p which are limits of points $\phi(t_i)$ for some sequence $t_i \to a$. If L_ω is empty, then $\phi(t)$ eventually leaves all compact sets. Suppose that ϕ does not leave some compact set C; that is, $\phi(t_i) \in C$ for some sequence $t_i \to a$. Then for some subsequence $\phi(t_i)$ converges to a $p \in C$ and p is in L_ω. To prove the theorem we must show that if L_ω is nonempty, it contains a unique point.

Any two points p and q in L_ω have the same first coordinate. We have sequences s_i and t_i converging to a such that $\lim_{i \to \infty} \phi(s_i)=p$ and $\lim_{i \to \infty} \phi(t_i)=q$. Therefore $\lim_{i \to \infty} \phi_1(s_i)=p_1$ and $\lim_{i \to \infty} \phi_1(t_i)=q_1$. Since ϕ_1 is eventually monotone, ϕ_1 converges to p_1 and q_1; thus $p_1=q_1$.

All points in L_ω are equilibrium points. Given p in L_ω there is a solution τ defined on $(-\varepsilon, \varepsilon)$ with $\tau(o)=p$. The points $\tau(t)$ for $t \in (-\varepsilon, \varepsilon)$ lie in L_ω. Since $\tau_1(t)$ is constant, $\dot{\tau}_1(t)$ is zero and $\bar{\sigma}(\dot{\tau}(t))$ is not defined on $(-\varepsilon, \varepsilon)$. According to the proposition this can only occur if τ is a constant solution. Thus p is an equilibrium point.

An elementary argument shows that L_ω is connected. If we can show that it is discrete, it will follow that it contains at most one point. This follows from our last assertion.

The set of equilibrium points with first coordinate c is discrete. Let p be an equilibrium point with first coordinate c. Let B be a convex neighborhood of p contained in Ω. We will show that B can contain no second equilibrium point with first coordinate c. Assume that $q \in B$ is a second such point. For some $i \geq 1$ we have $p_1=q_1, p_2=q_2 \cdots p_i=q_i, p_{i+1} \neq q_{i+1}$. If p and q are both equilibrium points, then $F_i(p)=F_i(q)=0$. We will show however that $F_i(p) \neq F_i(q)$.

Let $X(t)$ be the linear path with $X(0)=p$ and $X(1)=q$.

$$F_i(q)-F_i(p)=\int_0^1 \frac{d}{dt}F_i(X(t))\,dt=\int_0^1 \sum \frac{\partial F_i}{\partial X_j}\frac{dX_j}{dt}dt$$

$$=\int_0^1 \sum \frac{\partial F_i}{\partial X_j}(q_j-p_j)\,dt.$$

By assumption $q_j-p_j=0$ for $j\leq i$. Since the system F is tridiagonal, the only nonzero term $\partial F_i/\partial X_j$ with $j>i$ is $\partial F_i/\partial X_{i+1}$. Thus

$$F_i(q)-F_i(p)=(q_{i+1}-p_{i+1})\int_0^1 \frac{\partial F_i}{\partial X_{i+1}}dt.$$

Because F is competitive, the integrand is positive. By assumption, $q_{i+1}-p_{i+1}$ is nonzero; thus $F_i(q)-F_i(p)$ is nonzero. This completes the proof of the theorem.

REFERENCES

[1] M. W. Hirsch, *Systems of differential equations which are competitive or cooperative* I: *limit sets*, this Journal, 13 (1982), pp. 167–179.

[2] ______, *Systems of differential equations which are competitive or cooperative* II: *convergence almost everywhere*, to appear.

[3] ______, *Differential equations and convergence almost everywhere of strongly monotone semiflows*, to appear.

[4] H. Matano, *Convergence of solutions of one dimensional semilinear parabolic equations*. J. Math. Kyoto Univ. (1978), pp. 221–227.

[5] S. Smale, *On the differential equations of species in competition*, J. Math. Biol. 3 (1976), pp. 5–7.

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics
009

THE ASYMPTOTIC POINCARE LEMMA AND ITS APPLICATIONS*

RICHARD W. ZIOLKOWSKI† AND GEORGES A. DESCHAMPS‡

Abstract. An asymptotic version of Poincaré's lemma is defined and solutions are obtained with the calculus of exterior differential forms. They are used to construct the asymptotic approximations of multidimensional oscillatory integrals whose forms are commonly encountered, for example, in electromagnetic problems. In particular, the boundary and stationary point evaluations of these integrals are considered. The former is applied to the Kirchhoff representation of a scalar field diffracted through an aperture and simply recovers the Maggi–Rubinowicz–Miyamoto–Wolf results. Asymptotic approximations in the presence of other (standard) critical points are also discussed. Techniques developed for the asymptotic Poincaré lemma are used to generate a general representation of the Leray form. All of the (differential form) expressions presented are generalizations of known (vector calculus) results.

1. Introduction. Multidimensional integrals are encountered in many areas of physics and engineering. A combination of Poincaré's lemma and Stokes' theorem provides a means of reducing a multidimensional integral to a lower dimensional form, hence, constitutes an appealing approach to its evaluation. However, the expressions that represent solutions of Poincaré's lemma are cumbersome and often difficult to evaluate explicitly. Furthermore, in many practical problems (for instance,in electromagnetics at high frequencies) a large parameter is present and an asymptotic approximation of these integrals is quite adequate. An asymptotic version of Poincaré's lemma whose solutions are readily computed would render the Poincaré–Stokes approach very tractable in these cases.

In §2 the asymptotic Poincaré lemma (APL) is formulated, and its solutions are derived with the calculus of exterior differential forms [1]–[3]. (All differential form notation concurs with that defined in [1].) These results are utilized in §§3, 4 to construct, respectively, the boundary and stationary point approximations of a multidimensional oscillatory integral. The resultant differential form representations encompass the standard vector expressions given, for instance, in [4] and [5]. The boundary point technique is applied in §3 to the Kirchhoff representation of the diffraction of a scalar field by an aperture in a perfectly conducting screen. The Maggi–Rubinowicz–Miyamoto–Wolf expressions [6]–[8] and their properties are recovered. Several other critical point contributions are also considered in §4. The Leray form [9] is constructed in an appendix with the APL method of solution. This form is utilized in the asymptotic approach given in [10]. The results of this paper are summarized in §5.

2. Asymptotic Poincaré lemma. Consider on the domain X, a set diffeomorphic to some open set in $\mathfrak{R}^n$, a p-form of the type

$$e^{\nu\Gamma}\beta, \tag{2.1}$$

where over X the phase function Γ is smooth and real-valued and the amplitude p-form β is smooth and complex-valued. The constant ν equals ik, where k is a large real

*Received by the editors August 2, 1982, and in revised form February 5, 1983. This work was supported in part by the National Science Foundation under grant NSF-ENG-77-20820 and under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract W-7405-ENG-48.

†Electronics Engineering Department, L-156, Lawrence Livermore National Laboratory, Livermore, California 94550.

‡Department of Electrical Engineering, University of Illinois, Urbana, Illinois 61801.

parameter. For electromagnetic (quantum mechanical) problems k is 2π divided by the wavelength λ: $k=2\pi/\lambda$ ($k=2\pi/h$, where h is Planck's constant). A $(p-1)$-form, $e^{\nu\Gamma}\alpha$, of the same type as (2.1) is desired such that

$$d(e^{\nu\Gamma}\alpha)=e^{\nu\Gamma}\beta. \tag{2.2}$$

From Poincaré's lemma [1]–[3] it is known that a solution, $e^{\nu\Gamma}\alpha$, of (2.2) can exist only if the p-form $e^{\nu\Gamma}\beta$ is *closed*, i.e., only if

$$d(e^{\nu\Gamma}\beta)=e^{\nu\Gamma}(\nu\kappa+d)\beta=0, \tag{2.3}$$

where

$$\kappa=d\Gamma. \tag{2.4}$$

(Obviously,

$$d\kappa=0, \tag{2.5}$$

hence, κ is closed.) Condition (2.3) is satisfied if

$$(\kappa+D)\beta=0, \tag{2.6}$$

where

$$D=\nu^{-1}d. \tag{2.7}$$

On the other hand, because

$$D(e^{\nu\Gamma}\alpha)=e^{\nu\Gamma}(\kappa+D)\alpha, \tag{2.8}$$

(2.2) is equivalently represented as

$$\nu(\kappa+D)\alpha=\beta. \tag{2.9}$$

An asymptotic solution, α, of (2.9), when condition (2.6) is satisfied asymptotically, is constructed as follows. The result will be an *asymptotic solution of Poincaré's lemma.*

Consider a differential p-form of the type (2.1) when β has an asymptotic expansion

$$e^{\nu\Gamma}\beta=e^{\nu\Gamma}(\beta_0+\nu^{-1}\beta_1+\nu^{-2}\beta_2+\cdots), \tag{2.10}$$

where the β_j are p-forms. It is *asymptotically closed to range m* if the expression

$$(\kappa+D)\beta=(\kappa+D)(\beta_0+\nu^{-1}\beta_1+\nu^{-2}\beta_2+\cdots) \tag{2.11}$$

has its first $(m+1)$ terms (ordered in decreasing powers of ν) equal to zero; i.e., if the $(m+1)$ equations

$$\begin{aligned}&\kappa\beta_0=0,\\&\kappa\beta_1+d\beta_0=0,\\&\cdots\\&\kappa\beta_m+d\beta_{m-1}=0\end{aligned} \tag{2.12}$$

are satisfied. When these conditions hold, it is possible to find a $(p-1)$-form with an asymptotic expansion of a similar type

$$e^{\nu\Gamma}\alpha=\nu^{-1}e^{\nu\Gamma}(\alpha_0+\nu^{-1}\alpha_1+\nu^{-2}\alpha_2+\cdots), \tag{2.13}$$

such that the first $(m+1)$ terms of $d(e^{\nu\Gamma}\alpha)$ reproduce the first $(m+1)$ terms of $e^{\nu\Gamma}\beta$. This means an α can be found so that

(2.14)
$$\begin{gathered}\kappa\alpha_0=\beta_0,\\ \kappa\alpha_1+d\alpha_0=\beta_1,\\ \cdots\\ \kappa\alpha_m+d\alpha_{m-1}=\beta_m.\end{gathered}$$

The resulting $(p-1)$-form α, limited to terms of degree not greater than $(m+1)$ in ν^{-1},

(2.15)
$$\alpha=\nu^{-1}\sum_{j=0}^{m}\nu^{-j}\alpha_j,$$

is an *m-th range asymptotic solution of Poincaré's lemma.*

The relations (2.12) and (2.14), which specify an mth range asymptotic solution of Poincaré's lemma, are represented by the flow diagram given in Fig. 1. Each location is the sum of the contributions indicated by the arrows leading to it. The operator $\hat{\kappa}$ represents the exterior product by κ from the left:

(2.16)
$$\hat{\kappa}\colon \alpha\mapsto\kappa\alpha.$$

The fact that the equations at the $(p+1)$-form level are satisfied results from the identity

(2.17)
$$d\circ\hat{\kappa}+\hat{\kappa}\circ d=0,$$

a consequence of κ being closed and the Leibnitz derivative rule [1, (H.16)].

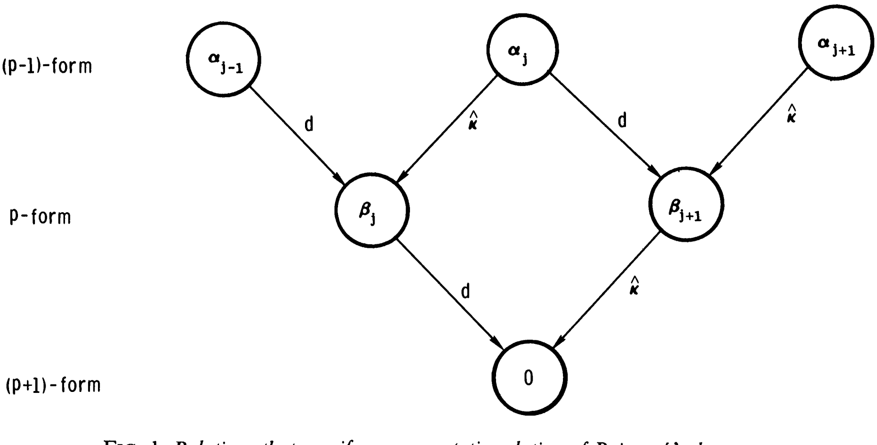


FIG. 1. *Relations that specify an asymptotic solution of Poincaré's lemma.*

Note that the expression (2.11) is automatically zero to any range if β is an n-form. Also, if the expression (2.10) consists of only the first term $e^{\nu\Gamma}\beta_0$; i.e., if $\beta=\beta_0$ and all other $\beta_j=0$, from (2.12) the terms of the expansion (2.13) of α are defined by the set of equations

(2.18)
$$\kappa\alpha_0=\beta,\quad \kappa\alpha_j=-d\alpha_{j-1},\quad \text{for } 1\leq j\leq m.$$

A solution of the system (2.14) when the conditions (2.12) are satisfied is based on the solution of an equation of the type

(2.19)
$$\kappa\alpha=\beta,$$

where the one-form κ and the p-form β are given and the $(p-1)$-form α is to be found. Its solution may be considered as a division of β by κ. Here α represents $\alpha_0, \alpha_1, \cdots, \alpha_m$ and correspondingly, β represents β_0, $\beta_1 - d\alpha_0, \cdots, \beta_m - d\alpha_{m-1}$. Multiplying (2.19) from the left by κ, one sees that a necessary condition for a solution to exist is that

$$\kappa\beta = 0. \tag{2.20}$$

Note that (2.19) and (2.20) are, respectively, the asymptotic (k large) approximations of (2.9) and (2.6). If $\kappa \neq 0$, let the one-form K be

$$K = (\kappa^*\kappa)^{-1}\kappa \equiv (\kappa\cdot\kappa)^{-1}\kappa = |\kappa|^{-2}\kappa. \tag{2.21}$$

The definitions of the star operator $*$ and the scalar product operation $\cdot$ are taken to be those given, respectively, in [1, Appendices E, F]. With (2.16) the operator "exterior product from the left by K" is simply

$$\hat{K} = (\kappa\cdot\kappa)^{-1}\hat{\kappa}. \tag{2.22a}$$

It is an operator of degree $+1$. Its adjoint, K^*, is the operator of degree -1 that equals

$$K^* = -*^{-1}\hat{K}*(-1)^p = (\kappa\cdot\kappa)^{-1}\left[-*^{-1}\hat{\kappa}*(-1)^p\right] \equiv (\kappa\cdot\kappa)^{-1}\kappa^*, \tag{2.22b}$$

when acting on a p-form. Applied to β it gives the $(p-1)$-form

$$K^*\beta = (\kappa\cdot\kappa)^{-1}\kappa^*\beta = \xi, \tag{2.23}$$

which is a solution of (2.19).

Proof. The operator κ^* satisfies the derivative property:

$$\kappa^*(\kappa\beta) = (\kappa^*\kappa)\beta - \kappa(\kappa^*\beta), \tag{2.24}$$

hence, the equivalent relation:

$$K^* \circ \hat{\kappa} + \hat{\kappa} \circ K^* = \mathrm{id}, \tag{2.25}$$

where id represents the identity operator. Consequently, (2.20) and (2.24) yield

$$\kappa(\kappa^*\beta) = (\kappa^*\kappa)\beta = (\kappa\cdot\kappa)\beta, \tag{2.26}$$

hence,

$$\kappa\xi = (\kappa\cdot\kappa)^{-1}\kappa(\kappa^*\beta) = (\kappa\cdot\kappa)^{-1}(\kappa\cdot\kappa)\beta = \beta. \tag{2.27}$$

□

The operator K^* is a (right) inverse of $\hat{\kappa}$, the operator product by κ; the solution ξ is an element of the kernel of κ^*: $\kappa^*\xi = 0$; i.e., $\kappa^*\kappa^* \equiv 0$. An interesting application of this inversion algorithm, the construction of the Leray form [9, §3.1], is given in Appendix A. The condition $\kappa \neq 0$ is satisfied except at those points in X at which the phase function Γ is stationary. Note, however, that this is only a sufficient condition. The p-form β may approach zero in regions where the operator K^* is singular (i.e., where $d\Gamma = \kappa = 0$) in such a manner that ξ given by (2.23) remains finite. This behavior is encountered in the stationary phase evaluation of an integral and will be discussed further in §4.

Consequently, solutions of (2.14) are

$$\alpha_0 = K^*\beta_0,\quad \alpha_1 = K^*(\beta_1 - d\alpha_0),\quad \dots\quad \alpha_m = K^*(\beta_m - d\alpha_{m-1}). \tag{2.28}$$

These relations are expressed more compactly as

$$\alpha_j = \sum_{p=0}^{j} \{-K^* d\}^{j-p}(K^*\beta_p). \tag{2.28'}$$

To justify that $\alpha_1, \cdots, \alpha_m$ are solutions, one must verify that conditions such as (2.20) apply to each of their equations; i.e., for α_j, that $\kappa(\beta_j - d\alpha_{j-1})$ is null.

Proof. From (2.17) and (2.12) one has, respectively, $\kappa d\alpha_{j-1} = -d(\kappa d_{j-1})$ and $\kappa\beta_j = -d\beta_{j-1}$. Thus, with (2.14)

$$\kappa(\beta_j - d\alpha_{j-1}) = d(\kappa\alpha_{j-1} - \beta_{j-1}) = -d(d\alpha_{j-2}) \equiv 0. \tag{2.29}$$

When $\beta = \beta_0$ (all $\beta_j = 0$ for $j > 0$) and $\kappa\beta = 0$, the solutions (2.28) and (2.28′) reduce to the expressions:

$$\alpha_0 = K^*\beta,\quad \alpha_j = -K^* d\alpha_{j-1} \quad \text{for } 1 \le j \le m, \tag{2.30}$$

and

$$\alpha_j = \{-K^* d\}^j (K^*\beta) \quad \text{for } 0 \le j \le m. \tag{2.30'}$$

These results can be summarized with a statement of the

Asymptotic Poincaré lemma (APL): *If a given p-form $e^{\nu\Gamma}\beta$ is asymptotically closed to range m over a domain X, a set diffeomorphic to some open set in $\mathcal{R}^n$, where $d\Gamma \neq 0$, it is asymptotically exact to range m over that domain; i.e., there exists a $(p-1)$-form $e^{\nu\Gamma}\alpha$ defined over X such that $d(e^{\nu\Gamma}\alpha) = e^{\nu\Gamma}\beta + O(\nu^{-(m+1)})$.*

The first $(m+1)$ terms of α are readily constructed from those of β with (2.28′). The construction is not valid in general at points where $d\Gamma = 0$. This restriction may be lifted for some particular β at some points where $d\gamma = 0$ as shown in §4.

The solution (2.23) of (2.19) is not unique. The general solution of (2.19) is actually

$$\alpha = \xi + \kappa\gamma, \tag{2.31}$$

where γ is an arbitrary $(p-2)$-form, since $\hat{\kappa} \circ \hat{\kappa} \equiv 0$. The expression (2.31) represents a gauge transformation of the solution (2.23). The freedom to include a gauge term, $\kappa\gamma$, in that solution occurs because (2.19) determines only the components of α "transverse" to κ. In particular, the operator $\hat{\kappa} \circ K^*$ is a projection operator that selects from the form β its component, $\kappa(K^*\beta)$, along κ; i.e., its "longitudinal" component. Hence, because the identity (2.25) means

$$\kappa(K^*\beta) + K^*(\kappa\beta) = \beta, \tag{2.32}$$

the projection operator $\{1 - \hat{\kappa} \circ K^*\}$ selects components transverse to κ. Therefore, from (2.19) one immediately obtains

$$K^*\beta = \{1 - \kappa \circ K^*\}\alpha = \xi; \tag{2.33}$$

i.e., the $(p-1)$-form ξ is the component of α transverse to κ. This is also reflected in the fact that ξ is an element of the kernel of K^*. Similarly, the solution, $e^{\nu\Gamma}\alpha$, of (2.2) is also not unique. If one replaces α by α':

$$\alpha\to\alpha'=\alpha+(\kappa+D)\gamma \tag{2.34}$$

where γ is any $(p-2)$-form, the equation

$$D(e^{\nu\Gamma}\alpha)=e^{\nu\Gamma}(\kappa+D)\alpha$$

is still satisfied since

$$(\hat{\kappa}+D)\circ(\hat{\kappa}+D)=(\hat{\kappa}+D)^2\equiv0. \tag{2.35}$$

The one-term asymptotic result (2.31) is clearly recovered in the limit $\kappa\to\infty$. The transformation $\alpha\to\alpha'$ of the general mth range APL solution given by

$$\begin{aligned}&\alpha_0\to\alpha_0'=\alpha_0+\kappa\gamma_0,\\&\alpha_j\to\alpha_j'=\alpha_j+(\kappa\gamma_j+d\gamma_{j-1})\quad\text{for } 1\le j\le m\end{aligned} \tag{2.36}$$

may be considered as an *asymptotic gauge transformation* of that solution.

Proof. The equations the gauge transformed solutions satisfy:

$$\kappa\alpha_j'=\beta_j-d\alpha_{j-1}', \tag{2.37}$$

reduce to (2.14) through the following sequence of relations:

$$\begin{aligned}&\kappa\alpha_j'=\kappa(\alpha_j+\kappa\gamma_j+d\gamma_{j-1})=\beta_j-d(\alpha_{j-1}+\kappa\gamma_{j-1}+d\gamma_{j-2})=\beta_j-d\alpha_{j-1}',\\&\kappa\alpha_j+\kappa d\gamma_{j-1}=\beta_j-d\alpha_{j-1}-d(\kappa\gamma_{j-1}),\\&\kappa\alpha_j=\beta_j-d\alpha_{j-1}.\end{aligned}$$

□

The gauge transformation (2.36) can be obtained from (2.34) with the $(p-2)$-form

$$\gamma=\nu^{-1}\sum_{j=0}^{m}\nu^{-j}\gamma_j \tag{2.38}$$

and with a truncation of the resultant expression for α' to degree $(m+1)$ in ν^{-1}. Thus, the asymptotic gauge transformation (2.36) is exact if γ_m is closed $(d\gamma_m=0)$.

Now consider the case where the p-form (2.1) is given exactly by the $(m+1)$-term expression:

$$e^{\nu\Gamma}\beta=e^{\nu\Gamma}(\beta_0+\nu^{-1}\beta_1+\cdots+\nu^{-m}\beta_m). \tag{2.39}$$

It is of interest to determine when the APL solution is actually an exact solution. Recall that an exact solution can exist only if the p-form (2.39) is closed. The equivalent condition (2.6) is satisfied if, in addition to β being asymptotically closed, β_m is closed:

$$d\beta_m=0. \tag{2.40}$$

Thus, if (2.12) and (2.40) are satisfied, there exists (locally) a $(p-1)$-form whose differential is $e^{\nu\Gamma}\beta$. This form is not necessarily given by the APL algorithm which searches for a particular expression $e^{\nu\Gamma}\alpha$, α given by (2.15) and (2.28′). However, if it is, the relation (2.9) requires

$$d\alpha_m=0 \tag{2.41}$$

in addition to the satisfaction of (2.14). Since the differential of the relation $\beta_m = \kappa\alpha_m + d\alpha_{m-1}$ gives

$$(2.42) \qquad d\beta_m = -\kappa d\alpha_m,$$

(2.40) is satisfied if (2.41) is. Conversely, (2.40) and (2.42) only imply that

$$(2.43) \qquad \kappa d\alpha_m = 0.$$

Hence, the fact that (2.39) is exactly closed does not imply that the APL solution is exact. With (2.28′) the condition (2.41) can also be represented as

$$(2.44) \quad d\alpha_m = d\left[\sum_{p=0}^{m} (-K^*d)^{m-p}(K^*\beta_p)\right] = \sum_{p=0}^{m} (-1)^{m-p}(d \circ K^*)^{m-p+1}\beta_p = 0.$$

Let the operator

$$(2.45) \qquad \hat{t} = -d \circ K^*.$$

Therefore, if

$$(2.46) \qquad \sum_{p=0}^{m} \hat{t}^{m-p+1}\beta_p = 0$$

the APL solution is exact.

The conditions for exactness when $m=0$, 1 will be employed in the next section. From (2.46) they are, respectively,

$$(2.47) \qquad \hat{t}\beta_0 = 0,$$

$$(2.48) \qquad \hat{t}(\hat{t}\beta_0 + \beta_1) = 0.$$

The $m=1$ condition (2.48) is satisfied if $\beta_1 = -\hat{t}\beta_0 = d\alpha_0$. In fact, since

$$(2.49) \qquad \sum_{p=0}^{m} \hat{t}^{m-p+1}\beta_p = \hat{t}(\beta_m - d\alpha_{m-1}),$$

the condition for exactness (2.46) is satisfied in the general case if $\beta_m = d\alpha_{m-1}$. In the $m=0$ case (2.12) and (2.40) are satisfied if the p-form $\beta = \beta_0$ is longitudinal and closed:

$$(2.50) \qquad \kappa\beta_0 = 0, \qquad d\beta_0 = 0,$$

conditions which are automatically satisfied if β_0 is an n-form. Subsequently, if V is the vector-field associated to the one-form K [1, App. E], (2.47) can also be written as

$$(2.51) \qquad \mathfrak{L}_V\beta_0 = 0$$

where the Lie derivative [1, App. L]

$$(2.52) \qquad \mathfrak{L}_V = d \circ K^* + K^* \circ d.$$

Thus, if β_0 is invariant along the flow defined by V [1, App. L] in the $m=0$ case, the APL algorithm is exact.

Figure 2 summarizes (2.12), (2.14) and (2.36), i.e., the expressions defining the general APL solution. It extends Fig. 1 by including the terms that can be added to the α's as expressed by (2.36). Parallel arrows designate the same operation $\hat{k}$ or d; each

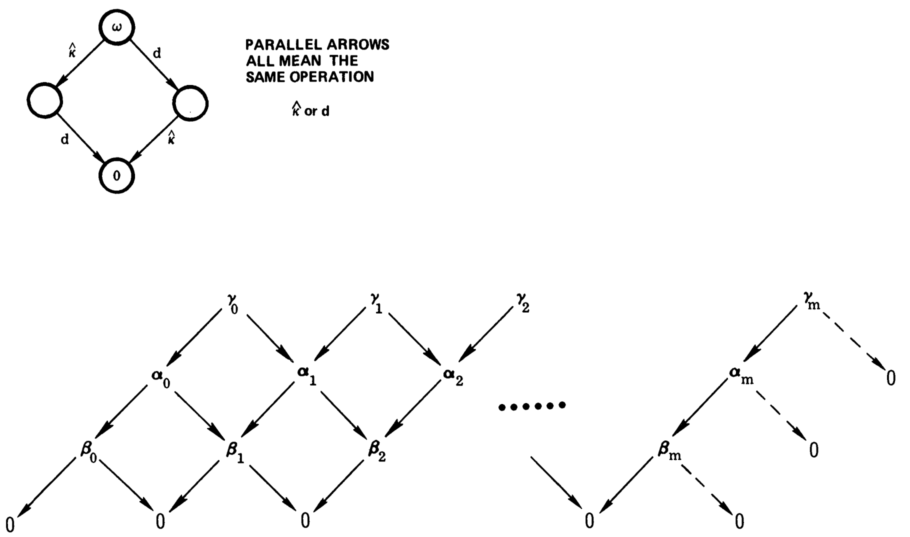


FIG. 2. *General relations characterizing an asymptotic solution of Poincaré's lemma.*

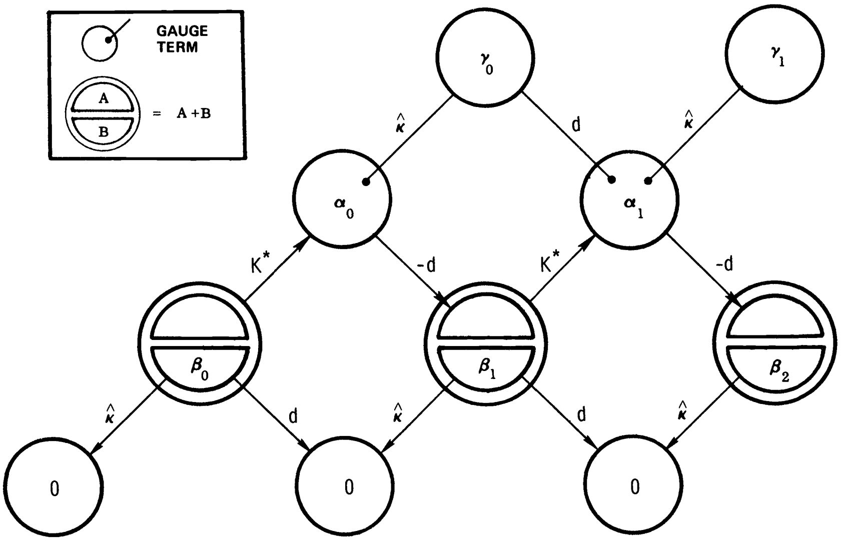


FIG. 3. *The general asymptotic solution of Poincaré's lemma.*

diamond pattern represents the identity (2.17). The dotted arrows on the right express the asymptotic satisfaction of those relations. Since they represent the expression $d\gamma_m = 0$ and (2.41) and (2.40), they also are the conditions which describe when the APL algorithm is exact. Similarly, the diagram given in Fig. 3 summarizes (2.12), (2.28) and (2.36), i.e., the general APL solution. The lines ending in solid dots represent the gauge terms of (2.36). Half-moon elements are added before the operator leading from the circle surrounding them is applied. When $\beta = \beta_0$, this diagram simplifies to the one shown in Fig. 4 which summarizes (2.30) and (2.36) and the condition $\kappa\beta = 0$. Note that with the representation of the operator $-K^* \circ d$ by the dotted lines, the diagram

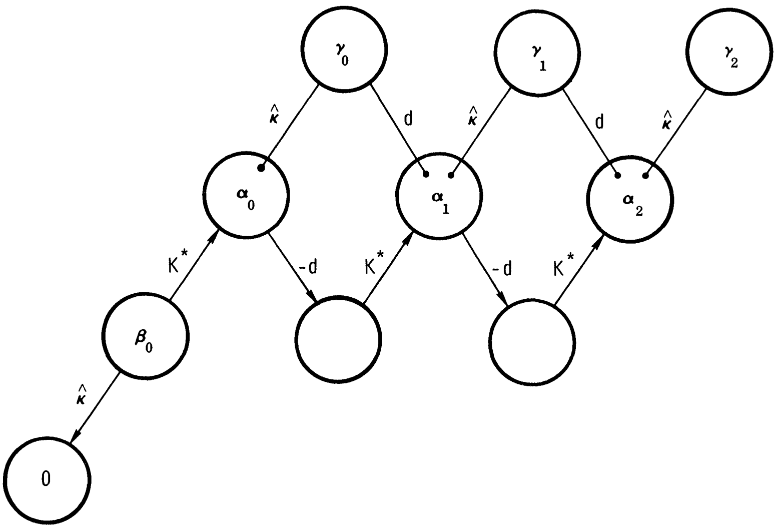


FIG. 4. *The asymptotic solution of Poincaré's lemma when $\beta=\beta_0$.*

suggests a geometrical progression. In particular, when $m=\infty$, the $(p-1)$-form α can be written as

$$\alpha=\nu^{-1}\sum_{j=0}^{\infty}\nu^{-j}(-K^*\circ d)^j\alpha_0=\nu^{-1}[1-\hat{s}]^{-1}\alpha_0, \tag{2.53}$$

where the operator

$$\hat{s}=\nu^{-1}(-K^*\circ d)\equiv -K^*\circ D. \tag{2.54}$$

3. Boundary point evaluation of an integral.

3A. General formalism. The contributions to the (asymptotic) evaluation of the integral

$$\int_{\mathcal{D}} e^{\nu\Gamma}\beta, \tag{3.1}$$

also denoted (see [1, p. 677])

$$e^{\nu\Gamma}\beta|\mathcal{D}, \tag{3.1$'$}$$

from the points on the boundary, Σ, of $\mathcal{D}$ are desired. The domain $\mathcal{D}$ is an oriented p-domain of finite extent in X; its boundary $\Sigma=\partial\mathcal{D}$. If (2.2) holds in $\mathcal{D}$, (3.1) can be evaluated immediately with Stokes' formula [1, App. J] so that

$$e^{\nu\Gamma}\beta|\mathcal{D}=d(e^{\nu\Gamma}\alpha)|\mathcal{D}=e^{\nu\Gamma}\alpha|\partial\mathcal{D}=e^{\nu\Gamma}\alpha|\Sigma. \tag{3.2}$$

The desired asymptotic results follow in a similar fashion from the APL.

It is assumed that Γ has no stationary points in $\mathcal{D}$; i.e., that $\kappa\neq 0$ in $\mathcal{D}$ so that K, hence, α is not singular. The given p-form $e^{\nu\Gamma}\beta$ can be replaced with an exact differential of a $(p-1)$-form $e^{\nu\Gamma}\alpha$ that is a range m APL solution plus a remainder:

$$e^{\nu\Gamma}\beta=d\left[\nu^{-1}\left(e^{\nu\Gamma}\sum_{j=0}^{m}\nu^{-j}\alpha_j\right)\right]-\nu^{-(m+1)}e^{\nu\Gamma}(d\alpha_m). \tag{3.3}$$

Consequently, with Stokes' formula the integral relation corresponding to (3.3) is

$$e^{\nu\Gamma}\beta|\mathcal{D}=\nu^{-1}\sum_{j=0}^{m}\nu^{-j}\left(e^{\nu\Gamma}\alpha_j|\Sigma\right)-\nu^{-(m+1)}\left(e^{\nu\Gamma}d\alpha_m|\mathcal{D}\right). \tag{3.4}$$

The original integral (3.1) over $\mathcal{D}$ has been replaced by an $(m+1)$-term asymptotic series defined over the boundary, Σ, of $\mathcal{D}$ and a remainder term defined over $\mathcal{D}$ that is of degree $(m+1)$ in ν^{-1}. Because this latter term is the same type as the original, it can be evaluated in a similar manner. Hence, it is asymptotically small compared to every other term in the sum. The desired (asymptotic) boundary point evaluation of the integral (3.1) is, therefore,

$$e^{\nu\Gamma}\beta|\mathcal{D}\sim\nu^{-1}\sum_{j=0}^{m}\nu^{-j}\left(e^{\nu\Gamma}\alpha_j|\Sigma\right)\equiv e^{\nu\Gamma}\alpha|\Sigma. \tag{3.5}$$

If $m=0$, this reduces to

$$e^{\nu\Gamma}\beta|\mathcal{D}\sim\nu^{-1}\left(e^{\nu\Gamma}\alpha_0|\Sigma\right). \tag{3.6}$$

Equation (3.5) is a generalization of the expressions given in [4, Chap. 8]. It is applicable to the vector as well as the scalar case. Furthermore, if the domain $\mathcal{D}$ is one-dimensional, (3.5) reduces to the standard endpoint evaluation of an integral given, for example, in [11] or [12]. The following examples demonstrate the utility of the APL-boundary point analysis.

3B. Kirchhoff approximation. Let the space $\mathcal{R}^3$ be divided into two regions V_1 and V_2 separated by an oriented surface $M=\partial V_1$, whose normal points toward V_2. The surface M is composed of a screen S and of an aperture $\mathcal{D}$ (whose edge is $\Sigma=\partial\mathcal{D}$) such that $M=\mathcal{D}\cup S$ and is assumed to lie on side V_2 of the screen [1, Fig. 2]. Consider in the absence of the screen two scalar field solutions u_j $(j=1,2)$ of the Helmholtz equation whose sources ρ_j are in V_j

$$\{\Delta+\kappa^2\}u_j=\rho_j. \tag{3.7}$$

As discussed in [1, §IV], if U_1 is the field due to ρ_1 in the presence of the screen and if ρ_2 is a point source at $\mathbf{s}_2$, the field U_1 at $\mathbf{s}_2$ can be represented in terms of the cross-flux of U_1 and u_2 through M as

$$U_1(\mathbf{s}_2)=(U_1*du_2-u_2*dU_1)|M. \tag{3.8}$$

The Kirchhoff approximation assumes that the field U_1 and its derivative along the normal of M (represented by the term $*dU_1$) are zero over S and are equal, respectively, to u_1 and its normal derivative over $\mathcal{D}$. The resultant representation of the field (3.8) is

$$U_1(\mathbf{s}_2)\cong\beta_{12}|\mathcal{D}\equiv(u_1*du_2-u_2*du_1)|\mathcal{D}. \tag{3.9}$$

The boundary point analysis will be applied to the integral $\beta_{12}|\mathcal{D}$ in several cases. The results represent a reduction of the Kirchhoff approximation of the field (3.9) to a line integral over the edge of the aperture.

1. *Plane wave-plane wave case.* Consider the plane waves $u_j(\mathbf{r})=\exp[\nu(\kappa_j|\mathbf{r})]$, where $(\kappa_j|\mathbf{r})$ is the duality product (see [1, App. D]) of κ_j, a constant unit propagation one-form (i.e., $d\kappa_j=0$ and $\kappa_j\cdot\kappa_j\equiv\kappa_j|\kappa_j\equiv\kappa_j^*\kappa_j=1$), and the position vector $\mathbf{r}$. In Cartesian coordinates, for example, $\kappa_j=\xi_j dx+\eta_j dy+\zeta_j dz$ and $\mathbf{r}=x\partial_x+y\partial_y+z\partial_z$ so that

$\kappa_j|\mathbf{r}=\xi_j x+\eta_j y+\zeta_j z$. Since

$$du_j=\nu\kappa_j u_j, \tag{3.10}$$

the cross-flux two-form

$$\beta_{12}=\nu u_1 u_2*(\kappa_2-\kappa_1)\equiv\nu(e^{\nu\Gamma}\beta), \tag{3.11}$$

where

$$\Gamma=(\kappa_1+\kappa_2)|\mathbf{r} \tag{3.12}$$

and

$$\beta=*(\kappa_2-\kappa_1). \tag{3.13}$$

One verifies that $\kappa\beta\equiv 0$ since $\kappa=d\Gamma=\kappa_1+\kappa_2$. The APL solution is

$$\alpha_0=K^*\beta=-\frac{*\kappa_1\kappa_2}{(1+\kappa_1\cdot\kappa_2)}, \tag{3.14}$$

all other $\alpha_j=0$ $(j=1,2,\cdots)$ because α_0 is a constant. Thus, if

$$\alpha_{12}=e^{\nu\Gamma}\alpha_0=-u_1u_2\left(\frac{*\kappa_1\kappa_2}{1+\kappa_1\cdot\kappa_2}\right), \tag{3.15}$$

the cross-flux integral

$$\beta_{12}|\mathscr{D}=\alpha_{12}|\Sigma. \tag{3.16}$$

Moreover, $d\beta=0$ (β is a constant); therefore, from the preceding section one realizes that (3.16) represents an exact result (hence, the equal sign). Note that (3.16) must be modified when $\kappa=0$; i.e., when $\kappa_1=-\kappa_2$, hence, when $\Gamma, \kappa_1\kappa_2$ and $(1+\kappa_1\cdot\kappa_2)$ are all zero.

2. *Spherical wave-spherical wave case.* Consider the two fields $u_i(\mathbf{r})=G(\mathbf{r}-\mathbf{s}_i)$ $(i=1,2)$ due to point sources at $\mathbf{s}_1$ and $\mathbf{s}_2$. These fields represent spherical waves originating at those points. The Green's function

$$G(\mathbf{r})=\frac{\exp(\nu r)}{4\pi r}, \tag{3.17}$$

where $r=|\mathbf{r}|$. If $r_i=|\mathbf{r}-\mathbf{s}_i|$ and $\kappa_i=dr_i$, the cross-flux two-form

$$\beta_{12}=u_1u_2*\left[\left(\nu-\frac{1}{r_2}\right)\kappa_2-\left(\nu-\frac{1}{r_1}\right)\kappa_1\right]=\nu(e^{\nu\Gamma}\beta), \tag{3.18}$$

where the phase

$$\Gamma=r_1+r_2, \tag{3.19}$$

so that

$$\kappa=d\Gamma=\kappa_1+\kappa_2 \tag{3.20}$$

and the two-forms

$$\beta_0=g_1g_2*(\kappa_2-\kappa_1) \tag{3.21}$$

and

$$\beta_1 = -g_1 g_2 * \left(\frac{\kappa_2}{r_2} - \frac{\kappa_1}{r_1} \right), \tag{3.22}$$

so that $\beta = \beta_0 + \nu^{-1}\beta_1$. The functions

$$g_i(\mathbf{r}) = \frac{1}{4\pi r_i}. \tag{3.23}$$

The corresponding APL solution $\alpha = \nu^{-1}(\alpha_0 + \nu^{-1}\alpha_1)$ is obtained from the system

$$\kappa\alpha_0 = \beta_0, \tag{3.24a}$$

$$\kappa\alpha_1 = \beta_1 - d\alpha_0. \tag{3.24b}$$

Since $\kappa\beta_0 = 0$, (3.24a) is satisfied by

$$\alpha_0 = \frac{(\kappa_1 + \kappa_2)^*\beta_0}{2(1 + \kappa_1 \cdot \kappa_2)} = -g_1 g_2 \left(\frac{{}^*\kappa_1 \kappa_2}{1 + \kappa_1 \cdot \kappa_2} \right). \tag{3.25}$$

However, as shown in Appendix B, this means

$$d\alpha_0 = \beta_1. \tag{3.26}$$

Therefore, (3.24b) together with the condition $\kappa^*\alpha_1 = 0$ yields

$$\alpha_1 \equiv 0. \tag{3.27}$$

Furthermore, (3.26) is the condition required for the exactness of the APL solution. Consequently, where $\kappa \neq 0$, *the differential of the one-form*

$$\alpha_{12} = e^{\nu\Gamma}\alpha_0 = -u_1 u_2 \left(\frac{{}^*\kappa_1 \kappa_2}{1 + \kappa_1 \cdot \kappa_2} \right), \tag{3.28}$$

is identical to the cross-flux two-form (3.18)

$$d\alpha_{12} \equiv \beta_{12}. \tag{3.29}$$

The one-form α_0 is a singular where $\kappa_1 = -\kappa_2$, hence, where $1 + \kappa_1 \cdot \kappa_2 = 0$. This occurs along the line segment $s_1 s_2$ connecting s_1 and s_2. Thus, if the line $s_1 s_2$ does not intersect $\mathfrak{D}$, the field at s_2 is

$$\beta_{12}|\mathfrak{D} = \alpha_{12}|\Sigma. \tag{3.30}$$

On the other hand, if it does intersect $\mathfrak{D}$ and because $\partial S = -\Sigma$, the field at s_2 is

$$\beta_{12}|\mathfrak{D} = \beta_{12}|M - \beta_{12}|S = u_1(\mathrm{s}_2) + \alpha_{12}|\Sigma, \tag{3.31}$$

where the first term follows from Green's theorem. Hence, the field at s_2 is now composed of the geometrical optics field (the first term) in addition to the diffracted field (the second term). The expression of the latter shows that it may be considered as originating on the edge Σ of the aperture, an interpretation that agrees with the viewpoint of the geometrical theory of diffraction (GTD). Furthermore, notice that the phase function Γ is stationary (i.e, $\kappa = 0$) along the line $s_1 s_2$, where $\kappa_1 = -\kappa_2$. In fact, with the results of the following section it can be shown that the geometrical optics term is recovered with the stationary phase approach. Finally, it must be re-emphasized that (3.30) and (3.31) are *exact representations* of the Kirchhoff approximation of the field resulting from the scattering of a spherical wave through an aperture.

3. *Plane wave-spherical wave case*. Consider the plane wave $u_1(\mathbf{r})=\exp[\nu(\kappa_1|\mathbf{r})]$ and the spherical wave $u_2(\mathbf{r})=G(\mathbf{r}-\mathbf{s}_2)$. The results for this case follow immediately from the preceding cases. If $\kappa_1\neq-\kappa_2$, the one-form

$$\alpha_{12}=e^{\nu\Gamma}\alpha_0=-u_1u_2\left(\frac{*\kappa_1\kappa_2}{1+\kappa_1\cdot\kappa_2}\right), \tag{3.32}$$

where

$$\Gamma=(\kappa_1|\mathbf{r})+r_2 \tag{3.33}$$

and

$$\alpha_0=-g_2\left(\frac{*\kappa_1\kappa_2}{1+\kappa_1\cdot\kappa_2}\right), \tag{3.34}$$

is an exact solution of the equation

$$\beta_{12}|\mathcal{D}=\alpha_{12}|\Sigma, \tag{3.35}$$

where the cross-flux two-form

$$\beta_{12}=u_1u_2*\left[\nu(\kappa_2-\kappa_1)-\frac{\kappa_2}{r_2}\right]. \tag{3.36}$$

When $\kappa_1=-\kappa_2$, the geometrical optics field $u_1(\mathbf{s}_2)$ must be added to the diffracted field in (3.35). The case dealing with a general incident field $u_1(\mathbf{r})$ can be handled (exactly) with these results by representing u_1 as a superposition of plane waves.

4. *Asymptotic field-spherical wave case*. Consider the field $u_1(\mathbf{r})=\exp[\nu\Phi(\mathbf{r})]A(\mathbf{r})$ which asymptotically satisfies the Helmholtz equation and the spherical wave $u_2(\mathbf{r})=G(\mathbf{r}-\mathbf{s}_2)$. The cross-flux two-form truncated to an asymptotic order corresponding to that of the incident field u_1 is

$$\beta_{12}\sim\nu u_1u_2*(\kappa_2-\kappa_1). \tag{3.37}$$

The subsequent asymptotic expression

$$\beta_{12}|\mathcal{D}\sim\alpha_{12}|\Sigma, \tag{3.38}$$

where

$$\alpha_{12}=e^{\nu\Gamma}\sum_{j=0}^{m}\nu^{-j}\alpha_j, \tag{3.39}$$

follows immediately from the APL-boundary point formalism. In particular, the phase function

$$\Gamma=\Phi+r_2, \tag{3.40}$$

and if $d\Phi=\kappa_1\neq\kappa_2$,

$$\alpha_0=K^*\beta_0=-Ag_2\left(\frac{*\kappa_1\kappa_2}{1+\kappa_1\cdot\kappa_2}\right), \tag{3.41}$$

so that

$$\alpha_j=\{-K^*d\}^j\alpha_0 \qquad (1\leq j\leq m). \tag{3.42}$$

The stationary point situation (where $d\Gamma=\kappa_1+\kappa_2=0$) is handled as shown in the following section.

The preceding results recover those given in [6]–[8]. However, the discussion has been appreciably simplified using differential forms. Moreover, it refutes the statement in [8, p. 210] that it is impossible to derive in a simple way the representation and the properties of the vector potential which corresponds to the one-form $e^{\nu\Gamma}\alpha$. In fact, in all of the above cases if $\bar{\kappa}_j=\hat{r}_j$ such that $\mathbf{r}_j=r_j\hat{r}_j$, the one-form solutions

$$(3.43)\qquad e^{\nu\Gamma}\alpha=-u_1u_2\left(\frac{*\kappa_1\kappa_2}{1+\kappa_1\cdot\kappa_2}\right)$$

are simply related to the vector potentials $\mathbf{W}$ of [8]

$$(3.44)\qquad \mathbf{W}=e^{\nu\Gamma}\bar{\alpha}=-u_1u_2\frac{\hat{r}_1\times\hat{r}_2}{1+\hat{r}_1\cdot\hat{r}_2}.$$

Furthermore, these results are readily extended to vector-field problems. A discussion of those problems is in preparation.

4. Stationary point contributions. The asymptotic evaluation of the integral

$$(4.1)\qquad e^{\nu\Gamma}\beta|X=e^{\nu\Gamma(x)}A(x)\,dx^N|X,$$

where $x=(x_1,x_2,\cdots,x_n)$ is a point in X and the volume n-form $dx^N=dx^1\,dx^2\cdots dx^n$, is desired when a nondegenerate stationary point, x_0, is the only critical point of Γ in X and the function A is smooth over X, is zero at infinity and is integrable. A point x_0 is a stationary point of Γ if

$$(4.2)\qquad d\Gamma(x_0)=0;$$

it is nondegenerate if, in addition,

$$(4.3)\qquad \det\left\|\left(\partial_{x_i}\partial_{x_j}\Gamma\right)(x_0)\right\|\neq 0,$$

where $\partial_{x_j}\Gamma$ means $\partial\Gamma/\partial x_j$. Note that only the case for which β is an n-form is considered explicitly. The general case in which β is a p-form integrated over a p-domain is handled in a similar manner; the generalizations of the following results to that case will be apparent. Also, if there is more than one stationary point of Γ in X and if they are not near to one another, each stationary point can be localized with neutralizers as shown, for instance, in [4] and treated like the present case. Problems involving the coalescing of stationary points, branch points, poles, and so on will not be discussed. The situation where the domain is an n-domain $\mathfrak{D}$ rather than the whole space X will be discussed at the end of this section.

In the vicinity of x_0 the Morse lemma [13] realizes a change of variables which reduces the phase to a quadratic form with coefficients (± 1). Denote the new variables by $u=(u_1,\cdots,u_n)$. They are functions of x in the vicinity of x_0; hence, they may be expressed as $u(x,x_0)$. Furthermore, they satisfy the relation

$$(4.4)\qquad \Gamma(x)-\Gamma(x_0)=\frac{1}{2}\sum_{j=1}^{n}\eta_j u_j^2(x,x_0)\equiv\frac{1}{2}\eta\cdot u^2(x,x_0),$$

where $\eta_j=\pm 1$. For each x_0 there is an associated map

$$(4.5)\qquad \mu_{x_0}\colon U\to X\colon u\mapsto x=\mu_{x_0}u\colon 0\mapsto x_0,$$

which expresses x in terms of the new variable u. It will be called the Morse map. The desired quadratic form:

$$Q(u)=\frac{1}{2}\sum_{j=1}^{n}\eta_j u_j^2 \tag{4.6}$$

is defined as

$$\mu_{x_0}^*[\Gamma(x)-\Gamma(x_0)]=Q(u), \tag{4.7}$$

where $\mu_{x_0}^*$ is the pullback through the Morse map, μ_{x_0}. To complete the integral (4.1), the change of variables or pullback $\mu_{x_0}^*$ must be applied to the amplitude n-form, β. This yields

$$\mu_{x_0}^*\beta=\mu_{x_0}^*[A(x)\,dx^N]=G(u,x_0)\,du^N, \tag{4.8}$$

where

$$G(u,x_0)=\mu_{x_0}^*\left[A(x)\left(\frac{du^N}{dx^N}\right)^{-1}\right]=A(\mu_{x_0}u)J(u,x_0) \tag{4.9}$$

and the Jacobian of the transformation from the u to the $x=\mu_{x_0}u$ coordinates:

$$J(u,x_0)=\left\{\mu_{x_0}^*\left[\det\|\partial_{x_j}u_i(x,x_0)\|\right]\right\}^{-1}=\det\|\partial_{u_j}(x=\mu_{x_0}u)_i\|. \tag{4.10}$$

Finally, with this change of variables the integral (4.1) becomes

$$e^{\nu\Gamma}\beta|X=e^{\nu\Gamma(x_0)}[e^{\nu Q(u)}G(u,x_0)\,du^N|U]\equiv I(\nu,x_0). \tag{4.11}$$

This expression can be rewritten immediately as

$$\begin{aligned}I(\nu,x_0)=&\,e^{\nu\Gamma(x_0)}G(0,x_0)[e^{\nu Q(u)}\,du^N|U]\\&+e^{\nu\Gamma(x_0)}\{e^{\nu Q(u)}[G(u,x_0)-G(0,x_0)]\,du^N|U\}.\end{aligned} \tag{4.12}$$

The first integral in (4.12) is simply [4], [14]

$$e^{\nu Q(u)}\,du^N|U=c_n c^{-1}=\left(\frac{k}{\pi}\right)^{-n/2}\exp\left[i\left(\frac{\pi}{4}\right)\operatorname{sgn}\eta\right], \tag{4.13}$$

where

$$c_n=2^{-n/2}\exp\left[i\left(\frac{\pi}{2}\right)\operatorname{ind}\eta\right] \tag{4.14}$$

and

$$c=\left(\frac{k}{2\pi}\right)^{n/2}\exp\left[-\frac{in\pi}{4}\right] \tag{4.15}$$

and where if n_+ and n_- are, respectively, the number of positive and negative η_j $(j=1,2,\cdots,n)$, then the signature $\operatorname{sgn}\eta=n_+-n_-$ and the index $\operatorname{Ind}\eta=n_-$. The results of the lemmas proved in Appendix C allow one to manipulate the second integral into a form suitable for an asymptotic evaluation. In particular, let the one-form

$$\rho=\sum_{i=1}^{n}\eta_i u_i\,du^i. \tag{4.16}$$

As shown in Appendix C, the amplitude n-form $G(u,x_0)\,du^N$ can be expressed in terms of its value at the stationary point $u=0$ of $Q(u)$ and a term linear in ρ

$$G(u,x_0)\,du^N=G(0,x_0)\,du^N+\rho H_0(u,x_0). \tag{4.17}$$

A suitable choice for the $(n-1)$-form H_0 is generated with the inversion algorithm introduced in §2

$$H_0=(\rho\cdot\rho)^{-1}\rho^*\{[G(u,x_0)-G(0,x_0)]\,du^N\}. \tag{4.18}$$

Comments on the nonuniqueness of this choice also follow from those given in §2. With (4.13) and (4.17) the expression (4.12) becomes

$$I(\nu,x_0)=(c_nc^{-1})e^{\nu\Gamma(x_0)}G(0,x_0)+e^{\nu\Gamma(x_0)}[e^{\nu Q(u)}\rho H_0(u,x_0)|U]. \tag{4.19}$$

The integral in (4.19) can be evaluated using the APL algorithm. Since

$$\beta_0=\rho H_0 \tag{4.20}$$

is an n-form and since

$$\kappa=d_uQ=\rho, \tag{4.21}$$

the condition

$$\kappa\beta_0=0 \tag{4.22}$$

is trivially satisfied. Although the (sufficient) condition $\kappa\neq 0$ was needed in the preceding section, we can still give meaning to the expression $\alpha_0=K^*\beta_0$. Because β_0, as well as κ, is zero at the stationary point $u=0$, the combination $K^*\beta_0$ has a finite limit there. Therefore, the $(n-1)$-form α_0 is well defined over U. Furthermore, since $\rho^*H_0=0$,

$$\alpha_0=K^*\beta_0=(\rho\cdot\rho)^{-1}\rho^*(\rho H_0)\equiv H_0(u,x_0). \tag{4.23}$$

With the relation

$$D_u[e^{\nu Q(u)}\alpha_0(u)]=e^{\nu Q(u)}\beta_0(u)+e^{\nu Q(u)}D_u\alpha_0, \tag{4.24}$$

one immediately obtains

$$e^{\nu Q(u)}\beta_0|U=\nu^{-1}[e^{\nu Q(u)}H_0|\partial U]+\nu^{-1}[e^{\nu Q(u)}G_1(u,x_0)\,du^N|U], \tag{4.25}$$

where

$$G_1(u,x_0)\,du^N=-d_uH_0. \tag{4.26}$$

Because the boundary, ∂U, of U is at infinity and the amplitude function A is zero there, the boundary integral in (4.25) is zero. The other integral has the same form as the original integral in (4.19); hence, this process can be repeated. After $(m+1)$ steps (4.12) becomes

$$\begin{aligned} I(\nu,x_0)={}&e^{\nu\Gamma(x_0)}(c_nc^{-1})\sum_{j=0}^{m}\nu^{-j}G_j(0,x_0)\\ &+\nu^{-(m+1)}e^{\nu\Gamma(x_0)}[e^{\nu Q(u)}G_{m+1}(u,x_0)\,du^N|U], \end{aligned} \tag{4.27}$$

where $G_0 = G$,

$$G_j(u,x_0)\,du^N = G_j(0,x_0)\,du^N + \rho H_j(u,x_0) \qquad (0 \le j \le m) \tag{4.28}$$

and

$$G_i(u,x_0)\,du^N = -d_u H_{i-1} \qquad (1 \le i \le m+1). \tag{4.29}$$

Truncating the remainder term in (4.27) which is of degree $(m+1)$ in ν^{-1}, the $(m+1)$-term stationary phase approximation of the integral (4.1) is

$$I(\nu,x_0) \sim e^{\nu\Gamma(x_0)}\left(c_n c^{-1}\right) \sum_{j=0}^{m} \nu^{-j} G_j(0,x_0). \tag{4.30}$$

Furthermore, as shown in Appendix D,

$$G_j(0,x_0) = \left(-\frac{1}{2}\right)^j \left\{\frac{1}{j!}\tilde{\Delta}^j_{u/0}\right\} G(u,x_0), \tag{4.31}$$

where the (modified) Laplacian operator

$$\tilde{\Delta}_u \equiv \sum_{j=1}^{n} \eta_j \partial^2_{u_j} \equiv \eta \cdot \partial^2_u \tag{4.32}$$

and

$$\tilde{\Delta}_{u/0} f(u) = \tilde{\Delta}_u f(u)\big|_{u=0}. \tag{4.33}$$

The standard stationary phase expressions are readily obtained from (4.30) and (4.31) [14]. Furthermore, the preceding derivation is a generalization of the ones given in [4] and [5] which employed vector identities (divergence theorem) and, hence, were restricted to Euclidean spaces. In contrast with [4] and [5], for example, the results of this section remain valid in cases where X is a manifold.

One can rewrite the expression (4.31) in a more manifest form [14]

$$cI(\nu,x_0) \sim e^{\nu\Gamma(x_0)} \sum_{j=0}^{m} \nu^{-j} T_j^{\Gamma} A(x_0), \tag{4.34}$$

where

$$T_j^{\Gamma} A(x_0) = \left\{c_n\left(-\frac{1}{2}\right)^j \left[\frac{1}{j!}\tilde{\Delta}^j_{u/0}\right] J(u,x_0)\mu^*_{x_0}\right\} A(x) \equiv c_n G_j(0,x_0), \tag{4.35}$$

each term in the braces being treated as operators and applied in order from right to left. The advantages are that the leading term of the series is independent of k and that the contributions to the expansion from the phase and the amplitude are apparent. The operators T_j^{Γ} depend on derivatives of the phase evaluated at the stationary point at most to the order $2(j+1)$ and contain derivatives which are applied to the amplitude and evaluated at the stationary point at most to the order $2j$. For instance, when $n=1$, if $\Gamma_i = \partial^i_{x/x_0}\Gamma$, then

$$T_0^{\Gamma} = \Gamma_2^{-1/2}\partial_{x/x_0}, \tag{4.36}$$

and

$$T_1^{\Gamma} = -\frac{1}{24\Gamma_2^{3/2}}\left[\left(5\Gamma_3^2 - 3\Gamma_4\Gamma_2\right)\partial^0_{x/x_0} - 12\Gamma_3\Gamma_2\partial^1_{x/x_0} + 12\Gamma_2^2\partial^2_{x/x_0}\right]. \tag{4.37}$$

Applications and ramifications of these results are given in [14] in connection with the asymptotic evaluation of the Fourier transform.

Now consider the integral $ce^{\nu\Gamma}\beta|\mathcal{D}$, where $\mathcal{D}$ is an n-domain in X. Let Γ have an isolated, nondegenerate stationary point, x_0, in the interior of $\mathcal{D}$; in particular, $d\Gamma \neq 0$ on $\Sigma = \partial\mathcal{D}$. Let $\overline{\mathcal{D}}$ be the complement of $\mathcal{D}$ in X whose boundary is oriented such that $\partial\overline{\mathcal{D}} = -\Sigma$. Thus,

$$ce^{\nu\Gamma}\beta|\mathcal{D} \equiv ce^{\nu\Gamma}\beta|X - ce^{\nu\Gamma}\beta|\overline{\mathcal{D}}. \tag{4.38}$$

The asymptotic evaluation of the first integral clearly yields the stationary phase contribution, $cI(\nu, x_0) \equiv F(\nu, x_0)$. Since $\overline{\mathcal{D}}$ is devoid of any stationary points, the second integral can be handled with the boundary point formalism. Thus, with the APL solution

$$\alpha = \nu^{-1} \sum_{j=0}^{m} \nu^{-j}\alpha_j \equiv \nu^{-1}\tilde{\alpha}, \tag{4.39}$$

it gives

$$-ce^{\nu\Gamma}\beta|\overline{\mathcal{D}} = ce^{\nu\Gamma}\alpha|\Sigma = (\nu^{-1}c)\left[e^{\nu\Gamma}\tilde{\alpha}|\Sigma\right] = \nu^{-1/2}\left[\tilde{c}e^{\nu\Gamma}\tilde{\alpha}|\Sigma\right] \equiv \nu^{-1/2}E(\nu, \Sigma), \tag{4.40}$$

where

$$\tilde{c} = -i(2\pi)^{-1/2}\left\{\left(\frac{k}{2\pi}\right)^{(n-1)/2} \exp\left[-i\left(\frac{\pi}{4}\right)(n-1)\right]\right\} \tag{4.41}$$

Therefore,

$$ce^{\nu\Gamma}\beta|\mathcal{D} \sim F(\nu, x_0) + \nu^{-1/2}E(\nu, \Sigma); \tag{4.42}$$

the total asymptotic expansion of the integral is defined in terms of the stationary and the boundary point contributions. This result assumes those contributions are independent. Modifications of (4.42) would be necessary if this condition was not satisfied.

Notice that the boundary integral can be reparameterized in terms of local coordinates on Σ. The resulting integral can then be evaluated asymptotically. If the resultant phase function is stationary at some point (a critical point of the second kind; the stationary point x_0 in the interior of $\mathcal{D}$ being a critical point of the first kind), the stationary phase formalism can be applied directly to that integral. In electromagnetics these terms account for the diffracted ray contributions to the total field—those rays generated from boundaries such as edges. The stationary points of the first kind, on the other hand, produce the geometrical optics terms. The characteristic $k^{-1/2}$ difference between these contributions is apparent in (4.42). However, in some instances every point on the boundary is a stationary point, and it becomes necessary to keep the nonlocal integral representation of the boundary point contributions. Similarly, if the boundary Σ has critical points such as discontinuities in its tangents (critical points of the third kind) or in its curvature (critical points of the fourth kind), the boundary Σ can be subdivided into regions over which the derivatives are continuous to a certain order and whose boundaries coincide with the points of discontinuity. The integrals over these subregions can now be treated with the boundary point formalism. Standard results given in [4] or [5] are readily recovered. Note, however, that the differential form expressions are especially suited to calculations of these types.

Finally, consider the case where x_0 is a nondegenerate stationary point of Γ and lies on the boundary Σ. Near x_0 a local coordinate system $u=(u',u_n)$ (where $u'=(u_1,\cdots,u_{n-1})$ defines a point on Σ and u_n is defined along the unit normal to Σ at x_0) can be constructed such that the transformed phase function, $\mu^*_{x_0}\Gamma$, is stationary at $u=(u',u_n)=0$ and takes the form $(\mu^*_{x_0}\Gamma)(u)=\Gamma'(u',x_0)+\eta u_n^2$, where $\eta=+1$ (-1) if u_n is positive (negative) for a point in the interior of $\mathfrak{D}$. Consequently, one has

$$(4.43)\qquad e^{\nu\Gamma}\beta|\mathfrak{D}=\int_0^\tau e^{\nu\eta u_n^2}g(u_n,x_0;\nu)\,d(\eta u_n),$$

where

$$(4.44)\qquad g(u_n,x_0;\nu)=e^{\nu\Gamma'(u',x_0)}G(u',u_n,x_0)\,du'|\Sigma$$

and where τ is a real positive constant. Note that with k large, the constant τ can be replaced with ∞, [4]. The asymptotic approximation of the original integral is obtained by applying the stationary phase approach first to the $(n-1)$-dimensional integral over Σ treating the u_n-variable as a parameter and then to the one-dimensional integral over u_n. The final result differs from (4.34) because of the form of the latter integration. Because that integration is only over the nonnegative reals, odd orders of derivatives and (referring to (4.13)) the coefficient $\frac{1}{2}$, characteristic of this case, appear. The expressions given in [4] and [5] are readily reproduced.

5. Conclusions. Exterior differential calculus techniques were used to formulate and to obtain asymptotic solutions of Poincaré's lemma. In particular, a new method of solution of a general type of differential form equation was developed. Several applications of these asymptotic Poincaré lemma results were presented. The boundary and stationary point contributions to the asymptotic approximation of a multidimensional integral were derived. Other critical point contributions and asymptotic techniques were also discussed. The boundary point approach was applied to the Kirchhoff representation of the diffraction of a scalar field through an aperture. A representation of the Leray form was synthesized that did not require the introduction of any local coordinate system. In all of these applications the resultant differential form expressions encompass, as special cases, standard vector calculus representations. Furthermore, in contrast with their vector counterparts the differential form expressions are easier to obtain and their properties are more transparent. The asymptotic Poincaré lemma and the associated techniques constitute a versatile approach to a large class of problems encountered in physics and engineering.

Appendix A: The Leray form. Consider an $(n-1)$-dimensional hypersurface S in X. A neighborhood V of a point on S can be defined by the equation $P(x_1,\cdots,x_n)=0$, where P is an infinitely differentiable function such that $dP\neq 0$ on V (i.e., there are no singular points on V). A form ω, which satisfies

$$(\mathrm{A}.1)\qquad dx^N=dP\omega,$$

is readily obtained with the inversion algorithm introduced in §2. In particular, since $dP\,dx^N\equiv 0$ (the volume form dx^N is an n-form), a solution of (A.1) is

$$(\mathrm{A}.2)\qquad \omega=\frac{dP^*\,dx^N}{dP^*\,dP}=(-1)^s\frac{*dP}{dP\cdot dP},$$

where s is the index of the metric associated with the space X. This $(n-1)$-form is called the *Leray form* [9, Chap. III, §1]. Note that we have taken

$$*dx^N=1 \tag{A.3}$$

and

$$*^{-1}=(-1)^s*w^{(n+1)}, \tag{A.4}$$

where w is the operator which applied to a p-form β gives $w\beta=(-1)^p\beta$. Clearly, the Leray form depends only on the function P by which V is represented. If $P(x)$ represents (up to higher order terms) the distance from x to V, the $(n-1)$-form ω reduces to the Euclidean element of area on S. Statements concerning uniqueness follow directly from those discussed in the APL. In particular, the form $\omega+(dP)\gamma$, where γ is any $(n-2)$-form, is also a solution of (A.1). Notice that if one assumes on V some $\partial_{x_j}P\neq 0$ such that $dP=(\partial_{x_j}P)\,dx^j$, (A.2) reduces to

$$\omega=(-1)^{j-1}\frac{dx^1\cdots dx^{j-1}dx^{j+1}\cdots dx^n}{\partial_{x_j}P}. \tag{A.5}$$

Similar arguments can be applied to an $(n-j)$-dimensional manifold defined locally by the equations: $P_1(x)=0$, $P_2(x)=0,\cdots,P_j(x)=0$. The $(n-j)$-form

$$\omega=\frac{(dP_1\,dP_2\cdots dP_j)^*\,dx^N}{(dP_1\cdots dP_j)^*(dP_1\cdots dP_j)}=\frac{(-1)^s*(dP_1\cdots dP_j)}{(dP_1\cdots dP_j)\cdot(dP_1\cdots dP_j)} \tag{A.6}$$

satisfies the relation

$$dx^N=dP_1\,dP_2\cdots dP_j\,\omega. \tag{A.7}$$

If γ is any $(n-2j)$-form, the form $\omega+(dP_1\cdots dP_j)\gamma$ is also a solution of (A.7). Furthermore, with the expression

$$dP_1\cdots dP_j=J(P_1,\cdots,P_j;x_1,\cdots,x_j)\,dx^1\cdots dx^j, \tag{A.8}$$

where the Jacobian

$$J(P_1,\cdots,P_j;x_1,\cdots,x_j)=\det\|\partial_{x_i}P_k\|_{(i,k=1,\cdots,j)}, \tag{A.9}$$

and with the adjoint operator identity $(\alpha_1\alpha_2)^*=\alpha_2^*\alpha_1^*$, the Leray form (A.6) becomes

$$\omega=\frac{dx^{j+1}\cdots dx^n}{J(P_1,\cdots,P_j;x_1,\cdots,x_j)}. \tag{A.10}$$

The preceding results coincide with those given in [9]. Note, however, that the present approach differs from the standard construct employed in [9]. For instance, (A.5) can be derived by introducing the local coordinates $(u_1,\cdots,u_n)=u$ such that $u_i=x_i$ for $i\neq j$ and $u_j=P_j$, hence, $J(x;u)=(\partial_{x_j}P)^{-1}$ and

$$dx^N=J(x;u)\,du^1\cdots du^{j-1}\,dP\,du^{j+1}\cdots du^n. \tag{A.11}$$

Equation (A.3) is recovered immediately from (A.1) and (A.11). On the other hand, the Leray form expressions (A.2) and (A.6) are globally valid and avoid the interjection of the local coordinate system.

Appendix B: The APL solution to the Kirchhoff diffraction of two spherical waves is exact. It will be shown that the 1-term APL solution to the Kirchhoff diffraction of two spherical waves is exact; i.e., that $d\alpha_0=\beta_1$. Let $\rho_j=r_j\kappa_j$. The one-form (3.25) can then be rewritten as

$$\alpha_0=-g_1g_2\frac{*\rho_1\rho_2}{r_1r_2+\rho_1\cdot\rho_2}. \tag{B.1}$$

Let

$$A=r_1r_2(r_1r_2+\rho_1\cdot\rho_2) \tag{B.2}$$

and

$$B=2r_1r_2+\rho_1\cdot\rho_2. \tag{B.3}$$

Since

$$d(\rho_1\cdot\rho_2)=\rho_1+\rho_2, \tag{B.4}$$

one has

$$dA=(r_1r_2)^{-1}\left[(B+r_1^2)r_2^2\rho_1+(B+r_2^2)r_1^2\rho_2\right] \tag{B.5}$$

so that

$$2A-(dA^*\rho_2)=-(r_1r_2)^{-1}r_2^2\left[B(\rho_1\cdot\rho_2)+(r_1r_2)^2\right]=-\left(\frac{r_2}{r_1}\right)(r_1r_2+\rho_1\cdot\rho_2)^2 \tag{B.6}$$

and

$$2A-(dA^*\rho_1)=-\left(\frac{r_1}{r_2}\right)(r_1r_2+\rho_1\cdot\rho_2)^2. \tag{B.7}$$

Consequently, with the identity

$$d^*(h\gamma)=hd^*\gamma-(dh)^*\gamma, \tag{B.8}$$

where h is a scalar function, γ is any q-form and the codifferential operator [1, F. 19]

$$d^*=*^{-1}d*(-1)^q, \tag{B.9}$$

when it is applied to a q-form, and with the relation

$$d^*(\rho_1\rho_2)=2(\rho_1-\rho_2), \tag{B.10}$$

one obtains

$$\begin{aligned}-(4\pi)^2d\alpha_0=*\left[d^*(A^{-1}\rho_1\rho_2)\right]&=A^{-2}*\left[2A(\rho_1-\rho_2)+dA^*(\rho_1\rho_2)\right]\\&=A^{-2}*\left[(2A-dA^*\rho_2)\rho_1-(2A-dA^*\rho_1)\rho_2\right]\\&=(r_1r_2)^{-1}*\left(\frac{\rho_2}{r_2^2}-\frac{\rho_1}{r_1^2}\right),\end{aligned} \tag{B.11}$$

so that

$$d\alpha_0=-g_1g_2*\left(\frac{\kappa_2}{r_2}-\frac{\kappa_1}{r_1}\right)\equiv\beta_1. \tag{B.12}$$

Appendix C: Linear representation of a p-form about a point. The inversion algorithm will be used in this appendix to generalize the following lemma to p-forms.

LEMMA C.1. *Let f be a C^∞ function in a convex neighborhood M of the point $x_0=(0,\cdots,0,x_{p+1},\cdots,x_n)$ in X. Then*

$$f(x)-f(x_0)=\sum_{j=1}^{p} x_j h_j(x) \tag{C.1}$$

for some suitable C^∞ functions h_j defined in M, with $(\partial_{x_j} f)(x_0)=h_j(x_0)$.

Proof.

$$f(x)-f(x_0)=\int_0^1 \frac{df}{dt}(tx_1,\cdots,tx_p,x_{p+1},\cdots,x_n)\,dt$$

$$=\int_0^1 \sum_{j=1}^{p} \frac{\partial f}{\partial x_j}(tx_1,\cdots,tx_p,x_{p+1},\cdots,x_n)x_j\,dt.$$

Therefore, set $h_j(x)=\int_0^1(\partial f/\partial x_j)\,(tx_1,\cdots,tx_p,x_{p+1},\cdots,x_n)\,dt$. □

Note that Lemma C.1 is a simple extension of [13, Lemma 2.1]. Now consider the following lemma.

LEMMA C.2. *In X let the p-form*

$$\beta(x)=a_J(x)\,dx^J, \tag{C.2}$$

where J is the multi-index of length p: $J=j_1j_2\cdots j_p$, so that

$$dx^J=dx^{j_1}dx^{j_2}\cdots dx^{j_p}.$$

With the subset $\mathfrak{J}=\{j_1,j_2,\cdots,j_p\}$ of the set $\{1,2,\cdots,n\}$, let the one-form

$$\theta=\sum_{j\in\mathfrak{J}} f_j\,dx^j, \tag{C.3}$$

where f_j is a scalar function. Then, if not all of the f_j are zero, the p-form (C.2) *has the linear representation*

$$\beta(x)=\theta\alpha(x), \tag{C.4}$$

where α is some suitable $(p-1)$-form.

Proof. The relation (C.4) follows immediately from the inversion algorithm discussed in §2. In particular, the necessary condition

$$\theta\beta(x)=0 \tag{C.5}$$

is trivially satisfied. Thus, since at least one $f_j\neq 0$, hence, $\theta\neq 0$, set

$$\alpha(x)=(\theta^*\theta)^{-1}[\theta^*\beta(x)]. \tag{C.6}$$ □

COROLLARY. *Let the one-form*

$$\theta=\sum_{j\in\mathfrak{J}} c_j x_j\,dx^j, \tag{C.7}$$

where c_j is a constant, and let $x_0=(x_{10},\cdots,x_{n0})$ be the point whose components $x_{j0}=0$ for $j\in\mathfrak{J}$. Then about x_0 one has the linear representation

$$\beta(x)-\beta(x_0)=\theta H(x). \tag{C.8}$$

for some suitable $(p-1)$-form H.

The corollary is clearly a special case of Lemma C.2; a suitable H is

$$H(x)=(\theta^*\theta)^{-1}\theta^*[\beta(x)-\beta(x_0)]. \tag{C.9}$$

Equation (C.8) is the desired generalization of (C.1).

Appendix D: A representation of the stationary phase amplitudes. The (modified) Laplacian

$$\tilde{\Delta}_u=\eta\cdot\partial_u^2\equiv L_u$$

applied to a p-form

$$\alpha=\sum_J a_J\,du^J,$$

where J is a multi-index of length p, acts only on its coefficients:

$$L_u\alpha=\sum_J (L_u a_J)\,du^J.$$

If the q-form

$$H_j=\sum_I h_{jI}\,du^I,$$

where I is a multi-index of length q and the one-form

$$\rho=\sum_{i=1}^n \eta_i u_i\,du^i$$

then

$$\begin{aligned}L_u(\rho H_j)&=\sum_{i=1}^n\sum_I\left[\eta_i L_u(u_i h_{jI})\right]du^i\,du^I \\ &=\sum_{i=1}^n\sum_I\left[\eta_i u_i(L_u h_{jI})\right]du^i\,du^I+2\sum_{i=1}^n\sum_{l=1}^n\sum_I\left(\eta_i\eta_l\partial_{u_l}u_i\partial_{u_l}h_{jI}\right)du^i\,du^I \\ &=\rho L_u H_j+2\sum_{i=1}^n\sum_I\left(\partial_{u_i}h_{jI}\right)du^i\,du^I=\rho L_u H_j+2d_u H_j.\end{aligned} \tag{D.1}$$

With the identity

$$d_u\circ L_u=L_u\circ d_u,$$

repeated applications of (D.1) give

$$\begin{aligned}L_u^m(\rho H_j)&=L_u^{m-1}(\rho L_u H_j+2d_u H_j)=L_u^{m-2}\left[\rho L_u^2 H_j+2d_u(L_u H_j)+2L_u(d_u H_j)\right] \\ &=L_u^{m-2}\left[\rho L_u^2 H_j+4L_u(d_u H_j)\right]=\cdots=\rho L_u^m H_j+2mL_u^{m-1}(d_u H_j).\end{aligned}$$

Thus,

$$\tilde{\Delta}^m_{u/0}(\rho H_j)=2m\tilde{\Delta}^{m-1}_{u/0}(d_u H_j). \tag{D.2}$$

Now consider the p-form $\beta(u,x_0)$ and the $(p-1)$-form $H_j(u,x_0)$ which satisfy the system

$$\begin{aligned}\beta_j(u,x_0)&=\beta_j(0,x_0)+\rho H_j(u,x_0) \qquad (0\le j\le m),\\ \beta_j(u,x_0)&=-d_u H_{i-1} \qquad (1\le i\le m+1).\end{aligned}$$

Equation (D.2) then gives

$$\tilde{\Delta}^m_{u/0}\beta_j(u,x_0)=-2m\tilde{\Delta}^{m-1}_{u/0}\beta_{j+1}(u,x_0).$$

Therefore,

$$\begin{aligned}\tilde{\Delta}^j_{u/0}\beta_0(u,x_0)&=-2j\tilde{\Delta}^{j-1}_{u/0}\beta_1(u,x_0)=(-2)^2 j(j-1)\tilde{\Delta}^{j-2}_{u/0}\beta_2(u,x_0)\\ &=\cdots=(-2)^j j!\beta_j(0,x_0),\end{aligned}$$

hence

$$\beta_j(0,x_0)=\left(-\frac{1}{2}\right)^j\left\{\frac{1}{j!}\tilde{\Delta}^j_{u/0}\right\}\beta_0(u,x_0). \tag{D.3}$$

Consequently, with the n-form

$$\beta_j(u,x_0)=G_j(u,x_0)\,du^N,$$

(D.3) yields

$$G_j(0,x_0)=\left(-\frac{1}{2}\right)^j\left\{\frac{1}{j!}\tilde{\Delta}^j_{u/0}\right\}G_0(u,x_0). \tag{D.4}$$

REFERENCES

[1] G. A. Deschamps, *Electromagnetics and differential forms*, Proc. IEEE, 69 (1981), pp. 676–696.

[2] F. W. Warner, *Foundations of Differential Manifolds and Lie Groups*, Scott, Foresman & Co., Glenview, IL, 1971.

[3] H. Flanders, *Differential Forms*, Academic Press, New York, 1963.

[4] N. Bleistein and R. A. Handelman, *Asymptotic Expansions of Integrals*, Holt, Rinehart and Winston, New York, 1975.

[5] M. V. Fedoryuk, *The stationary phase method in the multidimensional case. Contribution from the region boundary*, Zh. Vychisl. Mat. i Mat. Fiz., 10 (1970), pp. 286–299; USSR Comp. Math. Math. Phys., 10 (1970), pp. 5–23.

[6] K. Miyamoto and E. Wolf, *Generalization of the Maggi–Rubinowicz theory of the boundary diffraction wave-Part* I, J. Opt. Soc. Amer., 52 (1962), pp. 615–625.

[7] ———, *Generalization of the Maggi–Rubinowicz theory of the boundary diffraction wave-Part* II, J. Opt. Soc. Amer., 52 (1962), pp. 626–637.

[8] A. Rubinowicz, *The Miyamoto–Wolf diffraction wave*, Progress in Physics, Vol. IV, E. Wolf, ed., North-Holland, Amsterdam, 1965, pp. 201–240.

[9] I. M. Gelfand and G. E. Shilov, *Generalized Functions*, *Vol.* I: *Properties and Operations*, Academic Press, New York, 1964.

[10] D. S. Jones and M. Kline, *Asymptotic expansion of multiple integrals and the method of stationary phase*, J. Math. Phys., 37 (1958), pp. 1–28.

[11] A. Erdelyi, *Asymptotic Expansions*, Dover, New York, 1956.

[12] E. T. Copson, *Asymptotic Expansions*, Cambridge Univ. Press, Cambridge, 1965.

[13] J. Milnor, *Morse Theory*, Princeton Univ. Press, Princeton, NJ, 1969.

[14] R. W. Ziolkowski, *The Maslov method and the asymptotic Fourier transform: caustic analysis*, Ph. D. Thesis, Univ. Illinois, Urbana, September 1981.

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics

010

INTEGRAL REPRESENTATIONS FOR PRODUCTS OF LAMÉ FUNCTIONS BY USE OF FUNDAMENTAL SOLUTIONS*

HANS VOLKMER†

Abstract. In this paper we present integral representations for products of Lamé functions based on the theory of fundamental solutions. The kernels of these representations involve Legendre functions of the second kind. In particular, we generalize and improve integral representations for external ellipsoidal harmonics mentioned by Erdélyi, Magnus, Oberhettinger and Tricomi [*Higher Transcendental Functions* III, McGraw Hill, New York, 1955] and for Lamé functions of the second kind in terms of Lamé polynomials studied by Shail [SIAM J. Math. Anal., 11 (1980), pp. 702–723].

Introduction. Consider the partial differential equation

$$\frac{\partial^2 w}{\partial x^2}-\frac{\partial^2 w}{\partial y^2}+\nu(\nu+1)k^2(\mathrm{sn}^2 y-\mathrm{sn}^2 x)w=0, \tag{0.1}$$

where x,y are complex variables, ν is a complex number and k is the modulus of the Jacobian elliptic functions sn, cn, dn. Separating (0.1) we obtain, for each variable x,y, Lamé's equation

$$\frac{d^2U}{dz^2}+(\lambda-\nu(\nu+1)k^2\mathrm{sn}^2 z)U=0. \tag{0.2}$$

Now we proved in [5] that the Riemann function of (0.1) is given by

$$P_\nu(b(x,y,x_0,y_0)),$$

where P_ν is Legendre's function of degree ν and the function b is defined by

$$b(x,y,x_0,y_0)=k^2\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0\,\mathrm{sn}\,y\,\mathrm{sn}\,y_0-\frac{k^2}{k'^2}\mathrm{cn}\,x\,\mathrm{cn}\,x_0\,\mathrm{cn}\,y\,\mathrm{cn}\,y_0 +\frac{1}{k'^2}\mathrm{dn}\,x\,\mathrm{dn}\,x_0\,\mathrm{dn}\,y\,\mathrm{dn}\,y_0, \tag{0.3}$$

where k' is the complementary modulus. Starting from this observation and using Riemann's method for integrating partial differential equations we obtained in [5], [6] an integral formula for products of Lamé functions which contains practically all known linear integral relations for Lamé functions whose kernels involve Legendre's function P_ν. In this paper we shall show that a similar theory is also possible for the class of integral relations for Lamé functions whose kernels involve Legendre's function Q_ν.

As starting-point of our theory we show in the first section that

$$Q_\nu(b(x,y,x_0y_0))$$

is a fundamental solution of (0.1).

In the second section we use the well-known representation of solutions of elliptic partial differential equations by means of fundamental solutions to obtain the desired integral representation for products of Lamé functions; this representation is analogous

* Received by the editors May 17, 1982.

† Fachbereich 6-Mathematik, Universität Essen-GHS, Universitätsstrasse 3, 4300 Essen 1, West Germany.

to Cauchy's formula for analytic functions of a complex variable. In the second section we also show that many important special cases are contained in the general formula; in particular, we improve the results of Erdélyi, Magnus, Oberhettinger, Tricomi [2] and Shail [4]. These authors consider only the case when ν is a nonnegative integer, i.e., the situation where Lamé's equation (0.2) admits Lamé polynomials as solutions, whereas we allow ν to be arbitrary complex. Moreover, it is a problem in the literature to determine some multipliers in the representation formulae, but this problem does not arise if we base our results on the theory of fundamental solutions.

In §3 we shall demonstrate the value of our method by calculating some multipliers in Shail's formulae; the expressions for the multipliers which we get are considerably simpler than those given by Shail.

Let us specify some notation. We denote by $\mathbb{Z}, \mathbb{R}, \mathbb{C}$ respectively the set of integers, real numbers, complex numbers. Throughout this paper sn, cn, dn are the Jacobian elliptic functions with a fixed modulus k which belongs to the interval $]0,1[$. The complementary modulus k' and the complete elliptic integrals K and K' are defined as usual (see [7]).

1. Fundamental solutions for equation (0.1). Consider the elliptic partial differential equation

$$\frac{\partial^2 u}{\partial s^2}+\frac{\partial^2 u}{\partial t^2}+\nu(\nu+1)k^2(\operatorname{sn}^2 it-\operatorname{sn}^2 s)u=0, \tag{1.1}$$

which is obtained from (0.1) by setting

$$x=s, \quad y=it, \quad w(x,y)=u(s,t),$$

where s,t are real variables. The number ν is complex and it is sufficient to take the real part of ν greater than or equal to $-\frac{1}{2}$. We shall discuss equation (1.1) on the strip

$$S:=\{(s,t)\in\mathbb{R}^2\colon -K'<t<K'\}.$$

The coefficient $\operatorname{sn}^2 it-\operatorname{sn}^2 s$ which appears in (1.1) is (real-)analytic on this strip.

Now we assert that for each $(s_0,t_0)\in S$ the function

$$v(s,t):=Q_\nu(b(s,it,s_0,it_0)) \tag{1.2}$$

is a solution of (1.1) which has logarithmic singularities at several points in S which will be specified later. Here Q_ν is the Legendre function of degree ν of the second kind and b is the function defined in (0.3). To simplify the notation we do not indicate the dependence of v upon (s_0,t_0). This is allowable since we shall assume that

(s_0,t_0) is a point in S fixed throughout this section.

We shall now verify the assertion above, first proving the following lemma regarding the range of the function b.

LEMMA 1.3. (i) *If we denote by $L(s_0,t_0)$ the set of all $(s,t)\in S$ which satisfy $b(s,it,s_0,it_0)=1$ then*

$$L(s_0,t_0)=\{(s_0+4mK,t_0)\colon m\in\mathbb{Z}\}\cup\{(-s_0+4mK,-t_0)\colon m\in\mathbb{Z}\}.$$

(ii) *For all $(s,t)\in S$, $b(s,it,s_0,it_0)$ belongs to the interval* $[1,\infty[$.

Proof. The function b, considered as a meromorphic function of four complex variables, can be expressed by

$$b(x,y,x_0,y_0)=1+2\frac{(f(x+y)-f(x_0+y_0))(f(x-y)-f(x_0-y_0))}{(f(x-y)+f(x+y))(f(x_0-y_0)+f(x_0+y_0))}, \tag{1.4}$$

where $f(z)=k\operatorname{cn} z+\operatorname{dn} z$ is an even elliptic function of order 2 with periods $4K$ and $4iK'$ (see [6, (1.3)]). From this representation it follows that for $(s,t)\in S$

$$b(s,it,s_0,it_0)=1+\frac{1}{2}\frac{|f(s+it)-f(s_0+it_0)|^2}{\operatorname{Re} f(s+it)\cdot \operatorname{Re} f(s_0+it_0)}.$$

This proves (ii) since $|\operatorname{Im} z|<K'$ implies $\operatorname{Re} f(z)>0$. Moreover, we see that for $(s,t)\in S$ the equation $b(s,it,s_0,it_0)=1$ is equivalent to $f(s+it)=f(s_0+it_0)$. Now $f(s+it)=f(s_0+it_0)$ holds if and only if there exist integers m,n such that $s+it=s_0+it_0+4mK+4niK'$ or $s+it=-s_0-it_0+4mK+4niK'$. This yields (i).

Lemma 1.3 shows that the function v defined in (1.2) is analytic on $S\setminus L(s_0,t_0)$ since (the principal value of) the Legendre function Q_ν is analytic on the interval $]1,\infty[$. The function v is a solution of the partial differential equation (1.1) on $S\setminus L(s_0,t_0)$. This follows from the fact that every function of the form $K_\nu(b(\cdot,\cdot,x_0,y_0))$, where K_ν denotes any Legendre function of degree ν, satisfies the equation (0.1). This can be shown by direct calculation or, more satisfactorily, by use of suitable coordinate transformations of the three-dimensional wave equation (see [5]).

Now we examine the behavior of the function v at the points of $L(s_0,t_0)$. It suffices to consider $(s_0,t_0)\in L(s_0,t_0)$ since for all integers m we have

$$b(s,it,s_0,it_0)=b(s+4mK,it,s_0,it_0)=b(-s+4mK,-it,s_0,it_0).$$

The representation (1.4) of b shows that $b(x,y,x_0,y_0)=1$ if $(x-x_0)^2=(y-y_0)^2$. From this observation and from some direct calculations we see that for all x_0,y_0 which are different from $2mK+(2n+1)iK'(m,n\in\mathbb{Z})$ the function $b(\cdot,\cdot,x_0,y_0)$ near $x=x_0, y=y_0$ can be written in the form

(1.5)
$$b(x,y,x_0,y_0)=1+\big((x-x_0)^2-(y-y_0)^2\big)\big(\tfrac{1}{2}k^2(\operatorname{sn}^2 x_0-\operatorname{sn}^2 y_0)+b_1(x,y,x_0,y_0)\big),$$

where $b_1(\cdot,\cdot,x_0,y_0)$ is an analytic function of two complex variables defined in a neighborhood of (x_0,y_0) with the property that $b_1(x_0,y_0,x_0,y_0)=0$. Now from (1.5) we obtain

$$b(s,it,s_0,it_0)=1+r^2\big(\tfrac{1}{2}k^2(\operatorname{sn}^2 s_0-\operatorname{sn}^2 it_0)+b_1(s,it,s_0,it_0)\big), \tag{1.6}$$

where s,t are real, (s,t) is close to (s_0,t_0) and

$$r:=\big((s-s_0)^2+(t-t_0)^2\big)^{1/2}$$

denotes the distance between (s,t) and (s_0,t_0). First we suppose that for all integers m, (s_0,t_0) is different from $(2mK,0)$, since then $\frac{1}{2}k^2(\operatorname{sn}^2 s_0-\operatorname{sn}^2 it_0)$ is positive.

Now it is well known that the behavior of the Legendre function Q_ν near 1 can be expressed by the equation

$$Q_\nu(z)=-\frac{1}{2}P_\nu(z)\log(z-1)+H_\nu(z), \tag{1.7}$$

where $z\in\mathbb{C}\setminus]-\infty,1]$. Here P_ν is the Legendre function of degree ν of the first kind which is analytic on $\mathbb{C}\setminus]-\infty,-1]$, H_ν is a function which is analytic on the same domain and log denotes the principal value of the logarithm.

Now from (1.2), (1.6), (1.7) and $P_\nu(1)=1$ it follows that v can be written in the form

$$v(s,t)=v_1(s,t)\log\frac{1}{r}+v_2(s,t),\qquad v_1(s_0,t_0)=1, \tag{1.8}$$

where $(s,t)\in S$ is close to (s_0,t_0) but is different from (s_0,t_0) and v_1 and v_2 are functions which are analytic in a neighborhood of (s_0,t_0). The representation (1.8) shows that v is a fundamental solution of the partial differential equation (1.1) in a neighborhood of (s_0,t_0) (see [3, § 5.1]).

In the exceptional case that $(s_0,t_0)=(2mK,0)$ $(m\in\mathbb{Z})$ it can be shown by a similar analysis that

$$v(s,t)=v_1(s,t)\log\frac{1}{r}+v_2(s,t),\qquad v_1(2mK,0)=2, \tag{1.9}$$

with the same precise meaning as (1.8).

We summarize the results just obtained in the following

LEMMA 1.10. *The function v is a solution of* (1.1) *which is analytic on $S\setminus L(s_0,t_0)$ where $L(s_0,t_0)$ is specified in Lemma* 1.3(i) *and which has logarithmic singularities at every point of $L(s_0,t_0)$ in the sense of* (1.8) *if $(s_0,t_0)\neq(2mK,0)$ $(m\in\mathbb{Z})$ or in the sense of* (1.9) *if $(s_0,t_0)=(2mK,0)$ $(m\in\mathbb{Z})$.*

Now it is well known that the properties of v described in Lemma 1.10 lead to an integral representation for arbitrary solutions of (1.1) which is analogous to the Cauchy formula for analytic functions of a complex varıable.

THEOREM 1.11. *Let u be any complex-valued solution of* (1.1) *which is analytic on S. Let C be a closed path in $S\setminus L(s_0,t_0)$ with winding numbers γ_p with respect to the points p of $L(s_0,t_0)$. Then we have*

$$\int_C(u\partial_2v-v\partial_2u)\,ds+(v\partial_1u-u\partial_1v)\,dt=\sigma\cdot 2\pi\cdot\sum_{p\in L(s_0,t_0)}\gamma_p\cdot u(p),$$

where

$$\sigma=\begin{cases}1 & \text{if } (s_0,t_0)\neq(2mK,0)\quad (m\in\mathbb{Z}),\\ 2 & \text{if } (s_0,t_0)=(2mK,0)\quad (m\in\mathbb{Z}).\end{cases}$$

By ∂_1 and ∂_2 we denote the partial derivatives with respect to s and t, respectively.

For the benefit of readers not familiar with the notion of fundamental solutions we shall give a short sketch of the proof of Theorem 1.11. Since u and v are solutions of (1.1) we have $\partial_2(u\partial_2v-v\partial_2u)=\partial_1(v\partial_1u-u\partial_1v)$ on $S\setminus L(s_0,t_0)$. This implies (the proof is similar to that of the residue theorem) that

$$\int_C(u\partial_2v-v\partial_2u)\,ds+(v\partial_1u-u\partial_1v)\,dt$$
$$=\sum_{p\in L(s_0,t_0)}\gamma_p\oint_{|(s,t)-p|=\varepsilon_p}(u\partial_2v-v\partial_2u)\,ds+(v\partial_1u-u\partial_1v)\,dt,$$

where the ε_p are sufficiently small positive numbers. The value of

$$\oint_{|(s,t)-p|=\varepsilon}(u\partial_2v-v\partial_2u)\,ds+(v\partial_1u-u\partial_1v)\,dt, \tag{1.12}$$

which is independent of ε for sufficiently small positive values of ε, can easily be determined by setting $(s,t)=p+\varepsilon(\cos\varphi,\sin\varphi)$ $(\varphi\in[0,2\pi])$, using (1.8), (1.9) and taking the limit $\varepsilon\to 0$ (see [3, §5.1]). One gets, for the value of (1.12), $2\pi\sigma u(p)$ which establishes the theorem.

For later use we have to study the behavior of v for $K'>t\to K'$. From the definition of b and Lemma 1.3 we see easily that we can write

$$(1.13)\qquad b(s,it,s_0,it_0)=\frac{1}{K'-t}\tilde{b}(s,t)\qquad((s,t)\in S),$$

where $\tilde{b}\colon \mathbb{R}\times]-K',3K'[\to]0,\infty[$ is an analytic function with the property that $\tilde{b}(s,t)=\tilde{b}(s,2K'-t)$ $((s,t)\in\mathbb{R}\times]-K',3K'[)$. To simplify the notation we do not indicate the dependence of $\tilde{b}$ upon (s_0,t_0).

The behavior of the Legendre function Q_ν at infinity can be expressed by the equation

$$(1.14)\qquad Q_\nu(z)=z^{-\nu-1}q\left(\frac{1}{z}\right)\qquad(|z|>1),$$

where q is an analytic function on the unit disk of the complex plane. Now choose $\varepsilon>0$ so that the two sets $\mathbb{R}\times]K'-\varepsilon,K'[$ and $L(s_0,t_0)$ have no points in common. Then, using Lemma 1.3 and (1.13), we have

$$\frac{|K'-t|}{\tilde{b}(s,t)}<1$$

for all $(s,t)\in\mathbb{R}\times]K'-\varepsilon,K'+\varepsilon[$. Hence from (1.2), (1.13), (1.14) it follows that

$$v(s,t)=Q_\nu\left(\frac{1}{K'-t}\tilde{b}(s,t)\right)=(K'-t)^{\nu+1}\tilde{b}(s,t)^{-\nu-1}q\left(\frac{K'-t}{\tilde{b}(s,t)}\right)$$

for all $(s,t)\in\mathbb{R}\times]K'-\varepsilon,K'[$.

Thus we have shown the following

LEMMA 1.15. *Let $\varepsilon>0$ be such that the two sets $\mathbb{R}\times]K'-\varepsilon,K'[$ and $L(s_0,t_0)$ have no points in common. Then there exists an analytic function $\tilde{v}$ on the strip $\mathbb{R}\times]K'-\varepsilon,K'+\varepsilon[$ such that for all $(s,t)\in\mathbb{R}\times]K'-\varepsilon,K'[$ we have*

$$v(s,t)=(K'-t)^{\nu+1}\tilde{v}(s,t).$$

2. Integral representations for products of Lamé functions. Let U_1 and U_2 be solutions of Lamé's equation,

$$(2.1)\qquad \frac{d^2U}{dz^2}+(\lambda-\nu(\nu+1)k^2\mathrm{sn}^2z)U=0$$

corresponding to the same pair of complex parameters λ,ν, where U_1 is defined on the real axis and U_2 on the interval $]-iK',iK'[$. As in §1 we assume that the real part of ν is greater than or equal to $-\frac{1}{2}$.

Now the function u defined by

$$(2.2)\qquad u(s,t):=U_1(s)U_2(it)\qquad((s,t)\in S=\mathbb{R}\times]-K',K'[)$$

satisfies the partial differential equation (1.1). To this function we apply Theorem 1.11, choosing for the path of integration C the rectangle $C=C_1+C_2+C_3+C_4$ shown in Fig. 1.

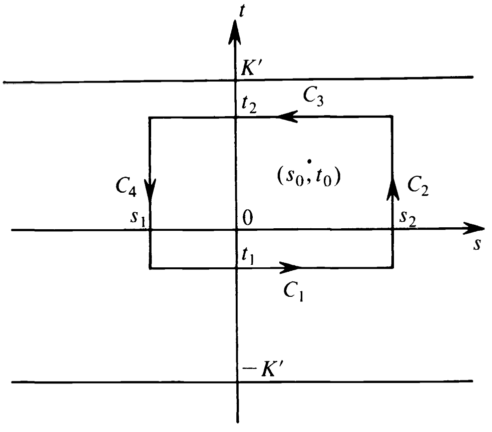


FIG. 1

We assume that

$$-t_0 < t_1 < t_0 < t_2 < K' \quad \text{and} \quad s_0 - 4K < s_1 < s_0 < s_2 < s_0 + 4K;$$

hence the winding number of C with respect to (s_0, t_0) is 1 and with respect to the other points of $L(s_0, t_0)$ is 0. Thus Theorem 1.11 implies that

$$2\pi U_1(s_0)U_2(it_0) = \int_C (u\partial_2 v - v\partial_2 u)\,ds + (v\partial_1 u - u\partial_1 v)\,dt, \tag{2.3}$$

where v corresponds to the point (s_0, t_0) as specified in (1.2). The line integral above can be decomposed in a sum of four integrals; $C = C_1 + C_2 + C_3 + C_4$. Each of these four line integrals reduces to an ordinary integral since in the line integrals along C_1 and C_3 the term $(v\partial_1 u - u\partial_1 v)dt$ can be omitted and in the line integrals along C_2 and C_4 the term $(u\partial_2 v - v\partial_2 u)ds$ can be omitted. Moreover, we could insert into (2.3) the definitions of v (see (1.2)) and u (see (2.2)). However, this would yield a somewhat lengthy formula.

For Lamé functions of special types the representation formula (2.3) can be simplified. First we consider (nontrivial) Lamé functions of period $4K$. It is well known that there are four classes of such Lamé functions U, namely,

Class I: U is even and of period $2K$,

Class II: U is odd and of period $2K$,

Class III: U is even and of half-period $2K$, i.e. $U(z+2K) = -U(z)$,

Class IV: U is odd and of half-period $2K$.

Of course such Lamé functions exist only for certain characteristic values of the parameters λ, ν. We shall denote a Lamé function of period $4K$ which is defined on the real axis (or on the strip $|\text{Im}\, z| < K'$) by E (of class $\cdots$). We do not use the usual notation of periodic Lamé functions (see [2, §15.5]) which depends on the Sturm–Liouville theory since this notation covers only real values of λ and $\nu(\nu+1)$ whereas we want to allow complex values of λ and ν.

We remark

LEMMA 2.4. *If $U_1 = E$ is of period $4K$ and $s_2 = s_1 + 4K$ then we have*

$$\int_{C_2 + C_4} (v\partial_1 u - u\partial_1 v)\,dt = 0.$$

Proof. For all $t\in]-K',K'[$ the functions $u(\cdot,t)=U_1(\cdot)U_2(it)$ and $v(\cdot,t)$ are of period $4K$; consequently

$$(v\partial_1 u-u\partial_1 v)(s_1,t)=(v\partial_1 u-u\partial_1 v)(s_2,t).$$

We next consider Lamé functions characterized by their simple behavior at the point iK'. Lamé's equation (2.1) has a regular singularity at $z=iK'$ with exponents $-\nu$ and $\nu+1$. Since we assume that $\operatorname{Re}\nu\geq-\frac{1}{2}$ we have $\operatorname{Re}(-\nu)\leq\operatorname{Re}(\nu+1)$. Hence Lamé's equation possesses a uniquely determined solution (up to a constant factor) of the form

$$F(z)=(z-iK')^{\nu+1}g(z),\qquad g(iK')\neq 0, \tag{2.5}$$

where g is an analytic function defined in a neighborhood of $z=iK'$. It should be noted that, in contrast to the periodic solutions E, such solutions F exist for all complex values of λ,ν ($\operatorname{Re}\nu\geq-\frac{1}{2}$). In the following F plays the role of the Lamé function U_2. Hence we have to consider F as a function defined on the interval $]-iK',iK'[$. At this stage it is not necessary to normalize the functions E and F and hence the branch of $(z-iK')^{\nu+1}$ used in (2.5) since both sides of (2.3) are bilinear in U_1, U_2.

We remark

LEMMA 2.6. *If $U_2=F$ is of the form* (2.5) *then we have*

$$\int_{C_3}(u\partial_2 v-v\partial_2 u)\,ds\to 0\quad \text{as } K'>t_2\to K'.$$

Proof. From Lemma 1.15 and (2.5) it follows that for sufficiently small $\varepsilon>0$ we can write

$$v(s,t)=(K'-t)^{\nu+1}\tilde{v}(s,t)\qquad((s,t)\in\mathbb{R}\times]K'-\varepsilon,K'[),$$
$$u(s,t)=(K'-t)^{\nu+1}\tilde{u}(s,t)\qquad((s,t)\in\mathbb{R}\times]K'-\varepsilon,K'[),$$

where $\tilde{v}$ and $\tilde{u}$ are analytic functions on $\mathbb{R}\times]K'-\varepsilon,K'+\varepsilon[$; therefore

$$u\partial_2 v-v\partial_2 u=(K'-t)^{2\nu+2}(\tilde{u}\partial_2\tilde{v}-\tilde{v}\partial_2\tilde{u}).$$

Since $\operatorname{Re}\nu\geq-\frac{1}{2}$ we have $\operatorname{Re}(2\nu+2)\geq 1$, hence $(u\partial_2 v-v\partial_2 u)(s,t_2)$ converges to 0 as $K'>t_2\to K'$ uniformly with respect to $s\in[s_1,s_2]$. This proves Lemma 2.6.

If we set $s_1=-2K$, $s_2=2K$, $t_2\to K'$ we obtain from (2.3) using Lemmas 2.4 and 2.6

THEOREM 2.7. *Let E and F be solutions of Lamé's equation* (2.1) *corresponding to the same pair of parameters λ,ν where E is of period $4K$ and F is of the form* (2.5). *Then we have, for all $s_0\in\mathbb{R}$, $t_0\in]0,K'[$ and $t_1\in]-t_0,t_0[$,*

$$2\pi iE(s_0)F(it_0)=iF(it_1)\int_{-2K}^{2K}\frac{\partial}{\partial t_1}Q_\nu(b(s,it_1,s_0,it_0))E(s)\,ds$$
$$+F'(it_1)\int_{-2K}^{2K}Q_\nu(b(s,it_1,s_0,it_0))E(s)\,ds$$

(*where b is defined in* (0.3)).

First Theorem 2.7 is valid only for $-2K<s_0<2K$, but then obviously also for all real s_0. In the case when $t_1=0$ the above formula can be simplified since $\partial_2 v(s,0)$ is an odd function with respect to s and $v(s,0)$ is an even function with respect to s. Hence, if E is even, i.e., of class I or III, the first integral on the right-hand side of the equation vanishes and, if E is odd, i.e., of class II or IV, the other integral vanishes. From Theorem 2.7 we get the following corollary which generalizes and improves an integral representation for external ellipsoidal harmonics mentioned in Erdélyi [2, p. 83].

COROLLARY 2.8. *Under the assumptions of Theorem* 2.7 *we have*

$$2\pi i E(s_0)F(it_0)E(it_1)=[E,F]\int_{-2K}^{2K} Q_\nu(b(s,it_1,s_0,it_0))E(s)\,ds,$$

where $[E,F]=E(z)F'(z)-E'(z)F(z)$ *denotes the Wronskian of* E *and* F.

Proof. Multiply the equation in Theorem 2.7 by $E(it_1)$ and note that

$$\int_{-2K}^{2K}[E(s)E(it_1)\partial_2 v(s,t_1)-iE(s)E'(it_1)v(s,t_1)]\,ds \tag{2.9}$$

is independent of t_1 for $t_1\in]-t_0,t_0[$. This follows from Theorem 1.11, where $u(s,t)=E(s)E(it)$ and C is a rectangle with corners at $(-2K,0)$, $(2K,0)$, $(2K,t_1)$, $(-2K,t_1)$. Now the integral (2.9) vanishes for $t_1=0$, hence vanishes for all $t_1\in]-t_0,t_0[$. This proves Corollary 2.8.

The above result is given in [2, p. 83] only in the case when ν is a nonnegative integer and E is a Lamé polynomial; the multiplier $[E,F]$ is not determined. Also we should be careful over the assumptions of Theorem 2.7 and Corollary 2.8. For example, if t_1 belongs to the interval $]t_0,K'[$ the formulae are false.

Theorem 2.7 and Corollary 2.8 (which are essentially equivalent) contain many important special cases. For instance, if we fix t_0 and t_1 we obtain integral equations for E. We consider only one example, namely $t_0=K'/2$, $t_1=0$ and E of class I or III.

COROLLARY 2.10. *Let E be of class* I *or* III. *Then we have for all real x*

$$2\pi i E(x)F\left(i\frac{K'}{2}\right)=F'(0)\int_{-2K}^{2K} Q_\nu\left(\frac{(1+k)^{1/2}}{k'^2}(-k^{3/2}\operatorname{cn}s\operatorname{cn}x+\operatorname{dn}s\operatorname{dn}x)\right)E(s)\,ds.$$

If we fix s_0 and t_1 in Theorem 2.7 or Corollary 2.8 we obtain representations for the Lamé function F in terms of the periodic Lamé function E; these representations generalize the results of Shail [4] who treated the case when ν is a nonnegative integer and E is a Lamé polynomial (see §3 of our paper). In the case when E is of class I or IV we set, in (2.7), $s_0=K$, $t_1=0$. In the case when E is of class II or III we first differentiate the equation in Theorem 2.7 (with $t_1=0$) with respect to s_0 and then we set $s_0=K$. In this way we get

COROLLARY 2.11. *If E is of class* I, II, III, IV, *respectively, then we have for* $y\in]0,iK'[$

$$\text{(I)}\quad 2\pi i E(K)F(y)=F'(0)\int_{-2K}^{2K} Q_\nu\left(\frac{1}{k'}\operatorname{dn}s\operatorname{dn}y\right)E(s)\,ds,$$

$$\text{(II)}\quad 2\pi i E'(K)F(y)=-\frac{k^4}{k'}F(0)\operatorname{cn}y\operatorname{sn}y\int_{-2K}^{2K}\operatorname{cn}s\operatorname{sn}s\,Q_\nu''\left(\frac{1}{k'}\operatorname{dn}s\operatorname{dn}y\right)E(s)\,ds,$$

$$\text{(III)}\quad 2\pi i E'(K)F(y)=\frac{k^2}{k'}F'(0)\operatorname{cn}y\int_{-2K}^{2K}\operatorname{cn}s\,Q_\nu'\left(\frac{1}{k'}\operatorname{dn}s\operatorname{dn}y\right)E(s)\,ds,$$

$$\text{(IV)}\quad 2\pi i E(K)F(y)=-k^2F(0)\operatorname{sn}y\int_{-2K}^{2K}\operatorname{sn}s\,Q_\nu'\left(\frac{1}{k'}\operatorname{dn}s\operatorname{dn}y\right)E(s)\,ds.$$

Of course, the factors $E(K)$ or $E'(K)$ which appear on the left-hand side of the equations do not vanish. In all cases the integrands are even functions of period $2K$ with respect to s. Hence we can replace the integral $\int_{-2K}^{2K}$ by 2 $\int_{-K}^{K}$ or by 4 $\int_0^K$.

Evidently, the range of validity of the representations in Corollary 2.11 can be enlarged by analytic continuation with respect to y. If $y \in]0, iK'[$ and s is real we know that $1/k' \operatorname{dn} s \operatorname{dn} y$ belongs to the interval $]1, \infty[$. Now one can ask, how far can we continue the equations in Corollary 2.11 if we want to use only the principal value of Q_ν which is defined on $\mathbb{C} \setminus]-\infty, 1]$? We can immediately answer this question since the value of $1/k' \operatorname{dn} s \operatorname{dn} y$ does not belong to the interval $]-\infty, 1]$ if s is real and y lies in the set

$$G := \{y \in \mathbb{C} : |\operatorname{Re} y| < K, 0 < \operatorname{Im} y < 2K'\} \setminus [iK', 2iK'[$$

(see [1, p. 61]), where the boundary of G corresponds to the cut $]-\infty, 1]$ above. In this sense the equations in Corollary 2.11 remain valid for all $y \in G$.

It is useful to consider, besides Lamé functions of period $4K$, also Lamé functions of periods $4iK'$ and $4(K+iK')$. For such Lamé functions there are similar formulae to those presented in this section for Lamé functions of period $4K$. The formulae corresponding to the other periods can be obtained in a way analogous to that which leads to the results of this section. For example, if we wish to treat Lamé functions of period $4iK'$ we set in (0.1) $x = K + is, y = t + iK'$ (s real, $t \in]0, 2K[$) and proceed as before. It is also possible to derive those formulae directly from the results of this section by use of suitable transformations of Lamé's equation involving a transformation of the modulus k. To treat Lamé functions of period $4iK'$ we use the transformation $k \to k'$ which is associated with Jacobi's imaginary transformation of sn, cn, dn (in [2, 15.5.2] this technique is applied). To treat Lamé functions of period $4(K+iK')$ we use the transformation $k \to 1/k$ which is associated with Jacobi's real transformation (see [1, p. 69]). However, since $1/k$ does not belong to the interval $]0, 1[$ this does not work directly. First we have to extend our results to arbitrary complex values of k. We shall not work out these ideas in detail, but state only the results which correspond to Corollary 2.8.

THEOREM 2.12. *Let E and F be solutions of Lamé's equation* (2.1) *corresponding to the same pair of parameters λ, ν where E is of period $4iK'$ and F is of the form* (2.5).

Then we have, for all x_0 with $\operatorname{Re} x_0 = K$ *and $y, y_0 \in]iK', 2K+iK'[$ with* $\operatorname{Re} y_0 < \operatorname{Re} y < 2K - \operatorname{Re} y_0$,

$$2\pi i E(x_0) F(y_0) E(y) = [E, F] \int_{K-2iK'}^{K+2iK'} Q_\nu(b(x, y, x_0, y_0)) E(x)\, dx.$$

THEOREM 2.13. *Let E and F be solutions of Lamé's equation* (2.1) *corresponding to the same pair of parameters λ, ν where E is of period $4(K+iK')$ and F is of the form* (2.5).

Then we have, for all x_0 which belong to the line $(K+iK') \cdot \mathbb{R}$ and $y, y_0 \in]-iK', iK'[$ with $-\operatorname{Im} y_0 < \operatorname{Im} y < \operatorname{Im} y_0$,

$$2\pi i E(x_0) F(y_0) E(y) = [E, F] \int_{-2(K+iK')}^{2(K+iK')} Q_\nu(b(x, y, x_0, y_0)) E(x)\, dx.$$

Now we can start with Theorems 2.12 and 2.13 and, working on the same lines as before, obtain results which correspond to Corollaries 2.10 and 2.11. This can be left to the reader.

3. Integral representations in case of Lamé polynomials. In this section we assume that ν is a nonnegative integer. Then it is well known that there exist $2\nu+1$ characteristic values of λ such that Lamé's equation (2.1) has a Lamé polynomial as a solution. We adopt the notation and normalization of Lamé polynomials and of the corresponding second solutions F explained in Shail's paper [4, §2]. We note explicitly that the normalization implies the following Wronskian relation

$$(3.1)\qquad E_\nu^m(z)F_\nu^{m'}(z)-E_\nu^{m'}(z)F_\nu^m(z)=(-1)^{\sigma+\tau}(2\nu+1)k^{2\tau+1}$$

(see [4, (7)]).

Each of the eight types of Lamé polynomials belongs to one of the classes of §2, specifically uE and dE to class I, scE and $scdE$ to class II, cE and cdE to class II and sE and sdE to class IV. Hence in each case Corollary 2.11 yields an integral representation for the second solution F in terms of a Lamé polynomial. Using the results which correspond to Corollary 2.11 for the other periods $4iK'$ and $4(K+iK')$ (see the remarks at the end of section 2) we obtain altogether $8\times 3=24$ such representations. This set of 24 representations was given by Shail [4, p. 709]. Shail derived these formulae from another method which requires two steps. In a first step the representations are determined up to characteristic values μ in the form $F(z)=\mu\cdot\int\cdots$. Then these characteristic values are calculated in a second step.

Now we shall demonstrate that the expressions for the characteristic values which we get from Corollary 2.11 are considerably simpler than those given by Shail. For instance, we look at the integral representation for uF_{2n}^m in terms of the Lamé polynomial uE_{2n}^m obtained from Corollary 2.11. This representation is

$$(3.2)\qquad 2\pi i uE_{2n}^m(K)uF_{2n}^m(y)=4uF_{2n}^{m'}(0)\int_0^K Q_{2n}\left(\frac{1}{k'}\operatorname{dn}s\operatorname{dn}y\right)uE_{2n}^m(s)\,ds,$$

where m,n are nonnegative integers with $m\leq n$. The variable y lies in the interval $]0,iK'[$ or in some larger domain but this is not of interest if we are concerned with the calculation of the characteristic value. Now we have from (3.1)

$$uF_{2n}^{m'}(0)=\frac{(4n+1)k}{uE_{2n}^m(0)}.$$

We insert this value of $uF_{2n}^{m'}(0)$ in (3.2) to arrive at

$$uF_{2n}^m(y)=\mu\cdot\int_0^K Q_{2n}\left(\frac{1}{k'}\operatorname{dn}s\operatorname{dn}y\right)uE_{2n}^m(s)\,ds,$$

where

$$(3.3)\qquad \mu=\frac{2(4n+1)k}{\pi i uE_{2n}^m(0)uE_{2n}^m(K)}.$$

The value of μ given by Shail [4, (43)] is

$$(3.4)\qquad \mu=\frac{(4n+1)kP_{2n}(0)}{iuE_{2n}^m(K+iK')}\int_0^K \operatorname{dn}s\,uE_{2n}^m(s)\,ds.$$

It is clear that the characteristic value μ is easier to determine by (3.3) than by (3.4) because by (3.3) it is not necessary to calculate the integral $\int_0^K\operatorname{dn}s\,uE_{2n}^m(s)\,ds$. We have just to calculate $uE_{2n}^m(0)$ and $uE_{2n}^m(K)$ which is very simple if we know uE_{2n}^m.

We close this paper with a list of the characteristic values for the eight integral representations which are obtained from Corollary 2.11. The designations of the formulae refer to Shail's list [4, p. 709].

formulae	characteristic value
I(iii)	$-\dfrac{2i(4n+1)k}{\pi uE_{2n}^{m}(0)uE_{2n}^{m}(K)}$
II(iii)	$-\dfrac{2i(4n+3)k^2}{\pi sE_{2n+1}^{m'}(0)sE_{2n+1}^{m}(K)}$
III(iii)	$\dfrac{2(4n+3)k^2}{\pi cE_{2n+1}^{m}(0)cE_{2n+1}^{m'}(K)}$
IV(i)	$\dfrac{2i(4n+3)k^3}{\pi dE_{2n+1}^{m}(0)dE_{2n+1}^{m}(K)}$
V(iii)	$\dfrac{2(4n+5)k^3}{\pi scE_{2n+2}^{m'}(0)scE_{2n+2}^{m'}(K)}$
VI(ii)	$\dfrac{2i(4n+5)k^4}{\pi sdE_{2n+2}^{m'}(0)sdE_{2n+2}^{m}(K)}$
VII(ii)	$-\dfrac{2(4n+5)k^4}{\pi cdE_{2n+2}^{m}(0)cdE_{2n+2}^{m'}(K)}$
VIII(iii)	$-\dfrac{2(4n+7)k^5}{\pi scdE_{2n+3}^{m'}(0)scdE_{2n+3}^{m'}(K)}$

REFERENCES

[1] P. Du Val, *Elliptic Functions and Elliptic Curves*, Cambridge Univ. Press, Cambridge, 1973.

[2] A. Erdélyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, *Higher Transcendental Functions* III, McGraw-Hill, New York, 1955.

[3] P. R. Garabedian, *Partial Differential Equations*, John Wiley, New York, 1964.

[4] R. Shail, *On integral representations for Lamé and other special functions*, this Journal, 11 (1980), pp. 702–723.

[5] H. Volkmer, *Integralrelationen mit variablen Grenzen für spezielle Funktionen der mathematischen Physik*, J. Reine Angew. Math., 319 (1980), pp. 118–132.

[6] ______, *Integral relations for Lamé functions*, this Journal, 13 (1982), pp. 978–987.

[7] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Cambridge Univ. Press, Cambridge, 1927.

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics
011

OSCILLATION PROPERTIES OF NONLINEAR HYPERBOLIC EQUATIONS*

KURT KREITH[†], TAKAŜI KUSANO[‡] AND NORIO YOSHIDA[§]

Abstract. A variety of oscillation properties are established for solutions of characteristic initial value problems for the nonlinear telegraph equation with a forcing term. Some analogous questions are considered for initial boundary value problems for the forced nonlinear wave equation. The principal tool is an averaging technique which enables one to establish such oscillation properties in terms of related ordinary differential inequalities.

Key words. characteristic initial value problems, convex function, nodal domain, initial boundary value problem, Jensen's inequality, timelike boundary

1. Introduction. This paper deals with several techniques for establishing oscillation properties associated with second order hyperbolic equations. In connection with characteristic initial value problems of the form

$$\begin{aligned} &u_{xy}+c(x,y,u)=0, \qquad x,y>0,\\ &u(x,0)=\varphi(x), \qquad u(0,y)=\psi(y), \end{aligned} \tag{1.1}$$

Yoshida [4] has used the function

$$U(t)=\int_0^t u(t-y,y)\,dy$$

to establish oscillation criteria. These considerations are extended in §2 of the present paper to the more general equation (2.1).

Initial boundary value problems are considered in §§3 and 4. First the averaging technique of [4] is adapted to equations of the form

$$u_{tt}-\Delta u+c(t,x,u)=f(t,x) \tag{1.2}$$

in cylindrical domains where $\partial u/\partial \nu$ is assigned on the lateral boundary. Then the same equation is studied in more general domains under the simpler boundary condition $u=0$ on the timelike part of the boundary.

2. Nonlinear characteristic initial value problem. In this section, we consider the nonlinear hyperbolic equation

$$u_{xy}+c(x,y,u)=f(x,y), \qquad (x,y)\in Q_\rho, \tag{2.1}$$

where

$$Q_\rho=\{(x,y)\in R^2:\ 0<x,y<\infty,\ \rho<x+y<\infty\} \qquad (\rho\geq 0).$$

*Received by the editors January 25, 1982, and in revised form April 15, 1983.

[†] Department of Mathematics, University of California, Davis, California 95616.

[‡] Hiroshima University, Hiroshima, Japan.

[§] Iwate University, Morioka, Japan.

Assuming throughout that $f(x,y)$ is continuous in Q_0 and the following conditions:

(A-I) $c(x,y,u)$ is real-valued and continuous in $Q_\rho \times R^1$;
(A-II) $c(x,y,\xi) \geq p(x+y)\varphi(\xi)$ for all $(x,y,\xi) \in Q_\rho \times (0,\infty)$, where p is continuous and positive in (ρ,∞) and φ is continuous, positive and convex in $(0,\infty)$;
(A-III) $c(x,y,-\xi) = -c(x,y,\xi)$ for all $(x,y,\xi) \in Q_\rho \times (0,\infty)$,

we investigate the oscillatory behavior of solutions of the characteristic initial value problem

$$\begin{aligned} &u_{xy} + c(x,y,u) = f(x,y), \qquad (x,y) \in Q_\rho, \\ &u_x(x,0) = g(x), \qquad x \in (\rho,\infty), \\ &u_y(0,y) = h(y), \qquad y \in (\rho,\infty), \end{aligned} \tag{2.2}$$

where $g(t)$ and $h(t)$ are continuous functions in (ρ,∞).

Associated with every function $u \in D(Q_\rho) \equiv C^2(Q_\rho) \cap C^1(\overline{Q_\rho})$, we define the function $z[u](t)$ by

$$z[u](t) = \frac{1}{t}\int_0^t u(t-\xi,\xi)\,d\xi, \qquad t > \rho. \tag{2.3}$$

LEMMA 2.1. *Assume that* (A-I) *and* (A-II) *hold. If $u \in D(Q_\rho)$ is a positive solution of the problem* (2.2) *in Q_{t_1} $(t_1 \geq \rho)$, then $z[u](t)$ given by* (2.3) *satisfies the ordinary differential inequality*

$$(tz[u](t))'' + tp(t)\varphi(z[u](t)) \leq g(t) + h(t) + \int_0^t f(t-\xi,\xi)\,d\xi, \qquad t > t_1. \tag{2.4}$$

Proof. From a result of Yoshida [4, Lemma 1] it follows that

$$\begin{aligned} (tz[u](t))'' &= u_x(t,0) + u_y(0,t) + \int_0^t u_{xy}(t-\xi,\xi)\,d\xi \\ &= g(t) + h(t) + \int_0^t u_{xy}(t-\xi,\xi)\,d\xi. \end{aligned} \tag{2.5}$$

By assumption (A-II) we get

$$\begin{aligned} \int_0^t u_{xy}(t-\xi,\xi)\,d\xi &= -\int_0^t c(t-\xi,\xi,u(t-\xi,\xi))\,d\xi + \int_0^t f(t-\xi,\xi)\,d\xi \\ &\leq -p(t)\int_0^t \varphi(u(t-\xi,\xi))\,d\xi + \int_0^t f(t-\xi,\xi)\,d\xi, \end{aligned} \tag{2.6}$$

while from Jensen's inequality we have

$$\int_0^t \varphi(u(t-\xi,\xi))\,d\xi \geq t\varphi(z[u](t)). \tag{2.7}$$

Combining (2.5)–(2.7) yields

$$(tz[u](t))'' \leq g(t) + h(t) - tp(t)\varphi(z[u](t)) + \int_0^t f(t-\xi,\xi)\,d\xi,$$

which is the desired inequality (2.4). □

THEOREM 2.2. *Assume that* (A-I) *through* (A-III) *hold. Every solution* $u\in D(Q_\rho)$ *of the characteristic initial value problem* (2.2) *is oscillatory in* Q_ρ *if the ordinary differential inequalities*

$$(tz)''+tp(t)\varphi(z)\le g(t)+h(t)+\int_0^t f(t-\xi,\xi)\,d\xi, \tag{2.8}$$

$$(tz)''+tp(t)\varphi(z)\le -g(t)-h(t)-\int_0^t f(t-\xi,\xi)\,d\xi \tag{2.9}$$

are oscillatory at $t=\infty$, *in the sense that neither* (2.8) *nor* (2.9) *has a solution which is positive on* $[t_0,\infty)$ *for any* $t_0>0$.

Proof. Suppose to the contrary that there exists a solution u of the problem (2.2) which has no zero in Q_{t_1} for some $t_1>\rho$. If $u>0$ in Q_{t_1}, we see from Lemma 2.1 that $z[u](t)$ is a positive solution of (2.8) in (t_1,∞). If $u<0$ in Q_{t_1}, $U\equiv -u$ is a positive solution which satisfies the problem

$$\begin{aligned} &U_{xy}+c(x,y,U)=-f(x,y), && (x,y)\in Q_{t_1},\\ &U_x(x,0)=-g(x), && x\in(t_1,\infty),\\ &U_y(0,y)=-h(y), && y\in(t_1,\infty). \end{aligned}$$

Hence, $z[U](t)$ is a positive solution of (2.9) in (t_1,∞). This contradicts the hypothesis and completes the proof. □

Using a result of Kusano and Naito [2, Thms. 2 and 3], we obtain the following results.

THEOREM 2.3. *Assume that* (A-I) *through* (A-III) *hold. Every solution* $u\in D(Q_\rho)$ *of the characteristic initial value problem* (2.2) *is oscillatory in* Q_ρ *if*

$$\liminf_{t\to\infty}\int_T^t\left(1-\frac{s}{t}\right)\left(g(s)+h(s)+\int_0^s f(s-\xi,\xi)\,d\xi\right)ds=-\infty,$$

$$\limsup_{t\to\infty}\int_T^t\left(1-\frac{s}{t}\right)\left(g(s)+h(s)+\int_0^s f(s-\xi,\xi)\,d\xi\right)ds=\infty$$

for all large T.

THEOREM 2.4. *Assume that* (A-I) *through* (A-III) *hold. Every solution* $u\in D(Q_\rho)$ *of the characteristic initial value problem* (2.2) *is oscillatory in* Q_ρ *if the ordinary differential inequality* $(tz)''+tp(t)\varphi(z)\le 0$ *has no eventually positive solution and if there exists a* C^2 *function* θ: $(\rho,\infty)\to R^1$ *with the following properties*:

(i) $\theta(t)$ *takes both positive and negative values for arbitrarily large* t;

(ii) $(t\theta(t))''=g(t)+h(t)+\int_0^t f(t-\xi,\xi)d\xi$, $t>\rho$;

(iii) $\lim_{t\to\infty} t\theta(t)=0$.

Example 1. Consider the problem

$$\begin{aligned} &u_{xy}+u=-(x+y)^{-1}2e^{x+y}+\log(x+y)\sin(x+y), && (x,y)\in Q_\rho,\\ &u_x(x,0)=e^x, && x\in(\rho,\infty),\\ &u_y(0,y)=e^y, && y\in(\rho,\infty), \end{aligned} \tag{2.10}$$

where ρ is some positive number. Since

$$\int_T^t\left(1-\frac{s}{t}\right)\left(e^s+e^s+\int_0^s\left(-\frac{2}{s}e^s+\log s\sin s\right)d\xi\right)ds$$
$$=\int_T^t\left(1-\frac{s}{t}\right)s\log s\sin s\,ds=-\log t\sin t+O(1)\qquad(t\to\infty),$$

we obtain

$$\liminf_{t\to\infty}\int_T^t\left(1-\frac{s}{t}\right)\left(e^s+e^s+\int_0^s\left(-\frac{2}{s}e^s+\log s\sin s\right)d\xi\right)ds=-\infty,$$
$$\limsup_{t\to\infty}\int_T^t\left(1-\frac{s}{t}\right)\left(e^s+e^s+\int_0^s\left(-\frac{2}{s}e^s+\log s\sin s\right)d\xi\right)ds=\infty.$$

Hence, every solution u of the problem (2.10) is oscillatory in Q_ρ. However, we note that the solution u of the problem

$$\begin{aligned}&u_{xy}+u=0, && (x,y)\in Q_0,\\ &u(x,0)=e^x, && x\in(0,\infty),\\ &u(0,y)=e^y, && y\in(0,\infty)\end{aligned}$$

satisfies $u\geq 1$ (see Pagan [3]).

Example 2. Consider the problem

$$\text{(2.11)}\qquad\begin{aligned}&u_{xy}+2u=-2e^{-x-y}\sin(x+y), && (x,y)\in Q_\rho,\\ &u_x(x,0)=e^{-x}\sin x, && x\in(\rho,\infty),\\ &u_y(0,y)=e^{-y}\sin y, && y\in(\rho,\infty).\end{aligned}$$

Since

$$\int_T^t\left(1-\frac{s}{t}\right)\left(e^{-s}\sin s+e^{-s}\sin s+\int_0^s(-2e^{-s}\sin s)\,d\xi\right)ds$$
$$=\int_T^t\left(1-\frac{s}{t}\right)(2e^{-s}\sin s-2se^{-s}\sin s)\,ds<\infty\qquad(t\to\infty),$$

Theorem 2.3 does not apply to the problem (2.11), but Theorem 2.4 does. Since the ordinary differential inequality $(tz)''+2tz\leq 0$ has no eventually positive solution, and $\theta(t)\equiv t^{-1}e^{-t}\sin t-e^{-t}\cos t$ satisfies the following:

(i) $\theta(t)$ satisfies condition (i) of Theorem 2.4;

(ii) $(t\theta(t))''=2e^{-t}\sin t-2te^{-t}\sin t$;

(iii) $\lim_{t\to\infty}t\theta(t)=0$,

Theorem 2.4 implies that every solution of the problem (2.11) is oscillatory in Q_ρ. In fact, the problem (2.11) has an oscillatory solution

$$u=-2^{-1/2}e^{-x-y}\sin\left(x+y+\frac{\pi}{4}\right).$$

3. Initial boundary value problems in cylinders. We now investigate the oscillatory properties of certain solutions of the hyperbolic equation

$$\text{(3.1)}\qquad u_{tt}-\Delta u+c(t,x,u)=f(t,x).$$

Let G be a bounded domain with smooth boundary ∂G, ν be the exterior normal vector to ∂G, and $\Omega \equiv (0,\infty)\times G$.

Associated with every function $u\in D(\Omega)\equiv C^2(\Omega)\cap C^1(\bar{\Omega})$, we define the function $z[u](t)$ by

$$z[u](t)=\frac{1}{|G|}\int_G u(t,x)\,dx, \qquad t\in(0,\infty),$$

where $|G|$ denotes the volume of G, i.e. $|G|=\int_G dx$.

We shall assume the following conditions:

(B-I) $c(t,x,u)$ is real-valued and continuous in $\Omega\times R^1$;

(B-II) $c(t,x,\xi)\geq p(t)\varphi(\xi)$ for all $(t,x,\xi)\in\Omega\times(0,\infty)$, where p is continuous and positive in $(0,\infty)$ and φ is continuous, positive and convex in $(0,\infty)$;

(B-III) $c(t,x,-\xi)=-c(t,x,\xi)$ for all $(t,x,\xi)\in\Omega\times(0,\infty)$.

We consider the problem

$$(3.2)\qquad \begin{aligned} &u_{tt}-\Delta u+c(t,x,u)=f(t,x), && (t,x)\in\Omega,\\ &\frac{\partial u}{\partial\nu}=g(t,x), && (t,x)\in(0,\infty)\times\partial G. \end{aligned}$$

Lemma 3.1. *Assume that* (B-I) *and* (B-II) *hold. Let $u\in D(\Omega)$ be a positive solution of the problem* (3.2) *in Ω_{t_1}, where $\Omega_{t_1}\equiv(t_1,\infty)\times G$. Then, $z[u](t)$ satisfies the ordinary differential inequality*

$$(3.3)\quad (z[u](t))''+p(t)\varphi(z[u](t))\leq\frac{1}{|G|}\int_{\partial G}g(t,x)\,d\sigma+\frac{1}{|G|}\int_G f(t,x)\,dx, \qquad t>t_1.$$

Proof. It is easy to see that

$$(3.4)\qquad (z[u](t))''=\frac{1}{|G|}\int_G u_{tt}\,dx=\frac{1}{|G|}\int_G[\Delta u-c(t,x,u)+f(t,x)]\,dx.$$

From Green's theorem we have

$$(3.5)\qquad \int_G \Delta u\,dx=\int_{\partial G}\frac{\partial u}{\partial\nu}\,d\sigma.$$

Using condition (B-II) and Jensen's inequality, we obtain

$$(3.6)\qquad \frac{1}{|G|}\int_G c(t,x,u)\,dx\geq p(t)\varphi(z[u](t)).$$

Combining (3.4)–(3.6) yields

$$\begin{aligned}(z[u](t))''&\leq\frac{1}{|G|}\int_{\partial G}\frac{\partial u}{\partial\nu}\,d\sigma-p(t)\varphi(z[u](t))+\frac{1}{|G|}\int_G f(t,x)\,dx\\ &=\frac{1}{|G|}\int_{\partial G}g(t,x)\,d\sigma-p(t)\varphi(z[u](t))+\frac{1}{|G|}\int_G f(t,x)\,dx,\end{aligned}$$

which is the desired inequality (3.3). □

THEOREM 3.2. *Assume that* (B-I) *through* (B-III) *hold. Every solution* $u \in D(\Omega)$ *of the problem* (3.2) *is oscillatory in* Ω *if the ordinary differential inequalities*

$$z''+p(t)\varphi(z)\le G(t)+F(t), \tag{3.7}$$
$$z''+p(t)\varphi(z)\le -G(t)-F(t), \tag{3.8}$$

are oscillatory at $t=\infty$, *where* $G(t)\equiv(1/|G|)\int_{\partial G} g(t,x)\,d\sigma$ *and* $F(t)\equiv(1/|G|)\int_G f(t,x)\,dx$.

THEOREM 3.3. *Assume that* (B-I) *through* (B-III) *hold. Every solution* $u \in D(\Omega)$ *of the problem* (3.2) *is oscillatory in* Ω *if*

$$\liminf_{t\to\infty}\int_T^t\left(1-\frac{s}{t}\right)(G(s)+F(s))\,ds=-\infty,$$
$$\limsup_{t\to\infty}\int_T^t\left(1-\frac{s}{t}\right)(G(s)+F(s))\,ds=\infty,$$

for all large T.

THEOREM 3.4. *Assume that* (B-I) *through* (B-III) *hold. Every solution* $u \in D(\Omega)$ *of the problem* (3.2) *is oscillatory in* Ω *if the ordinary differential inequality* $z''+p(t)\varphi(z)\le 0$ *has no eventually positive solution and if there exists a* C^2 *function* θ: $(0,\infty)\to R^1$ *with the following properties*:

(i) $\theta(t)$ *takes both positive and negative values for arbitrarily large* t;

(ii) $\theta''(t)=G(t)+F(t)$, $t>0$;

(iii) $\lim_{t\to\infty}\theta(t)=0$.

Example 1. Consider the problem

$$\begin{aligned} &u_{tt}-u_{xx}+2u=-2e^{-t}\cos t\cos x+3e^{-t}\sin t\cos x \quad \text{in } (0,\infty)\times\left(0,\frac{\pi}{2}\right),\\ &u_x\left(t,\frac{\pi}{2}\right)=-e^{-t}\sin t, \qquad t\in(0,\infty),\\ &-u_x(t,0)=0, \qquad t\in(0,\infty). \end{aligned} \tag{3.9}$$

We easily see that

$$G(t)=-\frac{2}{\pi}e^{-t}\sin t, \qquad F(t)=\frac{2}{\pi}(-2e^{-t}\cos t+3e^{-t}\sin t).$$

It is readily seen that $z''+2z\le 0$ has no eventually positive solution. Let $\theta(t)=(2/\pi)e^{-t}(\cos t+\sin t)$. Since

$$\theta(t)=\frac{2}{\pi}e^{-t}2^{1/2}\sin\left(t+\frac{\pi}{4}\right),$$
$$\theta''(t)=\frac{4}{\pi}e^{-t}(\sin t-\cos t)=G(t)+F(t),$$
$$\lim_{t\to\infty}\theta(t)=0,$$

$\theta(t)$ satisfies conditions (i)–(iii) of Theorem 3.4. Hence, every solution u of the problem (3.9) is oscillatory in $(0,\infty)\times(0,\pi/2)$. In fact, the problem (3.9) has an oscillatory solution $u=e^{-t}\sin t\cos x$.

Example 2. Consider the problem

$$u_{tt}-u_{xx}+\frac{2}{3}e^{-\sqrt{2}t}e^{4x}u^3=\left(-4\sin t+2^{1/2}\cos t-\frac{1}{6}\sin 3t\right)e^{2^{-1/2}t}e^{-2x}, \qquad (t,x)\in(0,\infty)\times(0,1), \tag{3.10}$$

$$u_x(t,1)=-2e^{2^{-1/2}t}(\sin t)e^{-2}, \qquad t\in(0,\infty),$$

$$-u_x(t,0)=2e^{2^{-1/2}t}\sin t, \qquad t\in(0,\infty).$$

It is easy to see that

$$G(t)=-2(e^{-2}-1)e^{2^{-1/2}t}\sin t,$$

$$F(t)=(e^{-2}-1)e^{2^{-1/2}t}\left(-2^{-1/2}\cos t+\frac{1}{12}\sin 3t\right).$$

A computation gives

$$\int_T^t\left(1-\frac{s}{t}\right)(G(s)+F(s))\,ds$$

$$=-(e^{-2}-1)e^{2^{-1/2}t}\left(3^{-1}2^{1/2}\sin(t+\alpha)-\frac{1}{114}\sin(3t+\alpha')\right)+B(t,T),$$

where α and α' are constants and $B(t,T)$ is bounded as $t\to\infty$. Hence, we conclude that the conditions of Theorem 3.3 are satisfied. In fact, the problem (3.10) has an oscillatory solution $u=e^{2^{-1/2}t}(\sin t)e^{-2x}$.

4. Noncylindrical domains. We continue to study the equation

$$u_{tt}-\Delta u+c(t,x,u)=f(t,x), \tag{4.1}$$

but now generalize our considerations to a domain $\Omega\subseteq(0,\infty)\times\mathbf{R}^n$ where $G_\tau\equiv\{(t,x)\in\Omega|t=\tau\}$ is a smooth nonempty domain in $\mathbf{R}^n$ for all $\tau>0$. Letting Γ denote $\{(t,x)\in\partial\Omega|t>0\}$, we consider (4.1) subject to the boundary condition

$$u=0 \quad \text{on } \Gamma. \tag{4.2}$$

If Γ is timelike in the sense of [1], then (4.1) and (4.2) are compatible, and a well-posed problem follows from the imposition of Cauchy data on $G_0=\{(t,x)\in\partial\Omega|t=0\}$.

Proceeding as in §3, we associate with every $u\in D_0(\Omega)\equiv\{u\in D(\Omega)|u=0 \text{ on } \Gamma\}$ the function

$$y[u](t)=\int_{G_t}u(t,x)\,dx.$$

By the multivariate form of Leibniz's rule and (4.2),

$$(y[u](t))''=\int_{G_t}u_{tt}(t,x)\,dx+\int_{\partial G_t}u_t(t,x)\tan\theta(t,x)\,d\sigma,$$

where θ is the complement of the angle between the inward unit normal to Γ and the positive t-axis. Denoting this $(n+1)$-dimensional unit normal by $\nu=(\nu_1,\cdots,\nu_{n+1})$, we have

$$\nu_{n+1}=\cos\left(\frac{\pi}{2}-\theta\right)$$

so that

$$\tan\theta = \tan\sin^{-1}\nu_{n+1} = \frac{\nu_{n+1}}{\sqrt{\nu_1^2 + \cdots + \nu_n^2}}.$$

Using Green's theorem to write

$$\int_{G_t} u_{tt}(t,x)\,dx = \int_{\partial G_t} \nabla u(t,x)\cdot n\,d\sigma - \int_{G_t} [c(t,x,y(t)) - f(t,x)]\,dx,$$

where

$$n = \frac{-1}{\sqrt{\nu_1^2 + \cdots + \nu_n^2}}(\nu_1, \cdots, \nu_n),$$

we get

$$(y[u](t))'' = \int_{\partial G_t} \frac{(\nabla u, u_t)\cdot(-\nu_1, -\nu_2 \cdots -\nu_n, \nu_{n+1})}{\sqrt{\nu_1^2 + \cdots + \nu_n^2}}\,d\sigma \tag{4.3}$$

$$-\int_{G_t} [c(t,x,y(t)) - f(t,x)]\,dx.$$

This formula leads to the proof of the following result.

THEOREM 4.1. *Suppose conditions* (B-I) *through* (B-III) *are satisfied in* Ω *and that* $\partial\Omega$ *is timelike. If neither*

$$\frac{1}{|G_t|} y'' + p(t)\varphi(y) \le -F(t) \tag{4.4}$$

nor

$$\frac{1}{|G_t|} y'' + p(t)\varphi(y) \le F(t) \tag{4.5}$$

has a solution which remains positive for large t, then for every $T>0$ *every solution of* (4.1), (4.2) *has a zero in* $\{(x,t)\in\Omega | t>T\}$.

Proof. Suppose to the contrary that $u(t,x)$ is positive in $\Omega_T \equiv \{(t,x)\in\Omega | t\ge T\}$. Since $u=0$ on Γ, we have for $(t,x)\in\Gamma$

$$(\nabla u, u_t) = g(t,x)\nu, \quad \text{where } g(t,x) = \sqrt{|\nabla u|^2 + u_t^2}\,.$$

Thus we have

$$(\nabla u, u_t)\cdot(-\nu_1, \cdots, -\nu_n, \nu_{n+1}) = g(t,x)\left[\nu_{n+1}^2 - \sum_{i=1}^n \nu_i^2\right],$$

and the assumption that Γ is timelike assures (see [1]) that the first integral in (4.3) is nonpositive. It therefore follows as in §3 that (4.4) is satisfied for $t>T$. If $u(t,x)$ is negative in Ω_T, then we similarly obtain that $U(t,x)$ is a positive solution of $U_{tt} - \Delta U + c(t,x,U) = -f(t,x)$ and that $y[U](t)$ satisfies (4.5). Thus if neither (4.4) nor (4.5) can have a solution which is positive for $T<t<\infty$, then every solution of (4.1), (4.2) must have a zero in Ω_T.

Specific oscillation criteria now follow by applying the previously cited criteria of Kusano and Naito [2]. In particular, Theorem 4.1 establishes the somewhat surprising fact that as long as Γ remains timelike and $u=0$ on Γ, the passage to noncylindrical domains tends to induce oscillation of solutions of (4.1) and (4.2) regardless of whether G_t is growing or shrinking.

REFERENCES

[1] K. Kreith, *A Sturm theorem for partial differential equations of mixed type*, Proc. Amer. Math. Soc., 81 (1981), pp. 75–78.

[2] T. Kusano and M. Naito, *Oscillation criteria for a class of perturbed Schrödinger equations*, Canad. Math. Bull., 25 (1982), pp. 71–77.

[3] G. Pagan, *An oscillation theorem for characteristic initial value problems in linear hyperbolic equations*, Proc. Roy. Soc. Edinburgh Sect. A, 77 (1979), pp. 265–271.

[4] N. Yoshida, *An oscillation theorem for characteristic initial value problems for nonlinear hyperbolic equations*, Proc. Amer. Math. Soc., 76 (1979), pp. 95–100.

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics
012

UNIFORM L^1 BEHAVIOR IN CLASSES OF INTEGRODIFFERENTIAL EQUATIONS WITH COMPLETELY MONOTONIC KERNELS*

KENNETH B. HANNSGEN† AND ROBERT L. WHEELER†‡

Abstract. We find conditions on a family $\mathcal{A}$ of completely monotone, locally integrable, nonconstant functions a which enable us to write the solution $u(t;\ a)$ of $u'(t)+\int_0^t a(t-s)u(s)\,ds=0$, $u(0)=1$, as $u(t;\ a)=-\int_{-\varepsilon}^0 e^{\sigma t}d\mu(\sigma;\ a)+u_1(t;\ a)$, where $\mu(\cdot;\ a)$ is a finite nonnegative measure on $[-\varepsilon,0]$ and $|u_1(t;\ a)|\leq Qe^{-\varepsilon t}$ with Q, ε positive constants independent of $a\in\mathcal{A}$. This formula is then utilized to give conditions on the collection $\mathcal{A}$ which ensure that $\rho(t)\sup_{a\in\mathcal{A}}|u(t;\ a)|\to 0(t\to\infty)$ and $\int_0^\infty\rho(t)\sup_{a\in\mathcal{A}}|u(t;\ a)|\,dt<\infty$, where ρ is a given weight function. These results can be combined with a resolvent formula to investigate the asymptotic behavior as $t\to\infty$ of solutions of certain integrodifferential equations in Hilbert space.

1. Introduction. We consider the solution $u(t)=u(t;\ a)$ of the problem

$$u'(t)+\int_0^t a(t-s)u(s)\,ds=0,\qquad u(0)=1,\tag{1.1}$$

where $a(t)$ is completely monotonic on $(0,\infty)$, and

$$\int_0^1 a(t)\,dt<\infty\quad\text{and}\quad 0\leq a(\infty)<a(0^+)\leq\infty.\tag{1.2}$$

In [7] we showed that for each such kernel $a(t)$ there exist finite numbers Q and ε and a finite nonnegative measure μ on $[-\varepsilon,0]$ such that

$$u(t)=-\int_{-\varepsilon}^0 e^{\sigma t}d\mu(\sigma)+u_1(t)\qquad(t\geq 0),$$

where

$$|u_1(t)|\leq Qe^{-\varepsilon t}\qquad(t\geq 0).$$

Here we show that Q and ε can be made uniform over certain classes of kernels. As a consequence, we obtain estimates of the form

$$\rho(t)\sup_{a\in\mathcal{A}}|u(t;\ a)|\to 0\qquad(t\to\infty)\tag{1.3}$$

and

$$\int_0^\infty\rho(t)\sup_{a\in\mathcal{A}}|u(t;\ a)|\,dt<\infty,\tag{1.4}$$

where $\mathcal{A}$ is a suitable family of kernels and ρ is a weight function.

When $\mathcal{A}=\{\lambda a_0(t):0<\lambda_0\leq\lambda<\infty\}$, (1.3) and (1.4) with $\rho(t)\equiv 1$ are true whenever $a=a_0$ satisfies (1.2), $a_0(t)$ is nonnegative, nonincreasing and convex, and $-a_0'(t)$ is convex. These results and similar ones were proved in [1], [2], [5], [6], where the technique of proof depends crucially on a deep theorem of D. F. Shea and S. Wainger

*Received by the editors October 19, 1982, and in revised form January 10, 1983. This material is based upon work supported by the National Science Foundation under grant MCS-8101618.

†Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

‡Support by the Mathematics Research Center for a short visit to Madison is gratefully acknowledged.

[12] which implies that $u(t;\ a_0)$ belongs to $L^1(0,\infty)$ for each such function a_0. This method of proof does not appear to be applicable to the more general families $\mathcal{A}$ considered in this paper.

The estimates (1.3) and (1.4) ($\rho\equiv 1$) were used in [1], [5], [6] to estimate the resolvent kernel

$$\mathbf{U}(t)=\int_{\lambda_0}^{\infty}u(t;\ \lambda a_0)\,d\mathbf{E}_\lambda$$

of the problem

$$\mathbf{y}'(t)=-\int_0^t a_0(t-s)\mathbf{L}\mathbf{y}(s)\,ds+\mathbf{f}(t),\qquad \mathbf{y}(0)=\mathbf{y}_0 \tag{1.5}$$

in a Hilbert space $\mathcal{H}$. Here $\mathbf{L}$ is a densely defined positive selfadjoint linear operator on $\mathcal{H}$ with spectral decomposition $\mathbf{L}=\int_{\lambda_0}^{\infty}\lambda\,d\mathbf{E}_\lambda$, and $\mathbf{y}_0$ and $\mathbf{f}(t)$ are prescribed elements of $\mathcal{H}$. Since $\|\mathbf{U}(t)\|\le\sup_{\lambda_0\le\lambda<\infty}|u(t;\ \lambda a_0)|$ $(t\ge 0)$, the estimates (1.3) and (1.4) with $\rho\equiv 1$ imply, respectively,

$$\|\mathbf{U}(t)\|\to 0\quad (t\to\infty),\qquad \int_0^{\infty}\|\mathbf{U}(t)\|\,dt<\infty, \tag{1.6}$$

so that the resolvent formula

$$\mathbf{y}(t)=\mathbf{U}(t)\mathbf{y}_0+\int_0^t\mathbf{U}(t-s)\mathbf{f}(s)\,ds \tag{1.7}$$

for (1.5) yields information about the asymptotic behavior of $\mathbf{y}(t)$ as $t\to\infty$.

In the same way, the results (1.3) and (1.4) for more general classes $\mathcal{A}=\{a(t;\ \lambda):\lambda\in\mathbb{R}\}$ provide estimates for the resolvent

$$\mathbf{U}(t)=\int_{-\infty}^{\infty}u(t;\ a(\cdot\,;\lambda))\,d\mathbf{E}_\lambda \tag{1.8}$$

for the problem

$$\mathbf{y}'(t)=-\int_0^t\mathbf{L}(t-s)\mathbf{y}(s)\,ds+\mathbf{f}(t),\qquad \mathbf{y}(0)=\mathbf{y}_0 \tag{1.9}$$

with

$$\mathbf{L}(t)=\int_{-\infty}^{\infty}a(t;\ \lambda)\,d\mathbf{E}_\lambda, \tag{1.10}$$

where $\{\mathbf{E}_\lambda\}$ is now some fixed resolution of the identity in $\mathcal{H}$. As in the earlier results, (1.3) and (1.4) ($\rho\equiv 1$) yield (1.6), which can be used with (1.7). (The proof of (1.7) for (1.9) with suitable $\mathbf{y}_0$ and $\mathbf{f}$ follows the same lines as that for (1.5) [1] and will not be given here.)

Our results for (1.9) include some operator kernels of the form

$$\mathbf{L}(t)=\sum_{k=0}^{N}a_k(t)\mathbf{L}_k. \tag{1.11}$$

(See Corollaries 2.2 and 2.3.) The requirement that the $\mathbf{L}_k$ have spectral decompositions with respect to a common resolution of the identity $\{\mathbf{E}_\lambda\}$ greatly restricts the applicability of the result, but we can obtain new results on the asymptotic behavior of solutions of some integrodifferential equations of interest. For example, we can take

$\mathbf{L}_k=\mathbf{L}_0^{k+1}$, $k=0,1,\cdots,N$, where $\mathbf{L}_0$ is a selfadjoint operator on $\mathcal{H}$. The equation

$$w_t(x,t)=\int_0^t\left[a_0(t-s)w_{xx}(x,s)-a_1(t-s)w_{xxxx}(x,s)\right]ds+f(x,t)$$

with selfadjoint boundary conditions has this form.

For another example, consider the problem

$$w_t(x,y,t)=\int_0^t\left[a_0(t-s)w_{xx}(x,y,s)+b_0(t-s)w_{yy}(x,y,s)\right]ds+f(x,y,t), \tag{1.12}$$

$$w(x,y,0)=w_0(x,y),\qquad \alpha<x<\beta,\quad \gamma<y<\delta,\quad t>0,$$

with boundary conditions

$$w(\alpha,y,t)=w(\beta,y,t)=w(x,\gamma,t)=w(x,\delta,t)=0. \tag{1.13}$$

This problem arises in a linear model for heat flow in a rectangular, orthotropic material with memory [3], [11] in which the axes of orthotropy are parallel to the edges of the rectangle. The corresponding problem in one space dimension is discussed in [5]. (See [6] for related problems with applications to viscoelasticity.)

Let

$$\mathbf{L}_0=-\frac{(\beta-\alpha)^2}{\pi^2}\frac{\partial^2}{\partial x^2}-\frac{(\delta-\gamma)^2}{\pi}\frac{\partial^2}{\partial y^2}.$$

Then $\mathbf{L}_0$ with boundary conditions (1.13) can be viewed as a densely defined selfadjoint operator on $L^2([\alpha,\beta]\times[\gamma,\delta])$ with orthonormal eigenfunctions

$$\mathbf{e}_{m,n}(x,y)=c_{m,n}\sin m\pi\frac{x-\alpha}{\beta-\alpha}\sin n\pi\frac{y-\gamma}{\delta-\gamma}$$

corresponding to the eigenvalues $\lambda(m,n)=m^2+\pi n^2$, $m,n=1,2,3,\cdots$, and spectral family $d\mathbf{E}_\lambda\mathbf{g}=\langle\mathbf{g},\mathbf{e}_{m(\lambda),n(\lambda)}\rangle\mathbf{e}_{m(\lambda),n(\lambda)}$. (Our choice of $\mathbf{L}_0$ ensures that $m=m(\lambda)$ and $n=n(\lambda)$ are uniquely determined by λ.) Now set

$$\mathbf{L}(t)=\int_{-\infty}^{\infty}\left[a_0(t)\left(\frac{m(\lambda)\pi}{\beta-\alpha}\right)^2+b_0(t)\left(\frac{n(\lambda)\pi}{\delta-\gamma}\right)^2\right]d\mathbf{E}_\lambda.$$

This puts (1.12) in the form (1.9), (1.10), and again Corollaries 2.2 and 2.3 can be used to study the asymptotic behavior of solutions of (1.12).

We emphasize that the restricted form and rectangular geometry of (1.12) were necessary for the application of our results regarding (1.9). The precise results available through (1.6) and (1.7) can perhaps serve to test other methods which apply to wider classes of problems.

2. Statement and discussion of results. A completely monotonic function on $(0,\infty)$ can be represented in [13] as the Stieltjes integral

$$a(t)=\int_0^\infty e^{-xt}d\alpha(x), \tag{2.1}$$

where $\alpha(0)=0$, α is nondecreasing, and $\alpha(x)=\alpha(x^-)$ for $0<x<\infty$. The conditions in (1.2) are equivalent, respectively, to

$$\int_0^\infty\frac{1-e^{-x}}{x}d\alpha(x)<\infty, \tag{2.2}$$

$$\alpha(\infty)>\alpha(0^+). \tag{2.3}$$

For each such function $a(t)$, we define

$$\phi(\sigma,\tau)=\int_0^\infty \frac{(x+\sigma)\,d\alpha(x)}{(x+\sigma)^2+\tau^2},\qquad \theta(\sigma,\tau)=\int_0^\infty \frac{d\alpha(x)}{(x+\sigma)^2+\tau^2}\qquad (\sigma+i\tau\notin(-\infty,0]),$$

$$\phi(\tau)=\phi(0,\tau),\qquad \theta(\tau)=\theta(0,\tau)\qquad (0<\tau<\infty),$$

$$\tilde{\phi}(\sigma)=\lim_{\tau\to 0+}\phi(\sigma,\tau)\qquad \text{(wherever the limit exists on } (-\infty,0)).$$

(We remark that [7, Lemma 2.1] ensures that $\tilde{\phi}(\sigma)$ is defined almost everywhere for $\sigma<0$.) Note that

$$\phi(\sigma,\tau)-i\tau\theta(\sigma,\tau)=\int_0^\infty \frac{d\alpha(x)}{x+\sigma+i\tau}\qquad (\sigma+i\tau\notin(-\infty,0])$$

$$=\int_0^\infty e^{-t(\sigma+i\tau)}a(t)\,dt\qquad (\sigma>0).$$

THEOREM 2.1. *Suppose* (2.1), (2.2) *and* (2.3) *hold and that there exist a positive number* x_0 *and a nonnegative function* β *such that*

$$\alpha'(x)\geq\beta(x)\quad a.e.\ on\ (0,x_0) \tag{2.4}$$

and

$$\int_0^{x_0}\frac{dx}{\beta(x)}<\infty. \tag{2.5}$$

Also, assume that there is a positive function B on $(0,\infty)$ *such that*

$$\frac{\theta(\tau)}{\phi(\tau)}\leq B(\delta)<\infty\quad if\ 0<\delta\leq\tau\ and\ \theta(\tau)\geq\frac{1}{4}. \tag{2.6}$$

Then there exist Q, $\varepsilon>0$ $(\varepsilon\leq x_0)$, *depending only on* x_0 *and the functions* β *and* B, *such that the solution* $u(t;\,a)$ *of* (1.1) *satisfies*

$$\left|u(t;\,a)+\int_{-\varepsilon}^0 e^{\sigma t}\frac{\alpha'(-\sigma)\,d\sigma}{[\sigma+\tilde{\phi}(\sigma)]^2+[\pi\alpha'(-\sigma)]^2}\right|\leq Qe^{-\varepsilon t}\qquad (t\geq 0). \tag{2.7}$$

For every fixed completely monotonic kernel a satisfying (1.2), (2.6) holds [1, Thm. 2.2 and Cor. 2.1]. An example below shows that the uniform condition (2.6) is needed for uniform bounds on Q and ε.

Next we consider a family of kernels $\mathcal{A}=\{a_j : j\in J\}$,

$$a_j(t)=\int_0^\infty e^{-xt}\,d\alpha_j(x)\qquad (t>0,\ \ j\in J),$$

where J is an arbitrary index set. If, for each j in J, a_j satisfies the hypotheses of Theorem 2.1, with x_0, β and B independent of j, then (2.7) together with the elementary inequality

$$\frac{\alpha_j'(-\sigma)}{[\sigma+\tilde{\phi}_j(\sigma)]^2+[\pi\alpha_j'(-\sigma)]^2}\leq\frac{1}{\pi^2\alpha_j'(-\sigma)}\leq\frac{1}{\pi^2\beta(-\sigma)} \tag{2.8}$$

($j\in J$, a.e. $\sigma\in(-\varepsilon,0)$) shows that (1.3) holds, provided we make the additional assumptions $\rho(t)=o(e^{\varepsilon t})$ $(t\to\infty)$ and

$$\rho(t)\int_0^{\varepsilon} e^{-xt}\frac{dx}{\beta(x)}\to 0 \qquad (t\to\infty). \tag{2.9}$$

In particular, the dominated convergence theorem and (2.5) show that (2.9) always holds when $\rho(t)\equiv 1$, so we have (1.3) with $\rho(t)\equiv 1$. On the other hand, to deduce (1.4), even with $\rho(t)\equiv 1$, for general families of kernels $\mathcal{A}$, we need to strengthen condition (2.5). (See the example provided by the family $\mathcal{A}_1$ following the next corollary.)

COROLLARY 2.2. *For each j in J, assume that* (2.1)–(2.3) *hold with $a=a_j$ and that the corresponding functions α_j, θ_j, and ϕ_j satisfy* (2.4), (2.5) *and* (2.6) *with x_0, β, and B independent of j. In addition, let $\rho(t)$ be a nonnegative locally bounded function on* $[0,\infty)$ *such that $\rho(t)=o(e^{\eta t})$ $(t\to\infty)$ for each $\eta>0$ and such that $\hat{\rho}(x)\equiv\int_0^\infty e^{-xt}\rho(t)\,dt$ $(x>0)$ satisfies*

$$\int_0^{x_0}\frac{\hat{\rho}(x)}{\beta(x)}\,dx<\infty. \tag{2.10}$$

Then (1.4) *holds.*

We note that the hypotheses of Corollary 2.2 are satisfied by families of the form

$$\mathcal{A}=\left\{\sum_{k=0}^{N}\lambda_{jk}a_k(t):\lambda_{jk}\ge 1\right\},\quad \text{where } a_0(t)=\int_0^\infty e^{-xt}\beta'(x)\,dx,$$

with β a differentiable function which satisfies (2.5) and (2.10), and a_k completely monotone on $(0,\infty)$ and satisfying (1.2) with $a=a_k$ for $1\le k\le N$. For (2.6) it is sufficient to note that each of the corresponding θ_k, ϕ_k $(0\le k\le N)$ satisfies (2.6) [1, Thm. 2.2 and Cor. 2.1], and then we use the form of the functions in $\mathcal{A}$. In particular, if $\mathbf{L}_k=\int_{-\infty}^{\infty}g_k(\lambda)\,d\mathbf{E}_\lambda$ with $g_k(\lambda)\ge 1$ $(0\le k\le N)$, we can apply our results to (1.9) with kernel (1.11) by taking $J=\mathbb{R}$, $a(t;\ \lambda)=\sum_{k=0}^{N}a_k(t)g_k(\lambda)$. Then $\{a(t;\ \lambda):\lambda\in\mathbb{R}\}\subset\{\sum_{k=0}^{N}\lambda_k a_k(t):\lambda_k\ge 1\}=\mathcal{A}$.

An example of the form $\mathcal{A}=\{\sum_{k=0}^{N}\lambda_{jk}a_k(t):\lambda_{jk}\ge 1\}$ for which (2.10) is satisfied when $\rho(t)\equiv 1$ (so that $\hat{\rho}(x)=1/x$, $x>0$) is provided by taking $a_0(t)=t^{-\gamma}=\int_0^\infty e^{-xt}\beta'(x)\,dx$ with $\beta(x)=x^{\gamma}/\gamma\Gamma(\gamma)$ $(0<\gamma<1)$.

When (2.10) fails to hold, one would like to use lower bounds on $\tilde{\phi}(\sigma)$ in (2.7) to get (1.4). However, an example showing that this is impossible for general families $\mathcal{A}$ is provided by taking $\mathcal{A}_1=\{a_c(t)=a_0(t)+c:0\le c\le 1\}$ with $a_0(t)=(1-e^{-t})/t$. Then $a_c(t)=\int_0^\infty e^{-xt}\,d\alpha_c(t)$ with

$$\alpha_c(t)=\begin{cases}0, & x=0,\\ c+x, & 0<x\le 1,\\ c+1, & 1<x<\infty.\end{cases}$$

It is shown in [8, §4] that (1.4) with $\rho\equiv 1$ is false for the family $\mathcal{A}_1$. On the other hand, (2.7) holds with Q, ε independent of a in $\mathcal{A}_1$, since the hypotheses of Theorem 2.1 are satisfied with $\beta(x)=1$, $0<x\le 1\equiv x_0$, and the existence of a finite $B(\delta)$ independent of c, $0\le c\le 1$, such that (2.6) holds, is readily checked.

The difficulty in relaxing (2.10) by using lower bounds on $\tilde{\phi}(\sigma)$ in (2.7) to get (1.4) is caused by the fact that $\tilde{\phi}(\sigma)$ is essentially the Hilbert transform of $d\alpha$ and is hard to estimate. Only restricted results like the following seem to be available.

COROLLARY 2.3. *Let* $\mathfrak{A}=\{\sum_{k=1}^{N}\lambda_{jk}a_k(t):\lambda_{jk}\geq 1,\ j\in J,\ 1\leq k\leq N\}$ *and assume that each of the corresponding functions* α_k *in* (2.1) *satisfies* (2.2), (2.3) *and*

$$\alpha_k(0^+)>0, \qquad 1\leq k\leq N. \tag{2.11}$$

In addition, assume that there exists $x_0>0$ *so that*

$$d\alpha_k(x)=\alpha_k'(x)\,dx, \qquad 0<x<x_0, \quad 1\leq k\leq N, \tag{2.12}$$

and so that (2.5) *holds with* $\alpha_k'=\beta$ *for at least one* k, $1\leq k\leq N$. *Finally, suppose that there exist positive constants* K *and* γ *such that for each* $\xi\in(0,x_0)$,

$$\alpha_k'(\xi+\delta)-\alpha_k'(\xi-\delta)\leq\frac{K\delta^{\gamma}}{\xi} \quad \textit{for a.e. } \delta\in(0,x_0-\xi), \quad 1\leq k\leq N. \tag{2.13}$$

Then both (1.3) *and* (1.4) *hold with* $\rho(t)=1+t$.

We remark that (2.11) is equivalent to $a_k(t)\to a_k(\infty)>0$ $(t\to\infty)$ for $1\leq k\leq N$. It was established in [9] that the solution u of (1.1) satisfies $(1+t)u(t)\in L^1(0,\infty)$ whenever a is a nonnegative, nonincreasing and convex function on $(0,\infty)$ such that (1.2) holds and $a(\infty)>0$. (Certain piecewise linear kernels are excepted.) Also, observe that condition (2.13) is satisfied whenever the functions $\alpha_k'(x)$, $1\leq k\leq N$, are nonincreasing on $(0,x_0)$, so that Corollary 2.3 can be applied to a fairly broad class of kernels $a_k(t)$.

We conclude this section with an example that shows the necessity of hypothesis (2.6) in Theorem 2.1. Let $a_c(t)=c+a_0(t)$, where $a_0(t)=\int_0^\infty e^{-xt}\,d\alpha(x)$ satisfies (2.2), (2.3) and $a_0(t)\to 0$ $(t\to\infty)$, and $1\leq c<\infty$. Set $\mathfrak{A}_2=\{a_c: 1\leq c<\infty\}$. Then (2.6) for a_c becomes

$$\frac{\theta_0(\tau)+c/\tau^2}{\phi_0(\tau)}\leq B(\delta) \quad \text{if } 0<\delta\leq\tau \text{ and } \theta_0(\tau)+\frac{c}{\tau^2}\geq\frac{1}{4},$$

where θ_0 and ϕ_0 correspond to a_0. Clearly there is no function B for which this inequality holds for $1\leq c<\infty$. Assume, in addition, that (2.10) holds with $\rho(t)\equiv 1$ and $\beta=\alpha'$ (so (2.5) holds, too). It follows that if the conclusion (2.7) of Theorem 2.1 is valid for a_c in $\mathfrak{A}_2$ with Q, ε independent of c, then (1.4) holds with $\rho(t)\equiv 1$. (See the proof of Corollary 2.2 in §4.) However, if we take Fourier transforms in (1.1), we see that the Fourier transform $u^*(\tau;\ a_c)$ satisfies

$$u^*(\tau;\ a_c)=\frac{1}{i\tau+c/i\tau+a_0^*(\tau)}, \qquad \tau>0,$$

where $a_0^*(\tau)=\int_0^\infty(i\tau+x)^{-1}\,d\alpha(x)$ $(\tau>0)$. By the dominated convergence theorem $a_0^*(\tau)\to 0$ $(\tau\to\infty)$, so $|u^*(\sqrt{c};\ a_c)|=|a_0^*(\sqrt{c})|^{-1}\to\infty$ $(c\to\infty)$. But then the elementary estimate

$$|u^*(\tau;\ a_c)|\leq\int_0^\infty|u(t;\ a_c)|\,dt, \qquad \tau>0, \quad 1\leq c<\infty,$$

shows that (1.4) $(\rho\equiv 1)$ must fail, so (2.7) cannot be true for a_c in $\mathfrak{A}_2$ with Q and ε independent of c.

3. Proof of Theorem 2.1.

Throughout this section K will denote a positive constant which depends only on x_0, β and B; the value of K may change from line to line. We introduce the notation

$$R(\sigma,\tau)=\sigma+\phi(\sigma,\tau), \qquad I(\sigma,\tau)=\tau[1-\theta(\sigma,\tau)],$$
$$D=R+iI \qquad (\sigma+i\tau\notin(-\infty,0]).$$

An integration by parts in (2.2) shows that

$$\frac{\alpha(x)}{x}\to 0 \quad (x\to\infty), \qquad \int_1^\infty \frac{\alpha(x)}{x^2}\,dx<\infty.$$

Also, using the definition of θ, we see that

$$(3.1) \qquad \theta'(\tau)<0 \quad (\tau>0), \qquad \infty=\int_0^\infty \frac{d\alpha(x)}{x^2}=\theta(0^+)>\theta(\tau)\downarrow 0 \quad (\tau\uparrow\infty).$$

(The first equality uses (3.5) below.)

We recall from [1, Lemmas 4.1 and 5.2] (see also [2]) that if (2.1), (2.2), and (2.3) hold, then there exist specific positive constants K_0, K_1 (independent of a) so that

$$(3.2) \qquad \frac{1}{K_0}\int_0^{1/\tau} ta(t)\,dt\le\theta(\tau)\le K_0\int_0^{1/\tau} ta(t)\,dt \qquad (\tau>0),$$

$$(3.3) \qquad |D(0,\tau)|\ge\begin{cases} K_1|\tau-\omega| & \left(\tau\ge\frac{\omega}{2}\right),\\ K_1\tau\int_0^{1/\tau} ta(t)\,dt & \left(\delta\le\tau\le\frac{\omega}{2}\right).\end{cases}$$

Here ω is the unique positive solution of $\theta(\omega)=1$, and $\delta=\delta(\beta)>0$ will be specified in Lemma 3.1 below. (We remark that here δ plays the role of ρ in [1]. The estimates in [1] are derived for $a(t)+d$ with $a(\infty)=0$ and appear to depend on $d/a(6/\rho)$. The proofs can easily be modified to show that (3.2) and (3.3) are valid.)

We now state and prove three lemmas needed for the proof of Theorem 2.1.

LEMMA 3.1. *Let* (2.1) *through* (2.5) *be satisfied. Then there exist positive constants* η *and* δ, *depending only on* β, *such that*

$$(3.4) \qquad \frac{I(\sigma,\tau)}{\tau}\le -1 \qquad (-\eta\le\sigma\le 0,\quad 0<\tau\le\delta).$$

Proof. The proof of Lemma 3.1 is essentially contained in [7, proof of Lemma 2.2]. Namely, by the Schwarz inequality

$$\int_0^{x_0}\frac{\beta(x)}{x^2}\,dx\int_0^{x_0}\frac{dx}{\beta(x)}\ge\left\{\int_0^{x_0}\frac{dx}{x}\right\}^2=\infty,$$

so (2.5) implies

$$(3.5) \qquad \int_0^{x_0}\frac{\beta(x)}{x^2}\,dx=\infty.$$

Choose $\eta>0$ so that $\int_\eta^{x_0}(\beta(x)/x^2)\,dx\ge 4$; then choose $\delta>0$ so that

$$\int_\eta^{x_0}\frac{\beta(x)}{x^2+\delta^2}\,dx\ge 2.$$

Now for $-\eta\le\sigma\le 0$, $0<\tau\le\delta$,

$$\int_0^{x_0}\frac{\beta(x)}{(x+\sigma)^2+\tau^2}\,dx\ge\int_\eta^{x_0}\frac{\beta(x)}{x^2+\delta^2}\,dx\ge 2.$$

Inequality (3.4) follows from this inequality since, by (2.4), we also have

$$\frac{I(\sigma,\tau)}{\tau}=1-\int_0^\infty \frac{d\alpha(x)}{(x+\sigma)^2+\tau^2}$$
$$\leq 1-\int_0^{x_0}\frac{\beta(x)}{(x+\sigma)^2+\tau^2}dx \qquad (\sigma+i\tau\notin(-\infty,0]). \qquad \square$$

LEMMA 3.2. *Let* (2.1), (2.2) *and* (2.3) *hold. Then*

$$\phi(\tau)\geq\frac{1}{2\tau^2}\int_0^\tau x\,d\alpha(x)+\frac{1}{2}\int_\tau^\infty\frac{d\alpha(x)}{x}>0 \qquad (\tau>0), \tag{3.6}$$

$$\frac{g(\tau)}{2}\leq\theta(\tau)\leq g(\tau) \qquad (\tau>0), \tag{3.7}$$

where

$$g(\tau)\equiv\int_0^\tau\frac{d\alpha(x)}{\tau^2}+\int_\tau^\infty\frac{d\alpha(x)}{x^2}=2\int_\tau^\infty\frac{\alpha(x)}{x^3}dx \qquad (\tau>0).$$

Proof. The inequality (3.6) is easily obtained from the definition of ϕ using elementary estimates. Similarly, the inequalities in (3.7) are consequences of the definitions of θ and g and easy estimates. The second equality in the definition of g is obtained by an integration by parts using $\alpha(0)=0$ and $\alpha(x)/x^2\to 0$ $(x\to\infty)$. □

The next lemma gives estimates for $\phi(\sigma,\tau)$ and $\theta(\sigma,\tau)$ in a strip to the left of the imaginary axis in terms of $\phi(\tau)$ and $\theta(\tau)$, respectively.

LEMMA 3.3. *Assume that* (2.1) *through* (2.3) *hold and let* $\delta>0$. *Then*

$$|\theta(\sigma,\tau)-\theta(\tau)|\leq\frac{24\varepsilon\theta(\tau)}{\tau} \qquad \left(\delta\leq\tau<\infty,\quad 0<-\sigma\leq\varepsilon<\frac{\delta}{2}\right). \tag{3.8}$$

Assume, in addition, that (2.6) *is satisfied. Then*

$$\phi(\sigma,\tau)\geq\frac{\phi(\tau)}{2} \qquad (\delta\leq\tau\leq\infty,\quad 0<-\sigma\leq\varepsilon) \tag{3.9}$$

provided that $\theta(\tau)\geq\frac{1}{4}$ *and* ε *satisfies*

$$\varepsilon\left(B(\delta)+\frac{4}{\delta}\right)<\frac{1}{4}. \tag{3.10}$$

Proof. To deduce (3.8), note that

$$\theta(\sigma,\tau)-\theta(\tau)=-\int_0^\infty\frac{2x\sigma+\sigma^2}{[(x+\sigma)^2+\tau^2](x^2+\tau^2)}d\alpha(x).$$

Thus, when $\delta\leq\tau<\infty$, $0<-\sigma\leq\varepsilon<\delta/2$, easy estimates, an integration by parts and (3.7) yield

$$|\theta(\sigma,\tau)-\theta(\tau)|\leq\frac{\varepsilon}{\tau^4}\int_0^\tau 2\tau\,d\alpha(x)+8\varepsilon\int_\tau^\infty\frac{d\alpha(x)}{x^3}$$
$$\leq 24\varepsilon\int_\tau^\infty\frac{\alpha(x)}{x^4}dx\leq 24\frac{\varepsilon}{\tau}\int_\tau^\infty\frac{\alpha(x)}{x^3}dx\leq\frac{24\varepsilon\theta(\tau)}{\tau},$$

and (3.8) is proved.

To establish (3.9), we use the definitions of $\phi(\sigma,\tau)$ and $\phi(\tau)$ and elementary algebra to write

$$\phi(\sigma,\tau)-\phi(\tau)=\int_0^\infty \frac{\sigma\tau^2-x\sigma(x+\sigma)}{\left[(x+\sigma)^2+\tau^2\right](x^2+\tau^2)}\,d\alpha(x).$$

Then for $\delta\leq\tau<\infty$, $0<-\sigma\leq\varepsilon$, with $\theta(\tau)\geq\frac{1}{4}$ and ε restricted by (3.10), the last expression, elementary estimates, (3.7), (3.6) and (2.6) can be employed to yield

$$\begin{aligned}\phi(\sigma,\tau)-\phi(\tau)&\geq-\varepsilon\tau^2\int_0^\infty\frac{d\alpha(x)}{\left[(x+\sigma)^2+\tau^2\right](x^2+\tau^2)}-\varepsilon^2\int_0^{-\sigma}\frac{x\,d\alpha(x)}{\left[(x+\sigma)^2+\tau^2\right](x^2+\tau^2)}\\&\geq-\frac{\varepsilon}{\tau^2}\int_0^\tau d\alpha(x)-4\varepsilon\tau^2\int_\tau^\infty\frac{d\alpha(x)}{x^4}-\frac{\varepsilon}{\delta^2\tau^2}\int_0^\varepsilon x\,d\alpha(x)\\&\geq-2\varepsilon\theta(\tau)-4\frac{\varepsilon}{\delta}\int_\tau^\infty\frac{d\alpha(x)}{x}-\frac{\varepsilon^2}{\delta^2\tau^2}\int_0^\tau x\,d\alpha(x)\\&\geq-\varepsilon\phi(\tau)\left(2B(\delta)+\frac{8}{\delta}\right)\geq-\frac{\phi(\tau)}{2},\end{aligned}$$

and the proof of (3.9) is complete. □

We are now in a position to set up the integral expression for $u(t;\,a)$ which is used to deduce (2.7). Let the hypotheses of Theorem 2.1 hold and let δ be as in Lemma 3.1. Fix $\varepsilon>0$ so that

$$\beta_0(\varepsilon)\equiv\lim_{\tau\to0^+}\frac{\tau}{\pi}\int_0^{x_0}\frac{\beta(x)}{(x-\varepsilon)^2+\tau^2}\,dx>0, \tag{3.11}$$

and so that the inequalities (3.16) and (3.20) below hold. This can be done since the expression on the left side of (3.11) is the Poisson integral of $\chi\beta$ ($\chi\equiv$characteristic function of $[0,x_0]$) and $\beta(x)>0$ almost everywhere on $[0,x_0]$.

Let T be the unique positive number such that $\theta(T)=\frac{1}{4}$. (The existence of T is ensured by (3.1).) Note that $T>\delta$. Using (3.8) we see that

$$1-\theta(\sigma,\tau)\geq\frac{1}{2}\qquad\left(T\leq\tau<\infty,\quad 0\leq-\sigma\leq\varepsilon<\frac{\delta}{48}\right). \tag{3.12}$$

The number T depends on the function $a(t)$, but by (2.6) we obtain

$$\phi(\tau)\geq\frac{1}{4B(\delta)}\qquad(\delta\leq\tau\leq T). \tag{3.13}$$

Thus, (3.9) implies

$$R(\sigma,\tau)\geq\frac{1}{4}\phi(\tau)\geq\frac{1}{16B(\delta)}\qquad(\delta\leq\tau\leq T,\quad 0\leq-\sigma\leq\varepsilon) \tag{3.14}$$

if (3.10) holds and $\varepsilon<1/16B(\delta)$. Then (3.4), (3.12) and (3.14) show that

$$D(\sigma,\tau)\neq0\qquad(0<\tau<\infty,\quad 0\leq-\sigma\leq\varepsilon) \tag{3.15}$$

provided ε satisfies

$$0<\varepsilon<\min\left\{\eta,\ \frac{\delta}{48},\ \frac{1}{16B(\delta)}\right\}. \tag{3.16}$$

(We remark that (3.16) implies (3.10).)

Now if ε is fixed and satisfies (3.11) and (3.16), then (3.15) and an examination of the proof of [7, Thm. 1] (see especially [7, (2.17) with 1 replaced by δ, and (3.2)]) yields the formula

$$\begin{aligned} u(t;\, a)= &-\int_{-\varepsilon}^{0} e^{\sigma t}\frac{\alpha'(-\sigma)}{[\sigma+\tilde{\phi}(\sigma)]^2+[\pi\alpha'(-\sigma)]^2}\,d\sigma \\ &+\frac{1}{\pi}e^{-\varepsilon t}\left\{\int_{0^+}^{\delta}\operatorname{Re}\left[\frac{e^{i\tau t}}{D(-\varepsilon,\tau)}\right]d\tau\right. \\ &\left.+\frac{1}{t}\operatorname{Re}\left[\frac{ie^{i\delta t}}{D(-\varepsilon,\delta)}+\int_{\delta}^{\infty}\frac{e^{i\tau t}D_\tau(-\varepsilon,\tau)}{iD^2(-\varepsilon,\tau)}\,d\tau\right]\right\} \qquad (t>0). \end{aligned} \tag{3.17}$$

(Here the first integral within the brackets exists as an improper Riemann integral at $\tau=0$.)

In order to estimate the first integral within the brackets in (3.17), choose $\delta_1\in(0,\delta)$ so that $\delta_1<\pi\beta_0(\varepsilon)/4$ and

$$\tau\int_0^{x_0}\frac{\beta(x)}{(x-\varepsilon)^2+\tau^2}\,dx>\frac{\pi}{2}\beta_0(\varepsilon) \qquad (0<\tau\le\delta_1).$$

Since

$$\tau\theta(-\varepsilon,\tau)\ge\tau\int_0^{x_0}\frac{\beta(x)}{(x-\varepsilon)^2+\tau^2}\,dx,$$

$$|I(-\varepsilon,\tau)|\ge\frac{\pi}{4}\beta_0(\varepsilon) \qquad (0<\tau\le\delta_1),$$

and, combining this with (3.4), we get

$$\left|\int_{0^+}^{\delta}\operatorname{Re}\left[\frac{e^{i\tau t}}{D(-\varepsilon,\tau)}\right]d\tau\right|\le\frac{4\delta_1}{\pi\beta_0(\varepsilon)}+\frac{\delta-\delta_1}{\delta_1}=K.$$

We also have, by (3.4), that $|D(-\varepsilon,\delta)|\ge\delta$. Thus since $|u(t;\, a)|\le 1$ $(0\le t<\infty)$ [10], [4], we can complete the proof of Theorem 2.1 by showing that

$$\int_{\delta}^{\infty}\left|\frac{D_\tau(-\varepsilon,\tau)}{D^2(-\varepsilon,\tau)}\right|d\tau\le K. \tag{3.18}$$

We now state and prove two more lemmas needed to deduce (3.18).

Lemma 3.4. *Suppose that* (2.1) *through* (2.3) *hold and that* $\delta>0$. *Then*

$$|D_\tau(-\varepsilon,\tau)-i|\le 11\theta(\tau) \qquad \left(\delta\le\tau<\infty,\ \ 0<\varepsilon\le\frac{\delta}{48}\right). \tag{3.19}$$

Proof. By differentiating the expression for $D(-\varepsilon,\tau)$ with respect to τ, we get

$$D_\tau(-\varepsilon,\tau)-i=-i\int_0^\infty \frac{(x-\varepsilon)^2-\tau^2-2i\tau(x-\varepsilon)}{\left[(x-\varepsilon)^2+\tau^2\right]^2}\,d\alpha(x),$$

so easy estimates yield

$$|D_\tau(-\varepsilon,\tau)-i|\le\int_0^\infty \frac{d\alpha(x)}{(x-\varepsilon)^2+\tau^2}+2\tau\int_0^\infty \frac{d\alpha(x)}{\left[(x-\varepsilon)^2+\tau^2\right]^{3/2}}.$$

Then, by the definition of θ, more elementary estimates (recall $\varepsilon\le\delta/48\le\tau/48$), an integration by parts, (3.7) and (3.8), we obtain

$$\begin{aligned}|D_\tau(-\varepsilon,\tau)-i|&\le\theta(-\varepsilon,\tau)+\frac{2}{\tau^2}\int_0^\tau d\alpha(x)+3\tau\int_\tau^\infty\frac{d\alpha(x)}{x^3}\\&\le\theta(-\varepsilon,\tau)+9\tau\int_\tau^\infty\frac{\alpha(x)}{x^4}dx\\&\le\frac{3}{2}\theta(\tau)+9\theta(\tau)\le 11\theta(\tau).\end{aligned}$$ □

LEMMA 3.5. *Let* (2.1) *through* (2.6) *hold and let* δ, η *be the constants given by Lemma* 3.1. *Assume that* ε *satisfies* (3.16) *as well as*

$$8\varepsilon<\min\left\{\frac{\delta}{12},\frac{\delta K_1}{24},\frac{\delta K_1}{12K_0}\right\},\tag{3.20}$$

where K_0 *and* K_1 *are the constants that occur in* (3.2), (3.3), *respectively. Then*

$$K|D(-\varepsilon,\tau)|\ge\begin{cases}\tau\theta(\tau) & \left(\delta\le\tau\le\frac{\omega}{2}\right),\\ |\tau-\omega| & \left(\max\left\{\delta,\frac{\omega}{2}\right\}\le\tau\le T,\quad |\tau-\omega|\ge\delta\right),\\ \phi(\tau) & (\delta\le\tau\le T,\quad |\tau-\omega|\le\delta),\\ \tau & (\tau\ge T).\end{cases}\tag{3.21}$$

Proof. The third and fourth estimates in (3.21) are easy consequences of (3.14) and (3.12), respectively.

To prove the first estimate in (3.21), note that if $\tau\in[\delta,\omega/2]$ and

$$\tau(\theta(\tau)-1)<\frac{K_1\tau}{2}\int_0^{1/\tau}ta(t)\,dt,$$

then (3.14), (3.3) and (3.2) give

$$|D(-\varepsilon,\tau)|\ge\frac{\phi(\tau)}{4}\ge\frac{K_1\tau}{8}\int_0^{1/\tau}ta(t)\,dt\ge\frac{K_1\tau\theta(\tau)}{8K_0}.$$

On the other hand, for other values of τ in $[\delta,\omega/2]$, we use (3.2), (3.8) and (3.20) to see that

$$\begin{aligned}\tau(\theta(\tau)-1)&\geq\frac{K_1\tau}{2}\int_0^{1/\tau}ta(t)\,dt\geq\frac{K_1}{2K_0}\tau\theta(\tau)\\&\geq\frac{\delta K_1\tau}{48\varepsilon K_0}|\theta(-\varepsilon,\tau)-\theta(\tau)|\geq2\tau|\theta(-\varepsilon,\tau)-\theta(\tau)|,\end{aligned}$$

and it follows that for these values of τ

$$-I(-\varepsilon,\tau)=\tau\{[\theta(\tau)-1]+[\theta(-\varepsilon,\tau)-\theta(\tau)]\}\geq\frac{\tau[\theta(\tau)-1]}{2}\geq\frac{K_1\tau\theta(\tau)}{4K_0}.$$

This completes the proof of the first inequality in (3.21).

Finally, the second inequality in (3.21) follows by (3.14) for those values of τ for which $\phi(\tau)\geq\frac{1}{2}K_1|\tau-\omega|$. For other values of τ in this range, (3.3) gives

$$\tau|\theta(\tau)-1|\geq\frac{K_1|\omega-\tau|}{2},\tag{3.22}$$

and the proof of the second estimate in (3.21) for such τ now splits into two cases.

Case 1. If $\omega/2\leq\tau\leq\omega-\delta$, then we use (3.8), (3.22) and (3.20) to get

$$\begin{aligned}\theta(-\varepsilon,\tau)&\geq\left(1-\frac{24\varepsilon}{\tau}\right)\theta(\tau)\\&\geq1+\left[\frac{K_1}{2}(\omega-\tau)-24\varepsilon\right]\frac{1}{\tau}-\frac{12\varepsilon K_1}{\tau^2}(\omega-\tau)\geq1+\frac{K_1(\omega-\tau)}{4\tau}.\end{aligned}$$

Thus, for such τ we have

$$|D(-\varepsilon,\tau)|\geq\frac{K_1|\tau-\omega|}{4}.\tag{3.23}$$

Case 2. If $\omega+\delta\leq\tau\leq T$, then by (3.22)

$$\theta(\tau)\leq1-\frac{K_1}{2\tau}(\tau-\omega),$$

so by (3.8) and (3.20) we get

$$\begin{aligned}\theta(-\varepsilon,\tau)&\leq\left(1+\frac{24\varepsilon}{\tau}\right)\theta(\tau)\\&\leq1-\left[\frac{K_1(\tau-\omega)}{2}-24\varepsilon\right]\frac{1}{\tau}-\frac{12\varepsilon K_1}{\tau^2}(\tau-\omega)\\&\leq1-\frac{K_1(\tau-\omega)}{4\tau}.\end{aligned}$$

Thus, (3.23) also holds for these values of τ, and the proof of the second inequality in (3.21) is complete. □

We can now prove inequality (3.18). Partition $[\delta,\infty)$ into five subsets depending on the function a:

$$\begin{aligned}
&E_1=[\delta,\infty)\cap\left[0,\frac{\omega}{2}\right),\\
&E_2=[\delta,\omega-\delta)\cap\left[\frac{\omega}{2},\omega-\delta\right),\\
&E_3=[\delta,T)\cap[\omega-\delta,\omega+\delta),\\
&E_4=[\omega+\delta,T),\quad E_5=[T,\infty).
\end{aligned}$$

(Of course some of these sets may be empty for particular functions a.)

Using (3.19), the first estimate in (3.21), and $\theta(\tau)>1$ on E_1, we obtain

$$\int_{E_1}\left|\frac{D_\tau(-\varepsilon,\tau)}{D^2(-\varepsilon,\tau)}\right|d\tau\le K\int_{E_1}\frac{1+\theta(\tau)}{\tau^2\theta(\tau)}d\tau\le K\int_\delta^\infty\frac{d\tau}{\tau^2}=\frac{K}{\delta}. \tag{3.24}$$

On E_2, we use (3.19), the second part of (3.21), (3.1), (3.2), and a change of variables to get

$$\int_{E_2}\left|\frac{D_\tau(-\varepsilon,\tau)}{D^2(-\varepsilon,\tau)}\right|d\tau\le K\int_{\omega/2}^{\omega-\delta}\frac{1+\theta(\tau)}{(\omega-\tau)^2}d\tau\le K\left[1+\int_0^{2/\omega}ta(t)\,dt\right]\int_\delta^\infty\frac{dy}{y^2}.$$

Now by the monotonicity of a, (3.2) and the definition of ω, we obtain

$$\begin{aligned}
\int_0^{2/\omega}ta(t)\,dt&\le\int_0^{1/\omega}ta(t)\,dt+a\left(\frac{1}{\omega}\right)\int_{1/\omega}^{2/\omega}t\,dt\\
&\le4\int_0^{1/\omega}ta(t)\,dt\le4K_0\theta(\omega)=4K_0,
\end{aligned}$$

and combining this with the previous inequality, we find that

$$\int_{E_2}\left|\frac{D_\tau(-\varepsilon,\tau)}{D^2(-\varepsilon,\tau)}\right|d\tau\le\frac{K}{\delta}. \tag{3.25}$$

For E_3, we use (3.19), the third part of (3.21) and (2.6) (recall that $\theta(\tau)\ge\frac{1}{4}$ on E_3) to obtain

$$\begin{aligned}
\int_{E_3}\left|\frac{D_\tau(-\varepsilon,\tau)}{D^2(-\varepsilon,\tau)}\right|d\tau&\le K\int_{E_3}\frac{1+\theta(\tau)}{\phi^2(\tau)}d\tau\le K\int_{E_3}\frac{1+B(\delta)\phi(\tau)}{\phi^2(\tau)}d\tau\\
&\le K\int_{E_3}(16+4)B^2(\delta)\,d\tau\le K\delta B^2(\delta).
\end{aligned} \tag{3.26}$$

Next, on E_4, use (3.19), (3.1) and the second inequality in (3.21) to see that

$$\int_{E_4}\left|\frac{D_\tau(-\varepsilon,\tau)}{D^2(-\varepsilon,\tau)}\right|d\tau\le K\int_{E_4}\frac{2}{(\tau-\omega)^2}d\tau\le2K\int_\delta^\infty\frac{dy}{y^2}\le\frac{K}{\delta}. \tag{3.27}$$

Finally, by using (3.19), (3.1) and the fourth inequality in (3.21), we get

$$\int_{E_5}\left|\frac{D_\tau(-\varepsilon,\tau)}{D^2(-\varepsilon,\tau)}\right|d\tau\le K\int_{E_5}\frac{2}{\tau^2}d\tau\le\frac{K}{\delta}. \tag{3.28}$$

Now, by combining (3.24)–(3.28) we obtain (3.18), and, as stated before inequality (3.18), the proof of Theorem 2.1 is complete. □

4. Proofs of Corollaries 2.2 and 2.3. Proof of Corollary 2.2. We use Theorem 2.1 to find fixed positive constants Q and ε ($\varepsilon \leq x_0$) so that the estimate (2.7) holds for each $a \in \mathscr{A}$, and then we combine (2.7) and (2.8) to write

$$\sup_{a\in\mathscr{A}} |u(t;\, a)| \leq \frac{1}{\pi^2}\int_{-\varepsilon}^{0} \frac{e^{\sigma t}}{\beta(-\sigma)}\,d\sigma + Qe^{-\varepsilon t} \qquad (t \geq 0).$$

Now we multiply this inequality by $\rho(t)$, integrate, interchange the order of integration, and use (2.8) to obtain

$$\int_0^\infty \rho(t) \sup_{a\in\mathscr{A}} |u(t;\, a)|\,dt \leq \frac{1}{\pi^2}\int_{-\varepsilon}^{0} \frac{\hat{\rho}(-\sigma)}{\beta(-\sigma)}\,d\sigma + Q\hat{\rho}(\varepsilon)$$
$$\leq \frac{1}{\pi^2}\int_0^{x_0} \frac{\hat{\rho}(x)}{\beta(x)}\,dx + Q\hat{\rho}(\varepsilon) < \infty.$$

This completes the proof of Corollary 2.2. □

Our proof of Corollary 2.3 requires the following estimate for the functions $\tilde{\phi}_j(\sigma)$.

LEMMA 4.1. *Let $a(t)$ satisfy* (2.1) *through* (2.3), *and assume that* (2.11), (2.12) *and* (2.13) *hold with $\alpha_k = \alpha$. Then there exists $\varepsilon > 0$ so that the function $\tilde{\phi}$ satisfies*

$$\tilde{\phi}(\sigma) \leq \frac{\alpha(0^+)}{2\sigma} \quad \textit{for a.e. } \sigma \in (-\varepsilon, 0). \tag{4.1}$$

Proof. Since

$$\phi(\sigma,\tau) = \frac{\alpha(0^+)\sigma}{\sigma^2+\tau^2} + \int_{(0,\infty)} \frac{(x+\sigma)}{(x+\sigma)^2+\tau^2}\,d\alpha(x) \qquad (\tau > 0),$$

it suffices to find $\varepsilon > 0$ so that

$$\tilde{\phi}_1(\sigma) \equiv \limsup_{\tau\to 0^+} \int_{(0,\infty)} \frac{(x+\sigma)}{(x+\sigma)^2+\tau^2}\,d\alpha(x)$$

satisfies

$$\tilde{\phi}_1(\sigma) \leq -\frac{\alpha(0^+)}{2\sigma} \quad \text{for } -\varepsilon \leq \sigma < 0. \tag{4.2}$$

To deduce (4.2), let $\sigma \in (-\frac{1}{2}, 0)$ satisfy $\sqrt{-2\sigma} < x_0$ and write

$$\int_{(0,\infty)} \frac{(x+\sigma)}{(x+\sigma)^2+\tau^2}\,d\alpha(x) \tag{4.3}$$
$$= \left\{\int_0^{-2\sigma} + \int_{-2\sigma}^{\sqrt{-2\sigma}} + \int_{\sqrt{-2\sigma}}^{x_0}\right\} \frac{(x+\sigma)\alpha'(x)}{(x+\sigma)^2+\tau^2}\,dx + \int_{x_0}^{\infty} \frac{(x+\sigma)}{(x+\sigma)^2+\tau^2}\,d\alpha(x)$$
$$\equiv I_1 + I_2 + I_3 + I_4.$$

To estimate $I_1(\sigma,\tau)$ for such σ, we use a change of variables and (2.13) with $\xi=-\sigma$, $\delta=-\sigma-x$, to obtain

$$
\begin{aligned}
(4.4)\qquad I_1(\sigma,\tau)&=\int_0^{-\sigma}\frac{(x+\sigma)}{(x+\sigma)^2+\tau^2}\left[\alpha'(x)-\alpha'(-2\sigma-x)\right]dx\\
&\le\frac{K}{-\sigma}\int_0^{-\sigma}\frac{(-x-\sigma)^{1+\gamma}}{(x+\sigma)^2+\tau^2}dx\\
&\le\frac{K}{-\sigma}\int_0^{-\sigma}(-x-\sigma)^{\gamma-1}dx=\frac{K}{\gamma}(-\sigma)^{\gamma-1}\qquad(\tau>0).
\end{aligned}
$$

We use elementary inequalities and the fact that $\alpha'\in L^1(0,x_0)$ to estimate I_2 and I_3, respectively, by

$$
\begin{aligned}
(4.5)\qquad -\sigma I_2(\sigma,\tau)&\le\int_{-2\sigma}^{\sqrt{-2\sigma}}\frac{(x+\sigma)^2}{(x+\sigma)^2+\tau^2}\alpha'(x)\,dx\\
&\le\int_{-2\sigma}^{\sqrt{-2\sigma}}\alpha'(x)\,dx=o(1)\qquad(\sigma\to0^-),
\end{aligned}
$$

$$
\begin{aligned}
(4.6)\qquad -\sigma I_3(\sigma,\tau)&\le-\sigma\int_{\sqrt{-2\sigma}}^{x_0}\frac{\alpha'(x)}{x+\sigma}dx\\
&\le\frac{-\sigma}{\sqrt{-2\sigma}+\sigma}\int_0^{x_0}\alpha'(x)\,dx=O\left(\sqrt{-\sigma}\right)\qquad(\sigma\to0^-).
\end{aligned}
$$

Finally, I_4 is estimated by

$$
(4.7)\qquad I_4(\sigma,\tau)\le\int_{x_0}^{\infty}\frac{d\alpha(x)}{(x+\sigma)}\le2\int_{x_0}^{\infty}\frac{d\alpha(x)}{x}<\infty\qquad(\tau>0).
$$

Thus, by (4.3)–(4.7) we see that (4.2) holds whenever $\varepsilon>0$ is chosen sufficiently small, and the proof of Lemma 4.1 is complete. □

Proof of Corollary 2.3. By Theorem 2.1 and Lemma 4.1, we can find fixed positive constants Q and ε $(\varepsilon\le x_0)$ so that (2.7) holds whenever $a\in\mathscr{A}$ and so that each $\tilde{\phi}_k$ corresponding to a_k satisfies (4.1) $(1\le k\le N)$. Since $\lambda_{jk}\ge1$ whenever $j\in J$, $1\le k\le N$, elementary estimates and (4.1) yield

$$
\begin{aligned}
&\sum\lambda_{jk}\alpha_k'(-\sigma)\left\{\left[\sigma+\sum\lambda_{jk}\tilde{\phi}_k(\sigma)\right]^2+\left[\pi\sum\lambda_{jk}\alpha_k'(-\sigma)\right]^2\right\}^{-1}\\
&\qquad\le\frac{\Sigma\lambda_{jk}\alpha_k'(-\sigma)}{\left(\Sigma\lambda_{jk}\tilde{\phi}_k(\sigma)\right)^2}\\
&\qquad\le\sum\frac{1}{\lambda_{jk}}\frac{\alpha_k'(-\sigma)}{\tilde{\phi}_k^2(\sigma)}\le\frac{4\sigma^2\Sigma\alpha_k'(-\sigma)}{\alpha_k^2(0^+)}\qquad(-\varepsilon<\sigma<0,\ \ j\in J).
\end{aligned}
$$

Here and below the sums are taken from $k=1$ to $k=N$. Thus, by (2.7), we obtain

$$
(4.8)\qquad \sup_{a\in\mathscr{A}}|u(t;\,a)|\le4\int_0^{\varepsilon}e^{-xt}x^2\sum\frac{\alpha_k'(x)}{\alpha_k^2(0^+)}dx+Qe^{-\varepsilon t}\qquad(t\ge0).
$$

To prove (1.3) with $\rho(t)=1+t$, integrate by parts and use the monotonicity of α_k to get the estimate

$$\begin{aligned}\int_0^\varepsilon e^{-xt}x^2\alpha_k'(x)\,dx&=\varepsilon^2\alpha_k(\varepsilon)e^{-\varepsilon t}+\int_0^\varepsilon x(xt-2)e^{-xt}\alpha_k(x)\,dx\\ &\le\varepsilon^2\alpha_k(\varepsilon)e^{-\varepsilon t}+\alpha_k(\varepsilon)t\int_{2/t}^\varepsilon x^2e^{-xt}\,dx\\ &\le\varepsilon^2\alpha_k(\varepsilon)e^{-\varepsilon t}+\alpha_k(\varepsilon)\int_2^\infty\frac{u^2e^{-u}\,du}{t^2}=O\left(\frac{1}{t^2}\right)\qquad(t\to\infty)\end{aligned}$$

which, together with (4.8), yields (1.3) ($\rho(t)=1+t$).

Finally, to establish (1.4) with $\rho(t)=1+t$, note that for this ρ, $\hat{\rho}(x)=(1+x)/x^2$ ($x>0$), so we can multiply (4.8) by $(1+t)$, integrate and interchange order of integrations to get

$$\int_0^\infty(1+t)\sup_{a\in\mathfrak{A}}|u(t;\,a)|\,dt\le4\int_0^\varepsilon(1+x)\sum\frac{\alpha_k'(x)}{\alpha_k^2(0^+)}\,dx+Q\int_0^\infty(1+t)e^{-\varepsilon t}\,dt<\infty$$

since $\alpha_k'\in L^1(0,x_0)$ for $1\le k\le N$. This completes the proof of Corollary 2.3. □

We remark that the proof of Corollary 2.3 shows that we can obtain this result for general weights $\rho(t)$ provided we assume, in addition to the hypotheses of Corollary 2.3, that

$$\rho(t)\int_0^{x_0}e^{-xt}x^2\alpha_k'(x)\,dx=o(1)\qquad(t\to\infty,\quad 1\le k\le N),$$

and

$$\int_0^{x_0}\hat{\rho}(x)x^2\alpha_k'(x)\,dx<\infty\qquad(1\le k\le N)$$

hold.

REFERENCES

[1] R. W. Carr and K. B. Hannsgen, *A nonhomogeneous integrodifferential equation in Hilbert space*, this Journal, 10 (1979), pp. 961–984.

[2] ______, *Resolvent formulas for a Volterra equation in Hilbert space*, this Journal, 13 (1982), pp. 459–483.

[3] H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids*, 2nd ed., Clarendon Press, Oxford, 1959.

[4] K. B. Hannsgen, *A Volterra equation with parameter*, this Journal, 4 (1973), pp. 22–30.

[5] ______, *The resolvent kernel of an integrodifferential equation in Hilbert space*, this Journal, 7 (1976), pp. 481–490.

[6] ______, *A linear integrodifferential equation for viscoelastic rods and plates*, Quart. Appl. Math., 41 (1983), pp. 75–84.

[7] K. B. Hannsgen and R. L. Wheeler, *Complete monotonicity and resolvents of Volterra integrodifferential equations*, this Journal, 13 (1982), pp. 962–969.

[8] ______, *A singular limit problem for an integrodifferential equation*, J. Integral Equations, to appear.

[9] G. S. Jordan and R. L. Wheeler, *Rates of decay of resolvents of Volterra equations with certain nonintegrable kernels*, J. Integral Equations, 2 (1980), pp. 103–110.

[10] J. J. Levin, *The asymptotic behavior of the solution of a Volterra equation*, Proc. Amer. Math. Soc., 14 (1963), pp. 534-541.

[11] R. C. MacCamy, *An integro-differential equation with application in heat flow*, Quart. Appl. Math., 35 (1977), pp. 1–19.

[12] D. F. Shea and S. Wainger, *Variants of the Wiener–Lévy theorem, with applications to stability problems for some Volterra integral equations*, Amer. J. Math., 97 (1975), pp. 312–343.

[13] D. V. Widder, *The Laplace Transform*, Princeton Univ. Press, Princeton, NJ, 1946.

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics
013

SERIES EXPANSIONS FOR RESOLVENTS OF VOLTERRA INTEGRODIFFERENTIAL EQUATIONS IN BANACH SPACE*

R. C. GRIMMER[†] AND F. KAPPEL[‡]

Abstract. This paper is concerned with the existence and properties of resolvent operators for linear Volterra integrodifferential equations in Banach space. Regularity of weak solutions given by the variation of parameters formula is also examined.

1. Introduction. In this paper, we consider the linear integrodifferential equation

$$\begin{aligned} (1.1) \qquad & x'(t)=Ax(t)+\int_0^t B(t-s)x(s)\,ds+f(t), \qquad t\geq 0, \\ & x(0)=x_0, \end{aligned}$$

in a Banach space X with norm $\|\cdot\|$. It is assumed throughout this paper that A generates an analytic semigroup $T(t)$ on X, that $B(t)x$ is Bochner integrable for each x in the domain of A and that f is continuous on $[0,\infty)$ into X.

Our concern here is to obtain a resolvent operator for (1.1) as defined in [6] and [8], which will then be used in a variation of parameters formula for (1.1). We shall then consider the problem of determining which x_0 and f in fact yield a solution of (1.1). We obtain the resolvent as an integral over an appropriate contour, and in this our work resembles earlier work by Da Prato and Iannelli [3], Friedman and Shinbrot [5] and Grimmer and Pritchard [8]. Our technique involves writing a series for what should be the Laplace transform of the resolvent operator. This allows us to consider equations not considered before and seems to yield additional flexibility in dealing with the problem studied in [8].

Other papers considering (1.1) include Grimmer [7], Grimmer and Schappacher [9], and Miller [11], while Carr and Hannsgen [1], [2], Sinestrari [12] and Webb [14] consider other equations in which the generator of an analytic semigroup plays a central role. Also see [7] for further references.

2. Preliminaries. Throughout this paper, it shall be assumed that A generates an analytic semigroup on X. Thus, the domain of A, $D(A)$, together with the graph norm is a Banach space which we shall denote Y. Further, for each $\delta>0$, the operator $(-A)^\delta$ is defined and $D((-A)^\delta)$ together with its graph norm yields a Banach space Y^δ. The norm on the space Y^δ we denote by $\|\cdot\|_\delta$. The space of bounded linear operators from a Banach space V to another Banach space W is denoted by $\mathfrak{B}(V,W)$ and the norm of $\mathfrak{B}(Y^\alpha,Y^\beta)$ is given as $\|\cdot\|_{\alpha,\beta}$. Note that $Y^0\equiv X$ and $Y^1\equiv Y$. As a convention, we will denote $\|\cdot\|_{0,0}$ by $\|\cdot\|$ and $\mathfrak{B}(V,V)$ as $\mathfrak{B}(V)$. (For material concerning semigroups and fractional powers, the reader is referred to Friedman [4], Hille and Phillips [10], and Tanabe [13].) We shall be using Laplace transforms frequently and the Laplace transform of a function $h(t)$ will be denoted $\hat{h}(\lambda)$.

*Received by the editors January 1, 1983, and in revised form March 9, 1983.

[†] Department of Mathematics, Southern Illinois University, Carbondale, Illinois 62901. The research of this author was partially supported by the National Science Foundation under grant MCS-8210322.

[‡] Institut für Mathematik, Karl-Franzens Universität Graz, Elisabethstrasse 16, A-8010 Graz, Austria. The research of this author was supported by the Fonds zur Förderung der Wissenschaftlichen Forschung under project no. P4534.

We can now state our basic hypotheses which will hold throughout this paper.

(H1). A generates an analytic semigroup and satisfies the estimate $\|(\lambda I-A)^{-1}\|\leq M/|\lambda|$, $\operatorname{Re}\lambda>0$, $M\geq 1$.

(H2). $\|B(t)\|_{1,0}\leq b(t)$ for some $b\in L^1_{\mathrm{loc}}(0,\infty)$ and $B(t)x$ is strongly measurable for each $x\in Y$.

(H3). For λ with $\operatorname{Re}\lambda>0$, $\hat{B}(\lambda)$ exists as an element of $\mathcal{B}(Y,X)$ and $\|\hat{B}(\lambda)\|_{1,0}\leq N/|\lambda|^{\beta}$ for some $\beta>0$ and $N\geq 1$.

The following lemma contains a number of estimates which shall be useful in the remainder of the paper.

LEMMA 2.1. *Assume* (H1)–(H3) *are valid. Then*

(a) $\|(\lambda I-A)^{-1}\|_{0,1}\leq 2M+1$, $\operatorname{Re}\lambda>1$.

(b) $\|(\lambda I-A)^{-1}\hat{B}(\lambda)\|_{1,1}\leq(2M+1)N/|\lambda|^{\beta}$, $\operatorname{Re}\lambda>1$.

(c) $\|\hat{B}(\lambda)(\lambda I-A)^{-1}\|_{0,0}\leq(2M+1)N/|\lambda|^{\beta}$, $\operatorname{Re}\lambda>1$.

(d) $\left\|\sum_{i=1}^{\infty}\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^{j}(\lambda I-A)^{-1}\right\|_{k,k}\leq 2MN(2M+1)/|\lambda|^{1+\beta}$ *if* $\operatorname{Re}\lambda\geq[2N(2M+1)]^{1/\beta}$, $k=0,1$.

(e) $\|(\lambda I-A)^{-1}\|_{\delta,1}\leq C(\delta)M/|\lambda|^{\delta}$, $0\leq\delta\leq 1$,

where $C(\delta)$ is a constant depending on δ.

Proof. (a) Let $x\in X$. Then we see

$$\begin{aligned}\|(\lambda I-A)^{-1}x\|_1&=\|A(\lambda I-A)^{-1}x\|+\|(\lambda I-A)^{-1}x\|\\&=\|(\lambda(\lambda I-A)^{-1}-I)x\|+\|(\lambda I-A)^{-1}x\|\\&\leq(M+1)\|x\|+M\|x\|/|\lambda|\\&\leq(2M+1)\|x\|\qquad(\operatorname{Re}\lambda\geq 1).\end{aligned}$$

(b) and (c) now follow from (a) and (H3). For (d), let $k=0$ and $(\operatorname{Re}\lambda)^{\beta}\geq 2(2M+1)N$. Then

$$\begin{aligned}&\left\|\sum_{j=1}^{\infty}\left((\lambda I-A)^{-1}\hat{B}(\lambda)\right)^{j}(\lambda I-A)^{-1}\right\|\\&\quad\leq\|(\lambda I-A)^{-1}\hat{B}(\lambda)\|_{1,0}\sum_{j=0}^{\infty}\|(\lambda I-A)^{-1}\hat{B}(\lambda)\|_{1,1}^{j}\|(\lambda I-A)^{-1}\|_{0,1}\\&\quad\leq\|(\lambda I-A)^{-1}\|\|\hat{B}(\lambda)\|_{1,0}\sum_{j=0}^{\infty}\left[(2M+1)N/|\lambda|^{\beta}\right]^{j}(2M+1)\\&\quad\leq 2MN(2M+1)/|\lambda|^{1+\beta}.\end{aligned}$$

If $k=1$, the argument is similar. For (e) the reader is referred to Tanabe [13, p. 39].

The concept of a solution of (1.1) seems to vary somewhat in the literature. Our definition of solution is patterned after the concept of solution in semigroup theory. (For work unifying many concepts, see [9].)

DEFINITION 2.2. A solution of (1.1) is a function $x\in C([0,\infty),Y)\cap C^1([0,\infty),X)$ with $x(0)=x_0$ which satisfies (1.1) for $t\geq 0$.

Later, we shall wish to relax this condition a bit at $t=0$. This will be the case in particular if $x_0 \in Y^{\delta}$, $0 \leq \delta < 1$. However, this concept of solution leads us to the definition of a resolvent operator for (1.1).

DEFINITION 2.3. $R(t)$ is said to be a resolvent operator for (1.1) if $R(t) \in \mathcal{B}(X)$, $0 \leq t < \infty$ and if it satisfies:

(a) $R(t)$ is strongly continuous for $t \geq 0$ with $R(0)=I$ and $\|R(t)\| \leq Me^{\beta t}$ for some constants β and $M \geq 1$.

(b) $R(t) \in \mathcal{B}(Y)$ and $R(t)$ is strongly continuous, $t \geq 0$, on Y.

(c) For each $x \in Y$, $R(t)x$ is continuously differentiable, $t \geq 0$, with

$$R'(t)x = AR(t)x + \int_0^t B(t-u)R(u)x\,du$$

and

$$R'(t)x = R(t)Ax + \int_0^t R(t-u)B(u)x\,du.$$

3. Main theorems. We are now prepared for the main results which concern the existence of a resolvent operator for (1.1) and the existence of solutions of (1.1) for various initial conditions x_0 and functions $f(t)$.

THEOREM 3.1. *Assume* (H1) *through* (H3) *are valid. Then* (1.1) *has a resolvent operator.*

Proof. Define $R(0)=I$ and for $t>0$ define $R(t)$ by

$$\begin{aligned} R(t)x &= T(t)x + (2\pi i)^{-1}\int_{\gamma - i\infty}^{\gamma + i\infty} e^{\lambda t} \sum_{j=1}^{\infty} \left((\lambda I - A)^{-1}\hat{B}(\lambda)\right)^j (\lambda I - A)^{-1} x\,d\lambda \\ &= T(t)x + R_1(t)x, \end{aligned}$$

where $T(t)$ is the semigroup generated by A, $\gamma^{\beta} > 2(2M+1)N$. First note that it follows from Lemma 2.1(d) that

$$\int_{\gamma - i\infty}^{\gamma + i\infty} e^{\lambda t} \sum_{j=1}^{\infty} \left((\lambda I - A)^{-1}\hat{B}(\lambda)\right)^j (\lambda I - A)^{-1} x\,d\lambda$$

converges in X for $x \in X$ and in Y for $x \in Y$. Thus $R(t) \in \mathcal{B}(X) \cap \mathcal{B}(Y)$ for $t \geq 0$. Also, from Lemma 2.1(d) we see that

$$\sum_{j=1}^{\infty} \left((\lambda I - A)^{-1}\hat{B}(\lambda)\right)^j (\lambda I - A)^{-1} \in H^1(\alpha, \mathcal{B}(X)) \cap H^1(\alpha, \mathcal{B}(Y)),$$

where $\alpha^{\beta} = 2N(2M+1)$ so that the definition is independent of $\gamma > \alpha$, (cf. Hille–Phillips [10, p. 230].) Further, $R(t)$ is continuous in $\mathcal{B}(X) \cap \mathcal{B}(Y)$ on $(0, \infty)$. This, of course, implies the strong continuity of $R(t)x$ on $(0, \infty)$ in X if $x \in X$ and in Y if $x \in Y$. To show strong continuity at $t = 0^+$, we need only show $R_1(t)x \to 0$ as $t \to 0^+$ in X if $x \in X$ and in Y if $x \in Y$. However, it follows from Lemma 2.1(d) that the integrand defining $R_1(t)x$ is bounded by $(2\pi)^{-1} 2MN(2M+1)\|x\|_k / |\lambda|^{1+\beta}$, $k=0$, 1. It follows now from standard arguments that $\|R_1(t)x\|_k \to 0$ as $t \to 0^+$, $k=0$, 1, so that $R(t)$ is strongly continuous at $t=0^+$.

Now, for $x \in Y$, it follows from Lemma 2.1(d) that

$$R'(t)x = T'(t)x + (2\pi i)^{-1}\int_{\gamma - i\infty}^{\gamma + i\infty} \lambda e^{\lambda t} \sum_{j=1}^{\infty} \left[(\lambda I - A)^{-1}\hat{B}(\lambda)\right]^j (\lambda I - A)^{-1} x\,d\lambda.$$

Also, for $x \in Y$,

$$\begin{aligned}\sum_{j=1}^{\infty}\lambda\big[(\lambda I-A)^{-1}\hat{B}(\lambda)\big]^{j}(\lambda I-A)^{-1}x\\ &=\sum_{j=1}^{\infty}\big[(\lambda I-A)^{-1}\hat{B}(\lambda)\big]^{j}\big[I+(\lambda I-A)^{-1}A\big]x\\ &=\sum_{j=0}^{\infty}\big[(\lambda I-A)^{-1}\hat{B}(\lambda)\big]^{j}(\lambda I-A)^{-1}\hat{B}(\lambda)x\\ &\qquad+\sum_{j=1}^{\infty}\big[(\lambda I-A)^{-1}\hat{B}(\lambda)\big]^{j}(\lambda I-A)^{-1}Ax.\end{aligned}$$

Applying Lemma 2.1(d) this implies

$$R'(t)x=R(t)Ax+(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda t}\hat{R}(\lambda)\hat{B}(\lambda)x\,d\lambda$$

as $\hat{R}(\lambda)=\Sigma_{j=0}^{\infty}[(\lambda I-A)^{-1}\hat{B}(\lambda)]^{j}(\lambda I-A)^{-1}$, [10, p. 230]. In a similar manner, it is seen that

$$R'(t)x=AR(t)x+(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda t}\hat{B}(\lambda)\hat{R}(\lambda)x\,d\lambda.$$

Now $R(t)$ is strongly continuous for $t\geq 0$, so

$$\int_0^t R(t-s)B(s)x\,ds \quad\text{and}\quad \int_0^t B(t-s)R(s)x\,ds$$

both exist as continuous functions in X. Arguing as in [8] using Hille–Phillips [10, p. 219] yields

$$\int_0^t R(t-s)B(s)x\,ds=(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda t}\hat{R}(\lambda)\hat{B}(\lambda)x\,d\lambda,$$

so that

$$R'(t)x=R(t)Ax+\int_0^t R(t-s)B(s)x\,ds,\qquad t>0.$$

Similarly,

$$R'(t)x=AR(t)x+\int_0^t B(t-s)R(s)x\,ds,\qquad t>0.$$

To determine the right derivative of $R(t)x$ at $t=0$, write

$$R'(t)x=R(t)Ax+(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda t}\sum_{j=0}^{\infty}\big[(\lambda I-A)^{-1}\hat{B}(\lambda)\big]^{j}(\lambda I-A)^{-1}\hat{B}(\lambda)x\,d\lambda.$$

Letting $\Gamma_n=\{\lambda:\lambda=\gamma+i\delta,\ -n\leq\delta\leq n\}$ and $C_n=\{\lambda:\lambda=\gamma+ne^{i\theta},\ -\pi/2\leq\theta\leq\pi/2\}$, one estimates on C_n,

$$\begin{aligned}&\left\|\int_{C_n}\sum_{j=0}^{\infty}\big[(\lambda I-A)^{-1}\hat{B}(\lambda)\big]^{j}(\lambda I-A)^{-1}\hat{B}(\lambda)x\,d\lambda\right\|_0\\ &\leq\int_{C_n}\sum_{j=0}^{\infty}\|(\lambda I-A)^{-1}\|_{0,0}\|\hat{B}(\lambda)(\lambda I-A)^{-1}\|_{0,0}^{j}\|\hat{B}(\lambda)x\|_0 d\lambda\\ &\leq\int_{C_n}2MN/|\lambda|^{1+\beta}d\lambda\|x\|_1,\end{aligned}$$

using Lemma 2.1(c) and (H3). Letting $n\to\infty$, we see that this last integral tends to zero and it follows from Cauchy's theorem that $\lim_{t\to 0^+} R'(t)x=R(0)Ax=Ax$ so that the right derivative of $R(t)x$ at $t=0$ is Ax. It is now clear that $R(t)$ is the desired resolvent operator.

Although the resolvent operator obtained in Theorem 3.1 is not an analytic resolvent operator as defined in [8], it has some of the properties of an analytic resolvent operator. In particular, we wish to examine if $R(t): X\to Y$ or $R(t): Y^\delta \to Y$ for $t>0$ and if $R(t)x$ satisfies (1,1) for $t>0$ in case $f\equiv 0$.

THEOREM 3.2. *Let* $0\le\delta<1$. *If* $\beta+\delta>1$, *then* $R(t): Y^\delta\to Y$, $t>0$, *and* $R(t)x$ *is differentiable for* $t>0$ *if* $x\in Y^\delta$. *If, in addition,* $b(t)$ *is bounded on intervals of the form* $0<T_1\le t\le T_2<\infty$ *and* $\delta>0$, *then*

$$R'(t)x=AR(t)x+\int_0^t B(t-s)R(s)x\,ds, \qquad t>0, \quad for\, x\in Y^\delta. \tag{3.1}$$

Proof. As $R(t)=T(t)+R_1(t)$ and $T(t): X\to Y$ for $t>0$, we consider the behavior of $R_1(t)$. If $x\in Y^\delta$,

$$\left\| \sum_{j=0}^{\infty} A(\lambda I-A)^{-1}\hat{B}(\lambda)\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^j(\lambda I-A)^{-1}x \right\|$$

$$\le \left\|A(\lambda I-A)^{-1}\hat{B}(\lambda)\right\|_{1,0} \sum_{j=0}^{\infty} \left\|(\lambda I-A)^{-1}\hat{B}(\lambda)\right\|_{1,1}^{j}\left\|(\lambda I-A)^{-1}\right\|_{\delta,1}\|x\|_\delta$$

$$\le 2(M+1)NC(\delta)M\|x\|_\delta/|\lambda|^{\beta+\delta}.$$

Thus, $R_1(t)x\in Y$ for $t>0$. Also, $\|R_1(t)x\|_1\le K\|x\|_\delta$, $0<t\le T<\infty$, for any $T>0$ where $K=K(T)$ and $R_1(t)$ is uniformly continuous in $\mathfrak{B}(Y^\delta, Y)$ for $t\ge 0$. As in the proof of the previous theorem, for $x\in Y^\delta$,

$$\sum_{j=1}^{\infty} \lambda\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^j(\lambda I-A)^{-1}x$$

$$=\sum_{j=1}^{\infty} A(\lambda I-A)^{-1}\left[\hat{B}(\lambda)(\lambda I-A)^{-1}\right]^{j-1}\hat{B}(\lambda)(\lambda I-A)^{-1}x$$

$$+\sum_{j=0}^{\infty}\left[\hat{B}(\lambda)(\lambda I-A)^{-1}\right]^j\hat{B}(\lambda)(\lambda I-A)^{-1}x.$$

Estimating the norm of this sum in X using Lemma 2.1, one sees that it is bounded above by a constant times $1/|\lambda|^{\beta+\delta}$. Hence, for $t>0$ and $x\in Y^\delta$,

$$R'(t)x=T'(x)+(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty} e^{\lambda t}A\sum_{j=1}^{\infty}\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^j(\lambda I-A)^{-1}x\,d\lambda$$

$$+(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty} e^{\lambda t}\hat{B}(\lambda)\sum_{j=0}^{\infty}\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^j(\lambda I-A)^{-1}x\,d\lambda.$$

That is,

$$R'(t)x=AR(t)x+(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty} e^{\lambda t}\hat{B}(\lambda)\hat{R}(\lambda)x\,d\lambda. \tag{3.2}$$

Our only remaining problem is to determine that the last term in (3.2) is in fact $\int_0^t B(t-s)R(s)x\,ds$. As $R_1(s)x$ is bounded and continuous in Y for $x\in Y^\delta$, $\int_0^t B(t-s)R_1(s)x\,ds$ is continuous for $t\geq 0$ in X and

$$\int_0^t B(t-s)R_1(s)x\,ds=(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda t}\hat{B}(\lambda)\hat{R}_1(\lambda)x\,d\lambda,$$

as (H3) and Lemma 2.1 yield

$$\left\|\hat{B}(\lambda)\sum_{j=1}^{\infty}\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^j(\lambda I-A)^{-1}x\right\|\leq K_1\|x\|_\delta/|\lambda|^{\beta+\delta}$$

for some constant K_1. Thus, for $x\in Y^\delta$ and $t>0$, as $\hat{R}(\lambda)=\hat{T}(\lambda)+\hat{R}_1(\lambda)$,

$$R'(t)x=AR(t)x+\int_0^t B(t-s)R_1(s)x\,ds+(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda t}\hat{B}(\lambda)\hat{T}(\lambda)x\,d\lambda.$$

Now, if $x\in Y^\delta$, $\delta>0$, then $\|T(t)x\|_1\leq Lt^{\delta-1}\|x\|_\delta$, L constant, so if $b(t)$ is bounded on intervals of the form $0<T_1\leq t\leq T_2$, then $\int_0^t B(t-s)T(s)x\,ds$ exists and is continuous for $t>0$. Arguing as in [8] this yields, for $t>0$,

$$\int_0^t B(t-s)T(s)x\,ds=(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda t}\hat{B}(\lambda)\hat{T}(\lambda)x\,d\lambda, \tag{3.3}$$

and we have

$$R'(t)x=AR(t)x+\int_0^t B(t-s)R(s)x\,ds,\qquad t>0$$

as desired.

COROLLARY 3.3. *If $\beta>1$, then $R(t):X\to Y$ for $t>0$ and $R(t)x$ is differentiable for $t>0$.*

COROLLARY 3.4. *If $\beta>1$ and $B(t):Y^\mu\to X$, $\mu<1$, with $\|B(t)\|_{\mu,0}\leq b_\mu(t)$, where $b_\mu(t)$ is locally integrable and bounded on t intervals of the form $0<T_1\leq t\leq T_2<\infty$, then, for $x\in X$, $R(t)x\in Y$, $t>0$, and*

$$R'(t)x=AR(t)x+\int_0^t B(t-s)R(s)x\,ds.$$

Proof. The only part of the proof of Theorem 3.2 that required $\delta>0$ was (3.3). However, (3.3) will be valid in this case.

Once a resolvent operator is known to exist then the solution of (1.1), if it exists, is given by

$$x(t)=R(t)x_0+\int_0^t R(t-s)f(s)\,ds \tag{3.4}$$

[6]. As in semigroup theory, the problem is to determine when (3.4) actually yields a solution of (1.1).

THEOREM 3.5. *Suppose $\delta>0$ and $\beta+\delta>1$. If $x_0\in Y^\delta$ and $f\in C([0,\infty),Y^\delta)$, then $x(t)$, given by* (3.4), *satisfies* (1.1) *for $t>0$ with $x\in C([0,\infty),X)\cap C((0,\infty),Y)$.*

Proof. It follows from Theorem 3.2 that we need only consider $\int_0^t R(t-s)f(s)\,ds$. Now, $R(t)=T(t)+R_1(t)$ and it is known that $\int_0^t T(t-s)f(s)\,ds\in C([0,\infty),X)\cap C((0,\infty),Y)$ and, for $t>0$, $u(t)=\int_0^t T(t-s)f(s)\,ds$ satisfies $u'(t)=Au(t)+f(t)$. Also, for $x\in Y^{\delta}$, $\|R_1(t)x\|_1$ is bounded on intervals of the form $0\leq t\leq T<\infty$ so that $w(t)=\int_0^t R_1(t-s)f(s)\,ds$ is in $C([0,\infty),X)\cap C((0,\infty),Y)$. Now for $t>0$,

$$w'(t)=\frac{d}{dt}\left[\int_0^t(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda(t-s)}\sum_{j=1}^{\infty}\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^j(\lambda I-A)^{-1}f(s)\,d\lambda\,ds\right]$$

$$=\int_0^t(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}\lambda e^{\lambda(t-s)}\sum_{j=1}^{\infty}\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^j(\lambda I-A)^{-1}f(s)\,d\lambda\,ds$$

$$+(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}\sum_{j=1}^{\infty}\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^j(\lambda I-A)^{-1}f(t)\,d\lambda.$$

However, it can be shown that the last integral is zero as in the proof of Theorem 3.1. Hence,

$$w'(t)=\int_0^t(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda(t-s)}A\sum_{j=1}^{\infty}\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^j(\lambda I-A)^{-1}f(s)\,d\lambda\,ds$$

$$+\int_0^t(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda(t-s)}\hat{B}(\lambda)\sum_{j=0}^{\infty}\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^j(\lambda I-A)^{-1}f(s)\,d\lambda\,ds$$

$$=Aw(t)+\int_0^t(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda(t-s)}\hat{B}(\lambda)\hat{R}(\lambda)f(s)\,d\lambda\,ds$$

$$=Aw(t)+\int_0^t\int_0^{t-s}B(t-s-u)R(u)f(s)\,du\,ds$$

$$=Aw(t)+\int_0^t B(t-u)\int_0^u R(u-s)f(s)\,ds\,du.$$

Hence, for $t>0$, if $v(t)=\int_0^t R(t-s)f(s)\,ds=u(t)+w(t)$,

$$v'(t)=Av(t)+\int_0^t B(t-s)v(s)\,ds+f(t).$$

We now wish to consider the initial value problem (1.1) when f is Hölder continuous, $\|f(t)-f(s)\|\leq K|t-s|^{\alpha}$, $0<\alpha\leq 1$.

THEOREM 3.6. *Suppose f is Hölder continuous and $\beta>1$. Then $\int_0^t R(t-s)f(s)\,ds$ is a solution of* (1.1).

Proof. From the proof of Theorem 3.2 we see that $\|R_1(t)\|_{0,1}$ is bounded on bounded t intervals $(0,T)$. Thus, $\int_0^t B(t-s)R_1(s)x\,ds$ exists as a continuous function, $t\geq 0$, in X and $\int_0^t B(t-s)R_1(s)x\,ds$ is given by

$$(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda t}\hat{B}(\lambda)\hat{R}_1(\lambda)x\,d\lambda$$

for each $x\in X$. Consider now

$$\int_0^t R(t-s)f(s)\,ds=\int_0^t T(t-s)f(s)\,ds+\int_0^t R_1(t-s)f(s)\,ds$$

$$=u(t)+w(t).$$

From Theorem 3.2, it follows that $w(t)\in Y$ for $t\geq 0$ and is continuous in Y, $t\geq 0$. Now,

$$\begin{aligned}
&u'(t)+w'(t)\\
&=Au(t)+f(t)+\int_0^t(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda(t-s)}A\sum_{j=1}^{\infty}\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^j(\lambda I-A)^{-1}f(s)\,d\lambda\,ds\\
&\quad+\int_0^t(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda(t-s)}\hat{B}(\lambda)\sum_{j=0}^{\infty}\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^j(\lambda I-A)^{-1}f(s)\,d\lambda\,ds\\
&=Au(t)+f(t)+Aw(t)+\int_0^t(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda(t-s)}\hat{B}(\lambda)\hat{R}(\lambda)f(s)\,d\lambda\,ds.
\end{aligned}$$

As we are only concerned with existence on $[0,T]$, T arbitrary, we may assume that f is bounded for $t\geq 0$ so that $\|\hat{f}(\lambda)\|$ is bounded for $\operatorname{Re}\lambda\geq\gamma$. Letting f_t be the function defined by $f_t(s)=f(t+s)$, we see that

$$\begin{aligned}
&\int_0^t(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda(t-s)}\hat{B}(\lambda)\hat{R}(\lambda)f(s)\,d\lambda\,ds\\
&\qquad=(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda t}\hat{B}(\lambda)\hat{R}(\lambda)\hat{f}(\lambda)\,d\lambda-(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}\hat{B}(\lambda)\hat{R}(\lambda)\hat{f}_t(\lambda)\,d\lambda.
\end{aligned}$$

The second integral is easily shown to be zero so,

$$=\int_0^t B(t-s)[R*f](s)\,ds.$$

Thus, if $v(t)=u(t)+w(t)$,

$$v'(t)=Av(t)+\int_0^t B(t-s)v(s)\,ds+f(t).$$

Frequently, $\hat{B}(\lambda)$ will have an analytic extension to a sector of the form $\Lambda=\{\lambda\in C\colon |\arg\lambda|<\pi/2+\delta\}$ where $\delta>0$. This type of assumption was made in Grimmer and Pritchard [8] and leads to an analytic resolvent operator. In this case, it is reasonable to replace (H1) and (H3) by

(H1)′ A generates an analytic semigroup, $\|(\lambda I-A)^{-1}\|\leq M/|\lambda|$, $\lambda\in\Lambda$, $M\geq 1$.

(H3)′ $\|\hat{B}(\lambda)\|_{1,0}\leq N/|\lambda|^{\beta}$, $\lambda\in\Lambda$, $N\geq 1$ for some $\beta>0$.

An examination of the proof of Lemma 2.1 shows that all of the inequalities are valid for $\lambda\in\Lambda$, $|\lambda|^{\beta}>2(2M+1)N$.

It is clear that (H1)′ and (H3)′ imply (H1) and (H3) respectively, so that if (H1)′, (H2), and (H3)′ are valid, Theorem 3.1 implies the existence of a resolvent operator for (1.1). We wish now to take advantage of the ability to define $R(t)$ with an integral over a contour extending into the left half plane as in [8]. In particular, choose Γ to be a contour in Λ with $\Gamma=\Gamma_1\cup\Gamma_2\cup\Gamma_3$, where

$$\begin{aligned}
\Gamma_1&=\left\{\lambda\in C:\lambda=re^{i\phi},\ r\geq R_0,\ \frac{\pi}{2}<\phi<\pi\right\},\\
\Gamma_2&=\left\{\lambda\in C:\lambda=R_0e^{i\theta},\ -\phi\leq\theta\leq\phi\right\},\\
\Gamma_3&=\left\{\lambda\in C:\lambda=re^{-i\phi},\ r\geq R_0\right\},
\end{aligned}$$

where $R_0^\beta > 2(2M+1)N$. It is easy to see that

$$R(t)=T(t)+(2\pi i)^{-1}\int_\Gamma e^{\lambda t}\sum_{j=1}^{\infty}\left[(\lambda I-A)^{-1}\hat{B}(\lambda)\right]^j(\lambda I-A)^{-1}d\lambda. \tag{3.5}$$

THEOREM 3.7. *Assume that* (H1)′, (H2) *and* (H3)′ *are valid and that* $b(t)$ *is bounded on intervals of the form* $0<T_1\le t\le T_2<\infty$. *If* f *is Hölder continuous, then* $\int_0^t R(t-s)f(s)\,ds$ *is a solution of* (1.1).

Proof. First, note that $R_1(t)=(2\pi i)^{-1}\int_\Gamma e^{\lambda t}\hat{R}_1(\lambda)\,d\lambda$. Now, let $x\in X$ and consider $AR_1(t)x$, $t>0$. As in the proof of Theorem 3.2, one estimates to obtain $\|A\hat{R}_1(\lambda)x\|\le 2N(M+1)(2M+1)\|x\|/|\lambda|^\beta$. Now, let $\lambda t=\gamma$, $d\lambda=t^{-1}d\gamma$ and use Cauchy's theorem to get

$$AR_1(t)x=(2\pi i)^{-1}\int_\Gamma e^\gamma A\hat{R}_1(\gamma t^{-1})t^{-1}x\,d\gamma.$$

This leads to the estimate $\|AR_1(t)x\|\le Kt^{\beta-1}$ for some constant K which is independent of t. This implies $R_1(t)x$ is continuous in Y, $t>0$, and that $w(t)=\int_0^t R_1(t-s)f(s)\,ds$ exists and is continuous in Y, $t\ge 0$. Hence, $\int_0^t B(t-s)w(s)\,ds$ exists and is continuous. Also,

$$\int_0^t B(t-s)R_1(s)x\,ds=(2\pi i)^{-1}\int_{\gamma-i\infty}^{\gamma+i\infty}e^{\lambda t}\hat{B}(\lambda)\hat{R}_1(\lambda)x\,d\lambda$$

so that the proof may be completed in the same manner as the proof of the previous theorem.

4. Remarks. As an example of an equation which satisfies (H1) through (H3) but not (H3)′, we need only consider

$$u_t=\Delta u+\int_0^t a(t-s)\Delta u(s)\,ds+f(t),$$
$$u(0)=u_0,$$

where Δ is the Laplacian on a smooth body Ω with Dirichlet boundary conditions. Then Δ generates an analytic semigroup on $L^2(\Omega)$. Now let $a(t)$ be a periodic function with period T defined by $a(t)=t$, $0\le t\le T$. Then $\hat{a}(\lambda)=T/(1-e^{-T\lambda})\lambda$ so that $\hat{B}(\lambda)=\hat{a}(\lambda)\Delta$ cannot be extended to a sector. Taking $a_1=a*a$, one obtains a function with $\beta=2$.

We give one further example to indicate how more complicated equations may be handled. If A has a bounded inverse, A^{-1}, and A_1 is *any* closed operator with domain containing the domain of A, an application of the closed graph theorem yields that A_1A^{-1} is a bounded operator on X. Thus, for the equation

$$x'(t)=Ax(t)+\int_0^t a(t-s)A_1x(s)\,ds,$$

we see that on Y, $a(t)A_1=a(t)A_1A^{-1}A$ and we may choose $b(t)=|a(t)|\|A_1A^{-1}\|$ in (H2) while $\|\hat{a}(\lambda)A_1\|_{1,0}\le|\hat{a}(\lambda)|\|A_1A^{-1}\|$ in (H3).

Further applications, where $B(t)$ is the sum of terms $a_i(t)A_i$, are handled in much the same way.

Also, we note that if $\|\hat{B}(\lambda)\|_{1,0}\le N/|\mathrm{Im}\lambda|^\beta$ in (H3), the main results are also valid because we are integrating on a vertical line.

5. Acknowledgment. This work was begun while R. Grimmer was a visiting professor at the University of Graz. He gratefully acknowledges the support from the Bundesministerium für Wissenschaft und Forschung which made the visit possible.

REFERENCES

[1] R. W. CARR AND K. B. HANNSGEN, *A nonhomogeneous integrodifferential equation in Hilbert space*, this Journal, 10 (1979), pp. 961–984.

[2] ———, *Resolvent formulas for a Volterra equation in Hilbert space*, this Journal, 13 (1982), pp. 459–483.

[3] G. DAPRATO AND M. IANNELLI, *Linear integrodifferential equations in Banach spaces*, Rend. Sem. Mat. Padova, 62 (1980), pp. 207–219.

[4] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart, Winston, New York, 1969.

[5] A. FRIEDMAN AND M. SHINBROT, *Volterra integral equation in Banach space*, Trans. Amer. Math. Soc., 126 (1967), pp. 131–179.

[6] R. C. GRIMMER, *Resolvent operators for integral equations in a Banach space*, Trans. Amer. Math. Soc., 273 (1982), pp. 333–349.

[7] ———, *Resolvents for integral equations in abstract spaces*, in Evolution Equations and Applications, F. Kappel and W. Schappacher, eds., Research Notes in Mathematics, Pitman, London, 1982, pp. 101–120.

[8] R. C. GRIMMER AND A. J. PRITCHARD, *Analytic resolvent operators for integral equations in Banach space*, J. Differential Equations, to appear.

[9] R. GRIMMER AND W. SCHAPPACHER, *Weak solutions of integrodifferential equations and applications*, J. Integral Equations, to appear.

[10] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-Groups*, AMS Colloquium Publications 31, American Mathematical Society, Providence, RI, 1957.

[11] R. K. MILLER, *Volterra integral equations in a Banach space*, Funkcial. Ekvac., 18 (1975), pp. 163–193.

[12] E. SINESTRARI, *Time-dependent integrodifferential equations in Banach spaces*, International Conference on Nonlinear Phenomena in Mathematical Sciences, Arlington, Texas, 1980.

[13] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.

[14] G. WEBB, *Abstract Volterra integrodifferential equations and a class of reaction diffusion equations*, Proc. Helsinki Symposium on Integral Equations, Lecture Notes in Mathematics, 737, Springer-Verlag, Berlin, 1979.

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics
014

HILBERT AND FOURIER TRANSFORMS ON A SPHERE*

MICHAEL B. SAXE[†] AND THOMAS O. SHERMAN[‡]

Abstract. If ξ is a complex-valued eigenfunction of the Laplace–Beltrami operator Δ on a compact Riemannian symmetric space S, and if every positive integer power of ξ is also an eigenfunction, we investigate the extent to which the negative integer powers of ξ, properly regularized, fail to define eigendistributions of Δ. We give a simple formula for this failure in the general case and a deeper contrasting formula when S is a sphere. The source of the contrast lies in the theory of a Hilbert-like transform on the sphere.

Introduction. In the theory of harmonic analysis on a Riemannian symmetric space S, a key role is played by certain functions

$$\xi : S \to \mathbf{C},$$

all of whose positive integer powers are eigenfunctions of the Laplace–Beltrami operator Δ. Helgason ([4], [5], [6]) has a true Fourier theory on symmetric spaces of noncompact type which is based on the use of such functions in the transform kernel. An analogue of Helgason's theory has been proposed ([8], [9]) for symmetric spaces of compact type. In this analogue one must also be concerned with negative integer powers of ξ. Since ξ takes the value 0 in this case, the important integral

$$\int_S \xi^{-n} f, \qquad (f \in C^\infty(S)) \tag{0.1}$$

is singular and must be specially defined or "regularized." We use a particular regularization described in Lemma 1.12. On the open set where $\xi \neq 0$, ξ^{-n} is an eigenfunction of Δ:

$$(\Delta - \lambda_{-n})\xi^{-n} = 0. \tag{0.2}$$

For the Helgason–Fourier theory on compact S, it is important to know the extent to which (0.1) defines an eigendistribution of Δ. Thus we consider the deficit

$$\int_S \xi^{-n}(\Delta - \lambda_{-n}) f. \tag{0.3}$$

This defines a distribution on S which is supported on the set where $\xi = 0$. In §1 we give a simple formula for (0.3). The argument requires only that S be a compact Riemannian manifold and that the function $\xi : S \to \mathbf{C}$ satisfy certain conditions with respect to Δ (especially that ξ and ξ^2 are eigenfunctions of Δ). However, these seemingly simple hypotheses already impose considerable structure on S, and it may be that if a compact S supports a separating family of such functions, then it must be a homogeneous space of a compact lie group.

The idea behind the simple formula for (0.1) is to project the problem, via ξ, to the set $\xi(S)$, which turns out to be a disk around 0 in $\mathbf{C}$. Then (0.3) corresponds to a distribution supported at 0 in $\mathbf{C}$, and thus is of the form

$$\phi \to (\mathfrak{D}\phi)(0)$$

* Received by the editors January 8, 1982, and in revised form August 26, 1982.

[†] Missile Systems Division, Raytheon Company, Bedford, Massachusetts 01730.

[‡] Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

for some differential operator $\mathfrak{D}$. We give an explicit expression for $\mathfrak{D}$ in Theorem 1.1. This depends, of course, on the regularization of (0.1) which is given in Lemma 1.15.

In §2 and §3 we are concerned with the special case of the unit sphere S in Euclidean space. Section 2 discusses some Hilbert-like transforms on S. In §3 we apply the results of §2 to establish for (0.3) a second formula (Theorem 3.3) which is quite different from that of §1. In particular, its statement and proof depend strongly on the rich structure of the sphere.

As we show in §4, Theorem 3.3 is of considerable importance for the theory of the spherical Fourier transform in [8] and may be regarded as the main result of this paper. It was worked out in the first author's thesis without benefit of the material in §1 or §2, which came later. Indeed the results of §2 were formed by working backward from Theorem 3.3 toward Theorem 1.1, in an attempt to understand the difference between the two very different formulae for (0.3). We have confined the difference to a result (Theorem 2.1) on the Hilbert-like transform.

Notation. We will write $A := B$ to mean that A is assigned the value B or that A is defined to be B.

If $f: M \to N$, we denote by $f^{\#}$ the set-theoretic inverse set function: if $N' \subset N$ then

$$f^{\#}(N') := \{m \mid f(m) \in N'\}.$$

The usual notation f^{-1} will be retained to mean $1/f$.

1. Generalities. Throughout this section, S is a compact, connected, C^{ω} Riemannian manifold with Laplace–Beltrami operator Δ and metric-induced measure σ. Assume $\sigma(S) = 1$. Suppose that on S we have a fixed nonconstant function $\xi: S \to \mathbb{C}$ such that both ξ and ξ^2 are eigenfunctions of Δ.

LEMMA 1.1. *There are real numbers α and β such that for all integers n,*

$$\Delta \xi^n = n(\alpha + \beta n)\xi^n \tag{1.1}$$

on all of S if $n \geq 0$, and on the set where $\xi \neq 0$ if $n < 0$. For convenience we write λ_n for the eigenvalue $n(\alpha + \beta n)$.

Proof. Let ∇ denote the gradient on S so that for any C^2 function f and any integer n,

$$\Delta f^n = n f^{n-1} \Delta f + n(n-1) f^{n-2} \nabla f \cdot \nabla f$$

wherever f^n is defined. Apply this with $n = 2$ to ξ to get

$$\nabla \xi \cdot \nabla \xi = \beta \xi^2 \quad \text{with } \beta := \frac{\lambda_2}{2} - \lambda_1.$$

Apply it again to get

$$\Delta \xi^n = n(\lambda_1 + (n-1)\beta)\xi^n \tag{1.2}$$

so that we may take $\alpha := \lambda_1 - \beta = 2\lambda_1 - \lambda_2/2$.

LEMMA 1.2. *For $n = 0, 1, \cdots$ the eigenvalues λ_n are strictly decreasing.*

Proof. $\lambda_n < 0$ for $n > 0$ because S is compact. Therefore $\beta \leq 0$. The assertion follows from (1.2):

$$\lambda_n = n(\lambda_1 + (n-1)\beta).$$

Now $\tau := \sigma \circ \xi^{\#}$ defines a probability Borel measure on $\mathbb{C}$ whose support is precisely $\xi(S)$.

LEMMA 1.3. *The measure τ is rotation invariant on* $\mathbf{C}$.

Proof. For any nonnegative integers $m \neq n$,

$$\int_{\mathbf{C}} z^m \bar{z}^n \, d\tau = \int_S \xi^m \bar{\xi}^n \, d\sigma = 0, \tag{1.3}$$

since ξ^n and ξ^m correspond to distinct eigenvalues of Δ.

Now suppose θ is a rotation of $\mathbf{C}$ such that $\tau \circ \theta \neq \tau$. Then for some polynomial p in z and z,

$$\int_{\mathbf{C}} p \circ \theta \, d\tau \neq \int_{\mathbf{C}} p \, d\tau, \tag{1.4}$$

and (1.4) must hold for some monomial $p = z^m \bar{z}^n$, necessarily with $m \neq n$. But by (1.3), both sides of (1.4) must be 0, a contradiction proving the nonexistence of such a θ.

COROLLARY. *$\xi(S)$ is either a disk, circle or annulus, and is centered at* 0.

Proof. $\xi(S)$ is compact, connected and rotation invariant since it is the support of τ.

LEMMA 1.4. *Let α, β be as in Lemma* 1.1. *Let $\mathcal{O}$ be open in $\mathbf{C}$ and h a harmonic function on $\mathcal{O}$. Then on the preimage $\xi^{\#}(\mathcal{O})$ of $\mathcal{O}$ in S, we have*

$$\Delta(h \circ \xi) = \left((\alpha R + \beta R^2)h\right) \circ \xi,$$

where $R := r\partial/\partial r$ and r is the radial coordinate on $\mathbf{C}$.

Proof. The result is local on $\mathbf{C}$ and by (1.1) is obvious for $h := z^n$ or $\bar{z}^n$. Then it is true for holomorphic and antiholomorphic h, hence for harmonic h.

LEMMA 1.5. *If $\xi(S)$ does not contain* 0, *then it is a circle, and* $\alpha = 0$.

Proof. In Lemma 1.4 let $h(z) := \ln|z|$. Then $\Delta \ln|\xi| = \alpha$. On a compact space, Δ takes only constants to constants, so $\ln|\xi|$ is constant, and $\alpha = 0$.

COROLLARY. *$\xi(S)$ cannot be an annulus.*

For the remainder of this paper, we restrict our attention to the case in which $\xi(S)$ is a disk and assume, without loss of generality, that it is in fact the unit disk $\mathbb{D}$.

Our method is to project problems on S down to $\mathbb{D}$ using ξ and the related map of functions E_ξ. We also need similar machinery when ξ is replaced by its real part, so we summarize the essential facts stated in mild generality. This sort of thing is well understood. For proof of Lemmas 1.6 and 1.7 specific to spheres, see [8, pp. 13–18], and for a more general formulation see [3, pp. 301–304].

Suppose V is a finite dimensional vector space and $\phi: S \to V$ a C^ω map. Let $\sigma_\phi := \sigma \circ \phi^{\#}$ be the measure on V carried over from σ by ϕ. (If $\phi = \xi$ then $\sigma_\phi = \tau$.)

LEMMA 1.6. *There is a unique bounded linear map*

$$E_\phi : L^1(S) \to L^1(V, \sigma_\phi)$$

such that for f in $L^1(S)$ and g in $L^\infty(V, \sigma_\phi)$, we have

$$\int_S f \, d\sigma = \int_V E_\phi(f) \, d\sigma_\phi,$$

$$E_\phi((g \circ \phi)f) = g E_\phi(f).$$

We use the notation E_ϕ, because this operator is essentially a conditional expectation.

Let S^r denote the regular points of ϕ in S and V^r the regular values of ϕ in V. These are open sets in S and V respectively. If ϕ is not constant, then, because it is analytic, $\phi^{\#}(V^r)$ is open and dense in S. By this and Sard's theorem,

$$\sigma(S-S^r)=0=\sigma_\phi(V-V^r).$$

σ_ϕ is smooth on V^r in that it is given by a C^ω function times Lebesgue measure. Moreover, we have

LEMMA 1.7. *If f is in $C^k(S)$, then $E_\phi(f)|V^r$ is in $C^k(V^r)$.*

LEMMA 1.8. *With $\phi\colon S\to V$ a nonconstant, C^ω map, suppose that for every polynomial p of degree ≤ 2 on V, there is a function $\tilde{p}$ on $\phi(V)$ such that*

$$\Delta(p\circ\phi)=\tilde{p}\circ\phi.$$

Then there is a unique second order differential operator Λ on $\phi(S)$ such that for all g in $C^2(V)$,

$$\Delta(g\circ\phi)=(\Lambda g)\circ\phi. \tag{1.5}$$

Moreover, Λ is C^ω on V^r and is symmetric with respect to the measure σ_ϕ. For g in $C^2(S)$,

$$E_\phi(\Delta f)=\Lambda E_\phi(f) \quad \text{on } V^r. \tag{1.6}$$

Proof. The map $\mathrm{p}\to\tilde{\mathrm{p}}$ is linear and so may be expressed as $\tilde{\mathrm{p}}=\Lambda\mathrm{p}$ for a unique second order differential operator Λ on $\phi(S)$. (1.5) is verified by showing that both sides vanish at a point s such that g vanishes to the second order at $\phi(s)$. The smoothness of Λ on V^r and symmetry with respect to σ_ϕ are routine, with the latter using (1.5) and the symmetry of Δ. To show (1.6), take g in $C_c(V^r)$ and show that

$$\int_{\phi(S)} g\Lambda E_\phi(f)\,d\phi_\phi=\int_{\phi(S)} gE_\phi(\Delta f)\,d\sigma_\phi,$$

by using symmetry to move Λ over to g, then Lemma 1.6 and (1.5) to get

$$\int_{\phi(S)} E_\phi((\Lambda g\circ\phi)f)\,d\sigma_\phi=\int_S \Delta(g\circ\phi)f\,d\sigma,$$

and finally using symmetry of Δ and Lemma 1.6 again.

We now specialize to the case $\phi:=\xi$ and recall that $\tau=\sigma_\phi$ is rotation invariant and $\phi(S)=\mathbb{D}$, the unit disk in $\mathbb{C}$. We also wish to add the following hypotheses on ξ:

(H1) 0 is a regular value for ξ.

(H2) There is a function ψ on $[0,1]$ such that $\Delta\xi\bar{\xi}=\psi(|\xi|)$.

Since $1,\xi,\bar{\xi},\xi^2$ and $\bar{\xi}^2$ are all eigenfunctions of Δ, (H2) is all that is needed to fulfull the hypothesis of Lemma 1.8. We then have the operator Λ on $\mathbb{D}$ satisfying (1.5) and (1.6). If Δ_c denotes the Laplacian on $\mathbb{C}$, i.e.,

$$\Delta_c=\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}=4\frac{\partial}{\partial z}\frac{\partial}{\partial \bar{z}}=\frac{1}{r^2}\left(R^2+\frac{\partial^2}{\partial\theta^2}\right),$$

then we find

$$\Lambda=\alpha R+\beta R^2+\gamma(r)\Delta_c, \tag{1.7}$$

where $R:=r\partial/\partial r$, and $\gamma(r):=(\psi(r)-(2\alpha+4\beta)r^2)/4$, and α,β are as in Lemmas 1.1 and 1.4.

Let ω denote the Radon–Nikodým derivative of τ with respect to Lebesgue measure: $d\tau=\omega\,dx\,dy$. Then ω is a radial function which is C^ω on the set $\mathbb{D}^r$ of regular values in $\mathbb{D}$.

LEMMA 1.9. *$\int_x^1 \alpha\omega(t)t\,dt = -\kappa(x)\omega(x)$ at x in $[0,1]\cap\mathbb{D}^r$. Here*

$$\kappa(x) := \beta x^2 + \gamma(x).$$

Proof. Let f, g be radial C^2 functions on $\mathbb{D}$. Using the symmetry of Λ from Lemma 1.8, we get

$$0 = \frac{1}{2\pi}\int_{\mathbb{D}} (f\Lambda g - g\Lambda f)\omega r\,dr\,d\theta = \int_0^1 (\alpha h(r) + \kappa(r)h'(r))\omega(r)r\,dr,$$

where $h(r) := r(f(r)\partial g/\partial r - g(r)\partial f/\partial r)$. The result now follows by letting h approximate the Heaviside function:

$$h_x(r) := \begin{cases} 0 & \text{if } r \le x, \\ 1 & \text{if } r > x. \end{cases}$$

COROLLARY. *$\kappa\omega$ is continuous and monotone on $[0,1]$ and vanishes at $x=1$. On $[0,1]\cap\mathbb{D}^r$ it is C^ω and satisfies*

$$(\kappa\omega)' = \alpha\omega.$$

COROLLARY. *On $\mathbb{D}^r$,*

$$\Lambda = \frac{1}{r^2}\left(\frac{1}{\omega}R\,\omega\kappa R + \gamma\frac{\partial^2}{\partial\theta^2}\right).$$

LEMMA 1.10. *For a C^1 function f on $\mathbb{D}$ define the 1-form*

$$\begin{aligned}\delta f &:= \left(\kappa(Rf)\,d\theta - \gamma\frac{\partial f}{\partial\theta}d\ln r\right)\omega \\ &= \left(\beta(Rf)(x\,dy - y\,dx) + \gamma\left(\frac{\partial f}{\partial x}dy - \frac{\partial f}{\partial y}dx\right)\right)\omega.\end{aligned}$$

Then for C^2 functions f, g on $\mathbb{D}^r$,

$$(f\Lambda g - g\Lambda f)\omega(r)r\,dr\,d\theta = d(f\delta g - g\delta f).$$

Proof. By the last corollary,

$$\begin{aligned}&(f\Lambda g - g\Lambda f)\omega(r)r\,dr\,d\theta \\ &= \frac{\partial}{\partial r}(\omega(r)\kappa(r)(fRg - gRf)) + \frac{\partial}{\partial\theta}\left((\gamma(r)\omega(r)/r)\left(f\frac{\partial g}{\partial\theta} - g\frac{\partial f}{\partial\theta}\right)\right)dr\,d\theta \\ &= d(f\delta g - g\delta f).\end{aligned}$$

The second expression for δf is computed from the first by routine use of

$$R = x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}, \quad \frac{\partial}{\partial\theta} = -y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y}, \quad r\,dr = x\,dx + y\,dy,$$

$$r^2\,d\theta = -y\,dx + x\,dy, \qquad \kappa(r) = \beta r^2 + \gamma(r).$$

LEMMA 1.11. *Let $\mathbb{O}$ be an open subset of $\mathbb{D}$ bounded by a piecewise smooth curve C lying in $\mathbb{D}^r$. Let f, g be C^2 in a neighborhood of the closure of $\mathbb{O}$. Then*

$$\int_{\mathbb{O}} (f\Lambda g - g\Lambda f)\,d\tau = \int_C (f\delta g - g\delta f). \tag{1.8}$$

Proof. Split f into the sum of functions f_1, f_2, such that f_1 has support inside $\mathbb{D}^r$ and f_2 has support in the interior of $\mathfrak{O}$. Then (1.8) holds with f replaced by f_1 by Lemma 1.10; and (1.8) holds with f replaced by f_2, because both sides are 0.

The following well-known fact (see, e.g., [2, p. 60]) will prove useful in several places:

LEMMA 1.12. *If $f\colon [0,a]\times[-b,b]\to\mathbb{C}$ is continuous and of class C^n near $(0,0)$, then*

$$\lim_{x\to 0^+}\int_{-b}^{b} f(x,y)(x+iy)^{-n}dy$$

exists, depends on $f(0,y)$ only and is denoted as

$$\int_{-b}^{b} f(0,y)(iy+0)^{-n}dy.$$

Moreover,

$$(1.9)\quad \int_{-b}^{b} f(0,y)\left((iy+0)^{-n}-(iy-0)^{-n}\right)dy=\frac{2\pi}{(n-1)!}\left(\frac{-i\partial}{\partial y}\right)^{n-1}f(0,y)|_{y=0}.$$

Finally we are in a position to define and analyse the integral

$$(1.10)\qquad \int_S f\xi^{-n}d\sigma.$$

For $t>0$ let $\mathbb{D}_t := \{x+iy\in\mathbb{D} \,|\, |x|>t\}$ and

$$S_t := \xi^{\#}(\mathbb{D}_t).$$

LEMMA 1.13. *For f in $C^n(S)$ the following limit exists:*

$$\lim_{t\to 0^+}\int_{S_t} f\xi^{-n}d\sigma.$$

This will be the meaning assigned to (1.10).

Proof. By Lemma 1.6,

$$\int_{S_t} f\xi^{-n}d\sigma=\int_{\mathbb{D}_t} E_\xi(f)z^{-n}d\tau=\int_{|x|>t}\int_{|y|<x'}(x+iy)^{-n}g(x,y)\,dy\,dx,$$

where $x' := (1-x^2)^{1/2}$ and $g := E_\xi(f)\omega$. The inner integral defines a continuous function of x on $[-1,1]$ (except for a possible simple discontinuity at $x=0$) by the corollary to Lemma 1.13.

When it is necessary to have a term for the definition of (1.10) just given, we refer to it as the *slit regularization* of (1.10) to distinguish it from, for example, the Gelfand–Shilov "canonical regularization" [2, p. 61].

THEOREM 1.1. *Let f be in $C^n(S)$ and let $\tilde{f} := E_\xi(f)$. Then*

$$(1.11)\quad \int_S((\Delta-\lambda_{-n})f)\xi^{-n}d\sigma$$

$$=\frac{-2\pi}{(n-1)!}\left(-i\frac{\partial}{\partial y}\right)^{n-1}\left(\frac{\partial}{\partial x}-i\frac{\partial}{\partial y}\right)\left(\omega(y)\gamma(y)\tilde{f}(x,y)\right)\Big|_{(x,y)=(0,0)}.$$

(*Recall that $(\Delta-\lambda_{-n})\xi^{-n}=0$ on the set $\xi\neq 0$.*)

Proof. For $t, u>0$ let

$$\mathbb{D}_{t,u} := \{x+iy \in D \mid (|x|>t) \text{ or } (|y|>u)\}$$

and

$$\mathbb{D}'_{t,u} := \mathbb{D} - \mathbb{D}_{t,u}.$$

The left side of (1.11) equals

$$\lim_{t\to 0^+}\int_{\mathbb{D}_t}(\Lambda-\lambda_{-n})\tilde{f}z^{-n}\,d\tau = \lim_{t\to 0^+}\int_{\mathbb{D}_{t,u}}(\Lambda-\lambda_{-n})\tilde{f}z^{-n}\,d\tau \tag{1.12}$$

for any fixed $u>0$. Choose t_1, $u>0$ small enough so that $\mathbb{D}'_{t_1,u}\subset\mathbb{D}^r$, the set of regular values. Fix u. For $t<t_1$ $\tilde{f}\omega$ is C^n in a neighborhood of $\mathbb{D}'_{t,u}$. Lemma 1.11 implies the right side of (1.12) equals

$$-\int_C z^{-n}\delta\tilde{f}-\tilde{f}\delta z^{-n},$$

where C is the boundary of $\mathbb{D}'_{t,u}$. Neglecting the "short" ends of the rectangle C (on $\mathbb{R}\pm iu$) and using the second form of δ in Lemma 1.10, we get two integrals (one from $-iu+t$ to $iu+t$ and one from $iu-t$ to $-iu+t$) of the one-form

$$\begin{aligned}-(z^{-n}\delta\tilde{f}-\tilde{f}\delta z^{-n}) = -\biggl(z^{-n}\omega(r)\biggl(\beta xR\tilde{f}(x,y)+\gamma(r)\frac{\partial\tilde{f}}{\partial x}(x,y)\biggr)\,dy \\ +n\tilde{f}(x,y)\omega(r)(\beta z^{-n}x+\gamma(r)z^{-n-1})\,dy\biggr),\end{aligned}$$

where $r=(x^2+y^2)^{1/2}$ and $x=t$ or $-t$, depending on which integral we are concerned with. Lemma 1.12 shows that these integrals remain bounded as $t\to 0^+$ and the terms containing the factor of x go to 0. The net result is (in the notation of Lemma 1.12)

$$\begin{aligned}-\int_{-u}^{u}\omega(r)\gamma(y)\biggl(\biggl(\frac{\partial\tilde{f}}{\partial x}(0,y)\biggr)((iy+0)^{-n}-(iy-0)^{-n}) \\ +n\tilde{f}(0,y)((iy+0)^{-n-1}-(iy-0)^{-n-1})\biggr)\,dy,\end{aligned}$$

which, by Lemma 1.12, equals the right side of (1.11).

2. Hilbert-like transforms. Throughout this section, S is the unit sphere in a real inner-product space V of finite dimension $d+1$. The inner product in V is denoted by a dot. For any s in S, we have the projection $\phi_s\colon S\to[-1,1]$, by

$$\phi_s(s') := s'\cdot s,$$

and the corresponding

$$E_s := E_{\phi_s}\colon L^1(S)\to L^1([-1,1],\omega_d(x)\,dx),$$

by Lemma 1.6, where $\omega_d(x) := (1-x^2)^{(d-2)/2}/B(d/2,1/2)$. Here B is the Eulerian beta function.

LEMMA 2.1. *For an integer $m>0$ and f in $C^m(S)$, the following limit exists*:

$$\lim_{t\to 0^+}\int_S f(s')(t+is\cdot s')^{-m}\,ds'; \tag{2.1}$$

it will be denoted $H_m^+(f)(s)$. *If instead we take* $\lim_{t\to 0^-}$, *we get* $H_m^-(f)(s)$ *and*

$$H_m^-(f)(s)=(-1)^m H_m^+(f)(-s).$$

Proof. By Lemma 1.12, (2.1) exists and may be written

$$\int E_s(f)(y)(iy+0)^{-m}\omega_d(y)\,dy,$$

$$H_m^-(f)(s)=\lim_{t\to 0^+}\int_S f(s')(-t+is\cdot s')^{-m}ds'=(-1)^m H_m^+(f)(-s).$$

LEMMA 2.2. *For f in* $C^m(S)$,

$$H_m^+(f)(s)-H_m^-(f)(s)=\frac{2\pi}{(m-1)!}\left(-i\frac{d}{dx}\right)^{m-1}(\omega_d(x)E_s(f)(x))_{x=0}.$$

Proof. This is immediate from Lemma 1.12 and the formula for $H_m^+(f)$ in the proof of Lemma 2.1.

We will now approach H_m^+ from an alternate direction. H_m^+ commutes with the action of the rotation group on $C^m(S)$. Let W_j denote the space of homogeneous, harmonic polynomials of degree j on S.

LEMMA 2.3. $H_m^+\mid W_j=c_{m,j}I$, *where*

$$c_{m,j}:=\frac{(-i)^j\Gamma\left(\frac{d+1}{2}\right)\Gamma\left(\frac{m+j}{2}\right)2^{m-1}}{(m-1)!\Gamma\left(\frac{d-m+j+1}{2}\right)}.$$

Proof. $H_m^+\mid W_j$ is a scalar multiple of the identity by an adaptation of Schur's lemma. To compute the value of $c_{m,j}$, we fix s_0 in S and define

$$f(s):=P_j(s\cdot s_0),$$

where

$$P_j(x):=\frac{Q_j(x)}{Q_j(1)},$$

where

$$Q_j(x):=(1-x^2)^{-\delta}\left(\frac{d}{dx}\right)^j(1-x^2)^{j+\delta}\qquad\left(\delta:=\frac{d-2}{2}\right).$$

The point is that f is in W_j and $f(s_0)=1$ so that

$$c_{m,j}=H_m^+(f)(s_0)=\int_{-1}^{1}P_j(x)(ix+0)^{-m}\omega_d(x)\,dx.$$

Integration by parts reduces this to a constant K times

$$\int_{-1}^{1}(1-x^2)^{j+\delta}(ix+0)^{-m-j}dx=\pi\left(\frac{\Gamma(j+\delta+1)}{\Gamma\left(\frac{m+j+1}{2}\right)\Gamma\left(\frac{j-m+3}{2}+\delta\right)}\right),$$

where K is given by

$$K=\frac{(i)^j(m+j-1)!}{(m-1)!Q_j(1)B\left(\frac{d}{2},\frac{1}{2}\right)}=\frac{\left(\frac{-i}{2}\right)^j\Gamma\left(\frac{d+1}{2}\right)(m+j-1)!}{\left(\Gamma\left(j+\frac{d}{2}\right)\Gamma\left(\frac{1}{2}\right)(m-1)!\right)}.$$

The lemma follows after using the duplication theorem for Γ and cancelling some factors. Recall that $\delta=(d-2)/2$. □

We now introduce the notation

$$x^{\langle t\rangle}:=\frac{\Gamma\left(x+\frac{t+1}{2}\right)}{\Gamma\left(x+\frac{1-t}{2}\right)}.$$

The heuristic is that $x^{\langle t\rangle}$ is in many ways like x^t. Specifically:

$$x^{\langle 0\rangle}=1,\quad x^{\langle 1\rangle}=x,\quad x^{\langle -t\rangle}=\left(x^{\langle t\rangle}\right)^{-1},\quad x^{\langle t\rangle}\sim x^t \quad \text{as } x\to\infty.$$

Moreover,

$$x^{\langle 1/2\rangle}=\left(x+O\left(\frac{1}{x}\right)\right)^{1/2},$$

and therefore,

$$x^{\langle -1/2\rangle}=\left(x+O\left(\frac{1}{x}\right)\right)^{-1/2}$$

as $x\to\infty$.

LEMMA 2.4. $x^{\langle 2k+t\rangle}=x^{\langle t\rangle}\prod_{n=1}^{k}(x^2-((t-1)/2+n)^2)$ *for any integer* $k>0$.

Proof. By induction on k using the case $k=1$ which is just,

$$x^{\langle 2+t\rangle}=x^{\langle t\rangle}\left(x^2-\left(\frac{t+1}{2}\right)^2\right).$$

LEMMA 2.5. *Let t be the number in $\{-\frac{1}{2},0,\frac{1}{2},1\}$ and k the integer such that*

$$m-\left(\frac{d+1}{2}\right)=2k+t.$$

Then assuming $k>0$ (i.e. $((d+1)/2)>1$), we get

$$\frac{\Gamma\left(\frac{m+j}{2}\right)}{\Gamma\left(\frac{d+j+1-m}{2}\right)}=\left(\frac{2j+d-1}{4}\right)^{\langle t\rangle}2^{-2k}\prod_{n=1}^{k}\left(\lambda_{2n-m}-\lambda_j\right)$$

(*where* $\lambda_n:=-n(n+d-1)$).

Proof. The left side equals $((2j+d-1)/4)^{\langle 2k+t\rangle}$, so the result follows from Lemma 2.4 and the observation that

$$\left(\frac{2j+d-1}{4}\right)^2-\left(\frac{t-1}{2}+k+n-1\right)^2=\frac{1}{4}\left(\lambda_{2n-m}-\lambda_j\right).$$

The summary of our results so far is

$$c_{m,j}=K_{d,m}(-i)^j\left(\frac{2j+d-1}{4}\right)^{\langle t\rangle}\prod_{n=1}^{k}(\lambda_{2n-m}-\lambda_j),$$

where $K_{d,m}:=\Gamma((d+1)/2)2^{t+(d-1)/2}/(m-1)!$ and k,t are as in Lemma 2.5.

We now introduce the operator Ω, defined first on the sum of the spaces W_j by $\Omega\,|\,W_j:=jI$. Ω extends by continuity to $C^\infty(S)$.An alternative definition is: for f on S, extend f to a harmonic function in the ball bounded by S. Then

$$(\Omega f)(s)=\frac{\partial}{\partial t}f(ts)\bigg|_{t=1},$$

provided the derivative exists. Ω is related to Δ by

$$\Delta=-\Omega(\Omega+d-1),$$

or equivalently,

$$\left(\frac{d-1}{2}\right)^2-\Delta=\left(\Omega+\frac{d-1}{2}\right)^2.$$

If ψ is some function on $\{0,1,\cdots\}$, then $\psi(\Omega)\,|\,W_j=\psi(j)I$ by definition.

THEOREM 2.1. *On $C^\infty(S)$, if $m\geq d/2$, then*

$$H_m^+=K_{d,m}(-i)^\Omega\left(\frac{2\Omega+d-1}{4}\right)^{\langle t\rangle}\prod_{n=1}^{k}(\lambda_{2n-m}-\Delta),$$

where $K_{d,m}$ is as above and k and t are as in Lemma 2.5:

$$m-\left(\frac{d+1}{2}\right)+2k+t,\qquad t\in\left\{-\frac{1}{2},0,\frac{1}{2},1\right\},$$

k is an integer, and $\lambda_n=-n(n+d-1)$. If $k=0$, the product $\prod_{n=1}^{k}$ is 1.

Proof. Both sides are $c_{m,j}$ on W_j, are linear and are continuous on $C^\infty(S)$.

COROLLARY 2.1. *Suppose d is odd; let $m:=(d+1)/2$. Then*

$$(-i)^\Omega=2^{1-m}H_m^+.$$

COROLLARY 2.2. *Suppose d is even; let $m:=(d+2)/2$. Then*

$$H_m^+=\left(m-\frac{3}{4}\right)^{\langle -1/2\rangle}2^{m-1}\left(\frac{2\Omega+d-1}{4}\right)^{\langle 1/2\rangle}(-i)^\Omega.$$

COROLLARY 2.3. *Suppose d is odd and $m=(d-1)/2$. If f is in $C^\infty(S)$ and $f(-s)=(-1)^m f(s)$, then*

$$((-i)^\Omega f)(s)=(-i)^m/(d-2)(d-4)\cdots5\cdot3\cdot1\left(\frac{d}{dx}\right)^m$$

$$\cdot\left((1-x^2)^{(d-2)/2}E_s(f)(x)\right)\bigg|_{x=0}.$$

Proof. Combine Corollary 2.1 with Lemma 2.2.

COROLLARY 2.4. *Suppose d is even and $m=d/2$. If f is in $C^\infty(S)$ and $f(-s)=(-1)^m f(s)$, then*

$$\left((-i)^{\Omega}\left(\frac{2\Omega+d-1}{4}\right)^{\langle 1/2\rangle} f\right)(s)$$

$$=\left(\pi^{1/2}(-i)^m/(d-2)(d-4)\cdots 4\cdot 2\right)\left(\frac{d}{dx}\right)^m\left((1-x^2)^{(d-2)/2}E_s(f)(x)\right)\Big|_{x=0}.$$

Proof. Combine Corollary 2.2 with Lemma 2.2.

Remark. If p is a harmonic polynomial restricted to S, then $(-i)^{\Omega}p(s)=p(-is)$. If, in addition, S is the unit sphere in $\mathbf{C}^{(d+1)/2}$ and p is holomorphic along some complex coordinates in $\mathbf{C}^{(d+1)/2}$ and antiholomorphic along the remaining ones, then $(-i)^{\Omega}$ effects a quarter rotation of p. Our heuristic regarding $(-i)^{\Omega}$ then is that it is a quarter rotation of the functions on S which is isotropic in that it commutes with all true rotations of S. There is, of course, no geometric rotation of S giving $(-i)^{\Omega}$ by composition. Nevertheless, $(-i)^{\Omega}$ behaves in some ways as if there were. Corollary 2.3 is an instance of this. Let

$$\operatorname{par}_n(f)(s):=\tfrac{1}{2}\left(f(s)+(-1)^n f(-s)\right)$$

so that $\operatorname{par}_n(f)$ has the parity (odd or even) of n. Then Corollary 2.3 says that if $d=2m+1$, then

$$f\to(-i)^{\Omega}\operatorname{par}_m(f)(s)$$

is supported on $s^{\perp}$ in S, just as it would be if $(-i)^{\Omega}$ came from a quarter rotation in S.

We find it interesting that $(-i)^{\Omega}$ has three such differing representations, as in Corollaries 2.1 and 2.3.

3. The sphere. In this section we apply the results of §1 and §2 to the Helgason–Fourier transform on a sphere S. We begin by briefly describing this transform and its theory.

Let S be the unit sphere in an inner-product space V of dimension $d+2$ so that $\dim S=d+1$. Fix a point s_0 in S and let $S':=s_0^{\perp}$ in S. (If s_0 is the north pole then S' is the equator.)

Let Ξ denote the set of real-linear maps ξ of V into $\mathbf{C}$ such that

$$\xi(s_0)=1, \tag{3.1}$$

$$\xi(S)=\mathbf{D}. \tag{3.2}$$

Then each ξ in Ξ satisfies all the hypotheses on ξ of §1, viz., the first paragraph and H1, H2 of §1 and (3.1), (3.2).

We note that Ξ is parametrized by S':

$$s'_{\xi}\leftrightarrow\xi \quad \text{where } \xi(s)=s\cdot s_0+is\cdot s'.$$

Note also that all ξ in Ξ have the same real part. Let $\eta\colon S\to\{-1,0,1\}$ be defined as the sign of this real part:

$$\eta(s):=\operatorname{sgn}(s\cdot s_0).$$

The transform dual space is the set $\Xi \times N$ where $N := \{0,1,2,\cdots\}$. The transform Tf of f in $C^\infty(S)$ is defined as

$$Tf(\xi,m) := \int_S f\eta^d \xi^{-m-d} d\sigma, \tag{3.3}$$

where the integral is defined using the slit regularization of Lemma 1.13. Although the discontinuous function η appears at first to complicate matters (at least if d is odd), examination of the argument in Lemma 1.13 shows that it does not interfere with convergence of the integral, so Tf is well defined.

The inverse transform formula is

$$f(s) = \sum_{m=0}^{\infty} \dim(W_m) \int_\Xi Tf(\xi,m)\xi^m(s)\, d\xi, \tag{3.4}$$

where W_n is the space of homogeneous, harmonic polynomials of degree n on S, the measure on Ξ is the rotation invariant probability measure on S', and the series converges in $C^\infty(S)$. For details see [8].

In the transform (3.3) the task of the integral is to pick out the W_m component f_m of f. In this connection note that $\xi^m \varepsilon W_m$ and that

$$f_m = \dim(W_m) \int_\Xi Tf(\xi,m)\xi^m d\xi.$$

Moreover, on the set $S-S'$, the function $\eta^d\xi^{-m-d}$ is an eigenfunction of Δ with the same eigenvalue

$$\lambda_m := -m(m+d)$$

as ξ^m. Unfortunately, (3.3) does not define an eigendistribution of Δ. A principal goal of this paper has been to measure that failure by computing (3.3) with f replaced by

$$(\Delta - \lambda_m)f.$$

THEOREM 3.1. *If S is a sphere of odd dimension, then for f in $C^\infty(S)$,*

$$T((\Delta-\lambda_m)f)(\xi,m) = -(d(m+d-1)!)^{-1}\left(-i\frac{\partial}{\partial y}\right)^{m+d-1}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right)$$
$$\cdot\left((1-y^2)^{(d-2)/2}E(f)(x,y)\right)\Big|_{(x,y)=(0,0)}.$$

Proof. In Theorem 1.1 take $n := m+d$, $\gamma := 1$, and $\omega(z) = (1-z\bar{z})^{(d-2)/2}/(2\pi d)$. Then the present result follows immediately from Theorem 1.1.

Remark. The functional on f in Theorem 3.1 is clearly a distribution supported at $\xi = 0$. Unfortunately the same is not true for spheres of even dimension, due to the discontinuity in η^d. Thus we cannot expect a strict analogue of Theorem 3.1 for such spheres. However, we have:

THEOREM 3.2. *For any sphere S let s_0 and S' be as above. Parametrize S by cylindrical coordinates*:

$$(-1,1)\times S' \leftrightarrow S - \{s_0, -s_0\},$$

$$(x,s') \leftrightarrow s = xs_0 + (1-x^2)^{1/2}s'.$$

Write $\partial_x f(s')$ *for* $\frac{\partial}{\partial x}f(x,s')|_{x=0}$. *Then for* f *in* $C^\infty(S)$ *and* ξ *corresponding to* s'_ξ *and with* H_n *defined relative to* S', *we have*

(3.5)

$$T((\Delta-\lambda_m)f)(\xi,m)=-\left(B\left(\frac{d+1}{2},\frac{1}{2}\right)\right)^{-1}$$
$$\cdot\left\{H^+_{m+d}(\partial_x f|_{S'})(s'_\xi)-(-1)^d H^-_{m+d}(\partial_x f|_{S'})(s'_\xi)\right.$$
$$\left.+(m+d)\left(H^+_{m+d+1}(f|_{S'})(s'_\xi)-(-1)^d H^-_{m+d+1}(f|_{S'})(s'_\xi)\right)\right\}.$$

Proof. The left side equals the limit, as $t\to 0^+$, of

$$\int_{S_t}(\Delta-\lambda_m)f\eta^d\xi^{-m-d}d\sigma=\int_{\partial S_t}(\xi^{-m-d}\nabla f-f\,\nabla\xi^{-m-d})\eta^d\cdot d\sigma',$$

where $d\sigma'$ is the boundary measure on ∂S_t times u, the unit normal vector to ∂S_t. This follows from $(\Delta-\lambda_m)\xi^{-m-d}=0$ on S_t. Note that the boundary measure is

$$\left(B\left(\frac{d+1}{2},\frac{1}{2}\right)\right)^{-1}ds'$$

when $t=0$; here ds' denotes the normalized measure on S': $\int_{S'}1\,ds'=1$. On ∂S_t

$$\xi=t+i(1-t^2)^{1/2}s'\cdot s'_\xi,\qquad \nabla\xi\cdot u=1-i\left(\frac{t}{(1-t^2)^{1/2}}\right)s'\cdot s'_\xi.$$

If we make these substitutions above, factor out $(1-t^2)^{-(m+d)/2}$ and observe that $t/(1-t^2)\to 0$ as $t\to 0$, we find that we obtain the right side of (3.5).

Remark. There is a close parallel between this proof and the proof of Theorem 1.1, in that both involve Green's theorem applied in related contexts. Indeed, with some reworking, we could get along with a single proof. However, there are some important economies possible with the sphere (regarding the set $\mathrm{Re}(\xi)=0$) that are not available in §1, most notably the theory of H^+_m.

Recall that par_n is the projection of a function on S' to its odd or even part, depending on the parity (odd or even) of n.

COROLLARY 3.1. *In the notation of Theorem* 3.2,

$$T((\Delta-\lambda_m)f)(\xi,m)=-\left(\frac{2}{B\left(\frac{d+1}{2},\frac{1}{2}\right)}\right)$$
$$\cdot\left\{H^+_{m+d}(\mathrm{par}_{m-1}(\partial_x f|_{S'}))(s'_\xi)+(m+d)H^+_{m+d+1}(\mathrm{par}_m f|_{S'})(s'_\xi)\right\}.$$

We wish to restate this by making use of Theorem 2.1. An objective of this restatement is to bring to the surface the derivatives lurking on the right side in Corollary 3.1, thereby making the right side more computable.

Let $\delta:=(d-1)/2$.

THEOREM 3.3. *Given the integer* $m\geq 0$, *we write uniquely*

$$m+\delta-2=t_2+2k_2,\qquad m+\delta-1=t_1+2k_1,$$

where k_2, k_1 are integers and t_2, t_1 are in $\{-\frac{1}{2},0,\frac{1}{2},1\}$. Then for any f in $C^\infty(S)$ (and with the notation $\partial_x f$ used above), we have

(3.6)

$$T((\Delta-\lambda_m)f)(\xi,m)=\frac{2^{\delta+1}\Gamma(\delta+3/2)(-i)^{\Omega'}}{(m+d-1)!\Gamma(\frac{1}{2})}$$

$$\cdot\left\{\left(2^{t_2}\left(\frac{\Omega'+\delta}{2}\right)^{\langle t_2\rangle}\left(\prod_{j=0}^{k_2}(\lambda'_{m-1-2j}-\Delta')\right)\operatorname{par}_{m-1}(\partial_x f)\right)(s'_\xi)\right.$$

$$\left.+\left(2^{t_1}\left(\frac{\Omega'+\delta}{2}\right)^{\langle t_1\rangle}\left(\prod_{j=0}^{k_1}(\lambda'_{m-2j}-\Delta')\right)\operatorname{par}_m(f|_{S'})\right)(s'_\xi)\right\}.$$

(*Here λ', Δ', Ω' are all defined with respect to S'.*)

Proof. This is an immediate consequence of Theorem 2.1 applied to Corollary 3.1, bearing in mind that d is the dimension of the S' of this section and the sphere S of §2, so that the results of §2 may be applied directly to S'. In particular the λ', Δ', Ω' used in this theorem are the λ,Δ,Ω of §2. The m of Theorem 2.1 must be replaced by $m+d$ or $m+d+1$. We must also use

$$\lambda'_{2j-m-d+1}=\lambda'_{m-2j}.$$

Remark. There is some inevitability about part of the products involving $(\lambda'_n-\Lambda')$ on the left in (3.6): according to [8, p. 26, (5)] the left side of (3.6) must vanish on polynomials f of degree $\leq m$, but then so must the right side. This is accomplished by the factors $(\lambda'_n-\Delta')$ (with $0\leq n\leq m$) together with the screen for parity.

We would like, but do not have, a similarly simple explanation for the remaining factors $(\lambda'_n-\Lambda')$ (with $n<0$) as well as $((\Omega'+\delta)/2)^{\langle t\rangle}$, etc. in (3.6). The explanation must take a different tack from the one just given since $((\Omega'+\delta)/2)^{\langle t\rangle}$ (with t in $\{-\frac{1}{2},0,\frac{1}{2},1\}$) and $(\lambda'_n-\Delta')$ (with $-d+1<n<0$) have bounded inverses on $L^2(S')$.

4. Applications to the transform *T*. We consider briefly three types of application of the results of this paper to computations and estimates involving $T(f)$.

The first is the most obvious: if p is a polynomial on $\mathbb{R}$, m is an integer ≥ 0 and q is the polynomial such that

$$p(x)=(x-\lambda_m)q(x)+p(\lambda_m),$$

then

$$T(p(\Delta)f)(\xi,m)=p(\lambda_m)T(f)(\xi,m)+T((\Delta-\lambda_m)g)(\xi,m),$$

where $g:=q(\Delta)f$. Now the term in g may be evaluated by Theorem 1.1 or 3.3. In this same spirit we find

$$2T(\nabla\xi\cdot\nabla f)(\xi,m)=(\lambda_m-\lambda_{m-1})T(f)(\xi,m-1)$$
$$+T((\Delta-\lambda_m)(\xi f))(\xi,m)+T((\Delta-\lambda_{m-1})f)(\xi,m-1),$$

where, again, we may evaluate the $\Delta-\lambda$ terms by our results.

A second type of application is the explicit computation of $T(f)(\xi,m)$ in special cases. The simplest example of this arises when f is an eigenfunction of Δ. If $\Delta f=\lambda_n f$

we may distinguish three cases: $n<m$, $n=m$, $n>m$. The first two are treated completely in [8, p. 26, (5)]. The third falls to our results by

$$T(f)(\xi,m)=(\lambda_n-\lambda_m)^{-1}T((\Delta-\lambda_m)f)(\xi,m).$$

We wish to elaborate slightly. If we use the cylindrical coordinates (x,s') of Theorem 3.2, then an arbitrary eigenfunction f may be written as a sum of functions

$$f(x,s')=(1-x^2)^{j/2}p(x)g(s'),$$

where the polynomial p is $P_{n-j,d+1+2j}$, and $P_{n,d}$ is the polynomial P_n of Lemma 2.3. Moreover, g is an eigenfunction of Δ' with eigenvalue λ'_j and $\Omega' g=jg$. From Corollary 3.1 and Lemma 2.3 we have

$$T(f)(\xi,m)=cg(s'_\xi),$$

where the constant c is 0 if m and n have different parity; otherwise we get two different formulae for c, depending on whether or not the common parity of m and n equals that of j. The computation requires the values

$$p(0)=\frac{\cos\left(\frac{\pi(n-j)}{2}\right)\Gamma\left(j+\frac{d+1}{2}\right)\Gamma\left(\frac{n-j+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n+j+d+1}{2}\right)},$$

$$p'(0)=\frac{2\sin\left(\frac{\pi(n-j)}{2}\right)\Gamma\left(j+\frac{d+1}{2}\right)\Gamma\left(\frac{n-j+2}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n+j+d}{2}\right)}$$

from [1, p. 174–5] but is otherwise routine. The conclusion is

$$c=-(\lambda_n-\lambda_m)^{-1}\cdot\left(\frac{(-i)^j 2^{m+d}\Gamma\left(\frac{d+2}{2}\right)\Gamma\left(j+\frac{d+1}{2}\right)}{(m+d-1)!\pi}\right)$$

$$\cdot\begin{cases} 0 \quad \text{if } m-n \text{ is odd or if } m\geq j, \\ \dfrac{(-1)^{(n-j-1)/2}\Gamma\left(\frac{m+j+d}{2}\right)\Gamma\left(\frac{n-j+2}{2}\right)}{\Gamma\left(\frac{n+d+j}{2}\right)\Gamma\left(\frac{j-m+1}{2}\right)} & \text{if } n-j \text{ and } m-j \text{ are odd,} \\ \dfrac{(-1)^{(n-j)/2}\Gamma\left(\frac{m+j+d+1}{2}\right)\Gamma\left(\frac{n-j+1}{2}\right)}{\Gamma\left(\frac{n+d+j+1}{2}\right)\Gamma\left(\frac{j-m}{2}\right)} & \text{if } n-j \text{ and } m-j \text{ are even.} \end{cases}$$

This result together with that of [8, p. 26, (5)] gives the complete matrix of the transform T in terms of the classical expansion of a function in terms of spherical harmonics.

A third type of application is to the estimation of the norm of

$$f\mapsto T(f)(\cdot,m)$$

as bounded operator between appropriate Banach spaces of differentiable functions on S and S'. Without going into detail on this, we remark that the impression conveyed by Theorem 3.1 is that roughly $f \in C^{m+d}(S)$ is needed for $T(f)(\cdot, m)$ to be in $C(S')$; this is improved considerably by estimates based on Theorem 3.3 which suggest that $m+d$ can be replaced by something more like $m+d/2$. Our best estimate of this kind is between Sobolev spaces $L^{2,r}(S)$ and $L^2(S')$, with $r=m+(d-2)/2$. Here

$$f \in L^{2,r}(S) \quad \text{if } (I+\Omega)^r f \in L^2(S).$$

Details of the second and third of these application types were worked out in [7].

Acknowledgment. We wish to thank Professors Richard Rasala and Woody Lichtenstein for asking us the right questions at a formative stage of this work.

REFERENCES

[1] A. Erdélyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, *Higher Transcendental Functions*, (The Bateman Manuscript Project), vol. 2, McGraw-Hill, New York, 1953.

[2] I. M. Gelfand and G. E. Shilov, *Generalized Functions*, vol. 1, Academic Press, New York, 1964.

[3] V. Guillemin and S. Sternberg, *Geometric Asymptotics*, Math. Surveys, no. 14, American Mathematical Society, Providence, RI, 1977.

[4] S. Helgason, *Radon-Fourier transforms on symmetric spaces and related group representations*, Bull. Amer. Math. Soc., 71 (1965), pp. 757–763.

[5] ______, *A duality in integral geometry on symmetric spaces*, Proc. U.S.-Japan Seminar in Differential Geometry, Kyoto, Japan, 1965, pp. 37–56.

[6] ______, *A duality for symmetric spaces with applications to group representations*, Adv. Math., 5 (1970), pp. 1–154.

[7] M. B. Saxe, *Analysis of the singular aspects of the Helgason-Fourier transform on the unit sphere in* $\mathbb{R}^{d+1}$, thesis, Northeastern University, Boston, 1980.

[8] T. Sherman, *Fourier analysis on the sphere*, Trans. Amer. Math. Soc., 209 (1975), pp. 1–31.

[9] ______, *Fourier analysis on compact symmetric space*, Bull. Amer. Math. Soc., 83 (1977), pp. 378–380.

SIAM J. MATH. ANAL.
Vol. 15, No. 3, May 1984

© 1984 Society for Industrial and Applied Mathematics
015

ORTHOGONAL FUNCTION SERIES EXPANSIONS AND THE NULL SPACE OF THE RADON TRANSFORM*

ALFRED K. LOUIS[†]

Abstract. The range of the Radon transform for compactly supported functions is spanned by products of Gegenbauer polynomials and spherical harmonics. The inverse transform of those basis functions is given for arbitrary dimensions and arbitrary Gegenbauer polynomials. The resulting inversion formula is applied to study the "ghosts", i.e. the functions in the null space of the transform for finitely many projections. They are characterized by their series expansions and results concerning the optimal choice of directions for the data acquisition are deduced. The application to nuclear magnetic resonance zeugmatography is discussed.

1. Introduction. The Radon transform of a real-valued function in R^N is defined as its integrals over all $(N-1)$-dimensional hyperplanes. The problem of retrieval of a function from those integrals was first solved by Radon [24] and has been studied for example by Helgason [8], Ludwig [18]. In recent years it has found a spectacular application in medical imaging, in two dimensions with transaxial x-ray tomography, and in three dimensions with nuclear magnetic resonance (NMR) zeugmatography; for overviews see Herman [9], Marr [20], Shepp–Kruskal [27], and for theoretical studies motivated by these applications see Natterer [22], Smith–Solmon–Wagner [28].

Besides the direct implementation of the inversion formula, methods which project the searched for function on a certain subspace play an important role. Finite element spaces lead to huge systems of linear equations which have to be solved iteratively because of the lack of any structure of the few nonzero elements. For these methods see Herman [9, Chaps. 11, 12] and the vast literature cited there. The expansion of the Radon transform in terms of orthogonal functions avoids the solution of systems of equations but necessitates the knowledge of the inverse of those basis functions. This problem has been solved in two dimensions for Chebyshev polynomials of the second kind, see Cormack [2], Marr [19]. These formulas have led to better understanding of theoretical but practically motivated questions as for consistency of projection data and nonuniqueness of the reconstruction; see Lewitt–Bates [13], Louis [17].

The purpose of this paper is to provide further results for arbitrary dimensions and the general class of Gegenbauer polynomials, which include the results of Cormack and Marr as a special case. Besides their direct application as inversion formula the results can be used as a basis for an extrapolation algorithm for the limited angle problem, where the data are known only in directions in a restricted range. See, for example, Louis [15], [16] for the two-dimensional case. Furthermore the expansion in terms of orthogonal functions allows one to consider the transform under different aspects than by Radon's original inversion formula, see [24]. The main application that we have in mind is the study of the null space of the transform. For a function to be uniquely determined, its projections have to be known for infinitely many directions. This implies that the transform has a nontrivial kernel for finitely many projections, which is the case in all practical applications. The consequence of this has been extensively studied in two dimensions. See for example Katz [11], Logan [14], Louis [17], Natterer

*Received by the editors September 1, 1981, and in revised form April 1, 1983. This research was supported by the National Institutes of Health under grant HL18968 while the author was with the Medical Image Processing Group, Department of Computer Science, State University of New York at Buffalo, Amherst, New York, 14226.

[†] Fachbereich Mathematik, Universität Kaiserslautern, D-6750 Kaiserslautern, West Germany.

[22]. The results of Leahy–Smith–Solmon [12] and Smith–Solmon–Wagner [28] are valid for arbitrary dimensions, but the functions in the null space are characterized as solutions of partial differential equations, which allows no conclusions besides their existence and smoothness. Here we characterize these ghosts by their expansion in terms of orthogonal functions and deduce results concerning the optimal choice of directions in actual measurements and concerning the resolution of the reconstructed image. The most important application in more than two dimensions is nuclear magnetic resonance (NMR) zeugmatography, a new medical technology; for technical details see for example Marr–Chen–Lauterbur [21]. Using time-dependent magnetic fields leads to an example in four dimensions. The nonuniqueness results are discussed for the three-dimensional case.

After a short introduction in §2 of the transforms used later on, the inverses of the general class of basis functions are given in §3. Consequential results are the inversion formula of §4 and the construction in §5 of the null space of the Radon transform for finitely many projections. Finally the conclusions for NMR are studied in §6.

2. Preliminaries on the Radon transform and on special functions. Let f be a real-valued function in R^N, $N \geq 2$, with compact support. Without loss of generality we assume that the support is lying in the unit ball Ω. We use S^{N-1} to denote $\partial\Omega$, the N-dimensional unit sphere, and elements $\omega \in S^{N-1}$ are called directions. Let $\omega^\perp$ be the hyperplane perpendicular to ω and containing 0. The function

$$\mathfrak{R}f : Z \to R, \qquad Z = R \times S^{N-1},$$

is defined as

$$\mathfrak{R}f(s,\omega) = \int_{\omega^\perp} f(s\omega + \theta)\, d\theta = \int f(x)\delta(s - x\cdot\omega)\, dx, \tag{2.1}$$

and is called the Radon transform of f. Further let

$$\hat{f}(\xi) = (2\pi)^{-N/2} \int f(x) e^{-ix\cdot\xi}\, dx \tag{2.2}$$

be the Fourier transform of f. The fundamental relationship between the Radon and Fourier transforms is the so-called projection theorem, see Ludwig [18, formula (1.3)],

$$\hat{f}(\sigma\cdot\omega) = (2\pi)^{-(N-1)/2} (\mathfrak{R}f)^\wedge(\sigma,\omega) \tag{2.3}$$

where $\sigma \in R$ and $(\mathfrak{R}f)^\wedge$ denotes the one-dimensional Fourier transform of $\mathfrak{R}f$ with respect to the first argument.

Let $G \subset R^m$. Then we denote by $H_0^\alpha(G)$ the Sobolev space $\tilde{H}^\alpha(G)$ of Triebel [30, Chap. 4.3.2], equipped with the norm

$$\|f\|^2_{H_0^\alpha(G)} = \int \left(1 + |\xi|^2\right)^\alpha |\hat{f}(\xi)|^2 d\xi.$$

On the manifold Z we use the Sobolev space $H^\alpha(Z)$ with the norm

$$\|g\|^2_{H^\alpha(Z)} = \int_{S^{N-1}} \|g(\cdot,\omega)\|^2_{H^\alpha(R)} d\omega.$$

Then the following result is a consequence of the projection theorem, see Smith–Solmon–Wagner [28].

LEMMA 2.1. *The Radon transform is a bounded operator from $H_0^\alpha(\Omega)$ into $H^{\alpha+(N-1)/2}(Z)$ for real α.*

The problem of inversion of the Radon transform over L_2 is ill-posed and the transform has a small range which is characterized by the consistency conditions of Helgason [8] and Ludwig [18], see Smith–Solmon–Wagner [28, Thm. 13.4]. We restate the consistency conditions.

LEMMA 2.2. *Let $\alpha \geq (1-N)/2$. $g \in H^{\alpha+(N-1)/2}(Z)$ is the Radon transform of $f \in H_0^\alpha(\Omega)$ if and only if*

i) *$g(s,\omega)=0$ for hyperplanes missing Ω,*

ii) *g is even on Z,*

iii) *for all nonnegative integers m*

$$\int s^m g(s,\omega)\,ds \tag{2.4}$$

is a homogeneous polynomial in ω of degree $\leq m$.

Analogously to Ludwig [18] we use spherical harmonics to study the polynomials defined in (2.4). Spherical harmonics are the restriction to S^{N-1} of homogeneous harmonic polynomials on R^N, see Erdélyi et al. [5], Seeley [25]. Let Y_l be a spherical harmonic of degree l; then

$$\int_{S^{N-1}} Y_l(\omega)Y_k(\omega)\,d\omega=0 \quad \text{for } l\neq k. \tag{2.5}$$

We can represent Y_l as $Y_l(\omega)=C_l^{N/2-1}(\omega\cdot\eta)$, $\eta\in S^{N-1}$, where C_l^ν is the Gegenbauer polynomial of degree l and index ν. These polynomials are orthogonal over $[-1,1]$ with respect to the weight function

$$w_\nu(s)=(1-s^2)^{\nu-1/2}. \tag{2.6}$$

Finally we denote with Y_{lk}, $l\geq 0$, $k=1,\cdots,M(N,l)$ with

$$M(N,l)=\frac{(2l+N-2)(N+l-3)!}{l!(N-2)!}=O(l^{N-2}), \tag{2.7}$$

an orthonormal basis for the spherical harmonics of degree l, see Seeley [25].

To simplify the notation we introduce weighted L_2 spaces. Let $G\subset R^m$, $m\geq 1$, and let w be a weight on G. We define

$$\langle f,g\rangle_w=\int_G w(x)f(x)g(x)\,dx \tag{2.8}$$

to be the scalar product on $L_2(G,w)$. On the manifold Z we use weights which are dependent only on s and are extended to all of Z by $w(s,\omega)=w(s)$, for all $\omega\in S^{N-1}$.

Now we are able to give series expansions of the Radon transform with respect to products of Gegenbauer polynomials and spherical harmonics; for two-dimensional examples see Cormack [2], Louis [16], [17], Marr [19], [20].

LEMMA 2.3. *Let $\mathcal{R}f\in L_2(Z,w_\nu^{-1})$ and $\Phi_{mlk}=w_\nu C_m^\nu Y_{lk}$. Then*

$$\mathcal{R}f=\sum_{m=0}^{\infty}\sum_{\substack{l=0\\ m+l\ \text{even}}}^{m}\sum_{k=1}^{M(N,l)} h_{m\nu}^{-1}\langle \mathcal{R}f,\Phi_{mlk}\rangle_{w_\nu^{-1}}\Phi_{mlk}, \tag{2.9}$$

or equivalently

$$\mathcal{R}f(s,\omega)=w_\nu(s)\sum_{m=0}^{\infty}C_m^\nu(s)q_m(\omega) \tag{2.10}$$

with

$$q_m(\omega)=\sum_{\substack{l=0\\ m+l\,\text{even}}}^{m}\ \sum_{k=1}^{M(N,l)} d_{mlk}Y_{lk}(\omega),$$

where

$$d_{mlk}=h_{m\nu}^{-1}\int_{S^{N-1}}Y_{lk}(\omega)\int_{-1}^{1}\mathcal{R}f(s,\omega)C_m^\nu(s)\,ds\,d\omega$$

and

$$h_{m\nu}=\int_{-1}^{1}w_\nu(s)[C_m^\nu(s)]^2ds.$$

Proof. The set of functions Φ_{mlk}, $m\geq 0$, $l\geq 0$, $k=1,\cdots,M(N,l)$ form a complete set of orthogonal functions on $L_2(Z,w_\nu^{-1})$ and therefore the series on the right side of (2.9) converges in the weighted L_2 norm to $\mathcal{R}f$. The expansion coefficients with respect to the C_m^ν,

$$q_m(\omega)=h_{m\nu}^{-1}\int_{-1}^{1}C_m^\nu(s)\mathcal{R}f(s,\omega)\,ds,$$

are polynomials of degree less or equal to m according to the consistency condition (2.4). The evenness (oddness) of the Gegenbauer polynomials for m even (odd) and the evenness of the Radon transform result in $d_{mlk}=0$ for $m+l$ odd. (2.10) is then a consequence of the orthonormality of the Y_{lk}.

3. Inverse Radon transform of the basis functions. According to Lemma 2.3 the range of the Radon transform is spanned by $w_\nu C_m^\nu Y_l$, $m\geq 0$, $0\leq l\leq m$, $m+l$ even. An expansion of the searched for density can then be given with the inverse Radon transform of those basis functions.

THEOREM 3.1. *Let $\nu>N/2-1$ and $m\geq 0$, $0\leq l\leq m$, $m+l$ even and $\Phi(s,\omega)=w_\nu(s)C_m^\nu(s)Y_l(\omega)$. Then $V=\mathcal{R}^{-1}\Phi$ is given by*

$$V(s\cdot\omega)=c(N,m,\nu,l)(1-s^2)^{\nu-N/2}Q_{m,l}^{\nu,N}(s)Y_l(\omega) \tag{3.1}$$

with

$$Q_{m,l}^{\nu,N}(s)=s^lP_{(m-l)/2}^{(\nu-N/2,l+N/2-1)}(2s^2-1) \tag{3.2}$$

and

$$c(N,m,\nu,l)=2^{1-2\nu}\pi^{1-N/2}\frac{\Gamma(m+2\nu)\Gamma((m-l+2)/2)}{\Gamma(m+1)\Gamma(\nu)\Gamma((m-l+2+2\nu-N)/2)}, \tag{3.3}$$

where $P_n^{(\alpha,\beta)}$ is the Jacobi polynomial of degree n and indices α,β.

Proof. The proof of this result is based on the projection theorem (2.3) and it is different from the techniques Cormack [2] and Marr [17] have used for the proof of the special case $N=2$, $\nu=1$. The Fourier transform of the function V is (see (2.3))

$$(3.4)\qquad \begin{aligned} \hat{V}(\sigma\cdot\omega) &= (2\pi)^{(1-N)/2}(\mathcal{R}V)\hat{}\,(\sigma,\omega) \\ &= (2\pi)^{(1-N)/2}(w_\nu C_m^\nu)\hat{}\,(\sigma)Y_l(\omega) \\ &= (2\pi)^{-N/2}(-1)^m c_1(m,\nu)\sigma^{-\nu}J_{m+\nu}(\sigma)Y_l(\omega) \end{aligned}$$

with

$$(3.5)\qquad c_1(m,\nu)=\pi^{1/2}2^\nu\Gamma(\nu+1/2)C_m^\nu(1),$$

and J_n is the Bessel function of the first kind; see Gradshteyn–Ryzhik [6, formula 7.321, p. 830]. The inverse Fourier transform in polar coordinates is

$$\begin{aligned} V(s\cdot\omega) &= (2\pi)^{-N/2}\int_{S^{N-1}}\int_0^\infty \hat{V}(\sigma\cdot\theta)e^{is\sigma\omega\theta}\sigma^{N-1}\,d\sigma\,d\theta \\ &= (2\pi)^{-N}(-1)^m c_1(m,\nu)\int_0^\infty \sigma^{N-1-\nu}J_{m+\nu}(\sigma)\int_{S^{N-1}}Y_l(\theta)e^{is\sigma\omega\theta}\,d\theta\,d\sigma \end{aligned}$$

after a change of the order of integration. For the evaluation of the inner integral we use the Funk–Hecke theorem (see Erdélyi et al. [5] and Seeley [25])

$$\int_{S^{N-1}}F(\omega\cdot\theta)Y_l(\theta)\,d\theta=\lambda_l Y_l(\omega)$$

with

$$\lambda_l=c_2(N,l)\int_{-1}^1 F(t)C_l^{N/2-1}(t)(1-t^2)^{(N-3)/2}dt$$

and

$$c_2(N,l)=\operatorname{vol}(S^{N-2})/C_l^{N/2-1}(1).$$

In our case $F(t)=e^{is\sigma t}$ and, again using [6, formula 7.321], we have

$$\lambda_l=c_2(N,l)c_1(l,N/2-1)(s\sigma)^{1-N/2}J_{l+N/2-1}(s\sigma)$$

with c_1 from (3.5). This leads to

$$(3.6)\qquad \begin{aligned} V(s\cdot\omega) &= (2\pi)^{-N}(-1)^m c_1(m,\nu)c_1(l,N/2-1)c_2(N,l)Y_l(\omega) \\ &\quad\times s^{1-N/2}\int_0^\infty \sigma^{N/2-\nu}J_{m+\nu}(\sigma)J_{l+N/2-1}(s\sigma)\,d\sigma. \end{aligned}$$

The Weber–Schafheitlin type integral on the right-hand side of (3.6), which together with $s^{1-N/2}$ represents the Hankel transform of $\sigma^{-\nu}J_{m+\nu}(\sigma)$, is finite for $\nu>N/2-1$ with

$$(3.7)\qquad \begin{aligned} &\int_0^\infty \sigma^{N/2-\nu}J_{m+\nu}(\sigma)J_{l+N/2-1}(s\sigma)\,d\sigma \\ &\quad = c_3(N,m,\nu,l)s^{N/2-1+l}{}_2F_1((N+m+l)/2,(l-m+N-2\nu)/2;N/2+l;s^2), \end{aligned}$$

where

$$c_3(N,m,\nu,l)=2^{N/2-\nu}\Gamma((N+m+l)/2)/[\Gamma(N/2+l)\Gamma((2\nu-N+m-l+2)/2)]$$

and ${}_2F_1$ is the hypergeometric series; see Abramowitz–Stegun [1, formula 11.4.33]. We now make use of the linear transformation [1, 15.3.3] and the representation of the Jacobi polynomials as hypergeometric series [1, 15.4.6] to get

(3.8)

$$\begin{aligned}
&{}_2F_1((N+m+l)/2,(l-m+N-2\nu)/2;N/2+l;s^2)\\
&\quad=(1-s^2)^{\nu-N/2}{}_2F_1(-(m-l)/2,(l+m+2\nu)/2;N/2+l;s^2)\\
&\quad=c_4((m-l)/2,N/2+l-1)(1-s^2)^{\nu-N/2}P_{(m-l)/2}^{(N/2+l-1,\nu-N/2)}(1-2s^2)\\
&\quad=c_4((m-l)/2,N/2+l-1)(-1)^{(m-l)/2}(1-s^2)^{\nu-N/2}P_{(m-l)/2}^{(\nu-N/2,l+N/2-1)}(2s^2-1)
\end{aligned}$$

(see [1, 22.4.1]), where

$$c_4(n,\alpha)=\Gamma(n+1)\Gamma(\alpha+1)/\Gamma(\alpha+1+n).$$

With (3.4)–(3.8) we finally come to

$$V(s\cdot\omega)=c(N,m,\nu,l)s^l(1-s^2)^{\nu-N/2}P_{(m-l)/2}^{(\nu-N/2,l+N/2-1)}(2s^2-1)Y_l(\omega)$$

with

$$\begin{aligned}
c(N,m,\nu,l)=(2\pi)^{-N}(-1)^m c_1(m,\nu)c_1(l,N/2-1)c_2(N,l)&\\
\times c_3(N,m,\nu,l)(-1)^{(m-l)/2}c_4((m-l)/2,N/2-1+l)&.
\end{aligned}$$

Gathering the constants leads to (3.3).

Finally we want to note the simplest case for each dimension.

COROLLARY 3.2 (special cases of Theorem 3.1).

i) *Let $N\geq 2$ and $\nu=N/2$:*

$$\begin{aligned}
V(s\cdot\omega)=\pi^{1-N/2}2^{1-N}\Gamma(m+N)/[\Gamma(m+1)\Gamma(N/2)]&\\
\times s^l P_{(m-l)/2}^{(0,l+N/2-1)}(2s^2-1)Y_l(\omega)&.
\end{aligned}$$

ii) *Let $N=2$ and $\nu=1$* (Cormack [2], Marr [17]):

$$V(s\cdot\omega)=((m+1)/2)s^l P_{(m-l)/2}^{(0,l)}(2s^2-1)Y_l(\omega).$$

iii) *Let $N=3$ and $\nu=\frac{3}{2}$:*

$$V(s\cdot\omega)=\pi^{-1}\binom{m+2}{2}s^l P_{(m-l)/2}^{(0,l+1/2)}(2s^2-1)Y_l(\omega).$$

The condition $\nu>N/2-1$ in Theorem 3.1 results in the fact that there are no inversion formulas for the C_m^ν with $\nu\leq N/2-1$, especially for the Chebyshev polynomials of the first kind in two dimensions and for the Legendre polynomials in three dimensions.

COROLLARY 3.3. *Let* $W_\lambda(x)=(1-|x|^2)^{\lambda-N/2}$, *and let* $\tilde{c}(N,\nu)=c(N,0,\nu,0)^{-1}$ *with* c *from* (3.3),

$$\tilde{c}(N,\nu)=\pi^{(N-1)/2}\Gamma(\nu-N/2+1)/\Gamma(\nu+1/2).$$

Then

$$RW_\nu=\tilde{c}(N,\nu)w_\nu.$$

This generalization of a result of Davison–Grünbaum [3] for $N=2$ follows from Theorem 3.1 with $m=l=0$.

Using the fact that the radial part of the function V from (3.1) is mapped into the part of $\mathcal{R}V$ which is dependent on s by the Gegenbauer transform (see Deans [4] and Ludwig [18]), we can interpret Theorem 3.1 in the following way.

COROLLARY 3.4. *Let* $\varphi^{\nu,N}_{m,l}(s)=(1-s^2)^{\nu-N/2}Q^{\nu,N}_{m,l}(s)$ *for* $0\le s\le 1$. *Then*

$$\mathcal{C}^N_l\varphi^{\nu,N}_{m,l}(s)=[c(N,m,\nu,l)]^{-1}w_\nu(s)C^\nu_m(s)$$

with the constant c *from* (3.3) *and the Gegenbauer transform* $\mathcal{C}^N_l$ *defined as*

$$\mathcal{C}^N_l\varphi(s)=\mathrm{vol}(s^{N-2})/C_l^{N/2-1}(1)\int_s^1\varphi(t)w_{N/2-1}(s/t)C_l^{N/2-1}(s/t)t^{N-2}\,dt.$$

4. Inversion formula. With the results from the preceding sections we are now able to give a representation of a function using its Radon transform. But in contrast to Radon's original inversion formula [24] which says that the inverse transform is a composition of a pseudo-differential operator and a back projection (see Ludwig [18, Thm. 1.1]), we consider here the Fourier series of f with respect to the orthogonal functions defined in (3.1). This enables us to study the behaviour of this transform under various aspects.

THEOREM 4.1. *Let* f *be in* $L_2(\Omega,W_\nu^{-1})$, $\nu>N/2-1$, *and let* $\mathcal{R}f$ *be its Radon transform. Let* $V_{mlk}=\mathcal{R}^{-1}\Phi_{mlk}$ *with* $\Phi_{mlk}=w_\nu C^\nu_m Y_{lk}$, $m\ge 0$, $0\le l\le m$ *with* $m+l$ *even*, $k=1,\cdots,M(N,l)$. *Then*

$$f=\sum_{m=0}^{\infty}h_{m\nu}^{-1}\sum_{\substack{l=0\\ m+l\ \mathit{even}}}^{m}\sum_{k=1}^{M(N,l)}\langle\mathcal{R}f,\Phi_{mlk}\rangle_{w_\nu^{-1}}V_{mlk}\tag{4.1}$$

or equivalently

$$f(s\cdot\omega)=(1-s^2)^{\nu-N/2}\sum_{m=0}^{\infty}\sum_{\substack{l=0\\ m+l\ \mathit{even}}}^{m}c(N,m,\nu,l)\times Q^{\nu,N}_{m,l}(s)\sum_{k=1}^{M(N,l)}d_{mlk}Y_{lk}(\omega)\tag{4.2}$$

with c, $Q^{\nu,N}_{m,l}$ *from Theorem* 3.1 *and* d_{mlk}, $h_{m\nu}$ *from Lemma* 2.3.

LEMMA 4.2. *The*

$$\{Q^{\nu,N}_{m,l}Y_{lk}\colon m\ge 0, 0\le l\le m, m+l\ \mathit{even}, k=1,\cdots,M(N,l)\}$$

form a complete set of orthogonal polynomials over Ω *with respect to the weight* $W_\nu(x)=(1-|x|^2)^{\nu-N/2}$. *Similarly the* $\tilde{V}_{ml}=W_\nu Q^{\nu,N}_{m,l}Y_{lk}$ *form a complete set of orthogonal functions over* Ω *with respect to the weight* W_ν^{-1}.

Proof. This result is a consequence of the completeness and the orthogonality of both the Jacobi polynomials and the spherical harmonics. Let $\lambda=\nu+(1-N)/2$,

$$\begin{aligned} I&=\int_{\Omega}\tilde{V}_{mlk}(x)\tilde{V}_{m'l'k'}(x)W_{\nu}^{-1}(x)\,dx \\ &=\int_{S^{N-1}}Y_{lk}(\omega)Y_{l'k'}(\omega)\,d\omega\int_0^1 w_{\lambda}(s)Q_{m,l}^{\nu,N}(s)Q_{m',l'}^{\nu,N}(s)s^{N-1}\,ds. \end{aligned}$$

From the orthonormality of the Y_{lk} it follows that

$$\int_{S^{N-1}}Y_{lk}(\omega)Y_{l'k'}(\omega)\,d\omega=\delta_{ll'}\delta_{kk'}.$$

With the representation of the $Q_{m,l}^{\nu,N}$, $l=l'$, and the change of variables $t=2s^2-1$ the second integral is transformed into

$$c'\int_{-1}^{1}(1+t)^{l+N/2-1}(1-t)^{-\lambda+1/2+2\nu-N}P_{(m-l)/2}^{(\alpha,\beta)}(t)P_{(m'-l)/2}^{(\alpha,\beta)}(t)\,dt$$

with $\alpha=\nu-N/2$ and $\beta=l+N/2-1$. The orthogonality of $P_n^{(\alpha,\beta)}$ on $[-1,1]$ with respect to $(1-t)^{\beta}(1+t)^{\beta}$ completes the proof.

Proof of Theorem 4.1. Theorem 4.1 follows from the linearity of the Radon transform, its continuity from $L_2(\Omega, W_{\nu}^{-1})$ into $L_2(Z, w_{\nu}^{-1})$, and the uniqueness of the Fourier coefficients together with Theorem 3.1 and Lemma 4.2.

COROLLARY 4.3. *The* V_{mlk} *of Theorem* 4.1 *are the eigenfunctions of* $\mathfrak{R}^*\mathfrak{R}$, *where* $\mathfrak{R}^*$ *is the adjoint operator of* $\mathfrak{R}: L_2(\Omega, W_{\nu}^{-1})\to L_2(Z, w_{\nu}^{-1})$:

$$\mathfrak{R}^*g(x)=W_{\nu}(x)\int_{S^{N-1}}w_{\nu}^{-1}(x\cdot\omega)g(x\cdot\omega,\omega)\,d\omega.$$

Proof. The Fourier coefficients of f with respect ot V_{mlk} are defined as

$$\langle f, V_{mlk}\rangle_{L_2(\Omega,W_{\nu}^{-1})}/\langle V_{mlk}, V_{mlk}\rangle_{L_2(\Omega,W_{\nu}^{-1})}=\langle \mathfrak{R}f, \mathfrak{R}V_{mlk}\rangle_{L_2(Z,w_{\nu}^{-1})}h_{m\nu}^{-1}$$

because of Theorem 4.1. The definition of the adjoint operator leads to

$$\langle f, V_{mlk}\rangle_{L_2(\Omega,W_{\nu}^{-1})}=c_{mlk}\langle f, \mathfrak{R}^*\mathfrak{R}V_{mlk}\rangle_{L_2(\Omega,W_{\nu}^{-1})}$$

for suitable constants c_{mlk}.

5. The null space of the Radon transform for finitely many projections. In the following we consider complete projections of a function f in the directions ω, i.e. we assume that $\mathfrak{R}f(\cdot,\omega)$ is known. In practice this can be approximated by sufficiently fine sampling in the radial direction. It is shown by Smith–Solmon–Wagner [28] that a function is uniquely determined by its Radon transform for any infinite set of directions which is not contained in a proper algebraic variety on S^{N-1}. Here it is important that complete projections of a function with compact support are given. In the case of hollow projections of a function without compact support, i.e. $\mathfrak{R}f(s,\omega)$ is known for all $|s|\geq c>0$, this result is no longer true. See Quinto [23] and Shepp–Kruskal [27].

The condition that the projections in an infinite number of directions are known is clearly not fulfilled in practice. The linearity of the Radon transform implies that the transform has a nontrivial null space for finitely many projections. The functions in this space are "invisible" from those directions and are therefore called "ghosts". In practice this means that the searched for density cannot be uniquely determined by any measurement, which led to the reformulation of the precise mathematical statement into: "A finite set of radiographs tells nothing at all", (Smith–Solmon–Wagner [28, Thm. 4.2]). The contradiction of this assertion to practical experience gave rise to the study of the possible deviations of the reconstructed distribution from the original. Leahy–Smith–Solmon [12] have proven another negative result, to wit that the uncertainty in the reconstruction cannot be reduced by imposing regularity conditions on the null space. Positive results are known in the two-dimensional case. To sketch them briefly we assume that p complete projections are given. The proof in [28] does not lead to characterizations of the functions in the null space, since it is based on the existence of solutions of a certain set of partial differential equations. First uniqueness results are known when the solution is an element of a finite dimensional space of pixel functions. This result has been proven by Frieder–Herman [6] in connection with ART, a reconstruction technique where the projection of the solution on a space of piecewise constant functions on a fixed grid is determined, see also Smith–Peters–Bates [29]. For the study of the general uniqueness problem these spaces form a relatively unnatural environment, see Katz [11].

The spectrum of the functions in the null space was considered by Logan [14]. He has shown that most of its power is lying outside a circle around zero with radius $p-\varepsilon$, $\varepsilon>0$. This means that the functions consist mostly of high frequency components. Nearest to reality is Natterer [22], who assumes that only approximations to $\mathfrak{R}f(s_j,\omega_j)$, $j=1,\cdots,q$ with accuracy ε are known. Restricted to the situation that we consider here, i.e., p complete projections are given, he shows that bounds on the derivatives on the functions in the null space reduce their magnitude. More precisely, let f be a ghost with H_0^α norm 1, then

$$(5.1)\qquad \|f\|_{L_2(\Omega)}\le cp^{-\alpha},\qquad c \text{ independent of } p.$$

Besides this he shows that there is a ghost with $H_0^\alpha(\Omega)$ norm 1 and

$$(5.2)\qquad \|f\|_{L_2(\Omega)}=c'p^{-\alpha},\qquad c' \text{ independent of } p,$$

which means that (5.1) is sharp. Finally, in Louis [17] the functions in the null space are constructed by giving their series expansion in terms of Zernike polynomials, i.e., case $N=2$, $\nu=1$ of the preceding section. The relation of the amount of data to the achievable resolution is discussed, and pictures of ghosts are given. In particular, a function which fulfills (5.2) if $f(s\cdot\omega)=(1-s^2)^{1/2}C_p^1(s)q_p(\omega)$, where q_p is a polynomial of degree p which vanishes in the given directions. The behaviour of the functions in the null space is essentially determined by the lowest degree in the expansion of the form (4.1) or (2.9) for the Radon transform. Separation of the variables allows study of the angular part $q_m(\omega)$. The question is which of the $q_m(\omega)$ are uniquely determined by the measurements, or equivalently, which of them are identically equal to zero for the ghosts. This corresponds to an interpolation problem on S^{N-1}. The main difference between the cases $N=2$ and $N\ge3$ is that in the former case it is sufficient that the p given directions are different in order to guarantee that $q_m\equiv0$ for $m<p$. In higher dimensions we also have, besides this obvious condition, geometric restrictions on the directions. Furthermore, the number of vanishing coefficients q_m is also dependent on the dimension. Now let A be the set of given directions

$$(5.3)\qquad A=\left\{\omega_j\in S^{N-1}, 1\le j\le p, \omega_j\ne\pm\omega_k \text{ for } j\ne k\right\}.$$

The Radon transform $\mathcal{R}f(\cdot,\omega)$ of an L_2 function with compact support in Ω is in $L_2(-1,1)$; see Smith–Solmon–Wagner [28]. Therefore we can define the null space as

$$\mathcal{N}_A=\{f\in L_2(\Omega):\mathcal{R}f(s,\omega)=0 \text{ for } s \text{ a.e. and for all } \omega\in A\}. \tag{5.4}$$

Using the representation (2.10) for the functions in the range of $\mathcal{R}$ with $\nu=N/2$

$$\mathcal{R}f(s,\omega)=w_\nu(s)\sum_{m=0}^{\infty}C_m^\nu(s)q_m(\omega),$$

we conclude that for $f\in\mathcal{N}_A$

$$q_m(\omega)=0, \quad \text{for } \omega\in A \tag{5.5}$$

because of the linear independence of the Gegenbauer polynomials. (5.5) is a system of p linear homogeneous equations for the coefficients of the q_m. The condition that as many q_m as possible vanish identically is equivalent to the fact that the matrix of the equations has full rank. In order to solve this problem we consider first the number of unknowns for fixed m.

LEMMA 5.1. *Let* $\mathcal{P}_m=\operatorname{span}\{Y_{lk}:0\le l\le m, l+m \text{ even}, 1\le k\le M(N,l)\}$. *Then*

$$\dim\mathcal{P}_m=\binom{m+N-1}{N-1}. \tag{5.6}$$

The proof is a straightforward induction from $M(N,l)$.

THEOREM 5.2. *Let* $n>0$ *and* $p\ge\dim\mathcal{P}_{n-1}=\binom{n+N-2}{N-1}$. *Let A not be contained in an algebraic variety of degree* $<n$; *i.e., there is no* $q\in\mathcal{P}_{n-1}$, $q\not\equiv0$ *with* $q(\omega)=0$ *for all* $\omega\in A$. *Let* $f\in\mathcal{N}_A$. *Then*

$$\mathcal{R}f(s,\omega)=w_\nu(s)\sum_{m=n}^{\infty}C_m^\nu(s)q_m(\omega) \tag{5.7}$$

where $q_m\in\mathcal{P}_m$ *with* $q_m(\omega)=0$ *for all* $\omega\in A$.

Proof. Let $m\ge0$. Then (5.5) is a system of p equations in $\binom{m+N-1}{N-1}$ unknowns. For $m<n$ there are at least as many equations as unknowns and the matrix has full rank if and only if the directions are not lying on an algebraic variety of order m, i.e. there is no polynomial in $\mathcal{P}_m$ such that $p(\omega)=0$ for all $\omega\in A$. Therefore the q_m vanish identically for $m\le n-1$. For $m\ge n$ there exist nontrivial solutions of the underdetermined system (5.5), which means that there are $q_m\not\equiv0$ with $q_m(\omega)=0$ for $\omega\in A$.

Finally applying Theorem 4.1 we can restate the above result for the functions in $\mathcal{N}_A$.

THEOREM 5.3. *Let* $n>0$ *and* $p\ge\dim\mathcal{P}_{n-1}$. *Let A not be contained in an algebraic variety of degree* $<n$. *Let* $f\in\mathcal{N}_A$. *Then*

$$f(s\cdot\omega)=(1-s^2)^{\nu-N/2}\sum_{m=n}^{\infty}\sum_{\substack{l=0\\ m+l\ \text{even}}}^{m}Q_{m,l}^{\nu,N}(s)\sum_{k=1}^{M(N,l)}Y_{lk}(\omega)e_{mlk} \tag{5.8}$$

with

$$\sum_{\substack{l=0\\ m+l\ \text{even}}}^{m}\sum_{k=1}^{M(N,l)}\frac{e_{mlk}}{c(N,m,\nu,l)}Y_{lk}(\omega)=0 \quad \text{for } \omega\in A,$$

where $Q_{m,l}^{\nu,N}$ *are the polynomials of degree m defined in* (3.2).

Analogously to the case $N=2$ in Louis [17] we have given an expansion of the functions in the null space in terms of the polynomials $Q_{m,l}^{\nu,N}$ which are the generalization of the Zernike polynomials $Q_{m,l}^{1,2}$. If the directions fulfill the optimality condition of Theorem 5.3, then the expansion starts with the polynomial of degree n where $n=O(p^{1/(N-1)})$ according to formula (5.6). This has the consequence that the expansion of the Fourier transform of these functions in terms of Bessel functions starts with the Bessel function of order $n+\nu$ with $\nu>N/2-1$. The fact that the Bessel function $J_k(s)$ of order $k\geq n+\nu$ are small for $|s|<m+\nu$ can be interpreted in the way that the Fourier transform of the ghosts is also very small in the circle around zero with radius $n+N/2-1$; thus we have a generalization of Logan's result [14] to higher dimensions. One way of expressing that is the resolution in the reconstructed image is $2/(n+N/2-1)$.

6. Applications to nuclear magnetic resonance. The general results in the preceding chapters are motivated by their applications in two and three dimensions. Because of its enormous impact in radiology the case $N=2$, transaxial x-ray tomography, has been extensively studied. In contrast, NMR zeugmatography is presently only at the stage of clinical trials for the first commercial scanners. The data measured in NMR pertain to integrals of the unknown density over planes in R^3. For the physical background see Marr–Chen–Lauterbur [21], who also report on reconstructions from actual measurements. For numerical tests on a mathematical phantom see Shepp [26]. The problem of nonuniqueness in NMR reconstruction is tied to the choice of the directions as indicated in the last chapter. Because of the rapidly increasing importance of this field we now apply our results to the case $N=3$.

We introduce the usual spherical coordinates θ and ϕ to label $\omega\in S^2$ so that

$$\omega=\omega(\theta,\phi)=(\sin\theta\cos\phi,\sin\theta\sin\phi,\cos\theta),$$

with $\theta\in[0,\pi[$, $\phi\in[0,2\pi[$. Here we have $M(3,l)=2l+1$ and

$$\dim\mathcal{P}_m=\binom{m+2}{2}=(m+1)(m+2)/2, \tag{6.1}$$

where $\mathcal{P}_m$ is defined in Lemma 5.1.

According to Theorem 5.3 we need $n(n+1)/2$ complete projections if we want to determine uniquely the first n coefficients q_m, $m=0,\cdots,n-1$, in the series expansion of the searched for density. Observing that

$$s\omega(\theta,\phi)=-s\omega(\pi-\theta,\phi+\pi), \tag{6.2}$$

we have to restrict one of the coordinates to avoid redundancy in the measurements.

The simplest case in which we can meet the conditions of Theorem 5.3 are

$$\begin{aligned} &\theta_j\in]0,\pi[, \quad && j=1,\cdots,(n+1)/2 \text{ (or } n/2), \quad && \theta_j\neq\theta_k, \\ &\phi_i\in]0,\pi[, && i=1,\cdots,n \text{ (or } n+1), && \phi_i\neq\phi_k \end{aligned} \tag{6.3}$$

$$\text{for } n \text{ odd (or } n \text{ even).}$$

Applying the results of the last section we can recapitulate.

THEOREM 6.1. *Let $n(n+1)/2$ complete projections be given in the directions $\omega_{ji}=\omega(\theta_j,\phi_i)$, where the θ_j, ϕ_i fulfill* (6.3) *and ω_{ji} must lie on no algebraic variety of degree $<n$.*

Then the expansion coefficients $q_m(\omega)$, $0\leq m<n$, in (2.10) *are uniquely determined and the expansion of the ghosts in terms of $Q_{m,l}^{\nu,N}$ starts with a polynomial of degree $\geq n$.*

In the papers of Marr–Chen–Lauterbur [21] and Shepp [26] the angles θ and ϕ have been chosen equidistributed, which means that the directions are more concentrated at the poles than at the equator. In the practical experiments of Marr–Chen–Lauterbur the angles have been $\theta_i=\phi_i=6^\circ+(i-1)12^\circ$, $i=1,\cdots,30$, i.e. both angles are equidistributed over $]0,2\pi[$. The two-fold redundancy is motivated by the improvement of the final signal-to-noise ratio. According to Theorem 6.1 the first 15 expansion coefficients are then uniquely determined, so that the Fourier transform is nearly exact in a circle around 0 with radius 15, see the discussion at the end of the last chapter.

The numerical experiments by Shepp [26] have been performed for $\theta_j=(j-\frac{1}{2})\pi/n$, $j=1,\cdots,n$, $\phi_i=i2\pi/n$, $i=0,\cdots,n-1$ with $n=25$, 69, 99. Because of the choice of an odd n, there is no redundancy in the data, the same projections can be characterized by $\phi_i=i\pi/n$, $i=0,\cdots,n-1$. In these examples the first n coefficients are uniquely determined, a result which can also be achieved by using only $(n+1)/2$ different values for θ.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] A. M. Cormack, *Representations of a function by its line integrals, with some radiological applications*, II, J. Appl. Physics, 35 (1964), pp. 2908–2913.

[3] M. E. Davison and F. A. Grünbaum, *Convolution algorithms for arbitrary projection angles*, IEEE Trans. Nucl. Sci., NS26 (1976), pp. 1670–1673.

[4] S. R. Deans, *Gegenbauer transforms via the Radon transform*, this Journal, 10 (1979), pp. 577–585.

[5] A. Erdélyi, W. Magnus, R. Oberhettinger and F. Tricomi, *Higher Transcendental Functions*, Vol. 2, McGraw-Hill, New York, 1953.

[6] G. Frieder and G. T. Herman, *Resolution in reconstructing objects from electron micrographs*, J. Theoret. Biol., 33 (1971), pp. 189–211.

[7] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series and Products*, Academic Press, New York, 1965.

[8] S. Helgason, *The Radon transform on Euclidean spaces, compact two-point homogeneous spaces, and Grassman manifolds*, Acta Math., 113 (1965), pp. 153–180.

[9] G. T. Herman, *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*, Academic Press, New York, 1980.

[10] G. T. Herman and F. Natterer, eds., *Mathematical Aspects of Computerized Tomography*, Lecture Notes in Medical Informatics, 8, Springer, Berlin, 1981.

[11] M. B. Katz, *Questions of Uniqueness and Resolution in Reconstruction from Projections*, Lecture Notes in Biomathematics 26, Springer, Berlin, 1978.

[12] J. K. Leahy, K. T. Smith and D. C. Solmon, *Uniqueness, nonuniqueness and inversion in the x-ray and Radon problems*, Proc. International Symposium on Ill-Posed Problems: Theory and Practice, Univ. Delaware, Newark, October 2–6, 1979.

[13] R. M. Lewitt and R. H. T. Bates, *Image reconstruction from projections, III: projection completion methods (theory)*, Optik, 50 (1978), pp. 189–204.

[14] B. F. Logan, *The uncertainty principle in reconstructing functions from projections*, Duke Math. J., 42 (1975), pp. 661–706.

[15] A. K. Louis, *Picture reconstruction from projections in restricted range*, Math. Meth. Appl. Sci., 2 (1980), pp. 209–220.

[16] ______, *Approximation of the Radon transform from samples in limited range*, in [10], pp. 127–139.

[17] ______, *Ghosts in tomography—the null space of the Radon transform*, Math. Meth. Appl. Sci., 3 (1981), pp. 1–10.

[18] D. Ludwig, *The Radon transform on Euclidean spaces*, Comm. Pure Appl. Math., 19 (1966), pp. 49–81.

[19] R. B. Marr, *On the reconstruction of a function on a circular domain from a sampling of its line integrals*, J. Math. Anal. Appl., 19 (1974), pp. 357–374.

[20] ______, *An overview of image reconstruction*, Proc. International Symposium on Ill-Posed Problems: Theory and Practice, Univ. Delaware, Newark, October 2–6, 1979.

[21] R. B. MARR, C.-N. CHEN AND P. C. LAUTERBUR, *On two approaches to* 3D *reconstruction in* NMR *zeugmatography*, in [8], pp. 225–240.

[22] F. NATTERER, *A Sobolev space analysis of picture reconstruction*, SIAM J. Appl. Math., 39 (1980), pp. 402–411.

[23] E. T. QUINTO, *Null spaces and ranges for the classical and spherical Radon transforms*, J. Math. Anal. Appl., 90 (1982), pp. 408–429.

[24] J. RADON, *Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten*, Ber. Verh. Sächs. Akad. Wiss., Leipzig, 69 (1917), pp. 262–277.

[25] R. T. SEELEY, *Spherical harmonics*, Amer. Math. Monthly, 73 (1966), pp. 115–121.

[26] L. A. SHEPP, *Computerized tomography and nuclear magnetic resonance*, J. Comp. Assist. Tomography, 4 (1980), pp. 94–107.

[27] L. A. SHEPP AND J. B. KRUSKAL, *Computerized tomography: the new medical x-ray technology*, Amer. Math. Monthly, 85 (1978), pp. 420–439.

[28] K. T. SMITH, D. C. SOLMON AND S. L. WAGNER, *Practical and mathematical aspects of the problem of reconstructing objects from radiographs*, Bull. AMS, 83 (1977), pp. 1227–1270.

[29] P. R. SMITH, T. M. PETERS AND R. H. T. BATES, *Image reconstruction from finite numbers of projections*, J. Phys. A., 6 (1973), pp. 361–382.

[30] H. TRIEBEL, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, 1978.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics
001

BOUNDARY AND INTERIOR LAYER PHENOMENA FOR SINGULARLY PERTURBED SYSTEMS*

WALTER G. KELLEY†

Abstract. Sufficient conditions are given for the existence of a solution to a singularly perturbed boundary value problem for a system of nonlinear equations which exhibits boundary or interior layer behavior for small positive values of the parameter. Examples are included to illustrate the results.

1. Introduction. In this paper, we establish sufficient conditions for the existence of solutions of boundary value problems

$$\varepsilon\mathbf{z}''=\mathbf{H}(t,\mathbf{z}), \tag{1}$$

$$\mathbf{z}(0)=\mathbf{A}, \qquad \mathbf{z}(1)=\mathbf{B}, \tag{2}$$

which exhibit interior or boundary layer behavior as $\varepsilon\to 0$. Here ε is a small positive parameter and $\mathbf{z},\mathbf{H},\mathbf{A}$ and $\mathbf{B}$ are vectors. In this setting, a solution of (1), (2) is said to exhibit interior (boundary) layer behavior if one or more of its components experiences a rapid transition in a neighborhood of a point interior to (on the boundary of) $[0,1]$ for small ε.

Howes [3] and Kelley [5] have examined the question of boundary layers for these problems. Their approach involves the comparison of the vector problem (1), (2) with a scalar problem which is known to have a solution exhibiting boundary layer behavior. The results of the present paper will apply to situations in which no such comparison is possible. The case of interior layer behavior does not seem to have been investigated for the problem (1), (2), except in the scalar case (see Howes [2]).

2. Interior layers. In order to make the discussion as simple as possible, we will consider only the case of two-dimensional vectors and write

$$\mathbf{H}=\begin{bmatrix} F \\ G \end{bmatrix} \quad \text{and} \quad \mathbf{z}=\begin{bmatrix} x \\ y \end{bmatrix}.$$

We will also assume in this section that there is a nonintersecting pair of curves in $(t,\mathbf{z})$ space where H vanishes and that one of the curves satisfies the boundary condition at $t=0$ and the other curve satisfies the boundary condition at $t=1$. The possibility of a boundary layer is thus excluded in this section but will be considered in §3. To simplify the geometry, let suitable t-dependent translations and rotations of $\mathbf{z}$ space be made so that $\mathbf{H}$ vanishes for $\mathbf{z}=\mathbf{0}$ and for $x=0$, $y=v(t)>0$, $0\leq t\leq 1$. These transformations introduce small perturbations into the problem which can be neglected in the analysis. Note that in (2) we now have $\mathbf{A}=\mathbf{0}$ and $\mathbf{B}=\mathbf{v}(1)$, where

$$\mathbf{v}(t)=\begin{bmatrix} 0 \\ v(t) \end{bmatrix}.$$

The following theorem gives sufficient conditions for the existence of a solution of (1), (2) which has an internal layer as $\varepsilon\to 0$.

THEOREM 1. *Assume*:

(a) $\mathbf{H}(t,\mathbf{0})=\mathbf{H}(t,\mathbf{v})=\mathbf{0}$ *for* $0\leq t\leq 1$;

* Received by the editors April 13, 1982, and in revised form May 19, 1983.

† Mathematics Department, University of Oklahoma, Norman, Oklahoma 73019.

(b) *there exist* $0<t_1<t_2<1$ *so that* $\int_0^u G(t_1,0,s)\,ds>0$ *for* $0<u\le v(t_1)$ *and* $\int_u^{v(t_2)} G(t_2,0,s)\,ds<0$ *for* $0\le u<v(t_2)$;

(c) *there is a* $D>0$ *and a class* C^2 *symmetric positive definite matrix function* $\mathbf{Q}(t)$ *so that*

$$\mathbf{z}^T\mathbf{Q}(t)\mathbf{J}(t,\mathbf{0})\mathbf{z}>D\mathbf{z}^T\mathbf{Q}(t)\mathbf{z},\qquad 0\le t\le t_1,$$
$$(\mathbf{z}-\mathbf{v})^T\mathbf{Q}(t)\mathbf{J}(t,\mathbf{v})(\mathbf{z}-\mathbf{v})>D(\mathbf{z}-\mathbf{v})^T\mathbf{Q}(t)(\mathbf{z}-\mathbf{v}),\qquad t_2\le t\le 1,$$

for all $\mathbf{z}$, *where* $\mathbf{J}$ *is the Jacobian matrix for* $\mathbf{H}$;

(d) $G_y(t,\mathbf{0})>0$ *and* $G_y(t,\mathbf{v})>0$ *for* $t_1\le t\le t_2$, $F_x(t,\mathbf{z})>0$ *for* $(t,\mathbf{z})\in R$ *and*

$$\frac{\max_R|F_y|\max_R|G_x|}{\min_R F_x}$$

is smaller than some computable positive number, where $R=\{(t,x,y): t_1\le t\le t_2,\ x$ *in some suitable bounded interval,* $0\le y\le v(t)\}$.

Then (1), (2) *has a solution* $\mathbf{z}(t,\varepsilon)$ *for small* $\varepsilon>0$ *so that* $\mathbf{z}(t,\varepsilon)=O(\varepsilon)$ *for* $0\le t\le t_1$ *and* $\mathbf{z}(t,\varepsilon)-\mathbf{v}(t)=O(\varepsilon)$ *for* $t_2\le t\le 1$.

Remarks.

(1) Suppose it is known that (1), (2) has a solution $\mathbf{z}(t,\varepsilon)$ with interior layer located asymptotically at $t=t^*$ as $\varepsilon\to 0$. Then it is not hard to show that the line integral $\int_C \mathbf{H}^T\cdot d\mathbf{z}$ is zero at $t=t^*$, where C is the limiting value of the solution curve at $t=t^*$ as $\varepsilon\to 0$. In Theorem 1, (b) requires the line integral of $\mathbf{H}$ along the line segment from $\mathbf{0}$ to $\mathbf{v}$ to change sign on the interval $[t_1,t_2]$. The interior layer will be located between t_1 and t_2, but the exact position may be difficult to pinpoint since the limiting value of the solution curve is unknown. We will explore this problem in Example 1.

(2) Hypothesis (c) is a stability assumption for the reduced solutions $\mathbf{0}$ and $\mathbf{v}$. For each t, there is a symmetric, positive definite matrix $\mathbf{Q}(t)$ for which the inequalities hold if, and only if, the real parts of the eigenvalues of $\mathbf{J}$ are all positive. (See Hirsch and Smale [1, pp. 145–149].)

Proof. The proof involves the definition of bounding surfaces for (1), (2) (see Knobloch and Schmitt [6] and Kelley [4]). These surfaces are defined first for $0\le t\le t_1$, where they are chosen to be the zeroes of the bounding functions $\varphi_1=y-\chi-\lambda$, $\varphi_2=-y-\lambda$, $\varphi_3=x-B$, $\varphi_4=-x-B$ and $\varphi=\mathbf{z}^T\mathbf{Q}(t)\mathbf{z}-r^2(t,\varepsilon)$; here χ is the unique increasing positive solution of

$$\varepsilon\chi''=G(t_1,0,\chi)-\rho\chi,$$
$$\chi(t_1,\varepsilon)=v(t_1),\qquad \lim_{\varepsilon\to 0}\chi(t,\varepsilon)=0\qquad (t<t_1),$$

r is of the form

$$r(t,\varepsilon)=P\exp\left[\sqrt{\frac{D-\eta}{\varepsilon}}\,(t-t_1)\right]+C\varepsilon,$$

and $\lambda, B, \rho, P, \eta$ and C are positive constants. The existence of such a χ follows from (b) when ρ is small.

In order for these to be bounding surfaces, they must satisfy certain differential inequalities which force all points of the surface bounding the region where all the bounding functions are negative to be egress points for (1). The general inequalities are given in Knobloch and Schmitt [6, §4], so here we merely display the form of these inequalities required for our special bounding functions. Our result is obtained by applying Corollary 3 of Theorem 4.1 in [6] with $\Omega=\{\phi<0\}\cap(\cap_{i=1}^{4}\{\phi_i<0\})$.

For φ_1 we require that

$$G(t,x,\chi+\lambda)-\varepsilon\chi''\geq 0,$$

when $|x|\leq B$ and $\varphi\geq 0$. Letting $\Gamma\equiv G(t,0,\chi)-\varepsilon\chi''$, we have by the mean value theorem

$$\begin{aligned}(3)\qquad G(t,x,\chi+\lambda)-\varepsilon\chi''&=\Gamma+G_x(t,*,\chi)x+G_y(t,0,\Delta)\lambda\\ &\geq\Gamma+\lambda G_y(t,0,\Delta)-|G_x(t,*,\chi)|B,\end{aligned}$$

where $*$ is between 0 and x, and Δ is between χ and $\chi+\lambda$. Thus the requirement on φ_1 is that (3)≥ 0 for $t_1-t=o(1)$, since for larger values of t_1-t, $\varphi=0$ is the bounding surface.

Similar calculations for φ_2,φ_3 and φ_4 result in the respective requirements

$$(4)\qquad \lambda G_y(t,0,\sim)-|G_x(t,\sim,-\lambda)|B\geq 0,$$

$$(5)\qquad BF_x(t,\sim,0)-|F_y(t,B,\sim)|(v(t_1)+\lambda)\geq 0,$$

$$(6)\qquad BF_x(t,\sim,0)-|F_y(t,-B,\sim)|(v(t_1)+\lambda)\geq 0,$$

for $t_1-t=o(1)$, where $\sim$ represents various intermediate choices of the variables.

The required inequality for φ is:

$$2\mathbf{z}^T\mathbf{QH}+\frac{\varepsilon}{2r^2}(\mathbf{z}^T\mathbf{Q}'\mathbf{z})^2+\varepsilon\mathbf{z}^T(\mathbf{Q}''-2\mathbf{Q}'\mathbf{Q}^{-1}\mathbf{Q}')\mathbf{z}+2\frac{\varepsilon}{r}\mathbf{z}^TQ'\mathbf{z}r'-2\varepsilon rr''\geq 0,$$

for $\mathbf{z}^T\mathbf{Qz}=r^2$. (This inequality is obtained from (4.4) and (4.12)′ in Knobloch and Schmitt [6].) We rewrite the inequality in the form

$$(7)\qquad r^{-1}\mathbf{z}^T\mathbf{QH}-\varepsilon r''+O(\varepsilon r)+O(\varepsilon r')\geq 0$$

for $\mathbf{z}^T\mathbf{Qz}=r^2$. From hypothesis (c) and the mean value theorem,

$$r^{-1}\mathbf{z}^T\mathbf{QH}=r^{-1}\mathbf{z}^T\mathbf{QJ}(\sim)\mathbf{z}\geq Dr^{-1}\mathbf{z}^T\mathbf{Qz}=Dr,$$

when $\mathbf{z}^T\mathbf{Qz}=r^2$ and $|\mathbf{z}|$ is sufficiently small, say $|\mathbf{z}|<E$. Then (7) is satisfied if

$$\begin{aligned}&Dr-\varepsilon r''+O(\varepsilon r)+O(\varepsilon r')\\ &\quad=\eta P\exp\left[\sqrt{\frac{D-\eta}{\varepsilon}}\,(t-t_1)\right]+DC\varepsilon+O(\varepsilon r)+O(\varepsilon r')>0,\end{aligned}$$

which is valid if ε is sufficiently small, C is sufficiently large, $|\mathbf{z}|<E$ and $\eta<D$.

We can choose small positive constants λ and μ so that:

$$(8)\qquad \Gamma+\lambda G_y(t,0,\Delta)>\lambda\mu,\quad t_1-t=o(1),\quad \chi<\Delta<\chi+\lambda,$$

$$(9)\qquad r^{-1}\mathbf{z}^T\mathbf{QH}>0,\qquad |\mathbf{z}|\leq\lambda\sqrt{1+M^{-2}\mu^2}\,,$$

$$(10)\qquad G_y(t_1,0,\Delta)>\mu,\qquad |\Delta|<\lambda,$$

where $M=\max_R|G_x|$. It is possible to satisfy (8) for sufficiently small λ and μ because $\Gamma=O(\rho\chi)$ is the dominant term and is positive for $\sqrt{\varepsilon}=O(\chi)$ and small λ, and G_y is positive for small χ.

To make hypothesis (d) precise, we now require

$$\frac{\lambda\mu}{v(t)+\lambda} > \frac{\max_R|F_y|\max_R|G_x|}{\min_R F_x}, \qquad t_1 \le t \le t_2. \tag{11}$$

Then we can choose B so that

$$\frac{\lambda\mu}{\max_R|G_x|} > B > \frac{\max_R|F_y|}{\min_R F_x}(v(t)+\lambda), \qquad t_1 \le t \le t_2. \tag{12}$$

From (8) and (12), (3)>0 for $t_1 - t = o(1)$, (4) follows from (10) and (5) and (6) follow from (12).

Recall that (7) has been established only in case $|\mathbf{z}| < E$. We must show that for $|\mathbf{z}| \ge E$, the surface of the region where all the bounding functions are negative does not intersect the zero set of φ. Equivalently, we show that the intersection of the zero set of φ with the zero set of each φ_i, $i = 1,2,3,4$, occurs where $|\mathbf{z}| < E$. From (9), (12) we can choose

$$E > \lambda\sqrt{1+M^{-2}\mu^2} > \sqrt{\lambda^2+B^2}.$$

It follows that the intersection of $\{\varphi=0\}$ with $\{\varphi_2=0\}$ and $\{\varphi_i=0, y\le 0\}$, $i=3,4$, occurs where $|\mathbf{z}|<E$. Furthermore, by selecting P sufficiently large in the definition of r and making $D-\eta > G_y(t_1,0,0)-\rho$, we can ensure that the intersection of $\{\varphi=0\}$ with $\{\varphi_1=0\}$ and $\{\varphi_i=0, y\ge 0\}$, $i=3,4$, occurs where $|\mathbf{z}|<E$. The construction is now complete for the interval $[0,t_1]$.

For the interval $[t_1,t_2]$, we define bounding functions φ_2, φ_3 and φ_4 as above and define $\varphi_1 = y - v(t) - \lambda$. Then inequalities (4), (5) and (6) must be satisfied, and for φ_1,

$$\begin{aligned} G(t,x,v(t)+\lambda) - \varepsilon v'' &= G(t,0,v(t)) + G_x(t,*,v(t))x + G_y(t,0,\Delta)\lambda + O(\varepsilon) \\ &\ge -|G_x(t,*,v(t))|B + G_y(t,0,\Delta)\lambda + O(\varepsilon), \end{aligned}$$

where $*$ is between 0 and x and Δ is between $v(t)$ and $v(t)+\lambda$. Choosing μ and λ smaller, if necessary, $G_y(t,0,\Delta)\ge\mu$ for $v(t)\le\Delta\le v(t)+\lambda$, so for φ_1 we require

$$\lambda\mu - |G_x(t,*,v(t))|B > 0. \tag{13}$$

Then (4), (5), (6) and (13) are consequences of (11) and (12).

The construction for the interval $[t_2,1]$ is very similar to that given above for $[0,t_1]$, so we omit the details. Q.E.D.

Example 1.

$$\begin{aligned} &\varepsilon x'' = x + \delta y(y-2), \qquad x(0)=x(1)=0, \\ &\varepsilon y'' = x + f(t,y), \quad y(0)=0, \quad y(1)=2, \end{aligned}$$

where $f(t,y) = 2y(y+t-\frac{3}{2})(y-2)$ and $\delta>0$. Here $v(t)=2$, $0\le t\le 1$. Note that $\int_0^2 f(t,s)\,ds = \frac{4}{3}(1-2t)$ is positive for $0\le t<\frac{1}{2}$ and is negative for $\frac{1}{2}<t\le 1$. Also,

$$f_y(t,y) = 6y^2 + (4t-14)y + \left(\frac{3}{2}-t\right)4,$$

so $f_y(t,0)$ and $f_y(t,2)$ are positive for $0\le t\le 1$. It is readily checked that J in assumption (c) of Theorem 1 has eigenvalues with positive real parts if δ is sufficiently small.

Thus Theorem 1 can be applied to this example, provided δ is small enough to satisfy (d). If we choose $t_1=.4$ and $t_2=.6$ so that the layer is located in the interval $(.4,.6)$, then one can show that $\lambda=.019$, $\mu=2.95$ satisfy (8), (9) and (10). From (11), we compute that (d) is satisfied if $\delta<.0139$.

Note that (12) yields an upper bound on the deflection of the x component from zero in the layer. In this case, the deflection is less than $(4.038)\delta$.

Finally, we will approximate the location of the layer for small values of δ. If we multiply the second differential equation by y' and integrate, we obtain

$$\frac{\varepsilon(y')^2}{2}=\int_0^t(x(\tau,\varepsilon)+f(\tau,y(\tau,\varepsilon)))y'(\tau,\varepsilon)\,d\tau.$$

Since y' is bounded outside the layer,

$$\lim_{\varepsilon\to 0}\int_0^1(x(\tau,\varepsilon)+f(\tau,y(\tau,\varepsilon)))y'(\tau,\varepsilon)\,d\tau=0. \tag{14}$$

If $\bar{t}$ is the asymptotic location of the layer,

$$\lim_{\varepsilon\to 0}\int_0^1 f(\tau,y(\tau,\varepsilon))y'(\tau,\varepsilon)\,d\tau=\int_0^2 f(\bar{t},y)\,dy=\frac{4}{3}(1-2\bar{t}). \tag{15}$$

We can use (14) and (15) to approximate $\bar{t}$, if we can approximate

$$\lim_{\varepsilon\to 0}\int_0^1 x(\tau,\varepsilon)y'(\tau,\varepsilon)\,d\tau.$$

Let $\xi=(t-\bar{t})/\sqrt{\varepsilon}$, $\bar{t}=t-\frac{1}{2}-\delta\tilde{t}-O(\delta^2)$, $x(\xi)=\delta\tilde{x}(\xi)+O(\delta^2)$ and $y(\xi)=\tilde{y}(\xi)+O(\delta)$ in the layer. Substituting these expressions into the differential equations and taking into account the matching conditions, we arrive at the boundary value problems

$$\frac{d^2\tilde{y}}{d\xi^2}=f\left(\frac{1}{2},\tilde{y}\right),\quad \tilde{y}(-\infty)=0,\quad \tilde{y}(\infty)=2, \tag{16}$$

and

$$\frac{d^2\tilde{x}}{d\xi^2}=\tilde{x}+\delta\tilde{y}(\tilde{y}-2),\qquad \tilde{x}(-\infty)=\tilde{x}(\infty)=0. \tag{17}$$

The solution of (16) is $\tilde{y}(\xi)=1+\tanh\xi$, and the corresponding solution of (17) is $\tilde{x}(\xi)=1-(\pi/2)e^{-\xi}-2\sinh\xi\tan^{-1}e^{\xi}$. Neglecting terms of order δ^2, we have

$$\begin{aligned}\lim_{\varepsilon\to 0}\int_0^1 x(\tau,\varepsilon)y'(\tau,\varepsilon)\,d\tau&=\int_{-\infty}^{\infty}x(\xi)\,\frac{dy}{d\xi}\,d\xi\\&=\delta\int_{-\infty}^{\infty}\left(1-\frac{\pi}{2}e^{-\xi}-2\sinh\xi\tan^{-1}e^{\xi}\right)\operatorname{sech}^2\xi\,d\xi\\&=\delta\left(-\frac{\pi^2}{2}\right).\end{aligned} \tag{18}$$

Combining (14), (15) and (18), we conclude that

$$\bar{t}=\frac{1}{2}-\frac{3}{16}\pi^2\delta+O(\delta^2).$$

Thus for this example, we not only can establish the existence of a solution with interior layer behavior for small $\varepsilon>0$, but also can use standard asymptotic methods to estimate both the location of the layer and the behavior of the solution in the layer.

3. Boundary layers. We assume now that there is a curve in $(t,\mathbf{z})$ space where $\mathbf{H}$ vanishes and that the curve satisfies one of the boundary conditions, say at $t=0$. As in §2, we can change coordinates in such a way that $\mathbf{H}$ vanishes at $\mathbf{0}$ and $\mathbf{B}$ has the form $[{0 \atop b}]$, $b>0$.

Sufficient conditions for the existence of a solution of (1), (2) with a boundary layer at $t=1$ are given by:

THEOREM 2. *Assume*:

(a) $\mathbf{H}(t,0)=\mathbf{0}$ *for* $0\le t\le 1$;

(b) $\int_0^u G(1,0,s)\,ds>0$ *for* $0<u\le b$;

(c) *there is a* $D>0$ *and a class* C^2 *symmetric positive definite matrix function* $\mathbf{Q}(t)$ *so that*

$$\mathbf{z}^T\mathbf{Q}(t)\mathbf{J}(t,\mathbf{0})\mathbf{z}>D\mathbf{z}^T\mathbf{Q}(t)\mathbf{z}$$

for $0\le t\le 1$ *and all* $\mathbf{z}$;

(d) $G_y(1,0)>0$, $F_x(1,\mathbf{z})>0$ *for* $(1,\mathbf{z})\in R'$ *and* $\max_{R'}|F_y|\max_{R'}|G_x|/\min_{R'}F_x$ *is smaller than some computable positive number*, *where* $R'=\{(1,x,y)$: *x in some suitable bounded interval*, $0\le y\le b\}$.

Then (1), (2) *has a solution* $\mathbf{z}(t,\varepsilon)$ *for small* $\varepsilon>0$ *so that* $\mathbf{z}(t,\varepsilon)=O(\varepsilon)$ *for* $0\le t\le 1-\gamma$, $\gamma>0$.

Proof. The proof is nearly identical to the construction given for the interval $[0,t_1]$ in the proof of Theorem 1. Q.E.D.

Remark. Theorem 2 can easily be extended to allow the occurrence of boundary layers at both endpoints. If $\mathbf{A}\neq[{0 \atop 0}]$, then a boundary layer will appear at $t=0$, provided we assume that the line integral of $\mathbf{H}$ along the line segment from $\mathbf{0}$ to $\alpha\mathbf{A}$ is positive for $0<\alpha\le 1$ and that a condition like (d) above holds at $t=0$. It is also possible to combine Theorems 1 and 2 to obtain solutions with both boundary and internal layers.

Example 2.

$$\begin{aligned}\varepsilon x''&=x+\delta y, \qquad x(0)=x(1)=0,\\ \varepsilon y''&=3x+g(y), \qquad y(0)=0, \qquad y(1)=b>0,\end{aligned}$$

where $g(y)=y-y^3$. Note that

$$\int_0^b(y-y^3)\,dy=\frac{b^2}{2}\left(1-\frac{b^2}{2}\right)>0,$$

if $b<\sqrt{2}$. Also, the Jacobian matrix for the system has eigenvalues with positive real parts if $\delta<\frac{1}{3}$, so condition (c) of Theorem 2 is satisfied (see the remark following Theorem 1). Thus the boundary value problem has a solution with a boundary layer at $t=1$, provided $b<\sqrt{2}$ and $|\delta|$ is small enough to satisfy (d) of Theorem 2.

This result cannot be obtained by comparing the system with a scalar problem. We would need to find a scalar comparison function $\phi(x^2+y^2)$ which is positive for x^2+y^2 sufficiently small but positive, and which satisfies:

$$(x,y)\cdot\begin{pmatrix}x+\delta y\\ 3x+g(y)\end{pmatrix}\ge\sqrt{x^2+y^2}\,\phi(x^2+y^2) \tag{19}$$

on a suitable region $\{(x,y): x^2+y^2 \leq r^2\}$ (see Howes [3]). However, the left-hand side of (19) is $(x+y)^2+(\delta+1)xy-y^4$, which is negative for $x=-y$ whenever $\delta \geq -1$ and $y \neq 0$.

Acknowledgments. The author gratefully acknowledges helpful and stimulating discussions with Fred Howes. He is also indebted to the hospitality of the University of California–Davis where much of this work was done.

REFERENCES

[1] M. W. Hirsch and S. Smale, *Differential Equations, Dynamical Systems and Linear Algebra*, Academic Press, New York, 1974.

[2] F. Howes, *Boundary-layer interactions in nonlinear singular perturbation theory*, Memoirs AMS, No. 203, 15, 1978.

[3] F. Howes, *Singularly perturbed semilinear systems*, Studies Appl. Math., 61 (1979), pp. 185–209.

[4] W. G. Kelley, *A geometric method of studying two point boundary value problems for second order systems*, Rocky Mountain J. Math., 7 (1977), pp. 251–263.

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

[6] H. W. Knobloch and K. Schmitt, *Nonlinear boundary value problems for systems of differential equations*, Proc. Royal Soc. Edinburgh, 78A (1977), pp. 139–159.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics
002

NONLINEAR BOUNDARY VALUE PROBLEMS AND A PRIORI BOUNDS ON SOLUTIONS*

P. W. ELOE[†] AND JOHNNY HENDERSON[‡]

Abstract. Results concerning the existence of solutions of multipoint boundary value problems are given. The results are based on a topological transversality method and rely on *a priori* bounds on solutions. Applications are made to conjugate type and focal type boundary value problems; the third order, 3-point boundary value problem $y'''=f(x,y,y',y'')$, $y(x_1)=r_1$, $y(x_2)=r_2$, $y(x_3)=r_3$, where $x_1<x_2<x_3$, is discussed.

1. Introduction. Let $I=[a,b]$ be a compact subinterval of the reals and let L be the nth order linear differential operator given by,

$$Ly=\sum_{i=0}^{n} a_i(x)y^{(i)}, \tag{1}$$

where $a_i(x)\in C(I)$, $0\le i\le n$, and $a_n(x)$ does not vanish on I. We shall be concerned with the existence of solutions of the boundary value problem (BVP)

$$Ly=f(x,y,y',\cdots,y^{(n-1)}), \tag{2}$$

$$U(y)=r, \tag{3}$$

where $f\colon I\times\mathbb{R}^n\to\mathbb{R}$ is continuous and $U\colon C^{n-1}(I)\to\mathbb{R}^n$ is a continuous linear operator. In establishing the existence of solutions of (2)–(3) we shall extend the methods employed in Granas, Guenther and Lee [4]. In particular, we shall exploit further a topological transversality method due to Granas [3] which is a generalization of the continuation theorem of Leray and Schauder found in [2] or [8, p. 153].

In §2, in order that this paper be self-contained, we shall state some definitions and results from Granas [3]. We shall then prove a general theorem, Theorem 2.3, concerning the existence of solutions of the BVP (2), (3). Although the boundary conditions given by (3) are more general than those considered by Granas, et al. [4] (in [4], of primary concern were linear homogeneous two-point boundary conditions), we show in the proof of Theorem 2.3 that their techniques carry over to problems of the form (2), (3). This main result in §2, depends on an a priori bound for solutions of a class of BVPs, and we shall provide applications in §3 where a priori bounds can be found. In particular, in §3, we shall consider multipoint BVPs of the conjugate and focal type, some of which satisfy general Nagumo-like estimates [5], [1].

2. Existence of solutions. The following definitions and Lemmas 2.1 and 2.2 are due to Granas [3].

DEFINITIONS. Assume all topological spaces are Hausdorff. Let Y be a topological space, $A\subset X\subset Y$ and A closed in X.

(i) A continuous mapping $f\colon X\to Y$ is *compact* if $\overline{f(X)}$ is compact.

*Received by the editors September 16, 1982.

[†] Department of Mathematics, University of Dayton, Dayton, Ohio 45469.

[‡] Department of Mathematics and Statistics, University of Missouri-Rolla, Rolla, Missouri 65401. The research of this author was partially supported by a grant from the Weldon Spring Endowment Fund (University of Missouri).

(ii) h: $[0,1]\times X\to Y$ is a *compact homotopy* if h is a homotopy and if for each $\lambda\in[0,1]$, $h|\lambda\times X\equiv h_\lambda$ is compact.

(iii) f: $X\to Y$ is *admissible with respect to* A if f is compact and $f|A$ is fixed point free. Let $M_A(X,Y)$ denote the class of admissible mappings with respect to A.

(iv) $f\in M_A(X,Y)$ is *inessential* if there exists $g\in M_A(X,Y)$ such that $f|A=g|A$ and g is fixed point free on X. Otherwise, $f\in M_A(X,Y)$ is *essential.*

(v) A compact homotopy h: $[0,1]\times X\to Y$ is *admissible* if for each $\lambda\in[0,1]$, h_λ is admissible. Two mappings $f,g\in M_A(X,Y)$ are *homotopic in* $M_A(X,Y)$, $f\sim g$, if there exists an admissible homotopy h: $[0,1]\times X\to Y$ such that $h_0=g$ and $h_1=f$.

(vi) F^* denotes the class of topological spaces which has the fixed point property for compact maps. We remark that a closed convex subspace of a Banach space is an F^* space by the Schauder fixed point theorem.

LEMMA 2.1. *Let Y be a connected space belonging to F^*, let $X\subset Y$ be closed, and let $A=\partial X$. If f: $X\to Y$ is a constant mapping, (that is, $f(x)=p$ for all $x\in X$), and $p\in X\setminus A$, then f is essential.*

LEMMA 2.2. *Let Y be a convex topological space and let A and X be as in Lemma* 2.1. *Assume $f\sim g$ in $M_A(X,Y)$. Then f is essential if and only if g is essential.*

Now for our purposes, while employing the techniques of Granas, et al. [4] to establish the existence of solutions of the BVP (2)–(3), we shall also be concerned with solutions of an associated family of boundary value problems

$$(4_\lambda)\qquad Ly=g(x,y,y',\cdots,y^{(n-1)},\lambda),\qquad 0\le\lambda\le1,$$

$$(3)\qquad U(y)=r,$$

where g: $I\times\mathbb{R}^n\times[0,1]\to\mathbb{R}$ is continuous, $g(x,y_1,\cdots,y_n,0)\equiv0$ and $g(x,y_1,\cdots,y_n,1)\equiv f(x,y_1,\cdots,y_n)$. We remark here that for the applications in §3 and for many of those in [4], it is the case that $g=\lambda f$. Moreover, note that the boundary value problems, (2), (3) and (4_1), (3) are equivalent.

Let $(C(I),|\cdot|_0)$ be the Banach space of continuous functions on I with supremum norm and let $(C^k(I),|\cdot|_k)$ be the Banach space of k-times continuously differentiable functions y with norm

$$|y|_k=\max\{|y|_0,|y'|_0,\cdots,|y^{(k)}|_0\}.$$

THEOREM 2.3. *Assume that $y\equiv0$ is the unique solution of the* BVP $Ly=0$, $U(y)=0$, *and assume there exists $R>0$ such that $|y|_{n-1}<R$, for all solutions y of the* BVP, (4_λ), (3), *for all* $0\le\lambda\le1$. *Then the* BVP (2), (3), *has at least one solution.*

Proof. Let $Y=C^{n-1}(I)$, let $X=\{y\in C^{n-1}(I)$: $|y|_{n-1}\le R\}$, and let $A=\partial X$. Since $y\equiv0$ is the unique solution of the BVP $Ly=0$, $U(y)=0$, there exists a Green's function $G(x,s)$ for the BVP $Ly=0$, $U(y)=0$. Define h: $[0,1]\times X\to Y$ by

$$h(\lambda,y)(x)=l_r(x)+S_a^bG(x,s)g(x,y(s),y'(s),\cdots,y^{(n-1)}(s),\lambda)\,ds,$$

where $l_r(x)$ is the unique solution of the BVP $Ly=0$, $U(y)=r$. For each λ, $0\le\lambda\le1$, h_λ: $X\to Y$ can be shown to be a compact map by an application of the Arzela-Ascoli theorem and since all solutions y of (4_λ), (3) satisfy $|y|_{n-1}<R$, each h_λ is admissible. Thus, h is an admissible homotopy and $h_0\sim h_1$.

Now, $h_0\equiv l_r$ and $l_r\in X/A$ since $|l_r|_{n-1}<R$. Thus, by Lemma 2.1, h_0 is essential; it follows from Lemma 2.2 that h_1 is essential. Since the BVPs (2), (3) and (4_1), (3), are equivalent, the BVP (2), (3) has at least one solution.

3. Some applications. The applicability of Theorem 2.3 depends upon the existence of an a priori $|\cdot|_{n-1}$-norm bound for solutions of the family of BVPs (4_λ), (3), which is independent of λ. In this section, we consider conditions under which such bounds exist for multipoint BVPs of the conjugate and focal types.

For conjugate type boundary conditions, let $2 \le k \le n$ be given and for $a \le x_1 < \cdots < x_k \le b$ and positive integers $m_1, \cdots, m_k$ such that $\Sigma_{i=1}^{k} m_i = n$, let U_1: $C^{n-1}(I) \to \mathbb{R}^n$ be the linear boundary operator defined by

$$U_1(y) = \left(r_{0,1}, \cdots, r_{m_1-1,1}, r_{0,2}, \cdots, r_{m_2-1,2}, \cdots, r_{0,k}, \cdots, r_{m_k-1,k}\right),$$

where $r_{i,j} = y^{(i)}(x_j)$, $0 \le i \le m_j - 1$, $1 \le j \le k$.

For focal type boundary conditions, let $x_1, \cdots, x_n \in I$ with at least two of the points distinct. Let U_2: $C^{n-1}(I) \to \mathbb{R}^n$ be the linear boundary operator defined by

$$U_2(y) = (r_1, r_2, \cdots, r_n),$$

where $r_i = y^{(i-1)}(x_i)$, $1 \le i \le n$.

We now apply Theorem 2.3 in order to establish existence results for solutions of the BVP

$$(5) \qquad y^{(n)} = f(x, y, y', \cdots, y^{(n-1)}),$$
$$(6_l) \qquad U_l(y) = 0,$$

with $l = 1$ or $l = 2$.

THEOREM 3.1. *Assume that there exists a positive real-valued function,* $\phi(t_1, \cdots, t_n)$, *defined for* $t_i \ge 0$, $1 \le i \le n$, *which is nondecreasing in each variable and such that*

$$|f(x, y_1, \cdots, y_n)| \le \phi(|y_1|, \cdots, |y_n|),$$

for all $(x, y_1, \cdots, y_n) \in I \times \mathbb{R}^n$. *If*

$$(7) \qquad \sum_{i=1}^{n} \frac{t_i}{\phi(t_1, \cdots, t_n)} \to +\infty \quad \text{as} \ \sum_{i=1}^{n} t_i \to +\infty,$$

then the BVP, (5), (6_l), $l = 1, 2$, *has a solution.*

Proof. Let $l \in \{1, 2\}$ be fixed and consider the associated family of BVPs

$$(8_\lambda) \qquad y^{(n)} = \lambda f(x, y, y', \cdots, y^{(n-1)}), \qquad 0 \le \lambda \le 1,$$
$$(6_l) \qquad U_l(y) = 0.$$

Since $y^{(n)} = 0$ is disconjugate and disfocal on any interval, $y \equiv 0$ is the unique solution of the BVP $y^{(n)} = 0$, $U_l(y) = 0$. By Theorem 2.3, the proof is complete if we exhibit $R > 0$ such that $|y|_{n-1} < R$ for all solutions y of the family of BVPs, (8_λ), (6_l), where R is independent of λ.

Let y be a solution of (8_λ), (6_l). By (6_l) (and repeated applications of Rolle's theorem, if necessary), there exists $x_i \in I$, $1 \le i \le n$, such that $y^{(i-1)}(x_i) = 0$. Note that for $x \in I$

$$|y^{(n-1)}(x)| = \left|S_{x_n}^{x} y^{(n)}(s)\, ds\right| \le |y^{(n)}|_0 (b-a), \cdots, |y(x)| \le |y^{(n)}|_0 (b-a)^n.$$

In particular, $\Sigma_{i=1}^{n}|y^{(i-1)}|_0 \leq K|y^{(n)}|_0 \leq K\phi(|y|_0, |y'|_0, \cdots, |y^{(n-1)}|_0)$, where $K>0$ is independent of the solution y and of λ. Thus,

$$\sum_{i=1}^{n} \frac{|y^{(i-1)}|_0}{\phi\left(|y|_0, |y'|_0, \cdots, |y^{(n-1)}|_0\right)} \leq K$$

for all solutions y of (8_λ), (6_l). By (7), this implies there exists $R>0$, independent of λ, such that $|y|_{n-1} < R$ for all solutions y of the family of BVPs (8_λ), (6_l).

COROLLARY 3.2. *Let $f: I\times\mathbb{R}^n \rightarrow \mathbb{R}$ be continuous. If there exist $K_1>0$, $K_2>0$, $\alpha_i \geq 0$, $0 \leq i \leq n-1$ such that $\Sigma_{i=0}^{n-1}\alpha_i < 1$ and such that*

$$|f(x,y_1,\cdots,y_n)| \leq K_1 + K_2 \prod_{i=1}^{n} |y_i|^{\alpha_{i-1}} \quad \textit{for all } (x,y_1,\cdots,y_n) \in I\times\mathbb{R}^n,$$

then each of the BVP*s* (5), (6_l), $l=1,2$, *has a solution.*

Remarks. 1) The Schauder fixed point theorem can be employed to obtain Theorem 3.1 for the BVP, (5), (6_1). The following theorem, stated for nonhomogeneous boundary conditions, is found in [6, Thm. 2.5, p. 109].

THEOREM 3.3. *Assume that $f(x,y,y',\cdots,y^{(n-1)})$ is continuous on $I\times\mathbb{R}^n$ and let N_i, $0\leq i\leq n-1$, be given positive constants. Then there exists $\delta=\delta(N_0,N_1,\cdots,N_{n-1})>0$ such that the* BVP, (5), (6_1), *has a solution provided $b-a\leq\delta$. In fact, let*

$$\delta(N_0,N_1,\cdots,N_{n-1}) = \min\left\{\left(N_i/(\gamma_i Q)\right)^{1/n-i} : 0\leq i\leq n-1\right\},$$

where $Q=\max\{|f(x,y_0,\cdots,y_{n-1})|: x\in I, |y_i|\leq N_i, 0\leq i\leq n-1\}$, and γ_i, $0\leq i\leq n-1$, is a constant such that

$$S_I\left|\frac{\partial^i}{\partial x^i}G(x,s)\right| ds \leq \gamma_i(b-a)^{n-i},$$

where $G(x,s)$ is the Green's function for the BVP $y^{(n)}=0$, $U_1(y)=0$.

To obtain Theorem 3.1 from Theorem 3.3, choose $M>0$ large enough such that

$$b-a \leq \min\left\{\left[M/\gamma_i\phi(M,\cdots,M)\right]^{1/n-i}, 0\leq i\leq n-1\right\}.$$

Let $N_0=N_1=\cdots=N_{n-1}=M$. Then $Q\leq\phi(M,\cdots,M)$ and $b-a\leq\delta(N_0,\cdots,N_{n-1})$.

2) Theorem 3.1 and Corollary 3.2 are valid for any boundary conditions (3) provided that given a solution y of the desired BVP for each $0\leq i\leq n-1$, $y^{(i)}$ vanishes at some point on I.

In our next application of Theorem 2.3, we consider the BVP (5), (3); that is, we consider $y^{(n)}=f(x,y,y',\cdots,y^{(n-1)})$, $U(y)=r$. We shall impose on f a Nagumo-like estimate as in Jackson [5] and Bebernes, et al. [1], in order to allow a faster growth rate on f than that allowed by (7) in Theorem 3.1. The following lemma can be found in [5].

LEMMA 3.4. *Let $y\in C^n(I)$. Then for each integer k, $0<k<n$,*

$$|y^{(k)}|_0 \leq C_{nk}|y|_0^{1-k/n}M_n^{k/n}$$

where $C_{nk}=4e^{2k}n^k k^{-k}$ and $M_n=\max\{|y^{(n)}|_0, 2^n n!, |y|_0(b-a)^{-n}\}$.

THEOREM 3.5. *Given any $M>0$, assume there exists a continuous, positive real-valued function, $\phi(t_1,\cdots,t_{n-1})$, defined for $t_i\geq 0$, $1\leq i\leq n-1$, which is nondecreasing in each variable, such that*

$$|f(x,y_1,\cdots,y_n)| \leq \phi\left(|y_2|,\cdots,|y_n|\right),$$

for all $(x,y_1,\cdots,y_n)\in I\times[-M,M]\times\mathbb{R}^{n-1}$, *and such that*

$$\sum_{i=1}^{n-1}\frac{t_i^{n/i}}{\phi(t_1,\cdots,t_{n-1})}\to+\infty \quad as \quad \sum_{i=1}^{n-1}t_i^{n/i}\to+\infty. \tag{9}$$

If there exists $N>0$ *such that* $|y|_0<N$ *for all solutions y of the associated family of* BVP's $y^{(n)}=\lambda f(x,y,\cdots,y^{(n-1)})$, $U(y)=r$ *for all* $0\leq\lambda\leq 1$, *then the* BVP (5), (3) *has a solution.*

Proof. Let y be a solution of the BVP $y^{(n)}=\lambda f(x,y,\cdots,y^{(n-1)})$, $U(y)=r$ and assume $|y|_0<N$. By Lemma 3.4, there exist constants $C_1(N)$ and $C_2(N)$, independent of λ and independent of y, such that

$$|y^{(k)}|_0^{n/k}\leq C_1(N)\leq\frac{C_1(N)}{\phi(0,\cdots,0)}\phi(|y'|_0,\cdots,|y^{(n-1)}|_0)$$

or

$$|y^{(k)}|_0^{n/k}\leq C_2(N)|y^{(n)}|_0\leq C_2(N)\phi(|y'|_0,\cdots,|y^{(n-1)}|_0).$$

Consequently, there exists a constant $C(N)$, independent of λ and independent of y, such that

$$\sum_{i=1}^{n-1}\frac{|y^{(i)}|_0^{n/i}}{\phi(|y'|_0,\cdots,y^{(n-1)}|_0)}\leq C(N).$$

By (9), it follows as in the proof of Theorem 3.1, that there exists $R>0$, independent of λ, such that $|y|_{n-1}<R$. As before, this completes the proof.

For our final result, we present as an example a corollary of Theorem 3.5. We employ the theory of subfunctions and third order differential inequalities, as developed by Jackson and Schrader [7], to find a priori bounds for solutions of 3-point conjugate type BVPs for third order differential equations. Consider the BVP

$$y'''=f(x,y,y',y''), \tag{10}$$

$$y(x_1)=r_1, y(x_2)=r_2, y(x_3)=r_3, \tag{11}$$

with $a<x_1<x_2<x_3<b$. Assume f: $(a,b)\times\mathbb{R}^2\to\mathbb{R}$ is continuous and satisfies a Nagumo condition (9).

COROLLARY 3.6. *Assume that solutions of* 3*-point conjugate type* BVPs *for* $y'''=\lambda f(x,y,y',y'')$, $0\leq\lambda\leq 1$, *have at most one solution on any interval* $(c,d)\subset(a,b)$ *and all solutions of initial value problems extend throughout* (a,b). *Assume there exist upper and lower solutions, v and w, respectively, of* (10) *satisfying*

$$w(x_1)=v(x_1)=r_1,\quad w(x_2)=v(x_2)=r_2,\quad w(x_3)=v(x_3)=r_3$$

and

$$w'''-\lambda f(x,w,w',w'')>0>v'''-\lambda f(x,v,v',v''),$$

$x\in(a,b)$, $0\leq\lambda\leq 1$. *Then the* BVP (10)–(11) *has a solution.*

Proof. Let $0\leq\lambda\leq 1$ be arbitrary, but fixed. Let $N>\max(|w|_0,|v|_0)$. By [7, Thm. 4.1], w is a subfunction and v is a superfunction with respect to solutions of $y'''=\lambda f(x,y,y',y'')$ on (a,b). By (11), any solution y satisfies $w(x_1)=y(x_1)=v(x_1)$, $w(x_2)=y(x_2)=v(x_2)$ and $w(x_3)=y(x_3)=v(x_3)$. Thus, $w(x)\geq y(x)\geq v(x)$, $x\in(x_1,x_2)$ and $w(x)\leq y(x)\leq v(x)$, $x\in(x_2,x_3)$. In particular, $|y|_0<N$ where N is independent of λ. Theorem 3.5 applies immediately and the proof is complete.

REFERENCES

[1] J. Bebernes, R. Gaines and K. Schmitt, *Existence of periodic solutions for third and fourth order ordinary differential equations via coincidence degree*, Ann. Soc. Sci. Bruxelles, Sér. I88 (1974), pp. 25–36.

[2] R. E. Gaines and J. L. Mawhin, *Coincidence Degree, and Nonlinear Differential Equations*, Lecture Notes in Mathematics 568, Springer-Verlag, New York, 1977, pp. 1–262.

[3] A. Granas, *Sur la méthode de continuité de Poincaré*, C. R. Acad. Sci. Paris, 282 (1976), pp. 983–985.

[4] A. Granas, R. B. Guenther and J. W. Lee, *Nonlinear boundary value problems for some classes of ordinary differential equations*, Rocky Mountain J. Math., 10 (1980), pp. 35–58.

[5] L. K. Jackson, *A Nagumo condition for ordinary differential equations*, Proc. Amer. Math. Soc., 57 (1976), pp. 93–96.

[6] ———, *Boundary value problems for ordinary differential equations*, in Studies in Ordinary Differential Equations, J. K. Hale, ed., MAA Studies in Mathematics vol. 14, Mathematical Association of America, Washington, DC, 1977.

[7] L. K. Jackson and K. Schrader, *Subfunctions and third order differential inequalities*, J. Differential Equations, 8 (1970), pp. 180–194.

[8] J. L. Mawhin, *Functional analysis and boundary value problems*, in Studies in Ordinary Differential Equations, J. K. Hale, ed., MAA Studies in Mathematics vol. 14, Mathematical Association of America, Washington, DC, 1977.

© 1984 Society for Industrial and Applied Mathematics

PERTURBATION OF PERIODIC BOUNDARY CONDITIONS*

LAWRENCE TURYN†

Abstract. We consider perturbations of the problem $(*) -x''+bx=\lambda ax$, $x(0)-x(1)=0=x'(0)-x'(1)$ both by changes of the boundary conditions and by addition of nonlinear terms. We assume that at $\lambda=\lambda_0$ there are two linearly independent solutions of the unperturbed problem $(*)$ and that $a(\cdot)$ is bounded away from zero. When only the boundary conditions are perturbed either the Hill's discriminant or the method of Lyapunov–Schmidt reduces the problem to $0=\det((\lambda-\lambda_0)A-\varepsilon H)+$higher order terms, where A and H are real 2×2 constant matrices. We analyse the existence of curves $(\lambda(\varepsilon),\varepsilon)$ of eigenvalues for this problem of linear perturbation and give as an example a heat problem with $H=\left(\begin{smallmatrix}0&1\\0&0\end{smallmatrix}\right)$.

The method of Lyapunov–Schmidt is used to analyse the full nonlinear problem. In a sequel to this paper we will analyse the bifurcation problem from a "generic" point of view and we will present some numerical examples.

1. Introduction. We consider boundary value problems which are perturbations of a linear boundary value problem with periodic boundary conditions. The first such problem we consider, in §2, is

$$\begin{aligned}&x''+(\lambda a-b)x=0,\\&x(0)-x(1)=\varepsilon\ (\text{terms linear in } x(1),x'(1)),\\&x'(0)-x'(1)=\varepsilon\ (\text{terms linear in } x(1),x'(1)).\end{aligned}$$

For this linear boundary value problem with a parameter ε we establish a condition for the local splitting of a double eigenvalue λ_0 for $\varepsilon=0$ into two curves $\lambda=\lambda^*(\varepsilon)$ for $\varepsilon\neq0$. This condition can be established either using the Hill's discriminant or the method of Lyapunov–Schmidt, and the boundary conditions can be permitted to be nonself-adjoint for $\varepsilon\neq0$. For problems where the boundary conditions remain self-adjoint for $\varepsilon\neq0$ some classical perturbation results of Rellich can be applied to the above situation, even though the differential operator has domain varying with ε.

In §3 we consider nonlinear perturbations, i.e.

$$x''+(\lambda a-b)x=(\text{nonlinear function of } x),$$

with the boundary conditions also having terms nonlinear in x. When the linearisation has for $\varepsilon=0$ a double eigenvalue λ_0 the method of Lyapunov–Schmidt is used to reduce the problem to a system of two equations in four unknowns, $\mathbf{u}\in\mathbb{R}^2$, $\lambda\in\mathbb{R}$, $\varepsilon\in\mathbb{R}$.

2. Linear perturbation of the periodic boundary value problem. In this section we consider the boundary value problem

$$\tau x=\lambda ax, \tag{2.1}$$

$$Mx=\varepsilon N(\varepsilon)x \tag{2.2}$$

* Received by the editors January 11, 1982, and in revised form July 22, 1982. This research was supported in part by the Natural Science and Engineering Research Council, Canada, and the University of Calgary. Portions of this paper appeared in the Proceedings of the University of Dundee Conference on Ordinary and Partial Differential Equations, Lecture Notes in Mathematics, Springer-Verlag, New York, 1983.

† Department of Mathematics and Statistics, Wright State University, Dayton, Ohio 45435.

where $\tau x = -x'' + b(t)x$, $' = d/dt$, $0 \le t \le 1$, $b(\cdot)$ and $a(\cdot)$ are continuous on $[0,1]$, $a(t) \ge a_0 > 0$ for $0 \le t \le 1$, $Mx = (x(0) - x(1), x'(0) - x'(1))^T$, ^{T} denoting transpose, and $N(\varepsilon)x = (H + \varepsilon \bar{H} + O(\varepsilon^2))(x(1), x'(1))^T$, where $H, \bar{H}$ are 2×2 real matrices, $H = (h_{ij})_{i,j=1,2}$. We will assume that $H \not\equiv 0, \varepsilon$ is a real parameter, and that $N(\varepsilon)$ is real-valued and two times continuously differentiable.

When $\varepsilon = 0$ (2.2) is called the periodic boundary conditions. Since (2.1)–(2.2) is linear in x we call this problem a linear perturbation of the periodic boundary value problem, although the periodic boundary conditions are perturbed by terms nonlinear in ε.

Denote by $X(t,\lambda)$ the principal fundamental matrix of solutions of (2.1), i.e. that matrix satisfying

$$X(0,\lambda) = I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and the differential equation $X' = A(t,\lambda)X$, where

$$A(t,\lambda) = \begin{pmatrix} 0 & 1 \\ b(t) - \lambda a(t) & 0 \end{pmatrix}.$$

Since $\operatorname{tr} A(t,\lambda) = 0$ for all t, λ it follows that $\det X(t,\lambda) = 1$ for all t, λ, in particular that $\det X(1,\lambda) = 1$ for all λ. Also, $X(1,\cdot)$ is analytic; see, for example, Hale [8, p. 82].

Define

$$\Delta(\lambda,\varepsilon) = \det(I - (I + \varepsilon N(\varepsilon))X(1,\lambda)).$$

This function is analytic in λ and three times continuously differentiable in ε. Given λ, ε the linear boundary value problem (2.1)–(2.2) has (a) no nontrivial solutions when $\Delta(\lambda,\varepsilon) \ne 0$, (b) at least one linearly independent solution when $\Delta(\lambda,\varepsilon) = 0$, (c) exactly two linearly independent solutions when $I - (I + \varepsilon N(\varepsilon))X(1,\lambda) \equiv 0$.

For $\varepsilon = 0$ the hypothesis $a(t) \ge a_0 > 0$ assures that (2.1)–(2.2) has an infinity of eigenvalues, i.e. values of λ for which there is at least one linearly independent solution. For this fact, see Birkhoff [1], Coddington and Levinson [6], or Magnus and Winkler [11]. Let us assume henceforth in this section that we are given an eigenvalue λ_0 for (2.1)–(2.2) at $\varepsilon = 0$. We will examine the question of the existence of a curve or curves $\lambda^*(\varepsilon)$ of eigenvalues for (2.1)–(2.2) passing through the point $(\lambda_0, 0)$.

Let $D_\lambda = \partial/\partial\lambda$, $D_\varepsilon = \partial/\partial\varepsilon$. From Birkhoff [1] or Magnus and Winkler [11] it is known that (2.1)–(2.2) at $(\lambda_0, 0)$ has (a) exactly one linearly independent solution when $\Delta(\lambda_0, 0) = 0 \ne D_\lambda \Delta(\lambda_0, 0)$, (b) exactly two linearly independent solutions when $\Delta(\lambda_0, 0) = 0 = D_\lambda \Delta(\lambda_0, 0)$. So, if at $(\lambda_0, 0)$ there is exactly one linearly independent solution then the implicit function theorem implies that there is a unique curve $(\lambda^*(\varepsilon), \varepsilon)$ in $\mathbb{R}^2$ passing through $(\lambda_0, 0)$ with $\Delta(\lambda^*(\varepsilon), \varepsilon) = 0$.

So let us consider case (b), i.e., $\Delta(\lambda_0, 0) = 0 = D_\lambda \Delta(\lambda_0, 0)$. By suitably modifying the calculations of Magnus and Winkler [11, p. 18] it follows that $\rho \overset{\text{def}}{=} D_{\lambda\lambda} \Delta(\lambda_0, 0) = 2(\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21})$ where

$$\sigma_{ij} \overset{\text{def}}{=} \int_0^1 a(t) x_i(t,\lambda_0) x_j(t,\lambda_0)\, dt, \qquad X(t,\lambda) = \begin{pmatrix} x_1(t,\lambda) & x_1(t,\lambda) \\ x_1'(t,\lambda) & x_2'(t,\lambda) \end{pmatrix}.$$

Since $\sigma_{12} = \sigma_{21}$ and $a(t) \ge a_0 > 0$, the Schwarz inequality and the linear independence of $x_1(\cdot, \lambda_0)$, $x_2(\cdot, \lambda_0)$ imply that $\rho > 0$. We will need some further notation:

$$\Xi \overset{\text{def}}{=} D_\lambda X(1, \lambda_0) = (\xi_{ij})_{i,j=1,2} \quad \text{and} \quad \Sigma = (\sigma_{ij})_{i,j=1,2}.$$

Again, by suitably modifying the calculations of Magnus and Winkler [11, p. 18] one can conclude that $\Xi = J\Sigma$ where $J = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. Let $\sigma = \det\Sigma$. We see then that $\rho = 2\sigma = 2\det\Sigma = 2\det\Xi$.

Let $\nu = \lambda - \lambda_0$. Since $X(1,\lambda) = I + \nu\Xi + O(\nu^2)$, one can calculate that

$$\Delta(\lambda,\varepsilon) = \det(I - (I+\varepsilon H)(I+\nu\Xi)) + O\big((|\varepsilon| + |\nu|)^3\big),$$
$$= \nu^2 \cdot \det\Xi + \varepsilon\nu\cdot\gamma + \varepsilon^2 \cdot \det H + O\big((|\varepsilon| + |\nu|)^3\big)$$

where

$$\gamma \overset{\text{def}}{=} h_{11}\xi_{22} + h_{22}\xi_{11} - h_{12}\xi_{21} - h_{21}\xi_{12}$$
$$= -h_{11}\sigma_{12} + h_{22}\sigma_{12} + h_{12}\sigma_{11} - h_{21}\sigma_{22},$$

since $\Xi = J\Sigma$. The validity of this asymptotic expansion follows from $\Delta(\lambda,\varepsilon)$ being analytic in λ and three times continuously differentiable in ε. Let $\delta = \det H$. The question of existence of a curve or curves passing through $(\nu,\varepsilon) = (0,0)$, i.e. $(\lambda,\varepsilon) = (\lambda_0, 0)$ is thus equivalent to the question of the existence of solutions to the equation

$$0 = \Delta = \sigma\nu^2 + \gamma\varepsilon\nu + \delta\varepsilon^2 + O\big((|\varepsilon| + |\nu|)^3\big). \tag{2.3}$$

This can be re-written directly as

$$0 = \Delta = \det(\nu\Sigma - \varepsilon JH) + O\big((|\varepsilon| + |\nu|)^3\big). \tag{2.3$'$}$$

The fact that the terms of second degree of Δ are equal to $\det(\nu\Sigma - \varepsilon JH)$ will also be derived, quite independently, in §3. There the method of Lyapunov–Schmidt will be used to find the bifurcation equations when the boundary value problem (2.1)–(2.2) is also subjected to nonlinear perturbation. When the nonlinear perturbation is taken to be identically zero, the bifurcation equations reduce to

$$(\nu\Sigma - \varepsilon JH)\mathbf{u} = O\big((|\varepsilon| + |\nu|)^2 |\mathbf{u}|\big) \tag{2.4}$$

where $\mathbf{u} \in \mathbb{R}^2$, $|\mathbf{u}| = |u_1| + |u_2|$. System (2.4) has solutions $\mathbf{u} \neq \mathbf{0}$ if and only if

$$0 = \det(\nu\Sigma - \varepsilon JH) + O\big((|\varepsilon| + |\nu|)^3\big).$$

Thus we see that the terms of second degree of Δ can be found by the method of Lyapunov–Schmidt just as well as by calculation of the Hill's discriminant.

THEOREM 2.1. *Assume that at* $(\lambda_0, 0)$ *there are two linearly independent solutions of* (2.1)–(2.2). *If* $\gamma^2 - 4\delta\sigma > 0$ *then there are two distinct continuously differentiable curves of eigenvalues* $\lambda = \lambda_0 + \nu^*(\varepsilon)$ *for all* $|\varepsilon|$ *sufficiently small, with* $\nu^*(0) = 0$.

Proof. This follows from (2.3) and $\sigma > 0$. □

In the above work no assumption was made about the self-adjointness of the boundary conditions for $\varepsilon \neq 0$. In fact we will give self-adjointness a privileged position only when applying the classical perturbation results of Rellich in Theorem 2.7 and in the following paragraphs.

From Coddington and Levinson [6, p. 297] it is known that (2.2) is self-adjoint if and only if

$$1 = \det(I + \varepsilon N(\varepsilon)) = 1 + \varepsilon \operatorname{tr} H + \varepsilon^2(\det H + \operatorname{tr} \overline{H}) + O\big(|\varepsilon|^3\big).$$

So, self-adjointness of the boundary conditions requires at least that $\operatorname{tr} H=0$. When $\operatorname{tr} H=0$, the 2×2 real matrix JH appearing in (2.4) is Hermitian. So, we see that self-adjointness of the boundary conditions implies self-adjointness of the matrix in (2.4) corresponding to the term of lowest order in ε.

Remark 2.2. *When* $\operatorname{tr} H=0$, $\gamma^2-4\delta\sigma\geq 0$.

Proof. Σ is always positive definite and self-adjoint because $a(t)\geq a_0>0$ by assumption. When $\operatorname{tr} H=0$, JH is self-adjoint. It follows that the eigenvalues β_1 of the generalised eigenvalue problem

$$(2.5) \qquad JH\mathbf{u}=\beta_1\Sigma\mathbf{u}\in\mathbf{C}^2$$

are real. Equivalent to β_1 being an eigenvalue is $0=\det(\beta_1\Sigma-JH)$; reality of the eigenvalues β_1 implies $\gamma^2-4\delta\sigma\geq 0$. □

So we see that when $\operatorname{tr} H=0$ the double eigenvalue λ_0 will usually split into two smooth curves of real eigenvalues for $\varepsilon\neq 0$; the exceptional case would be when $\gamma^2-4\delta\sigma=0$.

Example 2.3. When $a\equiv 1$, $b=0$, $\operatorname{tr} H=0$, and $\lambda_0=4\pi^2n^2$ for some positive integer n, we calculate $\sigma_{12}=0$ and $\gamma^2-4\delta\sigma=\frac{1}{4}((h_{12}+\lambda_0^{-1}h_{21})^2+\lambda_0^{-1}h_{11}^2)\geq 0$ since $H\not\equiv 0$ for nontriviality. It follows that Theorem 2.1 is applicable, except in the exceptional case $h_{11}=h_{22}=0$, $h_{12}+\lambda_0^{-1}h_{21}=0$.

As a specific sub-example, take $a\equiv 1$, $b\equiv 0$, $\lambda_0=4\pi^2n^2$ for some $n\geq 1$, and $H=\left(\begin{smallmatrix}1 & 0\\ 0 & -1\end{smallmatrix}\right)$. Theorem 2.1 is applicable; in fact, one can calculate explicitly $\Delta(\lambda,\varepsilon)=\varepsilon^2-2(1-\cos\lambda^{1/2})$ for $\lambda>0$. Explicitly, the curves are $\lambda=(\lambda_0^{1/2}\pm 2\arcsin(\varepsilon/2))^2$, which clearly cease to exist for $|\varepsilon|>2$. Note that the corresponding boundary conditions (2.2) are not self-adjoint for $\varepsilon\neq 0$, since $\det(I+\varepsilon H)=1-\varepsilon^2$.

Application 2.4. Consider a ring of metal obtained by joining the endpoints $\xi=0$, $\xi=1$. If the joining is not perfect then there will be some "contact resistance". Appropriate boundary conditions for the temperature $v(\xi)$ are then (see Özişik [12, p. 283] or Carslaw and Jaeger [3, p. 23])

$$u(0)-u(1)=\varepsilon u'(1),$$
$$u'(0)-u'(1)=0,$$

where $\varepsilon=k/h=(\text{Biot number})^{-1}$ can be taken to be small and positive if the heat transfer coefficient h is large. These boundary conditions have as a consequence a temperature drop across the join, this phenomenon being well known in practise. See Holman [10, pp. 45–48] for more details on the causes of contact resistance. These boundary conditions are self-adjoint for all ε, with $H=\left(\begin{smallmatrix}0 & 1\\ 0 & 0\end{smallmatrix}\right)$. Since $\gamma=\sigma_{11}>0$, Theorem 2.1 guarantees the existence of two curves $\lambda=\lambda_0+\nu_\pm^*(\varepsilon)$. In fact, for the example $a\equiv 1$, $b\equiv 0$ one can see that $\nu_+^*(\varepsilon)=0$ for all ε, since the second row of H is trivial, and further one can calculate that $D_\varepsilon\nu_-^*(0)=-2\lambda_0$.

It is useful to consider further the generalised eigenvalue problem (2.5). A *simple eigenvalue* β_1^0 for a pencil $(L_2;L_1)$ of two $n\times n$ matrices is a value of β_1 for which $\dim\mathfrak{N}(A)=1=\operatorname{codim}\mathfrak{R}(A)$ and $L_1\mathbf{z}\notin\mathfrak{R}(A)$ where $A=L_2+\beta_1^0L_1$ and $\mathbf{0}\neq\mathbf{z}\in\mathfrak{N}(A)$. The generalisation of the concept of simple eigenvalue to Banach space operators $(B;A_1,\cdots,A_N)$ originated in Hale [9]. Bibliographic references and extensive material on the use of simple eigenvalues in the analysis of linear and nonlinear problems can be found in Chow and Hale [4].

Remark 2.5. *If* $n\times n$ *matrices* L_2,L_1 *are Hermitian and* L_1 *is definite then* β_1^0 *is simple for* $(L_2;L_1)$ *whenever* $\dim\mathfrak{N}(L_2+\beta_1^0L_1)=1$.

Proof. Let $A = L_2 + \beta_1^0 L_1$ and $\mathbf{0} \neq \mathbf{z} \in \mathfrak{N}(A)$. The hypotheses imply that $\beta_1^0 = -\mathbf{z}^* L_2 \mathbf{z} / \mathbf{z}^* L_1 \mathbf{z}$ is real, so that A is also Hermitian. The Fredholm alternative implies that $\mathfrak{R}(A) = \{\mathbf{a} \in \mathbf{C}^n : \mathbf{z}^* \mathbf{a} = 0\}$; since L_1 is definite, $\mathbf{z}^* L_1 \mathbf{z} \neq 0$, so that $L_1 \mathbf{z} \notin \mathfrak{R}(A)$. □

With this background we can return to (2.5). If the discriminant $\gamma^2 - 4\delta\sigma > 0$, then whenever $0 = \det(-JH + \beta_1^0 \Sigma)$ necessarily $\dim \mathfrak{N}(-JH + \beta_1^0 \Sigma) = 1$. This and Remark 2.5 prove

Remark 2.6. *If* $\operatorname{tr} H = 0$ *and* $\gamma^2 - 4\delta\sigma > 0$ *then there are exactly two eigenvalues of* $(-JH; \Sigma)$ *and both are simple.*

The simplicity of the eigenvalues of $(-JH; \Sigma)$ is required for the application of the results of Chow and Hale [4, Chap. 7] for nonlinear bifurcation problems. In a sequel to this present paper we will consider such problems.

Remark. Given enough differentiability in ε for the original problem (2.1)–(2.2), Newton's polygon helps one to calculate solution(s) of $\Delta = 0$. Consider the case $\gamma = 0 = \delta$: If $D_\lambda D_\varepsilon^k \Delta(\lambda_0, 0) \neq 0$ or $D_\varepsilon^k \Delta(\lambda_0, 0) \neq 0$ for some $k \geq 3$ then Newton's polygon guarantees the existence of a curve of the approximate form $\lambda \sim \lambda_0 + c\varepsilon^p$ for some $p \geq \frac{3}{2}$. Of course, this assumes a sufficient amount of differentiability (in ε) of the boundary conditions.

Yet another approach, besides those utilising the Hill's discriminant or the method of Lyapunov–Schmidt, is to set problem (2.1)–(2.2) in a Hilbert space. This approach for the self-adjoint case will use the now-classic method of Rellich [13].

Let H be the Hilbert space of Lebesgue measurable functions $x: [0,1] \to \mathbf{C}$ constructed by completion of $C[0,1]$ with respect to the weighted inner product $(x, y) = \int_0^1 a x \bar{y}$. Define operators $T(\varepsilon): \mathfrak{U}(\varepsilon) \subset H \to H$ by $T(\varepsilon)x = (a^{-1}\tau - \lambda_0)$ on the domains $\mathfrak{U}(\varepsilon) = \{x \in H: T(\varepsilon)x \in H,\ x$ satisfies boundary conditions (2.2)$\}$. We will assume the self-adjointness of the boundary conditions, i.e. $1 = \det(I + \varepsilon H + \varepsilon^2 \overline{H} + \cdots)$, from which it follows that the operators $T(\varepsilon)$ on $\mathfrak{U}(\varepsilon)$ are symmetric.

Let $X(t, \lambda)$ denote the fundamental matrix for the differential equation (2.1) rewritten as a system. One can show that if the matrix $S_\lambda \overset{\text{def}}{=} X(1, \lambda) - I$ is invertible then the equation $(T(\varepsilon) + (\lambda - \lambda_0)I)x = p \in H$ has the unique solution

$$x_\varepsilon(t) = \int_0^t [-x_1(t)x_2(s) + x_2(t)x_1(s)]\, p(s)$$
$$+ (x_1(t), x_2(t))[S_\lambda - E_{\varepsilon,\lambda}]^{-1} E_{\varepsilon,\lambda} \begin{pmatrix} -\int_0^1 x_2 p \\ \int_0^1 x_1 p \end{pmatrix},$$

where $E_{\varepsilon,\lambda} = \varepsilon N(\varepsilon) X(1, \lambda)$, for all $|\varepsilon|$ sufficiently small. From study of Hill's equation one knows that in fact S_λ is invertible at all but discrete and isolated values of $\lambda \in \mathbb{R}$. In particular, $S_{\lambda_0 \pm i}$ is invertible, hence the operators $T(\varepsilon)$ are self-adjoint.

Therefore, from Rellich [13, pp. 71–72] we can conclude that $T(\varepsilon)$ on $\mathfrak{U}(\varepsilon)$ is a so-called regular family of self-adjoint operators. The next result follows from Rellich [13, p. 74].

THEOREM 2.7. *Assume that at* $(\lambda_0, 0)$ *there are two linearly independent solutions of* (2.1)–(2.2). *If the boundary conditions* (2.2) *are self-adjoint for all* ε *then there are two* (*counting multiplicity*) *real continuous curves* $\lambda = \lambda_0 + \nu^*(\varepsilon)$ *of eigenvalues for* (2.1)–(2.2) *with* $\nu^*(0) = 0$. *The case where there is a curve of double eigenvalues is not precluded.*

We remark that one could allow the parameter ε into the linear differential equation (2.1) without substantially altering any of the discussion in §2. The same cannot be said for allowing λ into the linear boundary conditions (2.2). Perturbation of problems with λ in separated boundary conditions has been considered in [17], and ε in the differential equation appeared in [16].

3. Nonlinear perturbation and Lyapunov–Schmidt. In this section we consider the boundary value problem

$$\tau x=\lambda a x+f_0(\varepsilon,\lambda;t,x,x'), \tag{3.1}$$

$$Mx=\varepsilon N(\varepsilon)x+\mathbf{f}(\varepsilon,\lambda;x), \tag{3.2}$$

where ε, a, M, and $N(\varepsilon)$ are as in §2 and $\mathbf{f}(\lambda,x)\in\mathbb{R}^2$ represents nonlinear contributions to the boundary conditions. Assume that both f_0 and $\mathbf{f}$ are $O(|\alpha|^2\|x\|+\|x\|^n)$ for some integer $n\geq 2$ as $|\alpha|$, $\|x\|\to 0$, where $\nu=\lambda-\lambda_0$, $\alpha=(\varepsilon,\nu)\in\mathbb{R}^2$, $|\alpha|=|\varepsilon|+|\nu|$, $\|x\|=|x|_\infty+|x'|_\infty+|x''|_\infty$, $|x|_\infty=\sup_{0\leq t\leq 1}|x(t)|$, the estimate on f_0 holding uniformly for $t\in[0,1]$. Here the symbol $O(s)$ for $s\in\mathbb{R}^+$ denotes any quantity $F(s)$ which satisfies $F(s)/s\to$ constant as $s\to 0+$. The term f may include things like $\int_0^1 w(t)x^2(t)\,dt$, but we do assume that $f_0(\cdot,\cdot;t,\cdot,\cdot), \mathbf{f}(\cdot,\cdot,\cdot)\colon(-\eta,\eta)\times(\lambda_0-\eta,\lambda_0+\eta)\times C^2[0,1]\times C^1[0,1]\to\mathbb{R}^2$ is $(n+1)$ times continuously differentiable for some $\eta>0$, this being true uniformly in $t\in[0,1]$ for f_0. Let us assume that for $\lambda=\lambda_0$, $\varepsilon=0\equiv f_0\equiv\mathbf{f}$ there are two linearly independent solutions of (3.1)–(3.2), as was assumed in the latter part of §2. As before, let $X(t,\lambda)$ denote the principal fundamental matrix of solutions.

To pose (3.1)–(3.2) as a bifurcation problem it will help to write that problem as a nonlinear equation in Banach spaces $\overline{Y}=C^2[0,1]$, $Z=C[0,1]\times\mathbb{R}^2$ with norms $\|x\|$, as above, and $\|(v;c,d)\|=|v|_\infty+|c|+|d|$, respectively. Then (3.1)–(3.2) is equivalent to the abstract problem

$$(B-\nu A-\varepsilon C)x=G(x)+O\big(|\alpha|^2\|x\|+|\alpha|\,\|x\|^n+\|x\|^{n+1}\big) \tag{3.3}$$

where $x\in\overline{Y}$, $Bx=(\tau x-\lambda_0 a(\cdot)x;Mx)$, $Ax=(a(\cdot)x,\mathbf{0})$, $Cx=(0;H(x(1),x'(1))^T)$, and $G(x)=(D_\lambda f_0(0,\lambda_0;\cdot,x,x'), D_\lambda\mathbf{f}(0,\lambda_0;x))$. Note that $G(x)=O(\|x\|^n)$ as $\|x\|\to 0$.

Recall that $x_1=x_1(\cdot,\lambda_0)$, $x_2=x_2(\cdot,\lambda_0)$ are linearly independent solutions of the linearisation of (3.3). One can show that $\mathcal{R}(B)=\{(v;c,d)\colon l_j(v;c,d)=0 \text{ for } j=1,2\}$ where $l_1(v;c,d)=-d+\int_0^1 vx_1$, $l_2(v;c,d)=c+\int_0^1 vx_2$ are linear functionals on Z. Setting $z_1=(x_1;c_1,0)$, $z_2=(x_2;0,c_2)$, $\alpha_j=\int_0^1 x_j^2$, one can define a projection $Q\colon Z\to\mathcal{R}(B)$ by

$$Qz=z-\alpha_1^{-1}l_1(z)z_1-\alpha_2^{-1}l_2(z)z_2.$$

The constants c_1,c_2 must be chosen in such a way as to assure that $l_j(Qz)=0$ for $j=1,2$ for all $z\in Z$, and this is equivalent to requiring $l_1(z_2)=0=l_2(z_1)$, a sort of Gram–Schmidt manipulation. One sees then that $c_2=\int_0^1 x_1x_2=-c_1$ satisfy this requirement. Further, let us define a projection $P\colon\overline{Y}\to\overline{Y}_0=$(linear span of x_1,x_2) by $Px=\alpha_1^{-1}(\int_0^1 xx_1)x_1+\alpha_2^{-1}(\int_0^1 xx_2)x_2$.

The method of Lyapunov–Schmidt consists of replacing (3.3) by the pair of equations

$$Q(B-\nu A-\varepsilon C)(Px+(I-P)x)=QG(x)+O\big(|\alpha|^2\|x\|+|\alpha|\,\|x\|^n+\|x\|^{n+1}\big), \tag{3.4}$$

$$(I-Q)(B-\nu A-\varepsilon C)(Px+(I-P)x)=(I-Q)G(x)+O\big(|\alpha|^2\|x\|+|\alpha|\,\|x\|^n+\|x\|^{n+1}\big). \tag{3.5}$$

Rewrite $Px=u_1x_1+u_2x_2$ for real numbers u_1, u_2. Since $Bx_i=0$ for $i=1,2$ and $QB(I-P)$: $\bar{Y}\to\mathcal{R}(B)$ is a linear operator with bounded inverse, equation (3.4) can be solved by $(I-P)x=w^*(\mathbf{u},\boldsymbol{\alpha})\in\bar{Y}$ for all sufficiently small $|\mathbf{u}|$, $|\boldsymbol{\alpha}|$, where $|\mathbf{u}|=|u_1|+|u_2|$. Furthermore $w^*=O(|\boldsymbol{\alpha}||\mathbf{u}|+|\mathbf{u}|^n)$ as $|\mathbf{u}|, |\boldsymbol{\alpha}|\to 0$, where n is the same integer n as in the estimate that both f_0 and $\mathbf{f}$ are $O(|\boldsymbol{\alpha}|^2\|x\|+\|x\|^n)$ as $|\boldsymbol{\alpha}|$, $\|x\|\to 0$.

Substitute $x=u_1x_1+u_2x_2+w^*$ into (3.5) to arrive at the bifurcation equation

$$(3.6)\qquad (I-Q)(B-\nu A-\varepsilon C)(u_1x_1+u_2x_2+w^*(\mathbf{u},\boldsymbol{\alpha}))$$
$$-(I-Q)G(u_1x_1+u_2x_2+w^*)=O\left(|\boldsymbol{\alpha}|^2|\mathbf{u}|+|\boldsymbol{\alpha}||\mathbf{u}|^{n+1}\right).$$

Now, $(I-Q)B\equiv 0$ by design of the projection Q. Using the linear independence of z_1, z_2 one can separate (3.6) into a system of two equations, after first multiplying through by -1:

$$(3.7)\qquad (\nu\Sigma-\varepsilon JH)\mathbf{u}+\mathbf{p}(\mathbf{u})=\mathbf{R}(\boldsymbol{\alpha},\mathbf{u})$$

where $\Sigma=(\sigma_{ij})_{i,j=1,2}$, $\sigma_{ij}=\int_0^1 ax_ix_j$, $J=\left(\begin{smallmatrix}0&1\\-1&0\end{smallmatrix}\right)$, $\mathbf{p}(\cdot)$ is homogeneous of degree n, $\mathbf{R}(\boldsymbol{\alpha},\mathbf{u})=O(|\boldsymbol{\alpha}|^2|\mathbf{u}|+|\boldsymbol{\alpha}||\mathbf{u}|^n+|\mathbf{u}|^{n+1})$, n as above, and H is as in §2 and the definition of the operator C.

As one can see, all of the information in §2 concerning the linear perturbation of the periodic boundary value problem is found also in (3.7). So, Lyapunov–Schmidt for an equation in Banach spaces correctly abstracts the linear problem. The author has shown [14] that the method of Fulton [7], Walter [15], Browne and Sleeman [2] et al., abstracting perturbations of the separated boundary conditions into the Hilbert space $L_2[0,1]\times\mathbf{C}^2$, fails to preserve the self-adjoint features of perturbation of the boundary value problem. Specifically, the analogue of the operator B for the $L_2[0,1]\times\mathbf{C}^2$ setting has either (i) codim $\mathbf{R}(B)\geq 2$ for *all* $\lambda_0\in\mathbf{C}$, not just eigenvalues, or (ii) B not self-adjoint. Cases (i), (ii) correspond to different definitions of $\mathcal{D}(B)$; one may recall that in Fulton et al. the dependence of the boundary conditions on a parameter is arranged by an efficacious choice of $\mathcal{D}(B)$. This is a definite distinction between the periodic and separated boundary conditions.

Note added in proof. See also Robert Magnus, *Topological equivalence in bifurcation theory*, in Lecture Notes in Mathematics 799, Springer-Verlag, 1980.

REFERENCES

[1] G. D. Birkhoff, *Existence and oscillation theorem for a certain boundary value problem*, Trans. Amer. Math. Soc., 10 (1909), pp. 259–270.

[2] P. J. Browne and B. D. Sleeman, *Regular multiparameter eigenvalue problems with several parameters in the boundary conditions*, J. Math. Anal. Appl., 72 (1979), pp. 29–33.

[3] H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids*, 2nd ed., Oxford, 1959.

[4] S.-N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, Berlin, 1983.

[5] S.-N. Chow, J. K. Hale and J. Mallet-Paret, *Applications of generic bifurcation*, II, Arch. Rational Mech. Anal., 62 (1976), pp. 209–235.

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

[7] C. T. Fulton, *Two-point boundary value problems with eigenvalue parameter contained in the boundary conditions*, Proc. Roy. Soc. Edinburgh A, 77 (1977), pp. 293–308.

[8] J. K. Hale, *Ordinary Differential Equations*, 2nd ed., Krieger, Huntington, NY, 1980.

[9] ——, *Bifurcation from simple eigenvalues for several parameter families*, Nonlinear Anal., 2 (1978), pp. 491–497.

[10] J. P. Holman, *Heat Transfer*, 4th ed., McGraw-Hill, New York, 1976.

[11] W. Magnus and S. Winkler, *Hill's Equation*, Interscience, New York, 1966.

[12] M. N. Özişik, *Boundary Value Problems of Heat Conduction*, International Textbook Co., Scranton, PA, 1968.

[13] F. Rellich, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, 1969.

[14] L. Turyn, unpublished note.

[15] J. Walter, *Regular eigenvalue problems with eigenvalue parameter in the boundary condition*, Math. Z., 133 (1973), pp. 301–312.

[16] L. Turyn, *Perturbations of periodic boundary conditions*, in Proceedings of the Seventh Conference on Ordinary and Partial Differential Equations, Dundee, 1982, Lecture Notes in Mathematics 964, Springer-Verlag, Berlin, 1982.

[17] ______, *Perturbation of two-point boundary value problems with eigenvalue parameter in the boundary conditions*, Proc. Roy. Soc. Edinburgh A, 94 (1983), to appear.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics
004

A PRÜFER TRANSFORMATION FOR LIÉNARD'S EQUATION*

DONALD C. BENSON†

Abstract. Conditions are given for oscillation and nonoscillation of solutions of Liénard's equation. In the oscillatory case, estimates are given for the decrement (i) of the magnitude of the solution from one zero of the derivative to the next, (ii) of the magnitude of the derivative from one zero of the solution to the next. In certain cases the estimates are sharp. The results are obtained by using a Prüfer transformation.

Introduction. This article is the third in a series of papers, including [2] and [3], by the author dealing with the Liénard equation, namely the ordinary differential equation

$$\frac{d^2x}{dt^2}+k(x(t))\frac{dx}{dt}+h(x(t))=0.$$

Since the present article concludes a chapter in this investigation, it seems appropriate to discuss the theme of this series of articles.

The starting point is oscillation and comparison theory for linear differential equations, which is expounded for example in [11] and [16]. The fact emerges that the success of this theory is largely based upon certain very powerful devices which appear to depend heavily on the special form of the linear second order differential equation. These devices are three in number: the Picone identity, the Riccati equation, and the Prüfer transformation.

As seen in [11] and [16], generalizations of these devices have generally been in the domain of *linear* equations. It would seem to be useful to find for some major class of nonlinear differential equations devices analogous to those mentioned above without resorting to linearization. It is the purpose of [2], [3] and the present paper to show that this can be done for the Liénard equation. In [2] and [3] are found analogues for the Liénard equation of the Picone identity and the Riccati equation, respectively; in the present paper, an analogue of the Prüfer transformation is developed.

The results obtained for the Liénard equation are only somewhat analogous to those which are obtained for the linear case. In some cases, roughly speaking, the role of the domain of the solutions for the linear case seems analogous to the role of the range of solutions for the Liénard equation. For example, in [2, Thm. 2] it is natural in considering two different Liénard equations to compare two solutions of the respective equations with the same range, whereas, in the linear case, it is natural to compare solutions with the same domain. Further, in [3] one does not look at the order of magnitude of the monotone solutions themselves for Liénard's equation (in the nonoscillatory case) but of the inverse functions of these solutions, whereas in the linear case one deals with the order of magnitude of the solutions themselves.

The analogies which have already been obtained suggest that the further study of Liénard's equation and its generalizations may yield a richness of results comparable to that which is known for the linear equation. Many other authors have studied the Liénard equation in various degrees of generality; see for example [6] and [8]. These two references have extensive bibliographies which cover work in the field up to 1970. More recent work includes [1], [4], [7], [13], [14], and [15].

*Received by the editors August 11, 1981, and in revised form February 4, 1983.

† Department of Mathematics, University of California, Davis, California 95616.

The Liénard equation may appear to be rather special to merit such a detailed investigation. However, the form of the equation given above involves two arbitrary functions; in this sense, the degree of generality is the same as that of the second order linear equation. Beyond this, Liénard's equation is studied because it gives a model for damped nonlinear vibrations (see [12]). The results which are obtained can be interpreted physically.

The principal tool of this article, the Prüfer transformation, is a device used in the study of certain linear ordinary differential equations, e.g., to obtain oscillation and comparison theorems. See [9], [10], and [11] for an account of the Prüfer transformation. A Prüfer transformation consists in giving a suitable polar coordinate representation of the trajectories of the solutions. In §1, a θ-coordinate is introduced and is used to prove a nonoscillation theorem. A phase plane different from the usual one (and different from the one used by Liénard [12, p. 105]) is used here.

In §2, a suitable r-coordinate is introduced. Among other things, it is shown that under certain conditions, Liénard's equation can have both oscillatory and nonoscillatory solutions. This is illustrated by means of an example.

In §3, an r-coordinate, different from the one of §2, is introduced and is used to prove an oscillation theorem.

Estimates of decrement for the magnitude of the solution and its derivative are given in §§2 and 3.

1. A nonoscillation theorem. Throughout this paper we will assume the following.

Condition A. *The functions* h,k: $\mathbb{R}\to\mathbb{R}$ *are continuous*; $k(X)$ *and* $Xh(X)$ *are positive unless* $X=0$; *a is a real number*; x: $[a,\infty)\to\mathbb{R}$ *is a solution, not identically zero, of Liénard's equation, namely the equation*

$$\frac{d^2x}{dt^2}+k(x(t))\frac{dx}{dt}+h(x(t))=0. \tag{1.1}$$

From [8, p. 35] it is known that for any real numbers x_0 and v_0 a solution of (1.1) satisfying $x(a)=x_0$ and $x'(a)=v_0$ exists on $[a,\infty)$. Moreover from [2, p. 258] this solution is unique. (No Lipschitz condition, or other condition beyond what is given above, is needed for uniqueness.)

Definition. The solution $x(\cdot)$, not identically zero, is said to be *oscillatory* if it has infinitely many zeros on $[a,\infty)$. Otherwise, $x(\cdot)$ is said to be *nonoscillatory.*

Remark. Unless $x(t)=0$ for all t in $[a,\infty)$, the uniqueness mentioned above implies that $x(\cdot)$ and $x'(\cdot)$ cannot vanish simultaneously. Hence $x(\cdot)$ must change sign at every isolated zero. Moreover, a finite subinterval of $[a,\infty)$ cannot contain infinitely many zeros unless $x(\cdot)$ is identically zero; if such were the case, the zeros of $x(\cdot)$ and $x'(\cdot)$ would have a common point of accumulation which, by continuity, would be a common zero of $x(\cdot)$ and $x'(\cdot)$.

Put

$$H(X)=\int_0^X h(\xi)\,d\xi \quad\text{and}\quad K(X)=\int_0^X k(\xi)\,d\xi. \tag{1.2}$$

If $x(\cdot)$ is not identically zero, then $K(x(t))$ and $x'(t)$ cannot vanish simultaneously. Hence we may define Θ: $[a,\infty)\to\mathbb{R}$ by the relation

$$e^{i\Theta(t)}=\frac{x'(t)+iK(x(t))}{\left(K(x(t))^2+x'(t)^2\right)^{1/2}}.$$

The further requirement that $\Theta(\cdot)$ is continuous and $0<\Theta(a)<2\pi$ determines $\Theta(\cdot)$ uniquely.

This construction has a simple geometrical interpretation. The solution $x(\cdot)$ determines a motion in the complex plane in which the real part is $x'(t)$ and the imaginary part is $K(x(t))$. In this phase plane the usual polar angle is $\Theta(t)$.

Using (1.1), we obtain

$$(1.3)\qquad \frac{d\Theta}{dt}=\begin{cases}\dfrac{k(x)\left(x'^2+x'K(x)+h(x)k(x)^{-1}K(x)\right)}{K(x)^2+x'^2} & \text{for } x\neq 0,\\ k(0) & \text{for } x=0.\end{cases}$$

Put

$$(1.4)\qquad M[x]=\sup\left\{h(x(t))k(x(t))^{-1}K(x(t))^{-1}: x(t)\neq 0, t\in[a,\infty)\right\}.$$

The following theorem generalizes the criteria for underdamping and overdamping of vibrations, well known in case h and k are constants.

THEOREM 1.1. *Let Condition* A *hold and let* $M[x]<\frac{1}{4}$. *Then* $x(\cdot)$ *has at most one zero in* $[a,\infty)$.

Proof. At any point where the trajectory of $x(\cdot)$ crosses the real axis, $d\Theta/dt=k(0)$ which is positive unless $k(0)=0$; even in that case $d\Theta/dt>0$ on the trajectory at least in some deleted neighborhood of the point at which the trajectory crosses the real axis because the expression in parentheses on the right side of (1.3) is continuous and positive on the real axis except at the origin.

On the other hand

$$\frac{d\Theta}{dt}\leq\frac{k(x)\left(x'^2+K(x)x'+M[x]K^2\right)}{K(x)^2+x'^2}.$$

Note that the expression in parentheses is a quadratic form in x' and $K(x)$. Since $M[x]<\frac{1}{4}$, an examination of the discriminant of this form shows that the form is indefinite. Hence there is at least one line l through the origin on which $d\Theta/dt$ is negative. Let t_1 be such that $x(t_1)=0$. The trajectory must cross the real axis in the counter-clockwise direction at $t=t_1$ because $d\Theta/dt$ must be positive at least in a deleted neighborhood of the crossing. Suppose there is another zero; let t_2 be the next zero. We must have $\Theta(t_2)=\Theta(t_1)+\pi$. But then for some t_3 in (t_1,t_2), the trajectory must cross the line l in the counter-clockwise direction; however, this is impossible because we must have $\Theta'(t_3)<0$. We conclude from this contradiction that $x(\cdot)$ can have at most one zero on $[a,\infty)$.

COROLLARY 1.1. *Let Condition* A *hold and let*

$$(1.5)\qquad \sup\left\{h(X)k(X)^{-1}K(X)^{-1}: X\in\mathbb{R}, X\neq 0\right\}<\tfrac{1}{4}.$$

Then $x(\cdot)$ *has at most one zero on* $[a,\infty)$.

Proof. If (1.5) holds, then $M[x]<\frac{1}{4}$, and the assertion follows from the theorem.

DEFINITION. We say that the zero solution of (1.1) on $[a,\infty)$ is *globally asymptotically stable* (g.a.s.) if every solution of (1.1) on $[a,\infty)$ satisfies $x(t)\to 0$ and $x'(t)\to 0$ as $t\to\infty$.

It is known [5] that the zero solution of (1.1) is g.a.s. if and only if

$$(1.6)\qquad |K(X)|+H(X)\to\infty \quad \text{as } X\to\pm\infty.$$

(See (1.2) for the definitions of $H(\cdot)$ and $K(\cdot)$.) The reference [5] makes the assumption $k(X)>0$, but the argument is valid even if $k(0)=0$ is allowed.

COROLLARY 1.2. *Let Condition* A *hold. Let the zero solution of* (1.1) *be g.a.s. Further let*

$$\limsup_{X\to 0} h(X)k(X)^{-1}K(X)^{-1}$$

be less than $\frac{1}{4}$. *Then* $x(\cdot)$ *is nonoscillatory on* $[a,\infty)$.

Proof. Let a' be chosen so that

$$\sup\{h(x(t))k(x(t))^{-1}K(x(t))^{-1}: t\in[a',\infty)\}$$

is less than $\frac{1}{4}$. We apply Theorem 1.1 with a replaced by a', and the desired conclusion follows.

2. The case $m[x]>\frac{1}{4}$, $M[x]<\infty$. Let $x(\cdot)$ be a solution of (1.1) on $[a,\infty)$. Define, in analogy with (1.4),

$$m[x]=\inf\{h(x(t))k(x(t))^{-1}K(x(t))^{-1}: x(t)\neq 0, t\in[a,\infty)\}.$$

This section is devoted to the study of solutions of (1.1) satisfying

$$m[x]>\tfrac{1}{4}. \tag{2.1}$$

As in §1, $M[x]$ is defined by (1.4). In this section, we assume also that $M[x]$ is finite. In §3, we obtain further results without this assumption.

Theorem 2.1 gives upper and lower bounds for $K(x(t))$ at a zero of $x'(t)$ when the value of $x'(t)$ at a zero of $K(x(t))$ is known and also gives bounds for $x'(t)$ at a zero of $K(x(t))$ when the value of $K(x(t))$ at a zero of $x'(t)$ is known. In either case, it is assumed that the interval between the two zeros in question contains no other zeros of $K(x(t))$ or $x'(t)$. In mechanical terms, we estimate the maximum displacement when the particle is initially at equilibrium with known velocity; and we estimate the velocity at equilibrium when the particle is initially at rest with known displacement.

Theorem 2.2 shows that, under certain conditions, equation (1.1) admits of both oscillatory and nonvanishing solutions. A criterion is given which in a particular example discriminates sharply between the two types of solutions.

We consider solutions of (1.1) with the same conditions on $h(\cdot)$ and $k(\cdot)$ which are described in §1. We define $\rho(t)=(x'(t)^2+K(x(t))^2)^{1/2}$. In the phase plane described in the §1, $\rho(t)$ is the distance from the origin. Define $\Theta(\cdot)$ as in the previous section and put $z(t)=\rho(t)e^{i\Theta(t)}$; the trajectory $z(t)$, $a\le t<\infty$ gives us a representation of the solution $x(\cdot)$.

We compute, using (1.1),

$$\frac{d(\rho^2)}{dt}=2K(x)k(x)x'+2x'(-x'k(x)-h(x)),$$

which may be written

(2.2)

$$\frac{1}{\rho}\frac{d\rho}{dt}=\begin{cases} k(x)\left(\sin\Theta\cos\Theta-\cos^2\Theta-\sin\Theta\cos\Theta h(x)k(x)^{-1}K(x)^{-1}\right) & \text{if } x\neq 0,\\ -k(0) & \text{if } x=0.\end{cases}$$

Also (1.3) may be written

$$(2.3)\qquad \frac{d\Theta}{dt}=\begin{cases} k(x)\left(\cos^2\Theta+\cos\Theta\sin\Theta+\sin^2\Theta h(x)k^{-1}(x)K^{-1}(x)\right) & \text{if } x\neq 0,\\ k(0) & \text{if } x=0.\end{cases}$$

From (2.1) follows

$$(2.4)\qquad \frac{d\Theta}{dt}\geq k(x)\left(\cos^2\Theta+\cos\Theta\sin\Theta+m[x]\sin^2\Theta\right).$$

The quadratic form on the right is positive definite as may be seen by computing its discriminant. Moreover, $k(x(t))$ can be zero only at isolated points unless $x(t)$ is identically zero. Hence $\Theta(t)$ is strictly increasing and we may introduce $\theta=\Theta(t)$ as a new independent variable. Define the function $r(\cdot)$ by the relation $r(\Theta(t))=\rho(t)$. We have, by (2.2) and (2.3),

$$(2.5)\qquad \frac{1}{r}\frac{dr}{d\theta}=\begin{cases} \dfrac{\left(\sin\theta\cos\theta-\cos^2\theta-\sin\theta\cos\theta h(x)k(x)^{-1}K(x)^{-1}\right)}{\cos^2\theta+\cos\theta\sin\theta+\sin^2\theta h(x)k^{-1}K(x)^{-1}} & \text{if } x\neq 0,\\ -1 & \text{if } x=0.\end{cases}$$

Below we will need to estimate $dr/d\theta$. We prove a lemma which prepares for this. On the right side of (2.3), put $h(x)k(x)^{-1}K(x)^{-1}=\alpha$; and put

$$\varphi(\theta,\alpha)=\frac{\sin\theta\cos\theta-\cos^2\theta-\alpha\sin\theta\cos\theta}{\cos^2\theta+\cos\theta\sin\theta+\alpha\sin^2\theta}.$$

LEMMA 2.1. *The function $\varphi(\theta,\alpha)$ is continuous in the halfplane $\{(\theta,\alpha)\in\mathbb{R}^2\colon \alpha>\frac{1}{4}\}$; $\partial\varphi/\partial\alpha$ is negative for θ in the first and third quadrants $(n\pi<\theta<n\pi+(\pi/2))$ and positive in the second and fourth quadrants $(n\pi-(\pi/2)<\theta<n\pi)$.*

Proof. The continuity follows from the fact that the denominator

$$\cos^2\theta+\cos\theta\sin\theta+\alpha\sin^2\theta$$

is positive definite for $\alpha>\frac{1}{4}$, as was already observed in connection with (2.3). Moreover, a calculation yields

$$\frac{\partial\varphi}{\partial\alpha}=-\frac{\sin\theta\cos\theta}{\left(\cos^2\theta+\cos\theta\sin\theta+\alpha\sin^2\theta\right)^2},$$

which shows immediately the assertions concerning the sign of $\partial\varphi/\partial\alpha$, and the proof of the lemma is concluded.

We now proceed to develop the estimates mentioned above. We prove the major lemma, upon which all the further results of §2 are based. First we define some terminology. Let m and M satisfy $\frac{1}{4}<m\leq M<\infty$. Put $\delta=(4m-1)^{-1/2}$, $\Delta=(4M-1)^{-1/2}$, $\sigma=\delta\arctan\delta-\Delta\arctan\Delta+\frac{1}{2}\log(M/m)$.

LEMMA 2.2. *Let Condition* A *hold and let m and M satisfy*

$$(2.6)\qquad \tfrac{1}{4}<m\leq m[x]\leq M[x]\leq M<\infty.$$

Then there exist positive, continuous functions $r_1(\cdot)$ and $r_2(\cdot)$ on $\mathbb{R}$ such that

$$(2.7)\qquad r_1(\Theta(t))\leq r(\Theta(t))=\rho(t)\leq r_2(\Theta(t))\quad \textit{for all } t \textit{ in } [a,\infty).$$

Moreover $r_1(\cdot)$ and $r_2(\cdot)$ satisfy the following:

$$r_1(\theta_0)=r(\theta_0)=r_2(\theta_0), \tag{2.8}$$

$$r_1(\theta+\pi)=r_1(\theta)\exp\left(-\frac{\pi}{2}(\Delta+\delta)-\sigma\right), \tag{2.9}$$

$$r_2(\theta+\pi)=r_2(\theta)\exp\left(-\frac{\pi}{2}(\Delta+\delta)+\sigma\right), \tag{2.10}$$

for all θ in $\mathbb{R}$ and $\theta_0=\Theta(a)$.

Furthermore, for any integer n the following relations hold:

$$r_1\left(n\pi+\frac{\pi}{2}\right)=r_1(n\pi)\exp\left(-\frac{\pi}{2}\Delta-\frac{1}{2}\log M+\Delta\arctan\Delta\right), \tag{2.11}$$

$$r_2\left(n\pi+\frac{\pi}{2}\right)=r_2(n\pi)\exp\left(-\frac{\pi}{2}\delta-\frac{1}{2}\log m+\delta\arctan\delta\right), \tag{2.12}$$

$$r_1(n\pi)=r_1\left(n\pi-\frac{\pi}{2}\right)\exp\left(-\frac{\pi}{2}\delta+\frac{1}{2}\log m-\delta\arctan\delta\right), \tag{2.13}$$

$$r_2(n\pi)=r_2\left(n\pi-\frac{\pi}{2}\right)\exp\left(-\frac{\pi}{2}\Delta+\frac{1}{2}\log M-\Delta\arctan\Delta\right). \tag{2.14}$$

Moreover, the functions

$$u_1(\theta)=r_1(\theta)\sin\theta \quad \text{and} \quad u_2(\theta)=r_2(\theta)\sin\theta$$

achieve local extrema only at points of the form $\theta=n\pi+(\pi/2)$ where n is an integer.

Furthermore,

$$\lim_{\theta\to\infty} r_1(\theta)=0; \tag{2.15}$$

and, if

$$\sigma<\frac{\pi}{2}(\Delta+\delta), \tag{2.16}$$

then

$$\lim_{\theta\to\infty} r_2(\theta)=0. \tag{2.17}$$

Proof. From (2.5) and Lemma 2.1,

$$\varphi(\theta,\alpha)\le\frac{1}{r}\frac{dr}{d\theta}\le\varphi(\theta,\beta), \tag{2.18}$$

where $\alpha=M$ and $\beta=m$ in the first and third quadrants, whereas $\alpha=m$ and $\beta=M$ in the second and fourth quadrants. Integrating these differential inequalities, certain conclusions may be drawn. In particular, let t_1 and t_2 $(t_2>t_1)$ be in $[a,\infty)$ and put $\theta_1=\Theta(t_1)$ and $\theta_2=\Theta(t_2)$; then

$$r(\theta_1)\exp\int_{\theta_1}^{\theta_2}\varphi(\theta,\alpha)\,d\theta\le r(\theta_2)\le r(\theta_1)\exp\int_{\theta_1}^{\theta_2}\varphi(\theta,\beta)\,d\theta. \tag{2.19}$$

Putting $\theta_0=\Theta(a)$, we see that (2.7) and (2.8) hold with

$$r_1(\theta)=r(\theta_0)\exp\int_{\theta_0}^{\theta}\varphi(\psi,\alpha)\,d\psi, \tag{2.20}$$

$$r_2(\theta)=r(\theta_0)\exp\int_{\theta_0}^{\theta}\varphi(\psi,\beta)\,d\psi.$$

For any θ_1 and θ_2

$$(2.21)\qquad r_1(\theta_2)=r_1(\theta_1)\exp\int_{\theta_1}^{\theta_2}\varphi(\psi,\alpha)\,d\psi,$$
$$r_2(\theta_2)=r_2(\theta_1)\exp\int_{\theta_1}^{\theta_2}\varphi(\psi,\beta)\,d\psi.$$

The integrals in (2.19) can be computed explicitly by using the substitution $u=\tan\theta$ and an expansion into partial fractions. In fact for *constant* γ we have the indefinite integral

$$(2.22)\qquad \int\varphi(\theta,\gamma)\,d\theta=-\frac{1}{2}\log(\cos^2\theta+\cos\theta\sin\theta+\gamma\sin^2\theta)$$
$$-(4\gamma-1)^{-1/2}\arctan\frac{2\gamma\tan\theta+1}{(4\gamma-1)^{-1/2}}.$$

We obtain (2.11) and (2.12) from (2.21) by putting $\theta_1=n\pi$ and $\theta_2=n\pi+(\pi/2)$ and $\gamma=\alpha=M$ and $\gamma=\beta=m$, respectively; similarly (2.13) and (2.14) are obtained from (2.21) by putting $\theta_1=n\pi-(\pi/2)$ and $\theta_2=n\pi$ and $\gamma=\alpha=m$ and $\gamma=\beta=M$, respectively. Moreover, (2.9) and (2.10) follow from the fact that $\varphi(\theta,\alpha)$ and $\varphi(\theta,\beta)$ are both periodic with period π, and hence the integral of $\varphi(\theta,\alpha)$ (similarly $\varphi(\theta,\beta)$) over an interval of length π does not depend on the choice of the endpoints; hence, for every θ,

$$(2.23)\qquad \int_{\theta}^{\theta+\pi}\varphi(\psi,\alpha)\,d\psi=\int_{-\pi/2}^{0}\varphi(\psi,m)\,d\psi+\int_{0}^{\pi/2}\varphi(\psi,M)\,d\psi=-\frac{\pi}{2}(\Delta+\delta)-\sigma,$$

and, similarly

$$(2.24)\qquad \int_{\theta}^{\theta+\pi}\varphi(\psi,\beta)\,d\psi=-\frac{\pi}{2}(\Delta+\delta)+\sigma.$$

Using (2.21), we see that (2.23) and (2.24) are equivalent to (2.9) and (2.10), respectively.

Now we establish the assertion concerning $u_1(\cdot)$ and $u_2(\cdot)$. Since

$$\frac{1}{r_1}\frac{dr_1}{d\theta}=\varphi(\theta,\alpha),\qquad \frac{1}{r_2}\frac{dr_2}{d\theta}=\varphi(\theta,\beta),$$

we have

$$\frac{du_1}{d\theta}=\frac{dr_1}{d\theta}\sin\theta+r_1\cos\theta=r_1(\varphi(\theta,\alpha)\sin\theta+\cos\theta)=\frac{r_1\cos\theta}{\cos^2\theta+\cos\theta\sin\theta+\alpha\sin^2\theta}.$$

Similarly,

$$\frac{du_2}{d\theta}=\frac{r_2\cos\theta}{\cos^2\theta+\cos\theta\sin\theta+\beta\sin^2\theta}.$$

It is clear that critical points occur only at points of the form $\theta=n\pi+(\pi/2)$. Since the derivatives change sign at these points, they are indeed extrema.

Finally, we must show (2.15) and (2.17). Let n be the largest integer not exceeding θ/π; put

$$A_1=\sup\{r_1(\theta)\colon 0\le\theta\le\pi\}$$

and

$$A_2=\sup\{r_2(\theta): 0\le\theta\le\pi\}.$$

From (2.9) and (2.10) we have

$$r_1(\theta)\le A_1\exp\left(-\frac{\pi n}{2}(\Delta+\delta)-\sigma n\right),$$

$$r_2(\theta)\le A_2\exp\left(-\frac{\pi n}{2}(\Delta+\delta)+\sigma n\right).$$

The first exponent is always negative; the second is negative if (2.16) holds. Thus, since $n\to\infty$ as $\theta\to\infty$, (2.15) and (2.17) hold as required. This concludes the proof of the lemma.

Formula (2.7) is an estimate of the type described in the second paragraph of this section. This is clarified by the following theorem.

THEOREM 2.1. *Let Condition* A *and* (2.6) *hold. Let* t_1 *and* t_2 *be in* $[a,\infty)$ *and* $x(t_1)=0$, $x'(t_2)=0$, *and let* $x(t)\neq 0$ *and* $x'(t)\neq 0$ *for all* t *between* t_1 *and* t_2.

A. *If* $t_1<t_2$, *then*

$$(2.25)\quad |x'(t_1)|\exp\left(-\frac{\pi}{2}\Delta-\frac{1}{2}\log M\right)\le|K(x(t_2))|$$
$$\le|x'(t_1)|\exp\left(-\frac{\pi}{2}\delta-\frac{1}{2}\log m+\delta\arctan\delta\right).$$

B. *If* $t_2<t_1$, *then*

$$(2.26)\quad |K(x(t_2))|\exp\left(-\frac{\pi}{2}\delta+\frac{1}{2}\log m-\delta\arctan\delta\right)$$
$$\le|x'(t_1)|\le|K(x(t_2))|\exp\left(-\frac{\pi}{2}\Delta+\frac{1}{2}\log M-\Delta\arctan\Delta\right).$$

Proof. We use (2.7) with $a=\min(t_1,t_2)$ and $t=\max(t_1,t_2)$. Since for some integer n, $\Theta(t_1)=n\pi$, $\Theta(t_2)=n\pi\pm(\pi/2)$ where the plus sign is taken if $t_1<t_2$ and the minus if $t_1>t_2$, and since $\rho(t_1)=|x'(t_1)|$ and $\rho(t_2)=|K(x(t_2))|$, (2.7) together with (2.11), (2.12), (2.13) and (2.14) imply (2.25) and (2.26). This concludes the proof.

The following result is included so that we may compare the results of this section with the results of the next section.

COROLLARY 2.1. *Let Condition* A *and* (2.6) *hold. Let* t_1 *and* t_2 $(t_1<t_2)$ *be consecutive zeros of* $x(\cdot)$. *Then*

$$|x'(t_1)|\exp\left(-\frac{\pi}{2}(\Delta+\delta)-\sigma\right)\le|x'(t_2)|\le|x'(t_2)|\exp\left(-\frac{\pi}{2}(\Delta+\delta)+\sigma\right).$$

Proof. Note that there is precisely one zero of $x'(\cdot)$ between t_1 and t_2. Indeed, if there were more than one zero of $x'(\cdot)$ in this interval then there would have to be two consecutive extrema which are either both relative maxima or both relative minima. Either case is impossible.

Let t_3 denote the unique zero of $x'(\cdot)$ between t_1 and t_2. We apply (2.26) to the interval $[t_1,t_3]$, and then apply (2.25) to the interval $[t_3,t_2]$. These inequalities imply the desired inequality. Alternatively, one can employ (2.9) and (2.10).

Recall that (1.6) is a necessary and sufficient condition for the zero solution of (1.1) to be g.a.s. It is interesting now to consider the behavior of solutions of (1.1) in case (1.6) does not hold. We shall illustrate results of this type by means of the equation

$$x''+\frac{1}{1+x^2}x'+C\frac{\arctan x}{1+x^2}=0 \tag{2.27}$$

with C constant and greater than $1/4$. In particular, we shall find a constant V such that the initial value problem of (2.27) subject to $x(a)=0$, $x'(a)=v_0$ has an oscillatory solution satisfying $x(t)\to 0$, $x'(t)\to 0$ if $v_0<V$; whereas if $v_0\geq V$, the initial value problem has a solution which is nonvanishing for $t>a$ and which satisfies $|x(t)|\to\infty$ as $t\to\infty$.

THEOREM 2.2. *Let Condition* A *and* (2.6) *hold. Let* $K(X)\to K^+<\infty$ *as* $X\to\infty$ *and* $K(X)\to -K^-\geq -\infty$ *as* $X\to -\infty$. *Put* $K^*=\min(K^+,K^-)$.

A. *If* (2.16) *holds and* $x(a)=0$, *and*

$$0<x'(a)<K^*m^{1/2}\exp\left(\frac{\pi}{2}\delta-\delta\arctan\delta\right), \tag{2.28}$$

then $x(\cdot)$ *is oscillatory and* $x(t)\to 0$ *and* $x'(t)\to 0$ *as* $t\to\infty$.

B. *If* $x(a)=0$ *and*

$$x'(a)\geq K^+M^{1/2}\exp\left(\frac{\pi}{2}\Delta-\delta\arctan\Delta\right), \tag{2.29}$$

then $x(t)>0$ *for all* $t>a$ *and* $x(t)\to\infty$ *as* $t\to\infty$.

Proof. A. Since (2.16) holds, $r_2(\theta)\to 0$ as $\theta\to\infty$. By (2.12), and since $r_2(0)=r(0)$,

$$r_2\left(\frac{\pi}{2}\right)=r(0)\exp\left(-\frac{\pi}{2}\delta-\frac{1}{2}\log m+\delta\arctan\delta\right).$$

Then, using $x'(a)=r(0)$ and (2.28), $r_2(\pi/2)<K^*$.

Next (2.16) and (2.10) show

$$r_2\left(\left(\frac{\pi}{2}\right)+n\pi\right)<r_2\left(\frac{\pi}{2}\right)<K^* \tag{2.30}$$

for all positive integers n. Further, Lemma 2.2 asserts that the relative extrema of $r_2(\theta)\sin\theta$ occur at $\theta=n\pi+(\pi/2)$ with n an integer. From this we conclude $r_2(\theta)\sin\theta<r_2(\pi/2)<K^*$ for all $\theta\geq 0$. Furthermore, $r(\Theta(t))\leq r_2(\Theta(t))$ for all $t\geq 0$. Geometrically, the solution trajectory must lie in a certain compact region in the $(x',K(x))$ phase plane which is contained in the strip

$$S=\{(u_1,v_1)\in\mathbb{R}^2:\ -K^-<v_1<K^+\}.$$

This implies that in the usual phase plane the trajectory

$$\{(u,v):\ u=x(t),\ v=x'(t),\ a\leq t<\infty\}$$

is contained in a compact subset of $\mathbb{R}^2$. (Note that $\mathbb{R}^2$ is mapped homeomorphically onto S by the mapping $u_1=v$, $v_1=K(u)$.)

Recall that $\Theta(t)$ is nondecreasing. Suppose $\Theta(t)\to\theta_\infty<\infty$. Hence the ω-limit set must be contained in the ray $\theta=\theta_\infty$ (in the $(x',K(x))$ phase plane). Further, the ω-limit set must be an invariant set. Since $d\Theta/dt>0$ unless $x=0$, the invariance requires that the ray be contained in the horizontal axis. (Since $k(0)$ is allowed to be zero,it is possible in view of (2.3) that $d\Theta/dt$ is zero on the horizontal axis of the $(x',K(x))$ phase plane.) However, the only invariant subset of the horizontal axis consists of the origin alone. Hence the ω-limit set is $\{(0,0)\}$, i.e.,

$$x(t)\to 0, \qquad x'(t)\to 0 \quad \text{as } t\to\infty. \tag{2.31}$$

But on the other hand $\rho(t)=r(\Theta(t))\geq r_1(\Theta(t))$ and hence $\rho(t)\geq \inf\{\mathrm{r}(\theta)\colon \Theta(a)\leq\theta\leq\theta_\infty\}>0$. But this contradicts (2.31). It follows that $\Theta(t)\to\infty$ as $t\to\infty$.

Now the relation $r_1(\Theta(t))\leq r(\Theta(t))$ implies, just as in the proof of Theorem 2.2, that $x(\cdot)$ has infinitely many zeros. Moreover, since (2.14) holds, Lemma 2.2 asserts that $r_2(\theta)\to 0$ as $\theta\to\infty$; and $r(\Theta(t))\leq r_2(\Theta(t))$ implies $x(t)\to 0$ and $x'(t)\to 0$ as $t\to\infty$. We have finished the proof of part A.

B. By (2.11), since $r_1(0)=r(0)$,

$$r_1\left(\frac{\pi}{2}\right)=r(0)\exp\left(-\frac{\pi}{2}\Delta-\frac{1}{2}\log M+\Delta\arctan\Delta\right).$$

Then, using (2.29),

$$r_1\left(\frac{\pi}{2}\right)\geq K^+. \tag{2.32}$$

Since $r(\Theta(t))\geq r_1(\Theta(t))$ and $K(x(t))<K^+$ for all t in $[a,\infty)$, (2.32) implies

$$\Theta(t)\to\theta_\infty\leq\frac{\pi}{2} \qquad \text{as } t\to\infty.$$

Therefore $x(t)>0$ for all t in (a,∞). Since $x(\cdot)$ is nondecreasing in the first quadrant, $x(t)\to X_\infty\leq\infty$ as $t\to\infty$. But X_∞ cannot be finite. Indeed, suppose $X_\infty<\infty$. Then $x'(t)\to K(X_\infty)\cot\theta_\infty$. Furthermore, in the standard phase plane $(X_\infty, K(X_\infty)\cot\theta_\infty)$ must be an invariant point, which is impossible. Hence $x(t)\to\infty$ as $t\to\infty$. This concludes the proof of the theorem.

Remark. From (1.6) it is clear that in order for (1.1) to have a solution satisfying $x(t)\to\infty$ as $t\to\infty$, as is asserted in part B of Theorem 2.2, it is necessary that either $\lim_{X\to\infty}H(X)$ or $\lim_{X\to-\infty}H(X)$ be finite. The theorem does not appear to make this assumption. However, since M is assumed to be finite,

$$h(x(t))\leq k(x(t))K(x(t))M$$

for $t\in[a,\infty)$. If $x(t)\to+\infty$, then

$$h(X)\leq k(X)K(X)M$$

for all sufficiently large X, say $X>X_0$. Integrating, we have

$$H(X)\leq H(X_0)+\frac{1}{2}K(X)^2M-\frac{1}{2}K(X_0)^2M.$$

Hence

$$\lim_{X\to\infty}H(X)\leq H(X_0)+\frac{1}{2}(K^+)^2M-\frac{1}{2}K(X_0)M;$$

in other words $\lim_{X\to\infty}H(X)<\infty$. Thus the theorem is consistent with condition (1.6).

Now we apply Theorem 2.2 to (2.27). Note the following:

$$k(X)=\frac{1}{1+X^2}, \qquad K(X)=\arctan X,$$

$$h(X)=C\frac{\arctan X}{1+X^2}, \qquad H(X)=C\frac{1}{2}(\arctan X)^2,$$

$$K(X)\to\pm\frac{\pi}{2},\qquad H(X)\to\frac{\pi^2}{8}\quad\text{as } X\to\pm\infty,$$
$$M=m=C,\quad \delta=\Delta=(4C-1)^{-1/2},\quad \sigma=0.$$

We apply Theorem 2.2 with $K^+=K^-=K^*=\pi/2$. The theorem states that if we put

$$V=\frac{\pi}{2}C^{1/2}\exp\left((4C-1)^{-1/2}\left(\frac{\pi}{2}-\arctan(4C-1)^{-1/2}\right)\right),$$

then the assertion immediately preceding Theorem 2.3 is correct.

3. The case $m[x]>\frac{1}{4}$, $M[x]\leq\infty$. It is somewhat unsatisfactory that in §2 we must assume $M[x]<\infty$. In this section we see that it is possible to find some conditions for oscillation without this assumption. Let $x(\cdot)$ be a solution of (1.1), not identically zero. We introduce a new phase plane in which to represent the solution. We define $\Theta(t)$ as in §1; however, instead of $\rho(\cdot)$, we make use of the function

$$\rho^*(t)=\frac{1}{2}x'(t)^2+H(x(t)). \tag{3.1}$$

Then $z^*=\rho^*(t)e^{e\Theta(t)}$, $t\in[a,\infty)$, is the representation of the solution which we wish to consider. If $m[x]>\frac{1}{4}$, then $\Theta(t)$ is strictly increasing, therefore, as in §2, we introduce $\theta=\Theta(t)$ as a new independent variable and define $r^*(\cdot)$ by the relation $r^*(\Theta(t))=\rho^*(t)$.

We now prove a lemma analogous to Lemma 2.2. Note that if $M<\infty$ and $\Delta=\delta$, Lemma 2.2 gives sharp results. We shall see that in this case, Lemma 3.1 does not give sharp estimates. However, Lemma 3.1 has the advantage of being free of the assumption $M<\infty$.

LEMMA 3.1. *Let Condition* A *and* (2.1) *hold. Then there exists a positive, continuous function $r_1^*(\cdot)$ on* $\mathbb{R}$ *such that*

$$r_1^*(\Theta(t))\leq r^*(\Theta(t))=\rho^*(t) \tag{3.2}$$

for all t in $[a,\infty)$ and for $\theta_0=\Theta(a)$. Moreover, putting $(4m[x]-1)^{-1/2}=\delta$, r_1^* satisfies the conditions

$$r_1^*(\theta_0)=r^*(\theta_0), \tag{3.3}$$

$$r_1^*(\theta+\pi)=r_1^*(\theta)\exp(-4\pi\delta). \tag{3.4}$$

Furthermore, for any integer n,

$$r_1^*\left(n\pi+\frac{\pi}{2}\right)=r_1^*(n\pi)\exp(-2\pi\delta+4\delta\arctan\delta), \tag{3.5}$$

$$r_1^*(n\pi)=r_1^*\left(n\pi-\frac{\pi}{2}\right)\exp(-2\pi\delta-4\delta\arctan\delta). \tag{3.6}$$

Proof. We compute, using (1.1),

$$\frac{d\rho^*}{dt}=-k(x(t))x'(t)^2. \tag{3.7}$$

Now, using (3.7) and (1.3), we obtain

$$\begin{aligned}\frac{dr^*}{d\theta}&=\frac{-x'^2\left(K(x)^2+x'^2\right)}{x'^2+x'K(x)+K(x)h(x)k(x)^{-1}}\\&\geq\frac{-x'^2\left(K(x)^2+x'^2\right)}{x'^2+x'K(x)+m[x]K(x)^2}\\&=\frac{-x'^2}{\cos^2\theta+\sin\theta\cos\theta+m[x]\sin^2\theta}\\&\geq\frac{-2r^*}{\cos^2\theta+\sin\theta\cos\theta+m[x]\sin^2\theta}.\end{aligned}\tag{3.8}$$

Put $\varphi^*(\theta,\alpha)=-2(\cos^2\theta+\sin\theta\cos\theta+\alpha\sin^2\theta)^{-1}$. Recall that $\Theta(t)$ is nondecreasing. From (3.8) follows

$$r^*(\Theta(t))\geq r^*(\Theta(a))\exp\int_{\Theta(a)}^{\Theta(t)}\varphi^*(\theta,m[x])\,d\theta,\tag{3.9}$$

for all t in $[a,\infty)$. Thus, in order to satisfy (3.2), we put

$$r_1^*(\theta)=\exp\int_{\Theta(a)}^{\Theta}\varphi^*(\psi,m[x])\,d\psi.\tag{3.10}$$

This integral can be computed explicitly; in fact, we have the indefinite integral

$$\int\varphi^*(\theta,m[x])\,d\theta=-4\delta\arctan(\delta(2m[x]\tan\theta+1)).\tag{3.11}$$

From (3.10) and (3.11) follow the properties (3.3), (3.4), (3.5), and (3.6). This concludes the proof.

We now prove a theorem analogous to Theorem 2.1.

THEOREM 3.1. *Let Condition* A *and* (2.1) *hold. Let* t_1 *and* t_2 *be in* $[a,\infty)$ *and* $x(t_1)=0$, $x'(t_2)=0$, *and let* $x(t)$ *and* $x'(t)\neq 0$ *for all* t *between* t_1 *and* t_2.

A. *If* $t_1<t_2$, *then*

$$\frac{1}{2}x'(t_1)^2\exp(-2\pi\delta+4\delta\arctan\delta)\leq H(x(t_2)).\tag{3.12}$$

B. *If* $t_2>t_1$, *then*

$$H(x(t_2))\exp(-2\pi\delta-4\delta\arctan\delta)\leq\frac{1}{2}x'(t_1)^2.\tag{3.13}$$

Proof. The result follows from (2.1), (3.1), (3.5), and (3.6).

The next corollary is included so that we may compare the estimates of this section with those of the previous section.

COROLLARY 3.1. *Let Condition* A *and* (2.1) *hold. Let* t_1 *and* t_2 $(t_1<t_2)$ *be consecutive zeros of* $x(\cdot)$. *Then*

$$|x'(t_1)|\exp(-2\pi\delta)\leq|x'(t_2)|.\tag{3.14}$$

Proof. As shown in the proof of Corollary 2.1, there is exactly one zero t_3 of $x'(\cdot)$ in the interval (t_1, t_2). We apply (3.12) to the interval (t_1, t_3), and we apply (3.13) to the interval (t_3, t_2). Combining these inequalities yields (3.14).

Remark. If $M = m = m[x]$, then Corollary 2.1 gives a much better result than Corollary 3.1. In fact, Corollary 2.1 yields

$$|x'(t_1)|\exp(-\pi\delta) \leq |x'(t_2)| \leq |x'(t_1)|\exp(-\pi\delta),$$

or, in other words,

$$|x'(t_2)| = |x'(t_1)|\exp(-\pi\delta),$$

which we compare with (3.14). However, if M, and hence also σ, are large, (3.14) is a better lower bound for $|x'(t_2)|$ than the one given in Corollary 2.1.

We conclude this section with an oscillation theorem. Note that a similar theorem could have been proved using Lemma 2.2. However, such a result would have been weaker because it would have to include the hypothesis $M[x] < \infty$.

THEOREM 3.2. *Let Condition* A *and* (2.1) *hold. Let the zero solution of* (1.1) *be g.a.s. Then* $x(\cdot)$ *has infinitely many zeros*; *i.e. all solutions of* (1.1) *oscillate.*

Proof. Since the zero solution is g.a.s., we must have $\rho^*(t) \to 0$ as $t \to \infty$. Suppose $|\Theta(t)| \leq M$ for some constant M. Then we would have

$$\rho^*(t) = r^*(\Theta(t)) \geq r_1^*(\Theta(t)) \geq \inf\{r_1^*(\theta) \colon |\theta| \leq M\} > 0,$$

which contradicts the fact that, since the zero solution of (1.1) is g.a.s., $\rho^*(t) \to 0$ as $t \to \infty$. Hence $\Theta(t)$ is unbounded and since $\Theta'(t) \geq 0$, we have $\Theta(t) \to +\infty$ as $t \to \infty$. A zero of $x(\cdot)$ occurs whenever $\Theta(t)$ is equal to a multiple of π. Hence $x(\cdot)$ has infinitely many zeros; i.e., $x(\cdot)$ is oscillatory. This concludes the proof.

The following corollary is a result previously obtained by the author using other methods [3].

COROLLARY 2.2. *Let Condition* A *hold, and let*

$$\liminf_{X \to 0} h(X)k(X)^{-1}K(X)^{-1} > \tfrac{1}{4}.$$

Let the zero solution of (1.1) *be g.a.s. Then* $x(\cdot)$ *has infinitely many zeros*; *i.e., all solutions of* (1.1) *oscillate.*

Proof. Since $x(t) \to 0$ as $t \to 0$, we may choose a' such that

$$\inf\{h(x(t))k(x(t))^{-1}K(x(t))^{-1} \colon x \in [a', \infty)\} > \tfrac{1}{4}.$$

The result follows by applying Theorem 3.2 with a replaced by a'.

Remark. Theorem 3.2 can be proved much more simply if the hypothesis

$$\inf\{k(x(t)) \colon t \in [a, \infty)\} \geq 0 \tag{3.15}$$

is included. In this case, note that (3.15) implies (1.6) and, therefore, the hypothesis is unnecessary that the zero solution of (1.1) is g.a.s. Let μ be the inf in (3.15), and let

$$\lambda = \inf\{\cos^2\Theta(t) + \cos\Theta + m[x]\sin^2\Theta\}.$$

Observe that (2.1) implies $\lambda > 0$; in fact, λ is the smaller of the two eigenvalues of the quadratic form. Then (2.4) implies $d\Theta/dt \geq \mu\lambda > 0$, and hence $\Theta(t) \to \infty$ as $t \to \infty$. As in the proof of Theorem 3.2, this implies that $x(\cdot)$ is oscillatory.

REFERENCES

[1] R. J. BALLIEU AND K. PEIFFER, *Attractivity of the origin for the equation* $\ddot{x}+f(t,x,\dot{x})|\dot{x}|^{\alpha}\dot{x}+g(x)=0$, J. Math. Anal. Appl., 65 (1978), pp. 321–332.

[2] DONALD C. BENSON, *Comparison and oscillation theory for Liénard's equation with positive damping*, SIAM J. Appl. Math., 24 (1973), pp. 251–271.

[3] ______, *Principal solutions for Liénard's equation*, this Journal, 12 (1981), pp. 398–412.

[4] T. T. BOWMAN, *Periodic solutions of Liénard systems with symmetries*, Nonlinear Anal., 2 (1978), pp. 457–464.

[5] T. A. BURTON, *The generalized Liénard equation*, SIAM J. Control, 3 (1965), pp. 223–230.

[6] T. A. BURTON AND C. G. TOWNSEND, *On the generalized Liénard equation with forcing term*, J. Differential Equations, 4 (1968), pp. 620–633.

[7] L. CESARI AND R. KANNAN, *Solutions in the large Liénard systems with forcing terms*, Ann. Mat. Pura Appl., (4) 111 (1976), pp. 101–124.

[8] J. R. GRAEF, *On the generalized Liénard equation with negative damping*, J. Differential Equations, 12 (1972), pp. 34–62.

[9] PHILIP HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[10] EINAR HILLE, *Lectures on Ordinary Differential Equations*, Addison-Wesley, Reading, MA, 1969.

[11] KURT KREITH, *Oscillation Theory*, Springer-Verlag, Berlin, 1973.

[12] NICHOLAS MINORSKY, *Nonlinear Oscillation*, Van Nostrand, Princeton, NJ, 1962.

[13] D. A. NEUMANN, *Periodic solution of Liénard systems*, J. Math. Anal. Appl., 62 (1978), pp. 148–156.

[14] SILVIO NOCILLA, *Sull'integrazione delle equazioni del tipo* $\ddot{x}+c(x)\dot{x}|\dot{x}|^{n-1}+k(x)=0$, Ann. Mat. Pura Appl., (4) 111 (1976), pp. 83–99.

[15] R. REISSIG, *Extension of some results concerning the generalized Liénard equation*, Ann. Mat. Pura Appl., (4) (1975), pp. 269–281.

[16] C. A. SWANSON, *Comparison and Oscillation Theory of Linear Differential Equations*, Academic Press, New York, 1968.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics
005

ON THE EXISTENCE OF A FREE BOUNDARY FOR A CLASS OF REACTION-DIFFUSION SYSTEMS*

J. ILDEFONSO DIAZ† AND JESUS HERNANDEZ‡

Abstract. Some nonlinear stationary reaction-diffusion systems involving nonlinear terms which may be discontinuous are considered. Such systems occur, for instance, in the study of chemical reactions, and the discontinuities correspond to reactions of order zero. In such concrete models, the set where the reactant vanishes plays an important role. Here we prove the existence of solutions for a general class of such systems satisfying Dirichlet or nonlinear boundary conditions. Necessary and sufficient conditions are given assuring that the reactant component vanishes on a set of positive measure. Estimates on the location of such set are given.

AMS-MOS subject classifications (1980). Primary 35J65, 35J15, 35B99

Key words. reaction-diffusion systems, discontinuous nonlinearity, nonlinear boundary conditions, dead core set, chemical reactions

Introduction. Many papers have been devoted during recent years to the study of reaction-diffusion systems which arise very often in applications such as, mathematical biology, chemical reactions, and combustion theory.

Here we consider a system describing a single, irreversible, nonisothermic stationary reaction of the form

$$\begin{aligned} &-\Delta u+\mu^2F(u)e^{\gamma-\gamma/v}=0 \quad \text{in } \Omega,\\ &-\Delta v-\nu\mu^2F(u)e^{\gamma-\gamma/v}=0 \quad \text{in } \Omega, \end{aligned} \tag{0.1}$$

$$\begin{aligned} &\frac{\partial u}{\partial n}+\varepsilon(u-1)=0 \quad \text{on } \partial\Omega,\\ &\frac{\partial v}{\partial n}+\zeta(v-1)=0 \quad \text{on } \partial\Omega, \end{aligned} \tag{0.2}$$

where Ω is a bounded open subset of $\mathbb{R}^N$, μ^2 is the Thiele number, ν is the Prater temperature and γ is the Arrhenius number (see [3]). Here ε and ζ (the Biot numbers) are positive, being in some cases infinity, in which case (0.2) is interpreted as the Dirichlet boundary conditions

$$u=1, \qquad v=1 \quad \text{on } \partial\Omega. \tag{0.3}$$

The function $F(u)$ is assumed to be nondecreasing and it is also assumed to satisfy $F(0)=0$, $F(1)=1$ and $F(s)>0$ if $s>0$. The unknowns u and v are nonnegative and represent, respectively, the concentration and the temperature of the reactant.

Very often F takes the simple form $F(u)=u^p$, where $p\geq 0$ is the *order* of the reaction (see [3, Vol I]). In the case of a reaction of order zero F is given by $F(0)=0$ and $F(s)=1$ if $s>0$ (thus F is a *discontinuous* function).

*Received by the editors February 19, 1982.

†Facultad de Mathemáticas, Universidad Complutense de Madrid, Madrid 3, Spain. The research of this author was sponsored by the U.S. Army under contract DAAG29-80-C-0041. This work was completed while this author was visiting the Mathematics Research Center at the University of Wisconsin-Madison, Madison, Wisconsin.

‡Division de Matematicas, Universidad Autonoma de Madrid, Madrid, Spain.

Existence and uniqueness results for the parabolic problem associated with (0.1), (0.2) (or (0.1),(0.3)) have been given by some authors (cf. e.g. [2],[4],[18]). Existence and, in some particular situations, uniqueness results for the elliptic problem can be found in [2], [21], [15] and [19] for $p\geq 1$. The case $0\leq p<1$ is considered in [3, p. 311] (see also [20]) but existence theorems are not given. It is shown in [3] and [20] that for $p=0$, if μ is large enough, no strictly positive solution can exist. It is also shown that, in some particular examples, the set $\Omega_0=\{x\in\Omega\colon u(x)=0\}$ (called the *dead core*) is not empty and has positive measure if $0\leq p<1$.

The main idea used in [20] and many other papers (cf. [3]) is to reduce (0.1), (0.2) to a nonlinear elliptic boundary value problem *for u alone*. Here we follow a different approach which allows us to obtain better results. Moreover, we are able to treat the case of nonlinear boundary conditions, which cannot be handled by the preceding device.

We shall consider the case of discontinuous functions $F(u)$ in the framework of maximal montone graphs in $\mathbb{R}^2$ (see [8]). For the reader's convenience, we recall that a maximal monotone graph α in $\mathbb{R}^2$ is always specified by a real nondecreasing function θ by $\alpha(r)=(-\infty,\theta(r-)]$ if $\theta(r-)=-\infty$, $\alpha(r)=[\theta(r-),\theta(r+)]$ if $-\infty<\theta(r-)\leq\theta(r+)<+\infty$ and $\alpha(r)=[\theta(r-),+\infty)$ if $\theta(r+)=+\infty$. We define $D(\alpha)=\{r\in R\colon \alpha(r)\neq\varnothing\}$ and the sections α^+ and α^- by

$$\alpha^+(r)=\max\{z\colon z\in\alpha(r)\}\ \text{if}\ r\in D(\alpha),$$
$$\alpha^-(r)=\min\{z\colon z\in\alpha(r)\}\ \text{if}\ r\in D(\alpha),$$
$$\alpha^+(r)=\alpha^-(r)=+\infty\ \text{if}\ r\notin D(\alpha),\ r\geq\sup D(\alpha),$$
$$\alpha^+(r)=\alpha^-(r)=-\infty\ \text{if}\ r\notin D(\alpha),\ r\leq\inf D(\alpha).$$

Finally, we define $\alpha^0(r)$ as the element of $\alpha(r)$ with minimal absolute value.

Through the paper we shall study the following general formulation including the system (0.1) as a particular case:

$$\text{(NLS)}\qquad \begin{aligned} -\Delta u+\alpha(u)f(v)&\ni 0 \quad \text{in}\ \Omega,\\ -\Delta v-\beta(u)g(v)&\ni 0 \quad \text{in}\ \Omega \end{aligned}$$

with the boundary conditions

$$\text{(DBC)}\qquad u=\varphi_1,\qquad v=\varphi_2\quad \text{on}\ \partial\Omega,$$

as well as the nonlinear boundary conditions

$$\text{(NBC)}\qquad \begin{aligned} Bu&\equiv\frac{\partial u}{\partial n}+b(u)=\psi_1 \quad \text{on}\ \partial\Omega,\\ Cv&\equiv\frac{\partial v}{\partial n}+c(v)=\psi_2 \quad \text{on}\ \partial\Omega \end{aligned}$$

where Ω is a bounded open subset of $\mathbb{R}^N$ with smooth boundary $\partial\Omega$. We also assume for the rest of the paper that

(0.4) α and β are maximal monotone graphs such that $0\in\alpha(0)\cap\beta(0)$.

(0.5) f and g are C^1 functions and $f(s)\geq 0$, $g(s)\geq 0$ if $s\geq 0$.

(0.6) φ_1, φ_2, ψ_1 and $\psi_2\in C^2(\partial\Omega)$

(0.7) b and c are C^2 nondecreasing real functions.

In particular, if α and β are single-valued (i.e. they are continuous real functions) then the set inclusion of (NLS) should be replaced by equality.

In the general situation, $(u,v)\in H^2(\Omega)\times H^2(\Omega)$ is a *solution* of (NLS) if there exists $a,d\in L^2(\Omega)$ such that $a(x)\in\alpha(u(x))$, $d(x)\in\beta(v(x))$ a.e. $x\in\Omega$ and

$$-\Delta u+af(u)=0, \qquad -\Delta v-dg(v)=0 \quad \text{in } \Omega.$$

We shall prove the following existence result, which extends in some sense those in [2], [21] and [15].

THEOREM A. *Assume*

(A.1) $D(\alpha)=D(\beta)=\mathbb{R}$,

(A.2) $f(s)\geq m_1\geq 0\ \forall s\in\mathbb{R}$ *and*

(A.3) $0\leq g(s)\leq m_2\ \forall s\in\mathbb{R}$.

Then there exists at least one solution (u,v) *of* (NLS) (DBC) (*resp.* (NLS) (NBC)). *Moreover* $u,v\in W^{2,p}(\Omega)$ *for any* p, $1\leq p<+\infty$.

We also consider the existence and nonexistence of a dead core Ω_0 where $u=0$ and consequently the existence of the free boundary $\partial\Omega_0$.[1] Roughly speaking, such a dead core for (0.1), (0.3) arises when it is impossible for diffusion to supply enough reactant from outside Ω to reach the central part of Ω. (cf. [20]). This may happen if the reaction rate $F(u)e^{\gamma-\gamma/v}$ remains high as the reactant concentration decreases. Thus (for (0.1), (0.3)) the existence of Ω_0 depends essentially on three things: the reaction order, the Thiele number and the size of Ω.

Our main result in this direction can be stated in the following general terms.

THEOREM B. *Assume that the hypotheses of Theorem A are satisfied. Then the following properties are true*:

i) *If* $\alpha(s)=\mu^2|s|^{p-1}s$ *and* (u,v) *is any solution of* (NLS) (DBC); *then a dead core may exist only if* $0\leq p<1$.[2]

ii) *Let* $\alpha(s)=\mu^2|s|^{p-1}s$ *with* $0<p<1$ *and let* (u,v) *be a solution of* (NLS) (DBC). *For* $\lambda>0$ *let*

$$\Omega_\lambda=\{x\in\Omega: f(v(x))\geq\lambda\}.$$

Then

$$\Omega_0\supset\left\{x\in\Omega_\lambda:\ d(x,\partial\Omega_\lambda-(\partial\Omega-\operatorname{supp}\varphi_1))\geq\left(\frac{M}{K_{\lambda,\mu}}\right)^{(1-p)/2}\right\} \tag{0.8}$$

where $M=\|\varphi_1\|_{L^\infty(\partial\Omega)}$ *and*

$$K_{\lambda,\mu}=\left[\frac{2N(1-p)+4p}{\lambda\mu^2(1-p)^2}\right]^{1/(p-1)}.$$

(iii) *Let* $\alpha(s)=\mu^2\cdot\text{sign}\ s$. *Then the estimate* (0.8) *holds if we replace* M *by* $M^*=\|z\|_{L^\infty(\Omega)}$, *where* z *satisfies* $\Delta z=\mu^2 m_1$ *in* Ω *and* $z=\varphi_1$ *on* $\partial\Omega$. *Furthermore, if* Ω *is convex, the above results are still valid for* (NLS) (NBC) *in the sense that if* $0\leq p<1$, *then* Ω_0 *has a positive measure for* μ *large and it is possible to estimate* Ω_0 (*see* (2.22)).

[1] There is a large literature about this subject in the case of a *single* nonlinear equation. See, e.g. the systematic study of [12].

[2] By convention $|s|^{-1}s=\operatorname{sign}s$ ($=-1$ if $s<0$, $=[-1,1]$ if $s=0$ and $=1$ if $s>0$).

The above theorem is specially meaningful if m_1 in (A.2) is strictly positive (this is true in the case of the combustion system (0.1)), (0.3): indeed, in this case $v>0$ on $\bar{\Omega}$ and then we have $\Omega_\lambda=\Omega$ for any $\lambda\in(0,m_1]$. So the estimate (0.8) reads

$$\Omega_0\supset\left\{x\in\Omega\colon d(x,\partial\Omega)\geq\left(\frac{M}{K_{m_1,\mu}}\right)^{(1-p)/2}\right\}. \tag{0.9}$$

From the definition of $K_{\lambda,\mu}$ in Theorem B we deduce that $K_{\lambda,\mu}\searrow 0$ when $\lambda\searrow 0$ or $\mu\searrow 0$, and that $K_{\lambda,\mu}\nearrow+\infty$ if $\lambda\nearrow+\infty$ or $\mu\nearrow+\infty$. Therefore for a fixed bounded Ω the existence of a dead core Ω_0 may only be guaranteed (by estimate (0.9)) if

$$\delta(\Omega)\geq\left(\frac{M}{K_{m_1,\mu}}\right)^{(1-p)/2}$$

where $\delta(\Omega)$ is the radius of the largest ball contained in Ω, assuming $0\leq p<1$. Then a critical value μ_c of μ can be found such that Ω_0 has positive measure if $\mu>\mu_c$. In fact, direct computations show (when $N=1$) that function u is strictly positive in Ω if $\mu<\mu_c$ (for $\varphi_1\geq 0$) and u vanishes only at one point if $\mu=\mu_c$. (see the proof of Lemma 2.1 and also [20]). Estimate (0.8) of Theorem B can be also written independently of the function v for other systems in which it is not difficult to estimate the set Ω_λ (for instance $\Omega_\lambda=\Omega$ if in (NLS) we assume $f(s)=s$ and $\varphi_2>\delta>0$ for δ large (or λ small) enough).

Through the paper we also remark on other more general formulations of (NLS). The parabolic problem associated with (NLS) will be studied in a forthcoming paper by the authors. The case of Ω unbounded will be also treated elsewhere.

1. Existence results. Consider first the problem

$$\text{(DP)}\qquad \begin{array}{ll} -\Delta u+\alpha(u)f(v)\ni 0 & \text{in }\Omega,\\ -\Delta v-\beta(u)g(v)\ni 0 & \text{in }\Omega,\\ u=\varphi_1,\ u=\varphi_2 & \text{on }\partial\Omega,\end{array}$$

where Ω is a bounded open subset of $\mathbb{R}^N$ with smooth boundary $\partial\Omega$ and $\alpha,\beta,f,g,\varphi_1$, φ_2 satisfy (0.4), (0.5) and (0.6). Set $X=(H^2(\Omega))^2$.

Definition 1. We shall say that $(u,v)\in X$ is a *solution* of (DP) if there exist functions $a,b\in L^2(\Omega)$ such that $a(x)\in\alpha(u(x))$, $b(x)\in\beta(v(x))$ a.e. in Ω and

$$\begin{array}{ll} -\Delta u(x)+a(x)\cdot f(v(x))=0 & \text{a.e. }x\in\Omega,\\ -\Delta v(x)-b(x)g(v(x))=0 & \text{a.e. }x\in\Omega,\end{array}$$

and the boundary conditions (DBC) are satisfied.

Definition 2. The pair $[(u_0,v_0),(u^0,v^0)]\in X\times X$ is a *sub-supersolution* of (DP) if $u_0\leq u^0$, $v_0\leq v^0$ a.e. on Ω and

$$-\Delta u_0+\alpha^-(u_0)f(v)\leq 0\leq-\Delta u^0+\alpha^+(u^0)f(v)\ \forall v\in\left[v_0,v^0\right], \tag{1.1}$$

$$-\Delta v_0-\beta^0(u)g(v_0)\leq 0\leq-\Delta v^0-\beta^0(u)g(v^0)\ \forall u\in\left[u_0,u^0\right], \tag{1.2}$$

$$(1.3)\qquad u_0 \le \varphi_1 \le u^0 \quad \text{on } \partial\Omega,$$

$$(1.4)\qquad v_0 \le \varphi_2 \le v^0 \quad \text{on } \partial\Omega,$$

where $[K,l]=\{h\in L^2(\Omega)|K(x)\le h(x)\le l(x) \text{ a.e. on } \Omega\}$ if $K,l\in L^2(\Omega)$.

Our main existence result for (DP) is the following

THEOREM 1.1. *Suppose that* $[(u_0,v_0),(u^0,v^0)]$ *is a sub-supersolution satisfying*

$$(\mathrm{H}_1)\quad u_0,v_0,u^0,v^0\in L^\infty(\Omega)$$

and that

$$(\mathrm{H}_2)\quad D(\alpha)=D(\beta)=\mathbb{R}.$$

Then there exists at least one solution (u,v) *of* (DP) *such that* $u_0\le u\le u^0$, $v_0\le v\le v^0$. *In addition* $u,v\in W^{2,p}(\Omega)$ *for any* $p, 1\le p<+\infty$.

Remark 1.1. This theorem generalizes results of [2], [15], [16] and [20].

To prove Theorem 1.1 we define $E=[L^2(\Omega)]^2$ and $K=[u_0,u^0]\times[v_0,v^0]$. It is clear that K is a convex, closed and bounded subset of E. Now we define a nonlinear operator T: $K\to E$ in the following way: for $(\bar u,\bar v)\in K$, $T(\bar u,\bar v)=(w,z)$ is the *unique* solution of the uncoupled system

$$\begin{array}{lll}
(1.5) & -\Delta w+\alpha(w)f(\bar v)+w\ni\bar u & \text{in } \Omega,\\
(1.6) & w=\varphi_1 & \text{on } \partial\Omega,\\
(1.7) & -\Delta z+M\cdot z=\beta^0(\bar u)g(\bar v)+M\cdot\bar v & \text{in } \Omega,\\
(1.8) & w=\varphi_2 & \text{on } \partial\Omega.
\end{array}$$

Here $M>0$ is such that the right-hand side of (1.7) is increasing in $\bar v$ (we can choose such a M because g is C^1 and (H_1) has been assumed). Indeed by (H_2) we can apply the results of [10] to obtain the existence of a unique solution w of (1.5), (1.6). Moreover, by (H_1), (H_2) and the L^p-regularity results (see e.g. [14]) $w\in W^{2,p}(\Omega)$ for any $p, 1\le p<+\infty$. A similar argument works for z.

The proof of Theorem 1.1 will follow from Schauder's fixed point theorem applied to the operator T. It is sufficient to check that T is compact and that it sends K into itself.

LEMMA 1.1. *T is compact.*

Proof. As K is bounded it is easy to show that

$$\|w\|_{H^1(\Omega)}\le C$$

with C independent of $(\bar u,\bar v)\in K$. Thus it is sufficient to recall the compactness of the imbedding $H^1(\Omega)\hookrightarrow L^2(\Omega)$ to see that T sends bounded subsets into relatively compact ones (the same for z). To prove that T is continuous, suppose that $(u_n,v_n)\to(u,v)$ in E. Then

$$\begin{array}{ll}
-\Delta(w-w_n)+\alpha(w)f(v)-\alpha(w_n)f(v_n)+w-w_n\ni u-u_n & \text{in } \Omega,\\
w-w_n=0 & \text{on } \partial\Omega.
\end{array}$$

Multiplying by $w-w_n$ and integrating by parts we obtain (for the case α single-valued for simplicity)

$$\begin{aligned}
&\int_\Omega |\nabla(w-w_n)|^2+\int_\Omega [\alpha(w)f(v)-\alpha(w_n)f(v_n)](w-w_n)+\int_\Omega |w-w_n|^2\\
&\quad=\int_\Omega |\nabla(w-w_n)|^2+\int_\Omega [\alpha(w)f(v)-\alpha(w_n)f(v)](w-w_n)\\
&\qquad+\int_\Omega \alpha(w_n)(f(v)-f(v_n))(w-w_n)+\int_\Omega |w-w_n|^2\\
&\quad=\int_\Omega (u-u_n)(w-w_n)\\
&\quad\geq\int_\Omega |\nabla(w-w_n)|^2+\int_\Omega \alpha(w_n)(f(v)-f(v_n))(w-w_n),
\end{aligned}$$

and by the Cauchy-Schwarz inequality it follows that

$$\begin{aligned}
\|w-w_n\|^2_{H^1(\Omega)}\leq \|\alpha(w_n)\|_{L^\infty(\Omega)}\|f(v)-f(v_n)\|_{L^2(\Omega)}\|w-w_n\|_{L^2(\Omega)}\\
+\|u_n-u\|_{L^2(\Omega)}\|w-w_n\|_{L^2(\Omega)}.
\end{aligned}$$

Now it is easy to conclude that $w_n\to w$ in $H^1(\Omega)$. A similar argument can be used for z. □

LEMMA 1.2. $T(K)\subset K$.

Proof. We first prove $u_0\leq w$, i.e. $(u_0-w)^+=0$, with $h^+=\max(h,0)$. For $v=\bar{v}$, (1.1) yields

$$0\geq -\Delta(u_0-w)+\alpha(u_0)f(\bar{v})-\alpha(w)f(\bar{v})+u_0-w.$$

(We again suppose α single-valued for simplicity in the notation.) Multiply this inequality by $(u_0-w)^+$, integrate over Ω and use Green's formula to obtain

$$\begin{aligned}
0&\geq\int_\Omega -\Delta(u_0-w)(u_0-w)^+ +\int_\Omega (\alpha(u_0)-\alpha(w))f(\bar{v})(u_0-w)^+\\
&\quad+\int_\Omega (u_0-w)(u_0-w)^+\\
&\geq\int_\Omega |\nabla(u_0-w)^+|^2
\end{aligned}$$

by the monotonicity of α. This gives $(u_0-w)^+=0$. A similar argument shows that $w\leq u^0$.

For the second component v we have, with $u=\bar{u}$ in (1.2),

$$0\geq -\Delta(v_0-z)+\beta^0(\bar{u})g(\bar{v})-\beta^0(\bar{u})g(v^0)+M(v_0-z).$$

Multiplying by $(v_0 - z)^+$ and integrating yields

$$0 \geq -\int_\Omega \Delta(v_0 - z)(v_0 - z)^+ + \int_\Omega \left[\beta^0(\bar{u})g(v_0) + Mv_0 - \beta^0(\bar{u})g(z) - Mz\right](v_0 - z)^+$$

$$\geq \int_\Omega \left|\nabla(v_0 - z)^+\right|^2$$

(by the choice of M the second integral is positive).

Then T has at least one fixed point (u,v) in K which is a solution of (DP). Moreover $u, v \in L^\infty(\Omega)$ and this implies that $u, v \in W^{2,p}(\Omega)$ for any p, $1 \leq p < \infty$.

Remark 1.2. It follows easily from Morrey's theorem ($W^{2,p}(\Omega) \hookrightarrow C^{1,r}(\bar{\Omega})$ if $p > N$ with $r = 1 - N/p$) that $u, v \in C^{1,\delta}(\bar{\Omega})$ for any $0 < \delta < 1$. On the other hand, if we suppose for instance that α and β are C^1 then $u, v \in C^{2,\delta}(\bar{\Omega})$ for every $0 < \delta < 1$. Indeed, in this case $\alpha(u)f(v)$, $\beta(u)g(v) \in C^\delta(\bar{\Omega})$ and we can apply Schauder theory ([14]).

The main conclusion of Theorem A (for the Dirichlet problem) follows from the next lemma.

LEMMA 1.3. *Suppose* (H$_1$), (H$_2$) *and*

(H$_3$) $0 \leq m_1 \leq f(s)$ $\forall s \in \mathbb{R}$.

Then if u_0, v_0, u^0, $v^0 \in H^2(\Omega)$ *satisfy*

(1.9) $-\Delta u_0 + m_1 \alpha^-(u_0) \leq 0 \leq -\Delta u^0 + m_1 \alpha^+(u^0)$ *in* Ω,

(1.10) $u^0 \leq \varphi_1 \leq u^0$ *on* $\partial\Omega$,

(1.11) $-\Delta v_0 - \beta^0\left(-\|\varphi_1\|_{L^\infty(\partial\Omega)}\right)g(v_0) \leq 0 \leq -\Delta v^0 - \beta^0\left(\|\varphi_1\|_{L^\infty(\partial\Omega)}\right)g(v^0)$ *in* Ω,

(1.12) $v_0 \leq \varphi_2 \leq v^0$ *on* $\partial\Omega$

the couple $[(u_0, v_0), (u^0, v^0)]$ *is a sub-supersolution for* (DP).

Proof. Let $u_0 \leq u^* \leq u^0$, $v_0 \leq v^* \leq v^0$. By the maximum principle we have

$$-\|\varphi_1\|_{L^\infty(\partial\Omega)} \leq u_0 \leq 0 \leq u^0 \leq \|\varphi_1\|_{L^\infty(\partial\Omega)}.$$

Then, by (1.9)

$$-\Delta u_0 + \alpha^-(u_0)f(v^*) \leq -\Delta u_0 + m_1\alpha^-(u_0) \leq 0,$$
$$-\Delta u^0 + \alpha^+(u^0)f(v^*) \geq -\Delta u^0 + m_1\alpha^+(u^0) \geq 0,$$

and also by (1.11)

$$-\Delta v_0 - \beta^0(u^*)g(v_0) \leq -\Delta v_0 - \beta^0\left(-\|\varphi_1\|_{L^\infty(\partial\Omega)}\right)g(v_0) \leq 0,$$
$$-\Delta v^0 - \beta^0(u^*)g(v^0) \geq -\Delta v^0 - \beta^0\left(\|\varphi_1\|_{L^\infty(\partial\Omega)}\right)g(v^0) \geq 0.$$

Moreover, a simple argument gives $v_0 \leq 0 \leq v^0$. □

Now the problem is to find u_0, u^0, v_0, $v^0 \in L^\infty(\Omega)$ satisfying (1.9)–(1.12). The fact that such u_0, u^0 exist follows from the results of [10] applied to α. It is easy to check that v_0 and v^0 can be taken as the (unique) solutions of the problems

$$\begin{aligned} -\Delta w &= am_2 && \text{in } \Omega, \\ w &= \varphi_2 && \text{on } \partial\Omega \end{aligned}$$

and

$$\begin{aligned} -\Delta w &= a^* m_2 && \text{in } \Omega, \\ w &= \varphi_2 && \text{on } \partial\Omega \end{aligned}$$

respectively, being $a=\beta^0(\|\varphi_1\|_\infty)$ and $a^*=\beta^0(-\|\varphi_1\|_\infty)$. This proves Theorem A.

Remark 1.3. It is clear that assumption (A.3) of Theorem A is only used to find v_0 and v^0. If for example, g is such that the nonlinear problem

$$\begin{aligned} -\Delta w &= \beta^0\big(\|\varphi_1\|_{L^\infty(\partial\Omega)}\big)g(w) && \text{in } \Omega, \\ w &= \varphi_2 && \text{on } \partial\Omega \end{aligned}$$

has a solution, then we can remove (A.3). There is a very extensive literature for this kind of problem with different assumptions on g, but we do not want to consider this point here (cf. e.g. [1] and the survey [17]).

Remark 1.4. If α is assumed to be single-valued and C^1 the hypothesis $f(s)\geq m_1\geq 0$ is not necessary (cf. [15]). On the other hand, if α and β are single-valued and α,β,f and g are C^1 with sufficiently "small" Lipschitz constants, then it can be shown (cf. [2], [15]) that the solution is unique.

It is very easy now to apply the preceding results to the particular example (0.1), (0.3) considered at the beginning of this paper. It is sufficient to take $f(v)=g(v)=e^{\gamma-\gamma/v}$, $\alpha(u)=\mu^2u^p$, $\beta(u)=\nu\mu^2u^p$, $p>0$ and $\varphi_1=\varphi_2=1$. A sub-supersolution is given by $u_0\equiv 0$, $u^0\equiv 1$, $v_0\equiv 0$ and v^0 the unique solution of

$$-\Delta v^0=\nu\mu^2e^\gamma \quad \text{in } \Omega, \qquad v^0=1 \quad \text{on } \partial\Omega.$$

The case of nonlinear boundary conditions can be handled in a very similar way. We only point out some differences. First, the definition of sub-supersolution is the same except that the boundary conditions

$$Bu_0\leq\psi_1\leq Bu^0, \qquad Cv_0\leq\psi_2\leq Cv^0$$

should be satisfied instead of (1.10), (1.12).

The main existence result is

THEOREM 1.2. *Suppose that* $[(u_0,v_0),(u^0,v^0)]$ *is a sub-supersolution satisfying* (H$_1$), (H$_2$). *Then there exists at least one solution* (u,v) *of*

$$\begin{aligned} -\Delta u+\alpha(u)f(v) &\ni 0 && \textit{in } \Omega, \\ -\Delta v-\beta(u)g(v) &\ni 0 && \textit{in } \Omega, \\ Bu=\psi_1,\ Cv &= \psi_2 && \textit{on } \partial\Omega, \end{aligned}$$

such that $u_0\leq u\leq u^0$, $v_0\leq v\leq v^0$. *Moreover*, $u,v\in W^{2,p}(\Omega)$ *for any* p, $1\leq p<+\infty$.

Proof (*sketch*). We just give the definition of the nonlinear operator T; the other details are very similar to those for the Dirichlet problem. For $(\bar u,\bar v)\in K$, define $T(\bar u,\bar v)=(w,z)$ to be the unique solution of the system

$$\begin{aligned} -\Delta w+\alpha(w)f(\bar v)+w &\ni \bar u && \text{in } \Omega, \\ Bw &= \psi_1 && \text{on } \partial\Omega, \\ -\Delta z+z &= \beta^0(\bar u)g(\bar v)+\bar v && \text{in } \Omega, \\ Cz &= \psi_2 && \text{on } \partial\Omega. \end{aligned}$$

The existence and uniqueness of w and z follows from [6, Thm. II.1] (for z we can also use the results of [10]). □

A result very close to Lemma 1.3 can also be proved for the boundary conditions (NBC).

Remark 1.6. The operator $-\Delta$ in (NLS) can be replaced by two (possible different) elliptic second order differential operators or even by nonlinear operators of the form

$$-\Delta_q u \equiv -\sum_{i=1}^{N} \frac{\partial}{\partial x_i}\left(\left|\frac{\partial u}{\partial x_i}\right|^{q-2}\frac{\partial u}{\partial x_i}\right)$$

with $1<q<\infty$. Indeed, in this case one can define a nonlinear operator T by using again [6] (cf. also [13]). The more involved situation of b and c maximal monotone graphs can also be studied by similar methods.

2. Existence of a "dead core". In this section we shall consider the existence of a "dead core" for solutions u of (NLS), i.e., we shall prove that the set $\Omega_0=\{x\in\Omega\colon u(x)=0\}$ has a strictly positive measure under adequate hypotheses on α and eventually on $\|\varphi_1\|_{L^\infty(\partial\Omega)}$ or $|\Omega|$. In fact much more precise information is obtained about Ω_0.

Our study will be carried out by using results concerning a single nonlinear equation but arguing in a different way than usual for the combustion example. Indeed, if (u,v) is *any* solution of (NLS) (DBC) [resp. (NLS) (NBC)] then u satisfies

$$-\Delta u+\tilde{f}(x)\alpha(u)\ni F(x) \quad \text{in } \Omega,\text{[3]} \tag{2.1}$$

$$u=\varphi_1 \quad \text{on } \partial\Omega, \tag{2.2}$$

[respectively,

$$\frac{\partial u}{\partial n}+b(u)=\psi_1 \quad \text{on } \partial\Omega], \tag{2.3}$$

where $F\equiv 0$ and $\tilde{f}(x)=f(v(x))$ a.e. on Ω.

The study of the subset Ω_0 corresponding to solutions of (2.1), (2.2) (or (2.1),(2.3)) has occupied the attention of many authors, but, as far as we know, all these results are given for the simplest case $\tilde{f}(x)\equiv$ constant. We recall the two different approaches in the literature:

a) $\Omega=\mathbb{R}^N$ [7] or Ω being an unbounded set [11];

b) α being multivalued at the origin [9], [5], [13], [22].

More recently, a systematic study has been made in [12] giving a unified view of both situations, but always for $\tilde{f}(x)$ constant. Our results, in this section, follow the ideas of [12].

2.1. Dirichlet problem. We now prove parts i), ii) and iii) of Theorem B. For this we begin with some useful lemmas.

LEMMA 2.1. *Let $F\in L^\infty(\Omega)$, $\varphi\in C^2(\partial\Omega)$ and suppose that $u\in H^2(\Omega)$ satisfies*

$$-\Delta u(x)+\mu^2\tilde{f}(x)|u(x)|^p\operatorname{sign} u(x)\ni F(x) \quad \text{in } \Omega, \tag{2.4}$$

$$u=\varphi \quad \text{on } \partial\Omega \tag{2.5}$$

where $\tilde{f}\in L^\infty(\Omega)$, $\tilde{f}\geq 0$ on Ω and $p\geq 0$.[4] *If $0\leq p<1$ and Ω_λ denotes the set*

$$\Omega_\lambda=\{x\in\Omega\colon \tilde{f}(x)\geq\lambda\}, \qquad \lambda>0,$$

[3] Equation (2.1) also appears in the study of a stationary isothermical single reaction (see [3, Chap. 3]).

[4] If $p=0$, (2.4) should be interpreted in the sense that there exists $w\in L^2(\Omega)$ such that $w(x)\in\operatorname{sign}(u(x))$ a.e. $x\in\Omega$ and $-\Delta u+\mu^2\tilde{f}w=F$ in Ω.

we have the estimate

$$(2.6)\qquad \begin{aligned}&\Omega_0\equiv\{x\in\Omega\colon u(x)=0\}\\ &\supset\left\{x\in\Omega_\lambda-\operatorname{supp}F\colon d(x,\partial(\Omega_\lambda-\operatorname{supp}F)-(\partial\Omega-\operatorname{supp}\varphi))\geq\left(\frac{\tilde{M}}{K_{\lambda,\mu}}\right)^{(1-p)/2}\right\}.\end{aligned}$$

Here

$$\tilde{M}=\max\left\{\left(\frac{\|F\|_{L^\infty(\Omega)}}{\lambda\mu^2}\right)^{1/p},\ \|\varphi\|_{L^\infty(\partial\Omega)}\right\}$$

for $p>0$ *and* $\tilde{M}=\|z\|_{L^\infty(\Omega)}$ (*with* $\Delta z=\mu^2\lambda$ *in* Ω, $z=\varphi$ *on* $\partial\Omega$) *for* $p=0$. $K_{\lambda,\mu}$ *is given by*

$$(2.7)\qquad K_{\lambda,\mu}=\left[\frac{2N(1-p)+4p}{\lambda\mu^2(1-p)^2}\right]^{1/(p-1)}.$$

Proof. If we denote by u_+ (resp. u_-) the solutions of (2.4), (2.5) corresponding to the data F^+, φ^+ (resp. F^-,φ^-) then by well-known comparison theorems we have $u_+\geq 0$ (resp. $u_-\leq 0$) and also $u_-(x)\leq u(x)\leq u_+(x)$ a.e. $x\in\Omega$. Hence it is clear that $\Omega_0\supset\{x\in\Omega\colon u_-(x)=0$ and $u_+(x)=0\}$. For the sake of simplicity we shall only consider the case $F=F^+$, $\varphi=\varphi^+$, the other case being analogous. Let $u_\lambda\in H^2(\Omega)$ such that

$$(2.8)\qquad \begin{aligned}&-\Delta u_\lambda+\lambda\mu^2|u_\lambda|^p\geq F &&\text{in } \Omega_\lambda,\\ &u_\lambda\geq\varphi &&\text{on } \partial\Omega_\lambda\cap\partial\Omega,\\ &u_\lambda\geq\|u\|_{L^\infty(\Omega)} &&\text{on } \partial\Omega_\lambda-\partial\Omega.\end{aligned}$$

We claim that $0\leq u(x)\leq u_\lambda(x)$ a.e. on Ω_λ. Indeed,[5] taking $\tilde{F}(x)=-\Delta u+\lambda\mu^2u^p$, it is clear that $\tilde{F}(x)=F(x)+\lambda\mu^2u^p-\tilde{f}(x)\mu^2u^p$ and hence $\tilde{F}\leq F$ on Ω_λ. Moreover, $-\Delta u+\lambda\mu^2u^p=\tilde{F}$ on Ω_λ and thus by the comparison results (cf. e.g. [12]) one has $0\leq u\leq u_\lambda$. Therefore the conclusion of the lemma will follow by *constructing* one of such functions u_λ and the set $\{x\in\Omega_\lambda\colon u_\lambda(x)=0\}$ will give the estimate (2.6) for Ω_0. We will choose $u_\lambda(x)=h(|x-x_0|)$ for some $x_0\in\Omega_\lambda$. First, note that for $h\in C^2(\mathbb{R})$ and any $\eta\in(0,1)$ we have

$$\begin{aligned}&-\Delta h(|x-x_0|)+\lambda\mu^2h(|x-x_0|)^p\\ &\quad=-h''(|x-x_0|)-\left(\frac{N-1}{|x-x_0|}\right)h'(|x-x_0|)+\lambda\mu^2h(|x-x_0|)^p\\ &\quad=-h''(|x-x_0|)+\eta\lambda\mu^2h(|x-x_0|)^p\\ &\qquad+(1-\eta)\lambda\mu^2h(|x-x_0|)^p-\frac{(N-1)}{|x-x_0|}h'(|x-x_0|).\end{aligned}$$

[5] We shall only prove this inequality for $p>0$. If $p=0$, a natural adaptation of the argument leads to the claim.

Now, for a fixed η, let h_η be a solution of the Cauchy problem

$$\begin{aligned} &h''_\eta(r)\in\eta\lambda\mu^2|h_\eta(r)|^p\operatorname{sign}(h_\eta(r)),\\ &h_\eta(0)=h'_\eta(0)=0. \end{aligned} \tag{2.9}$$

It is easy to check (recall that $0\leq p<1$) that

$$h_\eta(r)=L_\eta r^{2/(1-p)} \tag{2.10}$$

where

$$L_\eta=\left(\frac{2(1+p)}{\eta\lambda\mu^2(1-p)^2}\right)^{1/(p-1)} \tag{2.11}$$

is a solution of (2.9). We have

$$(1-\eta)\lambda\mu^2h_\eta(r)^p-\frac{(N-1)}{r}h'_\eta(r)=L_\eta r^{2p/(1-p)}\left[(1-\eta)\lambda\mu^2L_\eta^{p-1}-\frac{2(N-1)}{1-p}\right];$$

choosing η such that

$$\eta\leq\frac{p+1}{1+p+(N-1)(1-p)} \tag{2.12}$$

leads to

$$-\Delta h_\eta(|x-x_0|)+\lambda\mu^2h_\eta(|x-x_0|)^p\geq0$$

for any $x\in\Omega_\lambda$.

Finally, consider the set $\tilde\Omega=\Omega_\lambda-\operatorname{supp}F$. The considerations made above show that the function

$$u_\lambda(x)=K_{\lambda,\mu}|x-x_0|^{2/(1-p)}$$

with $K_{\lambda,\mu}$ given by (2.7) satisfies

$$\begin{aligned} &-\Delta u_\lambda+\lambda\mu^2u_\lambda^p\geq0=F(x) && \text{in } \tilde\Omega,\\ &u_\lambda\geq0=\varphi && \text{on } \partial\tilde\Omega\cap(\partial\Omega-\operatorname{supp}\varphi). \end{aligned}$$

Hence it is sufficient to have

$$u_\lambda\geq\max\{\varphi,\|u\|_{L^\infty(\Omega)}\}\quad\text{on } \partial\tilde\Omega-(\partial\tilde\Omega\cap(\partial\Omega-\operatorname{supp}\varphi)) \tag{2.13}$$

to obtain

$$0\leq u(x)\leq u_\lambda(x)\quad\text{on } \tilde\Omega.$$

But, by the maximum principle we know that $u(x)\leq\tilde M$ on Ω and this implies that (2.13) is satisfied if we choose x_0 such that

$$|x-x_0|\geq\left(\frac{\tilde M}{K_{\lambda,\mu}}\right)^{(1-p)/2} \tag{2.14}$$

for every $x \in \partial\tilde{\Omega} - (\partial\tilde{\Omega} \cap (\partial\Omega - \operatorname{supp}\varphi))$. The conclusion now follows trivially from (2.13) and (2.14) (we recall that $u_\lambda(x_0)=0$). □

Statements ii) and iii) of Theorem B follow immediately from the above lemma. We remark that the constant $K_{\lambda,\mu}$ given in (2.7) is such that $K_{\lambda,\mu} \searrow 0$ when $\lambda \searrow 0$ or $\mu \searrow 0$ and that $K_{\lambda,\mu} \nearrow +\infty$ if $\lambda \nearrow +\infty$ or $\mu \nearrow +\infty$. Then, if Ω_λ is bounded and not empty, estimate (2.6) shows that the measure of Ω_0 is positive at least if

$$\delta(\Omega_\lambda - \operatorname{supp} F) > \left(\frac{\tilde{M}}{K_{\lambda,\mu}} \right)^{(1-p)/2}$$

where $\delta(\Omega_\lambda - \operatorname{supp} F)$ is the radius of the largest ball contained in $\Omega_\lambda - \operatorname{supp} F$ (assuming $0 \le p < 1$). Therefore, if Ω_λ is given, Ω_0 "exists" if μ is large enough or $\tilde{M}$ is sufficiently small. In the simple case of problem (0.1), (0.3) with $f(s)=s^p$, $0 \le p < 1$, it is easy to find a critical value μ_c of μ (now depending on p, γ and Ω) such that $\mathring{\Omega}_0$ is not empty if $\mu > \mu_c$. When $N=1$ direct computations show that, for p,γ and Ω fixed, the function u is strictly positive if $\mu < \mu_c$ (see e.g. [3] and [20]).

We shall prove part i) of Theorem B. Indeed, we shall prove that if $p \ge 1$ then for any value of λ, μ and $\delta(\Omega)$ there exist functions (u,v) satisfying (DP) (with $\alpha(s)=|s|^{p-1}s$) such that $u(x)>0$ on Ω. To do this we shall consider the worst case, i.e., when $\delta(\Omega)=+\infty$ (for instance $N=1$ and $\Omega=(0,\infty)$) and even for a larger class of nonlinearities α.

LEMMA 2.2. *Let $u \in H^2(0,\infty)$ satisfying*

$$\text{(2.15)} \qquad \begin{aligned} &-u''(x)+\tilde{f}(x)\alpha(u(x)) \ni 0, \qquad x \in (0,\infty), \\ &u(0)=1, \end{aligned}$$

where α is a maximal monotone graph such that $0 \in \alpha(0)$ and the function $j(s)=\int_0^s \alpha^0(r)\,dr$ satisfies

$$\text{(2.16)} \qquad \int_0^1 \frac{ds}{\sqrt{j(s)}} = +\infty .$$

(These hypotheses are satisfied when $\alpha(s)=|s|^{p-1}s$, $p \ge 1$.) Assume that $\tilde{f} \in L^\infty(0,\infty)$ and $0 \le \tilde{f}(x) \le m_2$ a.e. $x \in (0,\infty)$, for some $m_2>0$. Then $u(x)>0$ for any $x \in [0,\infty)$.

Proof. We shall use some ideas of [7] and [13]. By reasoning as in the proof of Lemma 2.1 we can always suppose without loss of generality that $\alpha^{-1}(0)=0$ and that α is single-valued. By a comparison argument completely analogous to the ones in the proof of Lemma 2.1 we show that if $\underline{u} \in H^2(0,\infty)$ satisfies

$$-\underline{u}''(x)+m_2\alpha(u(x))=0 \quad \text{on } (0,\infty), \qquad \underline{u}(0)=1$$

then $\underline{u}(x) \le u(x)$ for any $x \in (0,+\infty)$. Thus it suffices to prove that $\underline{u}(x)>0$ for any $x \in (0,\infty)$. Suppose that $\underline{u}$ has compact support and we shall obtain a contradiction. The maximum principle implies $0 \le \underline{u}(x) \le 1$ and hence $\underline{u}'' \in L^\infty(0,\infty)$. Thus $\underline{u} \in C^1([0,\infty))$ with $\underline{u}'' \ge 0$. Let $R=\sup\{x\colon \underline{u}(x) \ne 0\}$ ($R>0$ and finite by assumption). As $u'(R)=0$ it is not difficult to see that $\underline{u}'(x)<0$ and $\underline{u}(x)>0$ on $(0,R)$ (it is a consequence of $\underline{u}'' \ge 0$). But (2.16) yields

$$\int_0^{\underline{u}(R)} \frac{ds}{\sqrt{j(s)}} = -\int_0^R \frac{\underline{u}'(r)}{\sqrt{j(\underline{u}(r))}}\,dr = +\infty$$

and we will get a contradiction by estimating $\underline{u}'(r)/\sqrt{j(\underline{u}(r))}$ on $(0,R)$. Defining $w(x)=(\underline{u}'(x))^2$ we have

$$(j(\underline{u}))'=\alpha(\underline{u})\underline{u}'=\frac{1}{m_2}\underline{u}'\underline{u}''=\left[\frac{1}{2m_2}(\underline{u}')^2\right]'.$$

But $w(R)=0$ and $j(\underline{u}(R))=0$. By integrating we get $j(\underline{u})=w/2m_2$, and, finally,

$$\int_0^R \frac{\underline{u}'(r)}{\sqrt{j(\underline{u}(r))}}\,dr=\sqrt{2m_2}\int_0^R ds<+\infty,$$

a contradiction. □

Remark 2.1. By arguing in a similar way as in [11] we can prove that if Ω is an unbounded subset of $\mathbb{R}^N$, the maximal monotone graph α satisfies

$$\int_0^1 \frac{ds}{\sqrt{j(s)}}<+\infty \qquad \left(j(s)=\int_0^s \alpha^0(r)\,dr\right), \tag{2.17}$$

and u satisfies

$$\begin{aligned} &-\Delta u+\tilde{f}(x)\alpha(u)\ni F && \text{in } \Omega \quad (\tilde{f}\ge\lambda),\\ &u=\varphi && \text{on } \partial\Omega, \end{aligned}$$

where F and φ are assumed with compact support, then u has compact support. We point out that the improper integral (2.17) converges when $\alpha(s)=|s|^p \operatorname{sign} s$ if and only if $0\le p<1$ and hence the compactness of the support of u is an obvious consequence of Lemma 2.1.

Remark 2.2. Lemma 2.1 (and then Theorem B) can also be obtained when the operator $-\Delta$ in (NLS) is replaced by other elliptic second order differential operators as in Remark 1.6. The new definition of the functions $u_\lambda(x)=h(|x-x_0|)$ in the proof of Lemma 2.1 can be found by the methods of [12].

2.2. Nonlinear boundary conditions. Statement iv) of Theorem B will follow as in the preceding section by considering the nonlinear equation

$$\begin{aligned} &-\Delta u+\mu^2\tilde{f}(x)|u|^p \operatorname{sign} u\ni F && \text{in } \Omega,\\ &Bu=\frac{\partial u}{\partial n}+b(u)=\psi && \text{on } \partial\Omega, \end{aligned} \tag{2.18}$$

where $\tilde{f}\in L^\infty(\Omega)$, $\tilde{f}\ge 0$, $0\le p<1$, b is C^1 nondecreasing with $b(0)=0$, $F\in L^\infty(\Omega)$ and $\psi\in C^2(\partial\Omega)$.

First, we remark that "interior estimates" for Ω_0 can be obtained as in Lemma 2.1. More precisely, we have

$$\Omega_0\supset\left\{x\in\Omega_\lambda-\operatorname{supp}F\colon d(x,\partial(\Omega_\lambda-\operatorname{supp}F))\ge\left(\frac{M^*}{K_{\lambda,\mu}}\right)^{(1-p)/2}\right\} \tag{2.19}$$

where now[6] $M^*=\|u\|_{L^\infty(\Omega)}$. To show this it is sufficient to choose x_0 in such a way that $u_\lambda\ge M^*$ on $\partial\tilde{\Omega}$, being $\tilde{\Omega}=\Omega_\lambda-\operatorname{supp}F$ in the proof of Lemma 2.1. It is clear that (2.19)

[6] One obtains estimates for M^* by means of comparison theorems (see e.g. Lemma 3 in [12]).

does not give any information about the behavior of Ω_0 near the boundary of $\Omega_\lambda - \operatorname{supp} F$.

To improve the estimate (2.19), we introduce the following notation: given a smooth curve Γ in $\mathbb{R}^N$ and $x_0 \in \mathbb{R}^N$, we define

$$O(x_0,\Gamma)=\inf\left\{\cos(\overrightarrow{n(x)},\overrightarrow{x-x_0}): x\in\Gamma\right\}, \tag{2.20}$$

where $\vec{n}(x)=(n_1(x),\cdots,n_N(x))$ is the unitary outward normal vector to Γ at x and $(\overrightarrow{n(x)}, \overrightarrow{x-x_0})$ denotes the angle between the vectors $\vec{n}(x)$ and $\overrightarrow{x-x_0}$. It is clear that the value of $O(x_0,\Gamma)$ depends essentially on the "geometry" of Γ. If for instance $\Gamma=\partial\Omega$ and Ω is a convex bounded set of $\mathbb{R}^N$ it is easy to see that $O(x_0,\Gamma)>0$ when $x_0\in\Omega$.

Lemma 2.3. *Assume that* $u\in H^2(\Omega)\cap L^\infty(\Omega)$ *satisfies* (2.18). *For* $\lambda>0$, *let* $\Omega_\lambda=\{x\in\Omega: \tilde{f}(x)\geq\lambda\}$. *Moreover, suppose* $0\leq p<1$ *and*

$$O(x_0,\partial(\Omega_\lambda-\operatorname{supp}F)\cap\partial\Omega)\geq 0 \ \forall x_0\in\Omega_\lambda-\operatorname{supp}F. \tag{2.21}$$

Define

$$\Gamma=\partial(\Omega_\lambda-\operatorname{supp}F)\cap\partial\Omega\cap\operatorname{supp}\psi.$$

Then

$$\Omega_0\supset\left\{x\in\Omega_\lambda-\operatorname{supp}F: d(x,\Gamma)\geq\left[\frac{(1-p)\|\psi\|_{L^\infty(\partial\Omega)}}{2K_{\lambda,\mu}O(x_0,\Gamma)}\right]^{(1-p)/(1+p)} \text{ and } d(x,\partial(\Omega_\lambda-\operatorname{supp}F)-\partial\Omega)\geq\left(\frac{M^*}{K_{\lambda,\mu}}\right)^{(1-p)/2}\right\}, \tag{2.22}$$

where $M^*=\|u\|_{L^\infty(\Omega)}$.

Proof. Arguing as in Lemma 2.1 we only consider the case $F\geq0$ and $\psi\geq0$. Let $\tilde{\Omega}=\Omega_\lambda-\operatorname{supp}F$. By again using comparison results (cf. e.g. [12]) it is not difficult to see that if u_λ satisfies

$$-\Delta u_\lambda+\lambda\mu^2u_\lambda^p\geq0 \quad \text{in } \tilde{\Omega}, \tag{2.23}$$

$$u_\lambda\geq M^* \quad \text{on } \partial\tilde{\Omega}-\partial\Omega, \tag{2.24}$$

$$\frac{\partial u_\lambda}{\partial n}\geq\|\psi\|_{L^\infty(\partial\Omega)} \quad \text{on } \Gamma=\partial\tilde{\Omega}\cap\partial\Omega\cap\operatorname{supp}\psi, \tag{2.25}$$

$$\frac{\partial u_\lambda}{\partial n}\geq0 \quad \text{on}\partial\tilde{\Omega}\cap(\partial\Omega-\operatorname{supp}\psi), \tag{2.26}$$

then $0\leq u(x)\leq u_\lambda(x)$ for $x\in\tilde{\Omega}$. From the proof of Lemma 2.1 we know that the function

$$u_\lambda(x)=K_{\lambda,\mu}|x-x_0|^{2/(1-p)},$$

satisfies

$$-\Delta u_\lambda+\lambda\mu^2u_\lambda^p\geq0 \quad \text{on } \tilde{\Omega}$$

for any $x_0 \in \tilde{\Omega}$. Condition (2.24) is satisfied if

$$|x-x_0| \geq \left(\frac{M^*}{K_{\lambda,\mu}}\right)^{(1-p)/2} \quad \forall x \in \partial\tilde{\Omega}-\partial\Omega. \tag{2.27}$$

On the other hand,

$$\frac{\partial u_\lambda}{\partial n}(x)=\sum_{i=1}^{N}\frac{\partial u_\lambda}{\partial x_i}(x)\cdot n_i(x)=K_{\lambda,\mu}\left(\frac{2}{1-p}\right)|x-x_0|^{(1+p)/(1-p)}\cos(\overrightarrow{n(x)},\overrightarrow{x-x_0})$$

$$\geq K_{\lambda,\mu}\left(\frac{2}{1-p}\right)|x-x_0|^{(1+p)/(1-p)}O(x_0,\partial\tilde{\Omega}\cap\partial\Omega).$$

Thus (2.26) is a consequence of (2.21), and (2.25) holds if we choose $x_0 \in \tilde{\Omega}$ satisfying

$$|x-x_0| \geq \left(\frac{(1-p)\|\psi\|_{L^\infty(\partial\Omega)}}{2K_{\lambda,\mu}O(x_0,\Gamma)}\right)^{(1-p)/(1+p)} \quad \forall x \in \Gamma.$$

This completes the proof. □

Remark 2.3. Part iv) of Theorem B follows from Lemma 2.3 if we set $F\equiv 0$; (2.21) holds easily if, for instance, $f(s) \geq m_1 > 0\ \forall s \in \mathbb{R}$ (as in the combustion system) and Ω is a convex set.

Addendum. After the completion of this work, the authors learned that C. Bandle, R. P. Sperb and I. Stakgold have recently obtained, in the paper *Diffusion-reaction with monotone kinetics*, results similar to our Lemma 2.1, by using different methods. Some results related to Remark 2.1 can be found in a paper (to appear) by M. Schatzman, *Stationary solutions and asymptotic behaviour of a quasilinear degenerate parabolic equation*.

REFERENCES

[1] H. Amann, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[2] ______, *Existence and stability of solutions for semilinear parabolic systems and applications to some diffusion-reaction equations*, Proc. Roy. Soc., Edinburgh, 81A (1978), pp. 35–47.

[3] R. Aris, *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts*, Clarendon Press, Oxford, 1975.

[4] J. Bebernes, K. N. Chueh and W. Fulks, *Some applications of invariance to parabolic systems*, Indiana Univ. Math. J., 28 (1979), pp. 269–277.

[5] A. Bensoussan, H. Brezis and A. Friedman, *Estimates on the free boundary for quasi-variational inequalities*, Comm. Partial Differential Equations, 2 (1977), pp. 297–321.

[6] Ph. Benilan, *Operateurs accretifs et semi-groupes dans les espaces $L^p(1 \leq p \leq +\infty)$*, in Functional Analysis and Numerical Analysis, H. Fujita, ed., Japan Society for the Promotion of Sciences, 1978, pp. 15–53.

[7] Ph. Benilan, H. Brezis and M. G. Crandall, *A semilinear elliptic equation in $L^1(\mathbb{R}^N)$*, Ann. Scuola Norm. Sup. Pisa., Serie IV-II (1975), pp. 523–555.

[8] H. Brezis, *Operateurs maximaux monotones et semigroupes de contractions dans les espaces de Hilbert*, North-Holland, Amsterdam, 1973.

[9] ______, *Solutions of variational inequalities with compact support*, Uspekhi Mat. Nauk., 129 (1974), pp. 103–108. (In Russian.)

[10] H. Brezis and W. Strauss, *Semilinear second-order elliptic equations in L^1*, J. Math. Soc. Japan, 25 (1973), pp. 565–590.

[11] J. I. DIAZ, *Soluciones con soporte compacto para algunos problemas semilineales*, Collect. Math. 30 (1979), pp. 141–179.

[12] ______, *Tecnica de supersoluciones locales para problemas estacionarios no lineales: applicacion al estudio de flujos subsonicos*, Memorias no. 14 de la Real Academia de Ciencias, Madrid, 1980.

[13] J. I. DIAZ AND M. A. HERRERO, *Estimates on the support of the solutions of some nonlinear elliptic and parabolic problems*, Proc. Roy. Soc. Edinburgh, 39A (1981), pp. 249–258.

[14] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.

[15] J. HERNÁNDEZ, *Some existence and stability results for solutions of reaction-diffusion systems with nonlinear boundary conditions*, in Nonlinear Differential Equations: Invariance, Stability and Bifurcation, P. de Mottoni and L. Salvadori, eds., Academic Press, New York, 1981, pp. 161–173.

[16] ______, *Existence and global stability results for reaction-diffusion equations with nonlinear boundary conditions*, to appear.

[17] P. L. LIONS, *On the existence of positive solutions of semilinear elliptic equations*, Tech. Summ. Rep. #2209, Mathematics Research Center, University of Wisconsin, Madison, 1981.

[18] T. NAGAI, *Estimates for the coincidence sets of solutions of elliptic variational inequalities*, Hiroshima Math. J., 9 (1979), pp. 335–346.

[19] C. V. PAO, *Asymptotic stability of reaction-diffusion systems in chemical reactor and combustion theory*, J. Math. Anal. Appl., to appear.

[20] I. STAKGOLD, *Estimates for some free boundary problems*, in Ordinary and Partial Differential Equations Proceedings, Dundee, Scotland, 1980, W. N. Everitt and B. D. Sleeman, eds., Lecture Notes in Mathematics 846, Springer-Verlag, New York, 1981.

[21] L. V. TSAI, *Existence of solutions of nonlinear elliptic systems*, Bull. Inst. Math. Acad. Sinica, 8 (1980), pp. 111–126.

[22] N. YAMADA, *Estimates on the support of solutions of elliptic variational inequalities in bounded domains*, Hiroshima Math. J., 9 (1979), pp. 7–16.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics
006

ON THE BEHAVIOR OF THE SOLUTIONS TO THE LAMM EQUATION OF THE ULTRACENTRIFUGE*

ATSUSHI YOSHIKAWA†

Abstract. Let $c(r,t)$ be the solution, i.e., concentration of the solute, to the Lamm equation of ultracentrifugal analysis: $\partial c/\partial t = r^{-1}(r\{D_0\partial c/\partial r - r\omega^2 s_0 c/(1+kc)\})/\partial r$, $0<r_a<r<r_b$, $t>0$, with the (nonlinear) boundary conditions $D_0\partial c/\partial r - r\omega^2 s_0 c/(1+kc)=0$ at $r=r_a$ and $r=r_b$, and the initial data $c=c_0(r)$ when $t=0$. Here D_0, s_0 and k are positive constants independent of c. We discuss the behaviors of $c(r,t)$ as $t\to+\infty$ or $D_0\to 0$. Studies are also made on the limit equations corresponding to the cases $D_0=0$ or $t=+\infty$. The main theorems are stated in the Introduction.

0. Introduction. We discuss mathematically the behavior of the solutions $c(r,t)$ of the Lamm equation of ultracentrifugal analysis [11]:

$$\frac{\partial c}{\partial t}=r^{-1}\frac{\partial}{\partial r}\left\{r\left(D_0\frac{\partial c}{\partial r}-r\omega^2 sc\right)\right\}, \tag{0.1}$$

$0<r_a<r<r_b$, $t>0$, with the (nonlinear) boundary condition

$$D_0\frac{\partial c}{\partial r}-r\omega^2 sc=0 \tag{0.2}$$

at $r=r_a$ and $r=r_b$, and the initial data

$$c=c_0(r) \quad \text{when } t=0. \tag{0.3}$$

Here $c(r,t)$ stands for the concentration of solute in a two-component system, D_0 is the diffusion coefficient, s the sedimentation coefficient, and ω the frequency of the rotor. We assume D_0 to be a positive constant independent of c as a first approximation while taking

$$s=s(c)=\frac{s_0}{1+kc}, \tag{0.4}$$

conforming to experiments, where s_0 and k are positive constants independent of c. For more details, see Fujita [6, Chapt. 1].

The initial concentration $c_0(r)$ naturally satisfies

$$c_0(r)\geq 0 \tag{0.5}$$

and

$$m_0=\int_{r_a}^{r_b}c_0(r)rdr>0. \tag{0.6}$$

The problem (0.1)–(0.6) makes sense mathematically. Namely, we have existence and uniqueness of solutions. Let $\mathfrak{L}^\infty$ stand for the space $C^0(r_a,r_b)$ of the continuous functions on the closed interval $[r_a,r_b]$, and $\mathfrak{L}^p$, $1\leq p<\infty$, for the space $L^p(r_a,r_b;\ rdr)$ of the p-summable functions over the interval (r_a,r_b) with respect to the measure rdr. The norms of $\mathfrak{L}^p$ are denoted by $|\cdot|_p$. Using the notion of a *generalized solution*

*Received by the editors June 6, 1982, and in revised form March 1, 1983.

†Department of Mathematics, Hokkaido University, Sapporo, Japan.

introduced in Definition 1.4 below, we prove in the next section, §1, the following:

THEOREM 1. *Assume $c_0(r)\in\mathfrak{L}^p$, $1\leq p\leq\infty$, satisfy* (0.5) *and* (0.6). *The initial boundary value problem* (0.1)–(0.4) *then has a unique generalized solution $c(r,t)$ in the large in t. The solution is nonnegative, the mapping $t\to c(r,t)\in\mathfrak{L}^p$ is continuous for $t\geq 0$, and $|c(\ ,t)|_p$ is uniformly bounded for $t\geq 0$. Mass is conserved*:

$$\int_{r_a}^{r_b} c(r,t)r\,dr=m_0 \quad \textit{for } t>0.$$

Moreover, if $c_0(r)$ is nondecreasing in r, then so is $c(r,t)$ for each $t>0$. Furthermore, if $c_0(r)$ is smooth and compatible with the boundary condition (0.2), *i.e.*,

$$D_0\frac{\partial c_0}{\partial r}-r\omega^2 s(c_0)c_0=0$$

at the boundary $r=r_a$, $r=r_b$, then the solution $c(r,t)$ is a smooth classical one.

In particular, we may write the solution $c(r,t)$ of the problem (0.1)–(0.4) as

$$c(r,t)=c\big(r,t;\ D_0,s_0\omega^2,k,r_a,r_b;\ c_0,m_0\big) \tag{0.7}$$

by making explicit all the parameters involved. The following homogeneity relation is easily established:

$$\begin{aligned}\alpha c\big(\beta r,\gamma t;\ D_0,s_0\omega^2,k,\beta r_a,\beta r_b;\ c_0,m_0\big)\\ =c\left(r,t;\ \frac{\gamma D_0}{\beta^2},\gamma s_0\omega^2,\frac{k}{\alpha},r_a,r_b;\ \alpha c_0,\frac{\alpha m_0}{\beta^2}\right)\end{aligned} \tag{0.8}$$

for any $\alpha>0$, $\beta>0$, $\gamma>0$.

We may rather schematically say that the classical Faxen solution [5] corresponds to the case $k=0$, $r_a=0$, $r_b=+\infty$, and the Archibald solution [1] to the case $k=0$.

Our principal purpose in the present article is a study of the behavior of the solution $c(r,t;\ D_0)=c(r,t;\ D_0,s_0\omega^2,k,r_a,r_b;\ c_0,m_0)$ when $t\to+\infty$ or $D_0\to 0$. Actually the equilibrium solution is important in the experiments, and so is the estimate of the convergence rate as $t\to+\infty$. The equations (0.1)–(0.4) yield to an equation of a kind of conservation laws (with a boundary condition) when $D_0\to 0$. Solutions to the latter equation, generally easier to handle, provide much information on the properties of solutions to the original Lamm equation, especially for small t.

The equilibrium solution c_E of the Lamm equation (0.1)–(0.4) is a solution to the following equation:

$$D_0\frac{\partial c_E}{\partial r}-r\omega^2 s(c_E)c_E=0, \tag{0.9}$$

$r_a<r<r_b$, with

$$\int_{r_a}^{r_b} c_E(r)r\,dr=m_0. \tag{0.10}$$

The problem (0.9)–(0.10) is essentially a two-point boundary value problem for a second order ordinary differential equation. The requirement (0.10) reflects (0.6) through the conservation of mass in the system (0.1)–(0.4).

In §2 we prove the following:

THEOREM 2. *The problem* (0.9)–(0.10) *has a unique solution* $c_E(r) = c_E(r;\ D_0, s_0\omega^2, k, r_a, r_b;\ m_0)$. $c_E(r)$ *is everywhere positive and strictly increasing in* r. *For* $r < r_b$, *we have*

$$(0.11) \qquad c_E(r) \to 0 \quad as \ \frac{s_0\omega^2}{D_0} \to +\infty.$$

For any Lipschitz continuous function $f(r)$, *we have*

$$(0.12) \qquad \int_{r_a}^{r_b} f(r) c_E(r) r\,dr \to m_0 f(r_b)$$

as $s_0\omega^2/D_0 \to +\infty$.

Note the homogeneity relation:

$$(0.13) \quad \alpha c_E(\beta r;\ D_0, s_0\omega^2, k, \beta r_a, \beta r_b;\ m_0) = c_E\left(r;\ \frac{D_0}{\beta^2}, s_0\omega^2, \frac{k}{\alpha}, r_a, r_b;\ \frac{\alpha m_0}{\beta^2}\right),$$

for any $\alpha > 0$, $\beta > 0$.

To discuss the convergence of the solution $c(r, t)$ of the Lamm equation (0.1)–(0.4) to the equilibrium solution $c_E(r)$ of (0.9)–(0.10), we introduce the following linear operator L_E: for $u(r) \in C^2(r_a, r_b)$ satisfying

$$(0.14) \qquad D_0 \frac{\partial u}{\partial r} - r s_0 \omega^2 s_1'(c_E(r)) u = 0$$

at $r = r_a$ and $r = r_b$, we put

$$(0.15) \qquad L_E u = -r^{-1} \frac{\partial}{\partial r}\left\{ r\left(D_0 \frac{\partial u}{\partial r} - r s_0 \omega^2 s_1'(c_E(r)) u \right)\right\},$$

$r_a < r < r_b$, where

$$(0.16) \qquad s_1(c) = \frac{c}{1 + kc} \quad \text{and} \quad s_1'(c) = \frac{ds_1(c)}{dc}.$$

The operator L_E with the boundary condition (0.14) is extended to a nonnegative selfadjoint operator $\hat{L_E}$ in the Hilbert space $L^2(r_a, r_b;\ q(r) r\,dr)$, where the density $q(r) r\,dr$ is related to

$$(0.17) \qquad q(r) = \exp\left\{ -D_0^{-1} s_0 \omega^2 \int_{r_a}^{r} s_1'(c_E(r')) r'\,dr' \right\}.$$

Let λ_E be the smallest nonzero eigenvalue of $\hat{L_E}$. In §3, we show the following:

THEOREM 3. *Let* $c_0(r)$ *be smooth, compatible with the boundary condition* (0.2) *and satisfy* (0.5)–(0.6). *Assume further that* $c_0(r)$ *is nondecreasing in* r. *Let* $c(r, t)$ *be the corresponding solution to* (0.1)–(0.4) *and* $c_E(r)$ *the equilibrium solution with the same mass as* c_0. *There is a positive constant* C *independent of* t, r *such that, for each* $t > 0$,

$$(0.18) \qquad |c(\ , t) - c_E|_\infty \leq C \exp(-\lambda_E t).$$

In the last section, §4 we discuss the equation with the vanishing diffusion coefficient: $D_0 = 0$, which is given by

$$(0.19) \qquad \frac{\partial c}{\partial t} + r^{-1} \frac{\partial}{\partial r}\{ r^2 \omega^2 s(c) c \} = 0$$

$r_a<r<r_b$, $t>0$, with the boundary condition

$$c=0 \quad \text{at } r=r_a \tag{0.20}$$

and the initial data

$$c=c_0(r) \quad \text{when } t=0. \tag{0.21}$$

Note that no boundary condition is imposed at $r=r_b$. $\tilde{c}(r,t)$ is said to be a generalized solution of (0.19)–(0.21) if, for any $T>0$, and for any continuously differentiable function $\phi(r,t)$ which vanishes at $r=r_b$ and at $t=T$, we have

$$\int_{r_a}^{r_b}\phi(r,0)c_0(r)r\,dr+\int_0^T dt\int_{r_a}^{r_b}\tilde{c}(r,t)\frac{\partial\phi(r,t)}{\partial t}r\,dr \\ +\int_0^T dt\int_{r_a}^{r_b} rs_0\omega^2 s_1(\tilde{c}(r,t))\frac{\partial\phi(r,t)}{\partial r}r\,dr=0. \tag{0.22}$$

The analogues of the Rankine–Hugoniot condition and the entropy condition are readily introduced in the present situation.

THEOREM 4. *Let $c_0(r)\in\mathfrak{L}^\infty$ be nonnegative and nondecreasing in r. Assume further that $c_0(r)$ vanishes for $r\leq r_a+\varepsilon$ for some $\varepsilon>0$. Then there is a generalized solution $\tilde{c}(r,t)$ to the problem* (0.19)–(0.21) *such that for each $t>0$, $\tilde{c}(r,t)$ is nonnegative almost everywhere in $r\in(r_a,r_b)$ and nondecreasing in r. Moreover, there is a family of classical solutions $c(r,t;\,D_0)$ of* (0.1)–(0.2) *with* (0.4) *for D_0 small enough such that $c(r,t;\,D_0)\to\tilde{c}(r,t)$ as $D_0\to0$ almost everywhere in r for each $t>0$. If $c_0(r)$ is smooth, then the same conclusion holds if we assume $c_0(r_a)=c_0'(r_a)=0$.*

The system (0.19)–(0.21) is not quite a conservation law because of the boundary condition. In fact, we have the following:

THEOREM 5. *Let $\tilde{c}(r,t)$ be the generalized solution of* (0.19)–(0.21) *as given by Theorem* 4. *For any $T>0$, we have*

$$\int_{r_a}^{r_b}c_0(r)r\,dr=\int_{r_a}^{r_b}\tilde{c}(r,T)r\,dr+s_0\omega^2r_b^2\int_0^T s_1(\tilde{c}(r_b-0,t))\,dt. \tag{0.23}$$

Moreover,

$$\int_{r_a}^{r_b}\tilde{c}(r,t)r\,dr\to0 \quad \text{as } t\to+\infty. \tag{0.24}$$

There remains much to investigate. The assumptions that the initial data are nondecreasing in r in Theorems 3–5 are stringent. As to Theorem 3, we include supplementary information in §3. Theorem 4 is certainly incomplete. We must weaken requirements on the initial data c_0 and complete Theorem 4 by evaluating the convergence rate as $D_0\to0$ of the solutions of the Lamm equations with the same initial data. The uniqueness question remains open for the moment.

In this respect, we recall that when the sedimentation coefficient $s(c)$ is given by

$$s(c)=1-kc, \tag{0.25}$$

(0.1)–(0.3) is linearized essentially by the Cole–Hopf transformation (Weiss [17]). The above questions have then naturally been settled mathematically in a rather complete way.

The ultracentrifuge, as a means to measure the weight of proteins, is said to have become less important recently. Probably owing to this situation, few comprehensive studies have recently been done, although physicochemical interest in the ultracentrifuge still prevails. (Some of the latest studies include Dyshon, Weiss and Yphantis [3] and Fujita [7].) On the other hand, in spite of the simple forms of (0.1)–(0.4), mathematical studies similar to the present one seem to have been rarely carried out.

1. Existence and uniqueness of generalized solutions in the large. Consider the initial boundary value problem for the equation:

$$\frac{\partial c}{\partial t}=r^{-1}\frac{\partial}{\partial r}\left(r\left\{D_0\frac{\partial c}{\partial r}-rS(c)c\right\}\right), \tag{1.1}$$

$r\in I=(r_a,r_b)$, $t>0$, with the boundary data

$$D_0\frac{\partial c}{\partial r}-rS(c)c=0 \tag{1.2}$$

at $r=r_a$, $r=r_b$, and the initial data

$$c=c_0(r) \tag{1.3}$$

when $t=0$. Here $S(c)$ is a slightly more general function than $\omega^2 s(c)$ of (0.4). Namely, we take $S(c)$ to be a twice continuously differentiable function of $c\in\mathbb{R}$ such that

$$\inf\{S(c);\, c\in\mathbb{R}\}>0, \tag{1.4}$$

$$M=\sup\left\{\left|S(c)+c\frac{dS(c)}{dc}\right|;\, c\in\mathbb{R}\right\}<+\infty, \tag{1.5}$$

$$2\frac{dS(c)}{dc}+\frac{cd^2S(c)}{dc^2}\leq 0 \quad \text{for } c\geq 0 \tag{1.6}$$

and

$$N=\sup\{cS(c);\, c\geq 0\}<+\infty. \tag{1.7}$$

Let $S_1(c)=cS(c)$. From (1.5), we have $|S_1(c)|\leq M|c|$, whence

$$|S_1(c(r))|_p\leq M|c(r)|_p \tag{1.8}$$

for all $c(r)\in\mathfrak{L}^p$, $1\leq p\leq\infty$.

Actually we take

$$S(c)=\omega^2 s(h(c)) \tag{1.9}$$

where $s(c)$ is that of (0.4) and $h(c)$ is a twice continuously differentiable function of c such that

$$h(c)=\begin{cases} c, & c\geq -\dfrac{1}{4k}, \\ -\dfrac{3}{4k}, & c\leq -\dfrac{1}{k}, \end{cases}$$

and $h'(c)\geq 0$ everywhere. Then (1.4)–(1.8) are easily verified.

If the initial data $c_0(r)$ is smooth and compatible with the boundary condition (1.2), the existence and uniqueness of a local classical solution to the problem (1.1)–(1.3) is known (see Ladyzhenskaya, Solonnikov and Ural'ceva [10, Chap. 5, Thm. 7.4]). However, we here argue a bit differently and rewrite (1.1)–(1.3) into an integral equation.

Let

$$L_0 u = -r^{-1}\frac{\partial}{\partial r}\left(rD_0\frac{\partial u}{\partial r}\right) \tag{1.10}$$

for $u\in D(L_0)=\{u\in C^2(I);\ \partial u/\partial r=0 \text{ at } \partial I\}$. The selfadjoint extension $\hat{L_0}$ of the operator L_0 to the Hilbert space $\mathfrak{L}^2$ is nonnegative definite and generates a contraction semigroup $W_0(t)$ of operators in $\mathfrak{L}^2$. Let $W_0(t,r,r')=W_0(t,r,r';\ D_0,I)$ be the kernel of the semigroup $W_0(t)$:

$$u(r,t)=W_0(t)u=\int_I W_0(t,r,r')u(r')r'\,dr'. \tag{1.11}$$

Then $u(r,t)$ satisfies the equation

$$\left(\frac{\partial}{\partial t}+L_0\right)u(r,t)=0, \qquad t>0, \quad r\in I, \tag{1.12}$$

with

$$\frac{\partial u(r,t)}{\partial r}=0, \qquad r\in I, \quad t>0, \tag{1.13}$$

$$u(r,0)=u(r), \qquad r\in I. \tag{1.14}$$

Note the homogeneity relation:

$$W_0(\gamma t,\beta r,\beta r';\ D_0,\beta I)=W_0\left(t,r,r';\ \frac{\gamma D_0}{\beta^2},I\right) \tag{1.15}$$

for $\beta>0$, $\gamma>0$, $\beta I=(\beta r_a,\beta r_b)$.

It is well known that $W_0(t,r,r')$ is expressed by means of the Bessel functions of order zero. In particular, we see that $W_0(t,r,r')$ determines a contraction semigroup, which we still denote by $W_0(t)$, of operators in each $\mathfrak{L}^p$, $1\le p\le\infty$. The following basic estimate is proved in the Appendix.

LEMMA 1.1. *Let* λ_0 *be the smallest nonzero eigenvalue of the operator* $\hat{L_0}$. *We can find a positive constant* C *independent of* D_0 *such that*

$$\sup_{r\in I}\int_I\left|\frac{\partial W_0(t,r,r')}{\partial r}\right|r'\,dr'\le C\left(1+D_0^{-1/2}t^{-1/2}\right)e^{-\lambda_0 t}$$

and

$$\sup_{r\in I}\int_I\left|\frac{\partial W_0(t,r,r')}{\partial r'}\right|r'\,dr'\le C\left(1+D_0^{-1/2}t^{-1/2}\right)e^{-\lambda_0 t}$$

for all $t>0$.

Remark. We shall omit $D_0>0$ in applying Lemma 1.1 unless a reference to D_0 is absolutely necessary.

Now we give the integral equation which replaces the initial boundary value problem (1.1)–(1.3).

PROPOSITION 1.2. *Let $c=c(r,t)$ be a smooth solution to the problem* (1.1)–(1.3). *Then c satisfies the integral equation*

$$c-R(c)=W_0(t)c_0, \tag{1.16}$$

where

$$R(c)(r,t)=\int_0^t dt' \int_I \frac{\partial W_0(t-t',r,r')}{\partial r'} S_1(c(r',t'))r'^2 dr'. \tag{1.17}$$

This follows from the next lemma. Recall that $W_p^m(I;\ r\,dr)$ denotes the Sobolev spaces of the functions $f(r)$, p-summable over I with respect to $r\,dr$ up to their mth derivatives $\partial_r^i f(r)$, $i\leq m$, and $\mathring{W}_p^m(I;\ r\,dr)$ the closure in $W_p^m(I;\ r\,dr)$ of $C_0^\infty(I)$.

LEMMA 1.3. *Let $g\in C^1([0,\infty);\ W_2^1(I;\ r\,dr))$ be given. Then the initial boundary value problem for the equation*:

$$\frac{\partial u}{\partial t}-r^{-1}\frac{\partial}{\partial r}\left(rD_0\frac{\partial u}{\partial r}\right)=r^{-1}\frac{\partial(r^2g(r,t))}{\partial r}, \tag{1.18}$$

$(r,t)\in I\times(0,\infty)$, *with the boundary datum*

$$D_0\frac{\partial u}{\partial r}+rg=0 \quad \textit{at } \partial I \tag{1.19}$$

and the initial data

$$u=u_0 \quad \textit{when } t=0, \tag{1.20}$$

is solved by

$$u=W_0(t)u_0+u_1, \tag{1.21}$$

where

$$u_1(r,t)=-\int_0^t dt' \int_I \frac{\partial W_0(t-t',r,r')}{\partial r'} g(r',t')r'^2 dr'. \tag{1.22}$$

Proof. Note that $u_1(r,t)$ is well defined because of Lemma 1.1. Let $G(r,t)= D_0^{-1}\int_{r_a}^r g(r',t)r'\,dr'$ and $u=v-G$. Then v is the solution of the problem: $\partial v/\partial t+\hat{L_0}v=\partial G/\partial t$, $v=u_0+G(\ ,0)$, in $\mathfrak{L}^2$. Thus, $u=W_0(t)u_0+u_2$, where

$$u_2=\int_0^t dt' \int_I W_0(t-t',r,r')\frac{\partial G(r',t')}{\partial t'} r'\,dr' + W_0(t)G(\ ,0)-G.$$

Since the definition domain $D(\hat{L_0})$ of the operator $\hat{L_0}$ and $W_2^1(I;\ r\,dr)$ are dense in $\mathfrak{L}^2$, we get, using integration by parts in t' and r', $u_1=u_2$ in $\mathfrak{L}^2$.

Remark. Let $g\in L^\infty([0,\infty);\ \mathfrak{L}^1)$ be measurable in (r,t). u_1 of (1.22) is then well defined because of Lemma 1.1. For $\phi\in C_0^1([0,\infty);\ D(\hat{L_0}))$, we have

$$-\int_0^\infty dt \int_I u(r,t)\frac{\partial\phi(r,t)}{\partial t} r\,dr$$

$$=\int_I u_0(r)\phi(r,0)r\,dr-\int_0^\infty dt \int_I (\hat{L_0}\phi)(r,t)\{u(r,t)+G(r,t)\}r\,dr.$$

In this sense, $u(r,t)$ of (1.21) is a *generalized solution* of the initial boundary value problem (1.18)–(1.20).

DEFINITION 1.4. We call $c(r,t)$ a generalized solution to the initial boundary value problem (1.1)–(1.3) if $c(r,t)$ is a solution to the integral equation (1.16)–(1.17).

Now we show the existence of the solution to the integral equation (1.16)–(1.17).

PROPOSITION 1.5. *Assume* (1.4), (1.5). *Then for any* $c_0(r)\in L^\infty(I;\ r\,dr)$, *the equation* (1.16), (1.17) *has a solution* $c(r,t)\in L^\infty([0,\infty);\ L^\infty(I;\ r\,dr))$. *If* $c_0\in\mathcal{L}^\infty$, *then* $c(r,t)\in C^0([0,\infty)\times I)$. *For any* $c_0\in\mathcal{L}^p$, $1\le p\le\infty$, *the equation* (1.16), (1.17) *has a solution* $c(r,t)\in C^0([0,\infty);\ \mathcal{L}^p)$.

Proof. We employ successive approximations. Let $c^0=W_0(t)c_0$ and $c^n=c^0+R(c^{n-1})$ for $n\ge1$. Then for $n\ge1$,

$$c^{n+1}-c^n=\int_0^t dt'\int_I\frac{\partial W_0(t-t',r,r')}{\partial r'}\{S_1(c^n(r',t'))-S_1(c^{n-1}(r',t'))\}r'^2\,dr'. \tag{1.23}$$

Let $c_0\in L^\infty(I;\ r\,dr)$. By Lemma 1.1 and (1.8),

$$|c^1(\ ,t)-c^0(\ ,t)|_\infty\le r_bM|c_0|_\infty B\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{3}{2}\right)^{-1}t^{1/2}, \tag{1.24}$$

for some constant B depending on D_0. Thus, for $n\ge2$, we get inductively from (1.23)

$$|c^n(\ ,t)-c^{n-1}(\ ,t)|_\infty\le|c_0|_\infty\left(BMr_b\Gamma\left(\frac{1}{2}\right)\right)^n\Gamma\left(1+\frac{n}{2}\right)^{-1}t^{n/2}, \tag{1.25}$$

where M is given by (1.5). If $c_0\in\mathcal{L}^\infty$, then $c^0(r,t)$ is continuous up to $t=0$ and so are c^n for $n\ge1$. If $c_0\in\mathcal{L}^1$, then (1.24) and (1.25) are replaced by

$$|c^1(\ ,t)-c^0(\ ,t)|_1\le r_bM|c_0|_1B\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{3}{2}\right)^{-1}t^{1/2}, \tag{1.26}$$

and for $n\ge2$

$$|c^n(\ ,t)-c^{n-1}(\ ,t)|_1\le|c_0|_1\left(BMr_b\Gamma\left(\frac{1}{2}\right)\right)^n\Gamma\left(1+\frac{n}{2}\right)^{-1}t^{n/2}, \tag{1.27}$$

respectively. Now, for each c^{n-1}, the right-hand side of (1.23) defines a nonlinear operator of $u=c^n-c^{n-1}$, which is quasilinear in the sense of Kree [14] on the space $\mathcal{L}^1+\mathcal{L}^\infty$ under the assumption (1.5). Hence, we can apply the real interpolation method to (1.23) (see also Komatsu [13]). Thus, if $c_0\in\mathcal{L}^p$, $1<p<\infty$, we can then derive the estimates

$$|c^1(\ ,t)-c^0(\ ,t)|_p\le M_pM|c_0|_pr_bB\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{3}{2}\right)^{-1}t^{1/2} \tag{1.28}$$

and

$$|c^n(\ ,t)-c^{n-1}(\ ,t)|_p\le M_p|c_0|_p\left(BMr_b\Gamma\left(\frac{1}{2}\right)\right)^n\Gamma\left(1+\frac{n}{2}\right)^{-1}t^{n/2} \tag{1.29}$$

for $n\ge2$. Here M_p is a positive constant. Since

$$\Phi(x)=\sum_{i\ge0}\frac{x^i}{\Gamma(1+i/2)}\le(1+ax)\exp(x^2) \tag{1.30}$$

for all $x\ge0$ with $a=\sup\{\Gamma(j+1)/\Gamma(j+3/2);\ j\ge0\}$, we see that $c(r,t)=c^0(r,t)+\sum_{i\ge1}(c^i(r,t)-c^{i-1}(r,t))$ converges in $\mathcal{L}^p$ uniformly with respect to t, $0\le t\le T$, for any $T>0$. $c(r,t)$ is then readily seen to be a solution of the integral equation (1.16), (1.17).

Proposition 1.6. *Let $c=c(r,t)$ and $c^*=c^*(r,t)$ be such that $c-R(c)=W_0(t)c_0$ and $c^*-R(c^*)=W_0(t)c_0^*$. If $c_0, c_0^* \in \mathfrak{L}^p$, $1\leq p\leq\infty$, then, for each $t>0$,*

$$|c(\ ,t)-c^*(\ ,t)|_p\leq|c_0-c_0^*|_p\Phi\left(C_pt^{1/2}\right), \tag{1.31}$$

where Φ is given by (1.30) *and C_p is a positive constant.*

Proof. We have $c-c^*-R(c)+R(c^*)=W_0(t)(c_0-c_0^*)$. Hence, again by the interpolation,

$$|c(\ ,t)-c^*(\ ,t)|_p\leq|c_0-c_0^*|_p+r_bMC_p'\int_0^t(t-t')^{-1/2}|c(\ ,t)-c^*(\ ,t')|_pdt'$$

with some constant C_p'. Now by a routine argument in proving Gronwall's inequality, we get (1.31) with $C_p=r_bMC_p'\Gamma(1/2)$.

The inequality (1.31) implies not only uniqueness of the solution to the equation (1.16), (1.17), but also its continuous dependence on the initial data. In this respect, we may write

$$c=c(c_0)=c(r,t;\ c_0) \tag{1.32}$$

for the solution $c(r,t)$ of the integral equation (1.16), (1.17). We have mentioned that we have unique local classical solutions for smooth initial data compatible with the boundary conditions. Thus, from what we have shown, we conclude that the problem (1.1)–(1.3) has a unique smooth solution $c(r,t)$ in the large for a smooth compatible initial datum. Any generalized solution can then be approximated by classical solutions in $\mathfrak{L}^p$.

Proposition 1.7. *Let $c_0\in\mathfrak{L}^p$, $1\leq p\leq\infty$. Then for $c(r,t)=c(r,t;\ c_0)$, $\int_I c(r,t)r\,dr=\int_I c_0(r)r\,dr$ for all $t>0$.*

Proof. We may assume c_0 to be a smooth compatible initial datum. Then the proposition is immediate from (1.1)–(1.3).

Proposition 1.8. *Let $c_0\in\mathfrak{L}^p$, $1\leq p\leq\infty$, be nonnegative. Then so is $c(r,t;\ c_0)$.*

Proof. We show the nonnegativity of a classical solution to (1.1)–(1.3) with a smooth compatible initial datum. We owe the following device to Hiroshi Matano. Let

$$v(r,c)=c\exp\left\{\frac{-r^2S(0)}{2D_0}+\int_0^c\frac{S(0)-S(c')}{S_1(c')}dc'\right\}.$$

Then $\partial v/\partial c=S(0)v/S_1(c)$, $v/c>0$, so that c is a function $c=c(r,v)$ with the same sign as v. Moreover, $S_1(c)^{-1}\partial c/\partial r=S(0)^{-1}v^{-1}\partial v/\partial r+rD_0^{-1}$, or $D_0\partial c/\partial r-rS_1(c)=D_0S_1(c)(S(0)v)^{-1}\partial v/\partial r$, and $\partial c/\partial t=S_1(c)(S(0)v)^{-1}\partial v/\partial t$. It then follows that

$$\frac{\partial v}{\partial t}=D_0\frac{\partial^2 v}{\partial r^2}+A(r,v)\frac{\partial v}{\partial r}+B(r,v)\left(\frac{\partial v}{\partial r}\right)^2$$

in $I\times(0,\infty)$ with the boundary condition $\partial v/\partial r=0$ at ∂I and the initial data $v=v_0=v(r,c_0(r))$ when $t=0$. Here, at $c=c(v,r)$, $A(r,v)=r^{-1}D_0+rS_1'(c)$, and $B(r,v)=S(0)^{-1}D_0v^{-1}c\{\int_0^1 S'(c\theta)\,d\theta+S'(c)\}$, $'$ denoting the differentiation in c. An immediate consequence of this rewriting is that for any classical solution $v(r,t)$:

$$\min_{r\in I} v(r,t)\geq\min_{r\in I} v(r,t')$$

and

$$\max_{r\in I} v(r,t)\le\max_{r\in I} v(r,t')$$

for $t\ge t'\ge 0$. If $c_0\ge 0$, then $v_0\ge 0$ so that $v(r,t)\ge 0$. Hence $c(r,t)\ge 0$.

Now we discuss the implications of the extra requirement (1.6) or (1.7).

PROPOSITION 1.9. *Assume* (1.7). *Let* $c_0(r)\in\mathfrak{L}^p$ *be nonnegative*, $1\le p\le\infty$. *Then, for* $c=c(c_0)$,

$$|c(\ ,t)|_p\le|c_0|_p+C_pD_0^{-1}N.$$

Here C_p *is a constant depending only on* p.

Proof. From (1.16) and (1.17), we have $|c(\ ,t)|_p\le|c_0|_p+|R(c)(\ ,t)|_p$. By Lemma 1.1 and (1.7), $|R(c)(\ ,t)|_p\le C_pN/D_0$ since λ_0 is homogeneous of degree 1 in D_0.

PROPOSITION 1.10. *Assume* (1.6). *Let* $c_0(r)\in\mathfrak{L}^p$ *be nonnegative and nondecreasing*, $1\le p\le\infty$. *Then* $c=c(r,t;\ c_0)$ *is a nondecreasing function of* r *for each* $t>0$.

Proof. We show $\partial c/\partial r\ge 0$ for the smooth solution to (1.1)–(1.3) when c_0 is a smooth compatible initial datum. In fact, $u=\partial c/\partial r$ satisfies the equation

$$\frac{\partial u}{\partial t}=D_0\frac{\partial^2 u}{\partial r^2}+\left(r^{-1}D_0-rS_1'(c)\right)\frac{\partial u}{\partial r}-\left(r^{-2}D_0+3S_1'(c)\right)u-rS_1''(c)u^2$$

in $I\times(0,\infty)$. At the boundary, $u=rS(c)$ at ∂I and $u=\partial c_0/\partial r$ when $t=0$. By the maximum principle, u cannot take negative minima for $t>0$ under assumption (1.6).

Remark. Under a milder assumption, Matano [15] has recently shown that the lap-number, or the minimum number of a solution's monotonicity intervals at each t, does not increase as $t\to\infty$.

Proof of Theorem 1. Recall (1.9) for the choice of $S(c)$. Since $s(h(c))=s(c)$ for $c\ge 0$, Theorem 1 follows from Propositions 1.5–1.10.

2. The equilibrium solution. Let $s_1(c)=c/(1+kc)$. Let $\alpha\ge 0$. Consider the problem:

$$\frac{\partial c_E(r)}{\partial r}-\alpha r s_1(c_E(r))=0 \tag{2.1}$$

in $I=(r_a,r_b)$ with

$$\int_I c_E(r)r\,dr=m_0, \tag{2.2}$$

m_0 being a given positive constant. In the original notation, $\alpha=s_0\omega^2/D_0$.

The problem (2.1)–(2.2) can be treated as a two-point boundary value problem for a second order nonlinear ordinary differential equation (consult Hartman [9], for instance). But we here argue in a more naive way. First we observe the following facts:

PROPOSITION 2.1. *The problem* (2.1)–(2.2) *has at most one solution. Any solution is positive and strictly increasing.*

Proof. We first show uniqueness of a solution. Let $c^*(r)$ be another solution to (2.1)–(2.2). Then $u=c_E-c^*$ satisfies the equation $\partial u/\partial r=\alpha r s_2(r)u$, where $s_2(r)=(s_1(c_E(r))-s_1(c^*(r)))/(c_E(r)-c^*(r))$. Then $u(r)=u_0\exp(\alpha\int_{r_a}^{r}s_2(r')r'\,dr')$. $u_0=0$ follows from $\int_I u(r)r\,dr=0$.

To show the positivity of a solution, we first modify the function $s_1(c)$ for negative c so that $s_1(c)=c/(1+kh(c))$, where $h(c)$ is the function appearing in (1.9). Let $v(r)$ be any function which satisfies (2.1) in $I=(r_a, r_b)$. Then $v(r)$ cannot take negative local minima nor positive local maxima at interior points of I. Observe $v(r)$ is strictly decreasing where $v(r)<0$ and strictly increasing where $v(r)>0$. Thus, $v(r_a)$ cannot be a negative minimum and $v(r_b)$ is a negative minimum provided $v(r)$ never takes positive values on I. We thus see any solution $c_E(r)$ to (2.1)–(2.2) is nonnegative. Since $c_E(r)\exp(kc_E(r)-\alpha r^2/2)=\text{constant}$, as is immediately verified, we conclude $c_E(r)>0$ and thus $c_E(r)$ is strictly increasing.

PROPOSITION 2.2. *For any* $\alpha\geq 0$, *the problem* (2.1)–(2.2) *has a solution* $c_E(r,\alpha)$. *Furthermore*, $c_E(r,\alpha)$ *is real analytic in* α.

Proof. Let $X=X(m_0)$ be the set for which the problem (2.1)–(2.2) has a solution. Note $0\in X$. We will show that X is an open and closed subset of the interval $[0,\infty)$, hence $X=[0,\infty)$. We first verify closedness. Let $\alpha_n\in X$, $\alpha_n\to\infty$. Since each $c_E(r,\alpha_n)$ is strictly increasing in r, $\sup_n c_E(r_a,\alpha_n)\leq m_0/\int_I r\,dr=C_1$ because of (2.2). From (2.1), $\partial c_E(r,\alpha_n)/\partial r\leq r_b k^{-1}\sup_n\alpha_n=C_2$. Therefore, $c_E(r,\alpha_n)\leq C_1+C_2r_b$. Now $\alpha\in X$ follows from the Ascoli–Arzela theorem.

We are now going to show openness of the set X and at the same time real analyticity in α of $c_E(r,\alpha)$. Note that we may reduce $k=1$ in the function $s_1(c)$ since we may take kc as a new unknown instead of c. So we assume $k=1$ in the rest of the proof. Let $\alpha\in X$. As a candidate for $c_E(r,\alpha+z)$, we consider a power series

$$\hat{c}(r,\alpha;z)=\sum_{i\geq 0}c_i(r,\alpha)z^i, \tag{2.3}$$

where $c_i(r,\alpha)$ are successively determined from the following relations:

$$c_0(r,\alpha)=c_E(r,\alpha), \tag{2.4}$$

i.e., the solution of (2.1)–(2.2), and for $i\geq 1$,

$$\frac{\partial c_i}{\partial r}-\alpha r s_1'(c_0)c_i=rs_{(i-1)}(c_0,\cdots,c_{i-1})+\alpha r\tilde{s}_{(i-1)}(c_0,\cdots,c_{i-1}), \tag{2.5}$$

$$\int_I c_i(r,\alpha)r\,dr=0. \tag{2.6}$$

Here $s_{(i)}$ and $\tilde{s}_{(i-1)}$ are given as follows:

$$s_{(i)}(c_0,\cdots,c_i)=\sum s_1^{(p)}(c_0)\frac{c(i)^{\nu(i)}}{\nu(i)!}, \tag{2.7}$$

where $c(i)=(c_1,\cdots,c_i)$ and the summation is taken over the multi-indices $\nu(i)=(n_1,\cdots,n_i)$ of nonnegative integers such that

$$n_1+\cdots+in_i=i,\qquad n_1+\cdots+n_i=p \tag{2.8}$$

and

$$\tilde{s}_{(i-1)}(c_0,\cdots,c_{i-1})=s_{(i)}(c_0,\cdots,c_i)-s_1'(c_0)c_i. \tag{2.9}$$

Namely,

$$\tilde{s}_{(i-1)}(c_0,\cdots,c_{-1})=\sum s_1^{(p)}(c_0)\frac{c(i)^{\nu(i)}}{\nu(i)!}, \tag{2.10}$$

where the summation is taken over the multi-indices $\nu(i)$ in (2.8) with $n_i=0$. In particular, $\tilde{s}_{(1)}=0$ (cf. for instance, Bourbaki [2, Chap. I, §3, Exercise 7]). We claim that the series $\hat{c}(r,\alpha;\, z)$ converges for small z at each $\alpha\in X$. From (2.6), (2.7), we have

$$(2.11)\quad c_i(r,\alpha)=\left\{\int_I s_1(c_0(r',\alpha))r'\,dr'\right\}^{-1}$$
$$\times\left\{\int_{r_a}^{r}F_i(r',\alpha)\int_{r_a}^{r'}s_1(c_0(r'',\alpha))r''\,dr''M(r,r',\alpha)r'\,dr'\right.$$
$$\left.-\int_{r}^{r_b}F_i(r',\alpha)\int_{r'}^{r_b}s_1(c_0(r'',\alpha))r''\,dr''M(r,r',\alpha)r'\,dr'\right\},$$

where

$$(2.12)\quad F_i(r,\alpha)=s_{(i-1)}(c_0(r,\alpha),\cdots,c_{i-1}(r,\alpha))+\alpha\tilde{s}_{(i-1)}(c_0(r,\alpha),\cdots,c_{i-1}(r,\alpha)),$$
$$M(r,r',\alpha)=\frac{s_1(c_0(r,\alpha))}{s_1(c_0(r',\alpha))}.$$

We then obtain the following estimates:

LEMMA 2.3. *Let* $B=\sup_{r\in I}\max\{s_1(c_0(r,\alpha)),\ s(c_0(r,\alpha))\}$ *and* $\delta=\sup_{r\in I}s(c_0(r,\alpha))$. *Then for* $i\geq 1$

$$|c_i(r,\alpha)|\leq a_i\frac{s_1(c_0(r,\alpha))K^i}{B\delta},\qquad K=\frac{B\delta(r_b^2-r_a^2)}{2}.$$

Here $a_1=1$ *and for* $i\geq 1$

$$a_{i+1}=\sum a(i)^{\nu(i)}\frac{p!}{\nu(i)!}+\alpha K\sum a(i)^{\nu(i)}\frac{p!}{\nu(i)!},$$

with $a(i)=(a_1,\cdots,a_i)$, *the first summation being taken over the multi-indices* $\nu(i)=(n_1,\cdots,n_i)$ *and nonnegative integers* p *such that* $n_1+\cdots+in_i=i$, $n_1+\cdots+n_i=p$, *and the second summation being taken over* $n_1+\cdots+in_i=i+1$, $n_1+\cdots+n_i=p$.

In fact, we have

$$(2.13)\qquad |c_1(r,\alpha)|\leq s_1(c_0(r,\alpha))\frac{(r_b^2-r_a^2)}{2}$$

from (2.11) and (2.12). Note that

$$|c_i(r,\alpha)|\leq s_1(c_0(r,\alpha))\int_I\frac{|F_i(r',\alpha)|}{s_1(c_0(r',\alpha))}r'\,dr'.$$

Thus, from (2.11), (2.12) and (2.7)–(2.10), Lemma 2.3 now follows by induction on i.

We return to the proof of Proposition 2.2. From the choice of a_i in Lemma 2.3, we have

$$(2.14)\qquad \sum_{j\geq 1}a_jx^j=\frac{\{1-\sqrt{(1-4(1+\alpha K)x)}\}}{\{2(1+\alpha K)\}}.$$

The right-hand side of (2.14) is holomorphic for $|x|<1/(4+4\alpha K)$. It then follows that the series $\hat{c}(r,\alpha;\ z)$ in (2.3) converges for $|z|<1/(4K(1+K\alpha))$. In particular, we conclude that each $\alpha\in X$ is an interior point of X, whence $X=[0,\infty)$. Note that we may take $B=\delta=1$ in Lemma 2.3. Thus, $c_E(r,\alpha)$ for

$$0\leq\alpha<\frac{1}{2(r_b^2-r_a^2)}$$

has the following convergent power series expression:

$$c_E(r,\alpha)=\sum_{i\geq 0}\alpha^i c_i(r,0),$$
$$c_0(r,0)=\frac{2m_0}{r_b^2-r_a^2},$$
$$c_1(r,0)=2^{-1}s_1(c_0(r,0))\left\{r^2-\frac{r_a^2+r_b^2}{2}\right\}$$

and so on. Actually we may take a sequence $\alpha^{(i)}\to\infty$, $i\to\infty$, by $\alpha^{(0)}=0$ and

$$\alpha^{(i)}=\alpha^{(i-1)}+4^{-1}\left(r_b^2-r_a^2\right)^{-1}\left\{1+\alpha^{(i-1)}\frac{r_b^2-r_a^2}{2}\right\}^{-1},$$

for $i\geq 1$ so that $c_E(r,\alpha)$ is analytic in α with

$$|\alpha-\alpha^{(i)}|<2(\alpha^{(i)}-\alpha^{(i-1)})$$

for each i. We have thus completed the proof of Proposition 2.2.

Now we try to give some idea how $c_E(r,\alpha)$ behaves when $\alpha\geq 0$ varies. Let

$$y=y(r,\alpha)=\exp\left(\frac{\alpha r^2}{2}-\frac{\alpha r_b^2}{2}\right) \tag{2.15}$$

and put

$$c^0(r,\alpha)=\frac{m_0\alpha y(r,\alpha)}{1-y(r_a,\alpha)}. \tag{2.16}$$

Then

$$\frac{dc^0(r,\alpha)}{dr}=\alpha r c^0(r,\alpha) \tag{2.17}$$

and

$$\int_I c^0(r,\alpha)r\,dr=m_0. \tag{2.18}$$

Let

$$c_E(r,\alpha)=c^0(r,\alpha)(1+w(r,\alpha)) \tag{2.19}$$

be the solution to (2.1)–(2.1). Let

$$\tilde{\alpha}=\frac{km_0\alpha}{1-y(r_a,\alpha)}. \tag{2.20}$$

Then regarding $w(r,\alpha)$ as a function of y: $W=w(y)=w(y,\alpha)$, we obtain the following differential equation for w:

$$\frac{dw(y)}{dy}=F(y,w), \tag{2.21}$$

$$F(y,w)=-\frac{\tilde{\alpha}(1+w)^2}{(1+\tilde{\alpha}(1+w)y)}, \tag{2.22}$$

with the vanishing mean value:

$$\int_{y_a}^{1} w(y)\,dy=0, \tag{2.23}$$

$y_a=y(r_a,\alpha)$, $1=y(r_b,\alpha)$.

We know in particular

$$w(y)>-1. \tag{2.24}$$

Since $F(y,w)<0$ for $w>-1$, we see $w(y)$ is strictly decreasing. Because of (2.23), we have

$$w(y_a)>0>w(1). \tag{2.25}$$

Let $y_0=y_0(\alpha)$ be the unique zero of

$$w(y_0)=0. \tag{2.26}$$

PROPOSITION 2.4. *Let*

$$\hat{w}(y,z)=-\log\left(\frac{1+\tilde{\alpha}y}{1+\tilde{\alpha}z}\right).$$

Then

$$\hat{w}(y,y_0)\leq w(y)\leq -1+(1-\hat{w}(y,y_0))^{-1}, \tag{2.27}$$

where the equality holds only when $y=y_0$.

Proof. Since $F(y,w)$ is strictly decreasing in $w>-1$, we have $dw(y)/dy>-\tilde{\alpha}/(1+\tilde{\alpha}y)$ for $y>y_0$, and $dw(y)/dy<-\tilde{\alpha}/(1+\tilde{\alpha}y)$ for $y<y_0$. On the other hand, let

$$F_1(y,w)=-\tilde{\alpha}\frac{(1+w)^2}{1+\tilde{\alpha}y}. \tag{2.28}$$

Then $(F_1(y,w)-F(y,w))w<0$ for $w\neq 0$, $w>-1$. Hence, $dw/dy<F_1(y,w)$ for $y>y_0$, and $dw/dy>F_1(y,w)$ for $y<y_0$. From these inequalities, (2.27) readily follows.

The trouble here is of course that we have no practical means to evaluate $y_0=y_0(\alpha)$ of (2.26). Yet we may say that Proposition 2.4 gives an approximation of $w(y)$ for small $\tilde{\alpha}$, and hence, of $c_E(r,\alpha)$ because of (2.19). A rather rough estimate of $y_0(\alpha)$ follows from (2.27). In fact, determine $\hat{y}_0=\hat{y}_0(\alpha)$ from $\int_{y_a}^{1}\hat{w}(y,y_0)\,dy=0$ or

$$\hat{y}_0(\alpha)=-\tilde{\alpha}^{-1}+e^{-\tilde{\alpha}}\left\{\frac{u(\tilde{\alpha})}{u(\tilde{\alpha}y_a)}\right\}^{\beta},$$

with $u(x)=(1+x)^{1+x}$, $\beta=(1-y_a)^{-1}\tilde{\alpha}^{-1}$. Then the left-hand side of (2.27) implies

$$y_a\leq y_0(\alpha)\leq\hat{y}_0(\alpha) \tag{2.29}$$

since $\partial\hat{w}(y,z)/\partial z>0$. The right-hand side of (2.27) does not seem to imply a better estimate.

LEMMA 2.5.

$$w(y)>(1+\tilde{\alpha}y-\tilde{\alpha}y_0)^{-1}-1, \quad y>y_0,$$
$$w(y)<(1+\tilde{\alpha}y-\tilde{\alpha}y_0)^{-1}-1, \quad y<y_0.$$

Proof. Note $F(y,w)>-\tilde{\alpha}(1+w)^2$ for $w>-1$. Then

$$0\le\liminf_{\alpha\to\infty}\tilde{\alpha}y_0(\alpha)\le\limsup_{\alpha\to\infty}\tilde{\alpha}y_0(\alpha)\le 1 \tag{2.30}$$

since $\tilde{\alpha}y_a\le\tilde{\alpha}y_0\le 1+\tilde{\alpha}y_a$ from the second inequality of Lemma 2.5. Since $\lim_{\alpha\to\infty}\hat{y}_0(\alpha)=1/\alpha$, (2.30) is sharper than (2.29) for large α. (2.27) and (2.30) now imply $0\le\liminf_{\alpha\to\infty}w(y_a)\le\limsup_{\alpha\to\infty}w(y_a)\le(1-\log 2)^{-1}\log 2$. From Proposition 2.4 and (2.24), we conclude

$$\lim_{\alpha\to\infty}w(1)=-1.$$

Proof of Theorem 2. Because of Propositions 2.1 and 2.2, what remains to be proven is (0.11) and (0.12). Returning to (2.19), we have

$$0\le c_E(r,\alpha)\le\frac{k^{-1}\tilde{\alpha}y(r,\alpha)}{1-\hat{w}(y,y_0)}.$$

(0.11) then follows immediately. Now for α large enough, $\alpha\ge\alpha_0>0$, we observe $|c_E(r,\alpha)(r_b-r)|\le C/(r_b+r)$, $r\in I$, for a positive constant C depending on α_0. If $f(r)$ is Lipschitz continuous, we have $|f(r)-f(r_b)|\le C_1|r-r_b|$ for some positive constant C_1 so that

$$|(f(r)-f(r_b))c_E(r,\alpha)|\le\frac{CC_1}{(r_a+r_b)}.$$

(0.12) is then clear.

3. Convergence to the equilibrium solution. Let $\mathcal{L}^p_+$, $1\le p\le\infty$, be the subset of $\mathcal{L}^p$, consisting of the nonnegative elements. Theorem 1 then means that the problem (0.1)–(0.4) defines a continuous mapping $U_p(t):\mathcal{L}^p_+\to\mathcal{L}^p_+$, $t\ge 0$, which assigns to each $c_0(r)\in\mathcal{L}^p_+$ the generalized solution $c(r,t;\,c_0)$ of (0.1)–(0.4). Then we have $U_p(t)\circ U_p(t')=U_p(t+t')$, $t\ge 0$, $t'\ge 0$, and $U_p(0)=$the identity operator in $\mathcal{L}^p_+$. In other words, we have a dynamical system $\mathfrak{D}_p=\{\mathcal{L}^p_+, U_p(t)\}$ for each p, $1\le p\le\infty$. Let

$$\omega_p(c_0)=\bigcap_{t'>0}\text{ closure of }\{U_p(t)c_0;\ t\ge t'\} \tag{3.1}$$

for $c_0\in\mathcal{L}^p_+$. Here the closure is taken in the space $\mathcal{L}^p$. We call the set $\omega_p(c_0)$ the ω-limit set of c_0 (see Hale and Infante [8]). Since $\mathcal{L}^p_+\subset\mathcal{L}^q_+$, $p>q$, we have $\omega_p(c_0)\subset\omega_q(c_0)$ if $c_0\in\mathcal{L}^p_+$. Let

$$J(c)=c_0\frac{\partial c}{\partial r}-rs_0\omega^2s_1(c) \tag{3.2}$$

for $c=c(r,t)$. Here $s_1(c)$ is given by (0.16).

PROPOSITION 3.1. *Let $c_0 \in \mathfrak{L}_+^p$ be smooth and compatible with the boundary condition* (0.2). *Assume either $c_0(r)$ is nondecreasing in r or*

$$|J(c_0)|_\infty \leq \frac{D_0^2}{s_0\omega^2 r_b^3 k}. \tag{3.3}$$

Then $\omega_p(c_0) = \{c_E\}$, $1 \leq p \leq \infty$, where c_E is the equilibrium solution with the same mass as c_0.

For a proof we employ a kind of Lyapunov functional $|J(c)|_2^2$. If $c(r,t)$ is a classical solution to the problem (0.1)–(0.4), then $J(c)$ satisfies the equation:

$$\frac{\partial J(c)}{\partial t} = D_0\frac{\partial^2 J(c)}{\partial r^2} + A(r,c)\frac{\partial J(c)}{\partial r} - B(r,c)J(c), \tag{3.4}$$

in $I \times (0,\infty)$ with

$$J(c) = 0 \quad \text{at } r = r_a, \quad r = r_b \tag{3.5}$$

and

$$J(c) = J(c_0) \quad \text{when } t = 0. \tag{3.6}$$

Here

$$A(r,c) = \frac{D_0}{r} - s_0\omega^2 s_1'(c)r, \tag{3.7}$$

$$B(r,c) = s_0\omega^2 s_1'(c) + \frac{D_0}{r^2}. \tag{3.8}$$

The maximum principle then yields the following:

LEMMA 3.2. *Let $c = c(r,t;\ c_0)$ be the classical solution to the problem* (0.1)–(0.4). *Then $|J(c(\ ,t))|_\infty$ is nondecreasing in $t \geq 0$.*

In particular,

$$|J(c(\ ,t))|_\infty \leq |J(c_0)|_\infty, \qquad t \geq 0.$$

LEMMA 3.3. *Let c_0 be as in Proposition* 3.1 *and $c(r,t) = c(r,t;\ c_0)$. Then there is a positive constant a such that*

$$|J(c(\ ,t))|_2^2 \leq e^{-at}|J(c_0)|_2^2.$$

Proof. By integration by parts,

$$2^{-1}\frac{d|J(c)|_2^2}{dt} = -D_0\left|\frac{\partial J(c)}{\partial r}\right|_2^2 + \int_I\left\{2^{-1}\left(\frac{\partial A}{\partial r} + r^{-1}A\right) + B\right\}J(c)^2 r\,dr.$$

Then (3.7) and (3.8) imply

$$2^{-1}\left(\frac{\partial A}{\partial r} + r^{-1}A\right) + B - D_0 r^{-2} = -2^{-1}s_0\omega^2 s_1''(c)\frac{\partial c}{\partial r}. \tag{3.9}$$

If $c_0(r)$ is nondecreasing in r, then $\frac{\partial c}{\partial r} \geq 0$. Since $s_1''(c) < 0$, the left-hand side of (3.9) is nonnegative. Now let c_0 satisfy (3.3). Since

$$\begin{aligned} 2^{-1}\left(\frac{\partial A}{\partial r} + r^{-1}A\right) + B &= D_0 r^{-2} - 2^{-1}\frac{(s_0\omega^2 r)^2 s_1''(c)s_1(c)}{D_0} - 2^{-1}\frac{s_0\omega^2 r s_1''(c)J(c)}{D_0} \\ &\geq -2^{-1}D_0^{-1}s_0\omega^2 r s_1''(c)\left(\frac{D_0^2}{s_0\omega^2 r_b^3 k} + J(c)\right) \\ &\geq -2^{-1}D_0^{-1}s_0\omega^2 r s_1''(c)\left(\frac{D_0^2}{s_0\omega^2 r_b^3 k} - |J(c)|_\infty\right), \end{aligned}$$

we see $2^{-1}(\partial A/\partial r + r^{-1}A) + B \geq 0$ because of Lemma 3.2. Note that (3.5) implies

$$|J(c)|_2^2 \leq C\left|\frac{\partial J(c)}{\partial r}\right|_2^2$$

with a positive constant C. In fact, this is a variant of Poincaré's inequality. Now the lemma follows immediately.

Proof of Proposition 3.1. For any function $\phi(r) \in C_0^1(I)$, we set

$$J(c,\phi) = \int_I \left\{c(r)D_0\frac{\partial\phi(r)}{\partial r} + r s_0\omega^2 s_1(c(r))\phi(r)\right\} dr,$$

so that $J(c,\phi) = -\int_I J(c)\phi(r)\,dr$ when $c \in C^1(I)$. Let $\tilde{c} \in \omega_p(c_0)$. There is a sequence $t_1 < t_2 < \cdots < t_N < \cdots \to \infty$ such that $c(r,t_i) \to \tilde{c}(r)$ in $\mathfrak{L}^p$. Then

$$J(\tilde{c},\phi) = \lim_{i\to\infty} J(c(\ ,t_i),\phi) = -\lim_{i\to\infty}\int_I J(c(r,t_i))\phi(r)\,dr = 0$$

by Lemma 3.3. A routine regularity argument shows that $\tilde{c}$ is the desired equilibrium solution.

Now we discuss the convergence rate of the solution as $t \to \infty$. Let $c_E(r)$ be an equilibrium solution, and recall the operator L_E of (0.14)–(0.15). L_E has a nonnegative selfadjoint extension $\hat{L}_E$ in $L^2(I;\ q(r)r\,dr)$, where $q(r)$ is given by (0.17). Let $W_E(t) = \exp(-t\hat{L}_E)$ be the contraction semigroup of operators in $L^2(I;\ q(r)r\,dr)$ and $W_E(t,r,r')$, the kernel of $W_E(t)$:

$$W_E(t)u = \int_I W_E(t,r,r')u(r')q(r')r'\,dr' \tag{3.10}$$

for $u \in L^2(I;\ q(r)r\,dr)$. Recall λ_E is the smallest nonzero eigenvalue of the operator $\hat{L}_E$. We prove in the Appendix the following:

LEMMA 3.4. *There is a constant $B > 0$ such that*

$$\sup_{r\in I}\int_I\left|\frac{\partial(W_E(t,r,r')q(r'))}{\partial r'}\right| r'\,dr' \leq B(1 + t^{-1/2})\exp(-\lambda_E t)$$

for all $t > 0$.

Let $c=c(r,t;\ c_0)$ be the solution to the problem (0.1)–(0.4) and $c_E(r)$, the equilibrium solution with the same mass as c_0. Then the difference $v=c-c_E$ satisfies the equation:

$$\frac{\partial v}{\partial t}=r^{-1}\frac{\partial}{\partial r}\left\{D_0\frac{\partial v}{\partial r}-rs_0\omega^2 s_1'(c_E)v-rs_0\omega^2 g\right\}$$

in $I\times(0,\infty)$ with the boundary condition

$$D_0\frac{\partial v}{\partial r}-rs_0\omega^2 s_1'(c_E)v-rs_0\omega^2 g=0$$

at $r=r_a$, $r=r_b$ and $v=v_0$ when $t=0$. Here $v_0=c_0-c_E$ and $g=v^2 g_1(r,t)$ with

$$g_1(r,t)=\int_0^1(1-\theta)s_1''(c_E(r)+\theta\{c(r,t)-c_E(r)\})\,d\theta.$$

Therefore, as in the proof of Lemma 1.3, we have

$$v(r,t)=W_E(t)v_0+\int_0^t dt'\int_I K(t-t',r,r',t')v(r',t')^2 r'\,dr', \tag{3.11}$$

where

$$K(t-t',r,r',t')=s_0\omega^2 g_1(r',t')r'\frac{\partial(W_E(t-t',r,r')q(r'))}{\partial r'}. \tag{3.12}$$

By Lemma 3.4 and Proposition 1.6, we see that (3.11) is valid for the generalized solution $c(r,t;\ c_0)\in\mathfrak{L}_+^\infty$ which is not necessarily classical. When we regard (3.11) as an equation for v, we note that the kernel $K(t-t',r,r',t')$ does not involve v since g_1 is determined from known functions c and c_E.

PROPOSITION 3.5. *Let c_E be an equilibrium solution. Then c_E is asymptotically stable. More precisely, there are $\delta>0$ and $C>0$ such that if $|c_0-c_E|_\infty<\delta$, $c_0\in\mathfrak{L}_+^\infty$, then $|U_\infty(t)c_0-c_E|\le C\exp(-\lambda_E t)$ for all $t>0$.*

Proof. We show from (3.11) that if $|v_0|_\infty<\delta$, then $|v|_\infty\le C\exp(-\lambda_E t)$ for a suitable choice of δ and C. Let us denote by $K(v)$ the second term in the right-hand side of (3.11). We claim that (3.11) is solved by successive approximations. Let $v^0=W_E(t)v_0$ and, for $n\ge 1$, $v^n=v^0+K(v^{n-1})$.

LEMMA 3.6. *There is a constant $A>0$ such that*

$$|v^n(\ ,t)|_\infty\le A\delta\exp(-\lambda_E t) \tag{3.13}$$

for all $t>0$, $n=0,1,2,\cdots$.

Proof. Since $\int_I v_0(r)q(r)r\,dr=0$, we get $|v^0(\ \ ,t)|_\infty\le C_0|v_0|_\infty\exp(-\lambda_E t)\le A_0\exp(-\lambda_E t)$, where $A_0=C_0\delta$. Since $|g_1(r,t)|$ is bounded, Lemma 3.4 yields to

$$\int_I K(t-t',r,r',t')r'\,dr'\le B\big(1+(t-t')^{-1/2}\big)\exp(-\lambda_E(t-t')). \tag{3.14}$$

Therefore, putting $a=A_0B(1+\lambda_E^{1/2}B(1/2,1/2)/\sqrt{(2e)})/\lambda_E$, we have $|v^1|_\infty\le A_0(1+a)\exp(-\lambda_E t)$. Let $A_n=1+A_{n-1}^2 a$. By induction on n, we obtain $|v^n|_\infty\le A_0A_n\exp(-\lambda_E t)$. Hence, if $a\le\frac14$, then $A_n\le 2/(1+\sqrt{(1-4a)})=A_\infty$ and (3.13) holds with $C=A_\infty A_0$. The requirement $a\le\frac14$ is always assured if we take δ small enough.

End of the proof of Proposition 3.5. Since

$$v^{n+1}-v^n=K(v^n)-K(v^{n-1})$$
$$=\int_0^t dt'\int_I K(t-t',r,r',t')(v^n+v^{n-1})(v^n-v^{n-1})r'\,dr',$$

we have

$$|v^{n+1}-v^n|_\infty\leq 2A_0A_\infty B\exp(-\lambda_E t)\int_0^t\left(1+(t-t')^{-1/2}\right)|v^n-v^{n-1}|_\infty dt',$$

for $n\geq 1$. The inequality $|v^1-v^0|_\infty\leq A_0 a\exp(-\lambda_E t)$ then implies $|v^{n+1}-v^n|_\infty\leq A_0a(2A_\infty a)^n\exp(-\lambda_E t)$. Thus, if $2A_\infty a<1$, we get $v=\Sigma_{n\geq 0}(v^{n+1}-v^n)+v^0$ in $\mathcal{L}^\infty$ and

$$|v|_\infty\leq A_0\left(1+\frac{a}{1-2A_\infty a}\right)\exp(-\lambda_E t).$$

We have just completed the proof of Proposition 3.5.

Proof of Theorem 3. Let c_0 satisfy the hypotheses of Proposition 3.1. Then since $\omega_\infty(c_0)=\{c_E\}$, we see for T large enough $|U_\infty(T)c_0-c_E|_\infty<\delta$, where δ is as in Proposition 3.5. Then for $t\geq T$, $|U_\infty(t)c_0-c_E|_\infty\leq C\exp(\lambda_E T)\exp(-\lambda_E t)$ while $|U_\infty(t)c_0-c_E|_\infty\leq C_T$ for $t\leq T$. Consequently, we conclude $|U_\infty(t)c_0-c_E|_\infty\leq C_1\exp(-\lambda_E t)$, taking $C_1=\max(C\exp(\lambda_E T),\ C_T\exp(\lambda_E T))$.

4. The case of the vanishing diffusion coefficient. Assume that the diffusion coefficient vanishes, i.e., $D_0=0$. We first study the relation between the solutions $c(r,t)$ of the problem (0.1)–(0.4) and $\tilde{c}(r,t)$ of the problem (0.19)–(0.21). The topic is very closely related to well-known discussions on equations of hyperbolic conservation law by the artificial viscosity method (cf. Lax [12]). Here, however, the present problem is posed on a finite interval with boundary condition (0.20), a deviation from classical theories.

Recall (0.22) for the definition of a generalized solution $\tilde{c}(r,t)$. Note that if $\tilde{c}(r,t)$ is smooth, then it is actually a classical solution of the problem (0.19)–(0.21), as seen from integrations by parts. If, on the other hand, $\tilde{c}(r,t)$ is piecewise continuous and if $r=r^*(t)$ is a smooth nonsingular discontinuity curve of $\tilde{c}(r,t)$, then we can derive from (0.22) an analogue of the Rankine–Hugoniot relation:

$$\frac{dr^*(t)}{dt}=s_0\omega^2r^*(t)\frac{\{s_1(c^+(t))-s_1(c^-(t))\}}{\{c^+(t)-c^-(t)\}},\tag{4.1}$$

where

$$c^\pm(t)=\tilde{c}(r^*(t)\pm 0,t)=\lim_{\varepsilon\downarrow 0}\tilde{c}(r^*(t)\pm\varepsilon,t)\tag{4.2}$$

and $s_1(c)$ is that of (0.16). These relations have in fact been introduced in Fujita [6] after a physicochemical argument.

To assure uniqueness of the (Cauchy) problem for (0.19), it is customary to impose the so-called entropy condition. In the present context, at every discontinuity point $(r^*(t),t)$, the entropy condition should be:

$$\dot{r}^-(t)>\frac{dr^*(t)}{dt}>\dot{r}^+(t).\tag{4.3}$$

Here $\dot{r}^{+}(t)$ and $\dot{r}^{-}(t)$ are the characteristic directions of the equation (0.19) respectively on the right side and on the left side of the discontinuity curve. Namely,

$$(4.4)\qquad \dot{r}^{\pm}(t)=s_0\omega^2 r^*(t)s_1'(c^{\pm}(t))$$

at $(r^*(t),t)$. (4.3) is immediately seen to be equivalent to

$$(4.5)\qquad \tilde{c}(r^*(t)-0,t)<\tilde{c}(r^*(t)+0,t)$$

at every discontinuity point. Although the above discussions are not entirely valid when $\tilde{c}(r,t)$ is not piecewise continuous, the condition (4.5) makes sense as long as values like $\tilde{c}(r_{\pm}0,t)$ are well defined. In particular, this is the case when $\tilde{c}(r,t)$ is nondecreasing in r.

LEMMA 4.1. *Let $c_0(r)\in\mathfrak{L}_+^\infty$ be continuously differentiable and nondecreasing in r. Assume*

$$(4.6)\qquad c_0(r_a)=c_0'(r_a)=0.$$

Then, for a suitable $\delta>0$, there is a family $c_0(r,D_0)\in\mathfrak{L}_+^\infty$, $\delta\geq D_0>0$, smooth, compatible with the boundary condition (0.2) *for each $\delta\geq D_0>0$, nondecreasing in r and in D_0:*

$$(4.7)\qquad \frac{\partial c_0(r,D_0)}{\partial r}\geq 0,\qquad \frac{\partial c_0(r,D_0)}{\partial D_0}\geq 0.$$

Furthermore, for some constant $C>0$,

$$(4.8)\qquad |c_0(\ ,D_0)|_\infty\leq|c_0|_\infty+C$$

and

$$(4.9)\qquad |c_0(\ ,D_0)-c_0|_1\leq CD_0$$

for all $\delta\geq D_0>0$.

Proof. Let $\phi(x)\in C^\infty(\mathbb{R})$ such that $\phi'(x)\geq 0$, $\operatorname{supp}\phi'=[-1,1]$, and $\phi'(0)>0$. Then $\operatorname{supp}\phi=[-1,+\infty)$ and $\phi(0)>0$. We set, for D_0 small enough, $c_0(r,D_0)=c_0(r)+B\phi((r-r_b)(1/D_0-c_0'(r_b)/B))$, where B is to be chosen so that $c_0(r,D_0)$ satisfies (0.2). Thus, $B\phi'(0)=r_bs_0\omega^2\{c_0(r_b)+B\phi(0)\}/\{1+k(c_0(r_b)+B\phi(0))\}$ at $r=r_b$. We have $r_bs_0\omega^2c_0(r_b)/(1+kc_0(r_b))>0$ for sufficiently small D_0 since $c_0(r_b)>0$. The inequalities (4.7) are then clear. The boundary condition at $r=r_a$ for $c_0(r,D_0)$ is obvious from (4.6). (4.8) holds with $C\leq(r_b-r_a)r_b\int_{-1}^{0}d(r)\,dr$. (4.9) is obvious.

Let $c_0(r)$ be as in Lemma 4.1. Let us denote by $c(r,t;\ D_0)$ the solution to the problem (0.1)–(0.4) with the initial data replaced by $c_0(r,D_0)$. We now apply the notion of potential $\Omega(r,t:D_0)$ of Rozhestvenskii and Yanenko [16, Chap. 4, §2.7], with necessary modifications.

LEMMA 4.2. *Put*

$$(4.10)\qquad \Omega(r,t;\ D_0)=rc(r,t;\ D_0)\,dr \\ +r\left\{D_0\frac{\partial c(r,t;\ D_0)}{\partial r}-rs_0\omega^2s_1(c(r,t;\ D_0))\right\}dt.$$

Then the potential

$$(4.11)\qquad \Phi(r,t;\ D_0)=\int_{(r_a,0)}^{(r,t)}\Omega(r',t';\ D_0)$$

does not depend on the choice of paths connecting the points $(r_a, 0)$ *and* (r,t), $r \in I$, $t \geq 0$. *Furthermore*, $\Phi(r,t;\ D_0)$ *are Lipschitz continuous in* r, t, *uniformly with respect to* D_0, $0 < D_0 \leq \delta$.

Proof. Since $c(r,t;\ D_0)$ is a classical solution of (0.1), $d\Omega = 0$, whence $\Phi(r,t;\ D_0)$ is well defined. We see

$$\frac{\partial \Phi}{\partial r} = rc, \qquad \frac{\partial \Phi}{\partial t} = r\left\{D_0 \frac{\partial c}{\partial r} - rs_0\omega^2 s_1(c)\right\}. \tag{4.12}$$

In particular, Φ satisfies the equation

$$\frac{\partial \Phi}{\partial t} = rD_0 \frac{\partial}{\partial r}\left(r^{-1}\frac{\partial \Phi}{\partial r}\right) - r^2 s_0\omega^2 s_1\left(r^{-1}\frac{\partial \Phi}{\partial r}\right), \tag{4.13}$$

for $(r,t) \in I \times (0, \infty)$, and

$$\Phi = 0 \quad \text{at } r = r_a, \tag{4.14}$$

$$\Phi = \int_I c_0(r, D_0) r\, dr \quad \text{at } r = r_b, \tag{4.15}$$

$$\Phi = \int_{r_a}^{r} c_0(r, D_0) r\, dr \quad \text{when } t = 0. \tag{4.16}$$

By Lemma 3.2, we have $|D_0 \partial c(r,t;\ D_0)/\partial r|_\infty \leq C_1$ for $0 < D_0 \leq \delta$. We now claim

$$|c(\ ,t;\ D_0)|_\infty \leq C, \tag{4.17}$$

independent of D_0 and t. In fact, since $c(r,t;\ D_0)$ are nonnegative and nondecreasing in r, we see by the maximum principle $|c(\ ,t;\ D_0)|_\infty \leq \mathrm{Max}\{|c_0(\ ,D_0)|_\infty, \max_{t' \leq t} c(r_b, t')\}$. Since $s_1(c)$ is a nondecreasing function of c, we get (4.17) from the boundary condition (0.2) and Lemma 3.2. Lemma 4.2 is thus proved.

LEMMA 4.3.

$$\frac{\partial \Phi(r,t;\ D_0)}{\partial D_0} \geq 0, \qquad 0 < D_0 \leq \delta.$$

Proof. $v = r^{-1}\partial\Phi/\partial D_0$. Then v satisfies the equation:

$$\frac{\partial v}{\partial t} = D_0 \frac{\partial^2 v}{\partial r^2} + \left(\frac{D_0}{r} + rs_0\omega^2 s_1'(c)\right)\frac{\partial v}{\partial r} - \left(\frac{D_0}{r} + s_0\omega^2 s_1'(c)\right)v + \frac{\partial c}{\partial r}$$

in $I \times (0, \infty)$ with $v \geq 0$ at the boundary $t = 0$, $r = r_a$ and $r = r_b$. Since $\partial c/\partial r \geq 0$, v cannot take a negative minimum in $t \leq T$ for any $T > 0$.

PROPOSITION 4.4. *Let* $c_0(r) \in \mathfrak{L}_+^\infty$ *be that of Lemma* 4.1 *and* $c(r,t;\ D_0)$, *the solution to* (0.1)–(0.4) *with the initial data* $c_0(r, D_0)$. *Then there is a generalized solution* $\tilde{c}(r,t)$ *of the problem* (0.19)–(0.21) *such that, for each* $t > 0$, *we have*

$$c(r,t;\ D_0) \to \tilde{c}(r,t) \tag{4.18}$$

as $D_0 \to 0$ *almost everywhere in* $r \in [r_a, r_b]$. $\tilde{c}(r,t)$ *is nonnegative almost everywhere, and for each* $t > 0$, *is nondecreasing in r almost everywhere.*

Proof. From Lemma 4.3,

$$\Phi(r,t;\ D_0) \to \Phi(r,t;\ 0) \quad \text{as } D_0 \to 0 \tag{4.19}$$

for some function $\Phi(r,t;\,0)$. $\Phi(r,t;\,0)$ is Lipschitz continuous in r, t: $|\Phi(r,t;\,0)-\Phi(r',t;\,0)|\leq A|r-r'|$ and $|\Phi(r,t;\,0)-\Phi(r,t';\,0)|\leq B|t-t'|$, A, B being positive constants independent of r, t. The convergence (4.19) is thus uniform in $r\in I$, $0\leq t\leq T$, for any $T>0$. Furthermore, for each $t>0$, $r^{-1}\partial\Phi(r,t;\,0)/\partial r$ exists almost everywhere in r. If c_0 is nondecreasing, $r^{-1}\partial\Phi(r,t;\,D_0)/\partial r$ are nondecreasing in r, uniformly bounded with respect to $t>0$ and $D_0>0$. This means that the set $\{r^{-1}\partial\Phi(r,t;\,D_0)/\partial r;\ \delta\geq D_0>0\}$ is bounded in the Sobolev space $W_1^1(I\colon\ r\,dr)$ uniformly with respect to t. We thus conclude that, for each $t>0$,

$$r^{-1}\frac{\partial\Phi(r,t;\,D_0)}{\partial r}\to r^{-1}\frac{\partial\Phi(r,t;\,0)}{\partial r}\tag{4.20}$$

as $D_0\to 0$ almost everywhere in r. Let

$$\tilde{c}(r,t)=r^{-1}\frac{\partial\Phi(r,t;\,0)}{\partial r}.\tag{4.21}$$

We claim that $\tilde{c}(r,t)$ is the desired generalized solution of (0.19). Let $w(r,t)$ be a smooth function of class C^1 such that $w(r,T)=0$. Multiply (4.13) by w. Using (4.14)–(4.16) and integration by parts, we obtain

$$\int_I w(r,0)\Phi(r,0;\,D_0)\,dr+\int_0^T dt\int_I w_t(r,t)\Phi(r,t;\,D_0)\,dr$$
$$+D_0\int_0^T\{w(r,t)\Phi_r(r,t;D_0)\}\Big|_{r=r_a}^{r=r_b}dt$$
$$=\int_0^T dt\int_I(rw(r,t))_r D_0 r^{-1}\Phi_r(r,t;\,D_0)\,dr$$
$$+s_0\omega^2\int_0^T dt\int_I w(r,t)s_1(r^{-1}\Phi_r(r,t;\,D_0))r^2\,dr,$$

where $w_t=\partial w/\partial t$, $w_r=\partial w/\partial r$, etc. Therefore, letting $D_0\to 0$, we have

$$-\int_I w(r,0)\Phi(r,0;\,0)\,dr-\int_0^T dt\int_I w_t(r,t)\Phi(r,t;\,0)\,dr$$
$$+s_0\omega^2\int_0^T dt\int_I w(r,t)s_1(\tilde{c}(r,t))r^2\,dr=0.$$

Let $v(r,t)=\int_r^{r_b}w(r',t)\,dr'$. Then $v(r_b,t)=0=v(r,T)$, and $v_r=-w$. Then again by integration by parts, we finally get the relation (0.22) with $\Phi(r,t)$ replaced by $v(r,t)$. Since $\Phi(r,t)$ of (0.22) can be approximated by functions like $v(r,t)$ in the C^1 topology, we conclude that $\tilde{c}(r,t)$ given by (4.21) is a generalized solution to the problem (0.19)–(0.21) with the required properties.

COROLLARY 4.5. $c(\ ,t)\in\mathfrak{L}_+^1$ *and the convergence* (4.18) *is in* $\mathfrak{L}^1$ *for each* $t>0$.

In fact, $c(r,t;\,D_0)$ are uniformly bounded for $t>0$, $\delta\geq D_0>0$.

PROPOSITION 4.6. *Let* $c_0(r)\in\mathfrak{L}_+^\infty$ *be nondecreasing in* r, *and vanish in* $r\leq r_a+\varepsilon$ *for some* $\varepsilon>0$. *Then the same conclusion as Proposition* 4.4 *holds.*

Proof. Extend $c_0(r)$ to the whole real line by setting $c_0(r)=c_0(r_b)$ for $r>r_b$ and $c_0(r)=0$ for $r<r_a$. Let

$$c_0^1(r,D_0)=D_0^{-1}\int_0^{D_0}c_0(r+r')\,dr'.$$

Then $c_0^1(r,D_0)$ are nondecreasing in each of r, D_0. Furthermore, $c_0^1(r_b,D_0)=c_0(r_b)$, $c_0^{1\prime}(r_b,D_0)=0$, and $c_0^1(r_a,D_0)=c_0^{1\prime}(r_a,D_0)=0$ for $D_0<\varepsilon$. Therefore, as in the proof of Lemma 4.1, we have a family of initial data $c_0(r,D_0)$, smooth, compatible with (0.2) for each $0<D_0\leq\min(\varepsilon,\delta)$, uniformly bounded, and nondecreasing in each of r, D_0. The conclusion now follows as in the proof of Proposition 4.4.

Proof of Theorem 4. Immediate from Propositions 4.4 and 4.6.

Proof of Theorem 5. Let $\psi(x)$ be a smooth nondecreasing function such that $\psi(x)=0$ for $x\leq-1$ and $c(x)=1$ for $x\geq0$. Let $\varepsilon>0$. Apply

$$\phi(r,t)=(T-t)\{\psi((r_b-r)/\varepsilon)-1\}$$

to the formula (0.22). Since

$$\int_0^T(T-t)\,dt\int_I\{-\varepsilon^{-1}\psi'((r_b-r)/\varepsilon)\}\{s_1(\tilde{c}(r,t))r^2-s_1(\tilde{c}(r_b-0,t))r_b^2\}\,dr$$

$$=\int_0^T(T-t)\,dt\int\psi'(\rho)\{s_1(\tilde{c}(r_b-\varepsilon\rho,t))(r_b-\varepsilon\rho)^2-s_1(\tilde{c}(r_b-0,t))r_b^2\}\,d\rho,$$

we see that this integral tends to zero as $\varepsilon\to0$ by Lebesgue's dominated convergence theorem. Therefore, letting $\varepsilon\to0$ in (0.22) with the above ϕ, we have

$$T\int_I c_0(r)r\,dr=\int_0^T dt\int_I\tilde{c}(r,t)r\,dr$$
$$+r_b^2s_0\omega^2\int_0^T(T-t)s_1(\tilde{c}(r_b-0,t))\,dt.$$

Differentiation in T leads to (0.23). Now (0.23) implies that, for some $\tilde{m}\geq0$, $\int_I\tilde{c}(r,t)r\,dr\to\tilde{m}$ as $T\to\infty$. If $\tilde{m}>0$, then for some $T_1>0$ and $\tilde{m}_1>0$, we would have $\tilde{c}(r_b-0,t)\geq\tilde{m}_1$ for $t\geq T_1$. But then $s_1(\tilde{c}(r_b-0,t))\geq\tilde{m}_2$ for $t\geq T_1$ with some $\tilde{m}_2>0$, contradicting (0.23). Therefore, we must have (0.24).

Appendix. Let $P(r)$ be a smooth (C^∞) positive function on the interval $I=[a,b]$, $0<a<b<\infty$, and let

$$M_pu=-P(r)^{-1}\frac{\partial}{\partial r}\left(P(r)\frac{\partial u}{\partial r}\right)\tag{A.1}$$

for $u\in\mathcal{D}_1=\{u\in C^2(I);\ \partial u/\partial r=0 \text{ at } \partial I\}$. M_p has a selfadjoint extension $\hat{M}_P$ in the Hilbert space $L^2(I;\ P(r)\,dr)$. $\hat{M}_P$ is nonnegative definite and generates a contraction semigroup $W_P(t)$ of bounded operators in $L^2(I;\ P(r)\,dr)$. Let $W_P(t,r,r')$ be the kernel of $W_P(t)$:

$$W_P(t)u(r)=\int_I W_P(t,r,r')u(r')P(r')\,dr'\tag{A.2}$$

for $u\in L^2(I;\ P(r)\,dr)$.

Proposition A.1. *Let μ be the smallest nonzero eigenvalue of the operator $\hat{M}_P$. Then there is a constant $C>0$ such that for all $t>0$ we have*

$$\int_I\left|\frac{\partial W_P(t,r,r')}{\partial r}\right|P(r')\,dr'\leq C(1+t^{-1/2})e^{-\mu t}$$

and

$$\int_I \left|\frac{\partial W_P(t,r,r')}{\partial r'}\right| P(r')\,dr' \le C(1+t^{-1/2})e^{-\mu t}.$$

Proof. Let $u^+(r,\lambda)$ and $u^-(r,\lambda)$ constitute a fundamental pair to the equation

$$-\lambda u + P^{-1}\frac{\partial}{\partial r}\left(P\frac{\partial u}{\partial r}\right)=0, \qquad \lambda\in\mathbf{C}. \tag{A.3}$$

We may choose $u^{\pm}(r,\lambda)$ so that they are entire analytic in λ and have the asymptotic expansions:

$$u^{\pm}(r,\lambda)=(4\lambda)^{-1/4}P(r)^{-1/2}(1+O(\lambda^{-1/2}))\exp\{\pm r\lambda^{1/2}\} \tag{A.4}$$

for $|\lambda|$ large enough outside the negative real axis, $|\arg\lambda|\le\pi-\varepsilon$, for any $\varepsilon>0$. Note then the wronskian:

$$w(r,\lambda)=P(r)\{u_r^+(r,\lambda)u^-(r,\lambda)-u_r^-(r,\lambda)u^+(r,\lambda)\}=1, \tag{A.5}$$

where $u_r^+=\partial u^+/\partial r$ etc. Let

$$\tilde{u}(r,r',\lambda)=u^+(r,\lambda)u^-(r',\lambda)-u^+(r',\lambda)u^-(r,\lambda)$$

and

$$\tilde{v}(r,r',\lambda)=\frac{\partial\tilde{u}(r,r',\lambda)}{\partial r'}.$$

Then $w(r,\lambda)=-P(r)\tilde{v}(r,r,\lambda)=1$. Let $\sigma(-\hat{M_P})$ be the spectrum of the operator $-\hat{M_P}$. For $\lambda\in\sigma(-\hat{M_P})$, let $G_P(\lambda)$ be the resolvent $(\lambda+\hat{M_P})^{-1}$ and $G(r,r',\lambda)$ the kernel of $G_P(\lambda)$:

$$G_P(\lambda)u(r)=\int_I G(r,r',\lambda)u(r')P(r')\,dr' \tag{A.6}$$

for $u\in L^2(I;\ P(r)\,dr)$. Let us set

$$u_a(r,\lambda)=\tilde{v}(r,a,\lambda) \quad\text{and}\quad u_b(r,\lambda)=\tilde{v}(r,b,\lambda).$$

Then $(\lambda+M_P)u_a=0$ in I with $\partial u_a/\partial r=0$ at $r=a$ and $(\lambda+M_P)u_b=0$ in I with $\partial u_b/\partial r=0$ at $r=b$. Since $u_a(r,\lambda)\partial u_b(r,\lambda)/\partial r-r_a(r,\lambda)\partial u_b(r,\lambda)/\partial r=\Delta(\lambda)/P(r)$, where $\Delta(\lambda)=\{u_r^+(b,\lambda)u_r^-(a,\lambda)-u_r^+(a,\lambda)u_r^-(b,\lambda)\}$, we have

$$\sigma(-\hat{M_P})=\{\lambda\in\mathbf{C};\ \Delta(\lambda)=0\}. \tag{A.7}$$

It then follows that, for $\lambda\in\sigma(-\hat{M_P})$, we have

$$G_P(r,r',\lambda)=\begin{cases}\dfrac{u_b(r,\lambda)u_a(r',\lambda)}{\Delta(\lambda)}, & r>r',\\[2ex] \dfrac{u_a(r,\lambda)u_b(r',\lambda)}{\Delta(\lambda)}, & r<r'.\end{cases} \tag{A.8}$$

Consequently, for $t>0$, the semigroup kernel is given by

$$W_P(t,r,r')=(2\pi i)^{-1}\int_\Gamma e^{\lambda t}G_P(r,r',\lambda)\,d\lambda, \tag{A.9}$$

where Γ is a path from $\infty e^{-i\theta}$ to $\infty e^{i\theta}$, $\pi/2<\theta<\pi$, encircling the negative real axis counterclockwise. Since $\operatorname{Re}\lambda^{1/2}>0$ for $\lambda\in\Gamma$ with $|\lambda|$ large enough, we see from (A.4) that

$$\Delta(\lambda)=-2^{-1}\lambda^{1/2}\{P(a)P(b)\}^{-1/2}\{1+O(\lambda^{-1/2})\}\exp\{(b-a)\lambda^{1/2}\},$$
$$u_b(r,\lambda)u_a(r',\lambda)=4^{-1}\{P(r)P(r')P(a)P(b)\}^{-1/2}$$
$$\times\{-1+O(\lambda^{-1/2})\}\exp\{(r'-a+b-r)\lambda^{1/2}\}$$

for $r>r'$ and

$$u_a(r,\lambda)u_b(r',\lambda)=4^{-1}\{P(r)P(r')P(a)P(b)\}^{-1/2}$$
$$\times\{-1+O(\lambda^{-1/2})\}\exp\{(r-a+b-r')\lambda^{1/2}\}$$

for $r'<r$. Therefore,

$$G_P(r,r',\lambda)=2^{-1}\{\lambda P(r)P(r')\}^{-1/2}\{1+O(\lambda^{-1/2})\}\exp\{-|r-r'|\lambda^{1/2}\} \tag{A.10}$$

for $r\neq r$, $\lambda\in\Gamma$, $|\lambda|\to\infty$. It follows immediately from (A.9) and (A.10) that

$$|W_P(t,r,r')|\leq Ct^{-1/2}\exp\left\{-\frac{|r-r'|^2}{4t}\right\}$$

for small $t>0$, $r\neq r'$. Differentiating (A.10) in r or r', we obtain similarly

$$\int_I\left|\frac{\partial W_P(t,r,r')}{\partial r}\right|P(r')\,dr'\leq Ct^{-1/2} \tag{A.11}$$

and

$$\int_I\left|\frac{\partial W_P(t,r,r')}{\partial r'}\right|P(r')\,dr'\leq Ct^{-1/2} \tag{A.12}$$

for small t, $0<t\leq t_0$. Let $\Gamma(N)$ be a contour in the complex plane which coincides with Γ for $|\lambda|$ large and the first N eigenvalues $\lambda(i)$, $i=0,\cdots,N-1$, lie outside the region encircled by $\Gamma(N)$. Then if

$$W_P^{(N)}(t,r,r')=(2\pi i)^{-1}\int_{\Gamma(N)}e^{\lambda t}G_P(r,r',\lambda)\,d\lambda,$$

we have

$$W_P(t,r,r')=\sum_{i<N}e^{-\lambda(i)t}u_i(r)u_i(r')+W_P^{(N)}(t,r,r').$$

Here $u_i(r)$ are the orthonormal eigenvectors of $\hat{M_P}$ corresponding to $\lambda(i)$. Note $\lambda(0)=0$ and $u_0(r)=(\int_I P(r)\,dr)^{-1}$. We then have

$$\int_I\left|\frac{\partial W_P(t,r,r')}{\partial r}\right|P(r')\,dr'\leq Ce^{-\lambda(1)t} \tag{A.13}$$

and

$$\text{(A.14)} \qquad \int_I \left| \frac{\partial W_P(t,r,r')}{\partial r'} \right| P(r')\,dr' \le Ce^{-\lambda(1)t}$$

for t large enough. Since $\lambda(1)=\mu$, Lemma A.1 is thus proved.

Remark. For a derivation of (A.4), consult Yosida [18] and Erdelyi [4]. If $P(r)=r$, the result is classical.

Proof of Lemma 1.1. If $D_0=1$, we verify Lemma 1.1 immediately from Proposition A.1 by taking $P(r)=r$, $a=r_a$, $b=r_b$. For general $D_0>0$, we apply the homogeneity (1.15).

Proof of Lemma 3.4. Let $D_0=1$. Since, by (0.15),

$$L_E u = -r^{-1}\frac{\partial}{\partial r}\left\{ rq(r)^{-1}\frac{\partial}{\partial r}(q(r)u) \right\},$$

we see $q(r)^{-1}W_P(t,r,r')=W_E(t,r,r')q(r')$ if $P(r)=rq(r)^{-1}$, $a=r_a$, $b=r_b$. Since $q(r)>0$ on I, Lemma 3.4 follows at once from Lemma A.1.

Acknowledgments. We acknowledge Dr. Fumio Arisaka of the Pharmaceutical Department, Hokkaido University, for the introduction to us of the Lamm equation as well as related literature. We owe much to Professor Hiroshi Matano of the Department of Mathematics, Hiroshima University, for removing our original restrictions in the initial data in stating Theorem 1. We are also grateful to Professor Hiroshi Fujita of the Department of Polymer Science, Osaka University, for relevant information.

REFERENCES

[1] W. J. Archibald, *The process of diffusion in a centrifugal field of force*, I, II, Phys. Rev. 53 (1938), pp. 746–752; 54 (1938), pp. 371–374.

[2] N. Bourbaki, *Fonctions d'une variable réelle. Théorie elémentaire*, Hermann, Paris, 1976.

[3] M. Dyshon, G. H. Weiss and D. A. Yphantis, *Numerical solutions of the Lamm equation*, I, II, III, Biopolymers, 4 (1966), pp. 449–455, pp. 457–468; 5 (1967), pp. 697–713.

[4] A. Erdelyi, *Asymptotic Expansions*, Dover, New York, 1960.

[5] H. Faxen, *Ueber eine Differentialgleichung aus der physikalischen Chemie*, Ark. Mat. Astron. Fys., 21B (No. 3) (1929).

[6] H. Fujita, *Mathematical Theory of Sedimentation Analysis*, Academic Press, New York, 1962.

[7] ______, *Foundations of Ultracentrifugal Analysis*, John Wiley, New York, 1975.

[8] J. K. Hale and E. F. Infante, *Extended dynamical systems and stability theory*, Proc. Nat. Acad. Sci. U.S.A., 58 (1967), pp. 405–409.

[9] P. Hartman, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[10] O. A. Ladyzhenskaya, V. A. Solonnikov and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[11] O. Lamm, *Die Differentialgleichung der Ultrazentrifugierung*, Ark. Mat. Astron. Fys., 21B (No. 2) (1929).

[12] P. D. Lax, *Hyperbolic systems of conservation law* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[13] H. Komatsu, *A general interpolation theorem of Marcinkiewicz type*, Tôhoku Math. J. 2nd. Ser., 33 (1981), pp. 383–393.

[14] P. Kree, *Interpolation d'espaces qui ne sont ni normés, ni complets. Applications*, Ann. Inst. Fourier, 17 (Fasc. 2) (1967), pp. 137–174.

[15] H. Matano, *Nonincrease of the lap-number of a solution for a one-dimensional semilinear parabolic equation*, J. Fac. Sci. Univ. Tokyo Sec. 1A Math., to appear.

[16] B. L. Rozhestvenskii and N. N. Yanenko, *Systems of Quasilinear Equations and its Applications to Gas Dynamics*, 2nd ed., Nauka, Moscow, 1978. (In Russian).

[17] G. H. Weiss, *An Archibald type solution to a nonlinear Lamm equation*, Nature, 202 (1964), pp. 792–793.

[18] K. Yosida, *Lectures on Differential and Integral Equations*, Interscience, New York, 1960.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics
007

ON THE EIGENVALUES OF A CERTAIN INTEGRAL EQUATION*

B. F. LOGAN†

Abstract. It is shown that the integral equation

$$\int_0^\infty f(t)\frac{\sin(x-t)}{\pi(x-t)}dt=\lambda f(x)\qquad (x>0)$$

has a solution f for any (complex) λ, excluding the real numbers $\lambda\le 0$, $\lambda>1$. The closure of the set of eigenvalues, i.e., the spectrum of the integral operator (over all functions in its domain) is then the entire complex plane.

The integral equation

$$(1)\qquad \int_0^\infty f(t)\frac{\sin(x-t)}{\pi(x-t)}dt=\lambda f(x)\qquad (x>0)$$

was shown by Krein and Nudel'man [1] to have a solution f for any λ in $(0,1)$. They remarked that $[0,1]$ constitutes the spectrum of the integral operator. Actually, this statement should be qualified in the context of $L^2(0,\infty)$. The linear transformation Af defined by the integral (1) carries $L^2(0,\infty)$ into $L^2(0,\infty)$ and the norm of the transformation is clearly 1. However the eigenfunctions do not belong to $L^2(0,\infty)$ because the transformation is not "completely continuous" (See Riesz-Sz. Nagy [2, pp. 231–232]) Landau and Pollak [3] have shown that (1) has "approximate solutions" in $L^2(0,\infty)$ in the sense that for any λ in $(0,1)$ and any $\varepsilon>0$ there exists f in $L^2(0,\infty)$ such that $\|Af-\lambda f\|<\varepsilon$. The purpose of this note is to point out that (1) has a solution f (such that the integral is absolutely convergent) for any (complex) λ excluding the real numbers $\lambda\le 0$, $\lambda>1$. The closure of the set of eigenvalues, i.e., the spectrum of the integral operator (over all functions in its domain) is then the entire complex plane. This is rather interesting, particularly in view of the fact that the integral kernel is positive definite.

Now we will show that

$$(2)\quad f(t)=e^{-it}{}_1F_1(\gamma;1;2it)=\frac{\sin\pi\gamma}{\pi}\int_{-1}^{1}\frac{e^{ixt}dx}{(1-x)^\gamma(1+x)^{1-\gamma}}\qquad (0<\operatorname{Re}\gamma<1)$$

satisfies (1) with $\lambda=(1-e^{-2\pi i\gamma})^{-1}$.

These functions are Fourier transforms of the eigenfunctions of the finite Hilbert transform, which apparently have not been pointed out in their entirety. Those in [1] correspond to $\operatorname{Re}\gamma=\frac{1}{2}$.

Suppose $f(t)$ and $(\operatorname{sgn} t)f(t)$ have Fourier transforms:

$$(3)\qquad F(x)=\int_{-\infty}^{\infty}f(t)e^{-ixt}dt,$$

$$(4)\qquad G_a(x)=\int_{-\infty}^{\infty}(\operatorname{sgn} t)f(t)e^{-a|t|}e^{-ixt}dt\qquad (a>0).$$

* Received by the editors March 12, 1982.

† Bell Laboratories, Murray Hill, New Jersey 07974.

We find that

$$G_a(x)=i\int_{-\infty}^{\infty}F(t)\frac{t-x}{\pi\left[(t-x)^2+a^2\right]}\,dt, \tag{5}$$

$$\lim_{a\to 0}G_a(x)=-i\fint_{-\infty}^{\infty}\frac{F(t)}{\pi(x-t)}\,dt=-i\tilde{F}(x) \tag{6}$$

where $\tilde{F}$ (often $-\tilde{F}$) is called the Hilbert transform of F. (The cut in the integral sign indicates a Cauchy principal value.)

We suppose in (1) that $f(t)$ is defined for negative argument by the integral; i.e., $f(x)$ is the restriction to the real line of an entire function of order 1, type 1, and if f has a Fourier transform it must vanish outside $[-1,1]$. Then writing

$$\int_0^{\infty}f(t)e^{-ixt}\,dt=\frac{1}{2}\int_{-\infty}^{\infty}(1+\operatorname{sgn}t)f(t)e^{-ixt}\,dt$$

we infer from (1) that

$$\frac{1}{2}\{F(x)-i\tilde{F}(x)\}=\lambda F(x) \qquad (-1<x<1) \tag{7}$$

or

$$\tilde{F}(x)=i(2\lambda-1)F(x) \qquad (-1<x<1), \tag{8}$$

i.e.,

$$\fint_{-1}^{1}\frac{F(t)}{\pi(x-t)}\,dt=\mu F(x) \qquad (-1<x<1), \tag{9}$$

where

$$\mu=i(2\lambda-1).$$

For $\mu=0$, a nontrivial solution of (9) is

$$F(t)=(1-t^2)^{-1/2},$$

which gives $f(t)=J_0(t)$ a solution of (1) for $\lambda=\frac{1}{2}$.

Generally one can evaluate finite Hilbert transforms explicitly (then by an indirect method) only for certain elementary functions; for example,

$$\int_{-1}^{1}\frac{P_n(t)}{(1-t)^{\gamma}(1+t)^{1-\gamma}}\frac{dt}{\pi(x-t)} \qquad (0<\operatorname{Re}\gamma<1),$$

where $P_n(t)$ is a polynomial in t, can be evaluated using the theory of functions analytic in the upper half-plane. We will find that

$$\fint_{-1}^{1}\frac{1}{(1-t)^{\gamma}(1+t)^{1-\gamma}}\frac{dt}{\pi(x-t)}=\frac{1}{\tan\pi\gamma}\frac{1}{(1-x)^{\gamma}(1+x)^{1-\gamma}} \tag{10}$$

$$(-1<x<1,\ 0<\operatorname{Re}\gamma<1),$$

which with (3) and (9) shows that (2) satisfies (1).

To obtain the result (10) we use the fact that for functions $H(z)$ analytic in the upper half-plane, satisfying a local integrability condition on the real axis, and a condition at infinity, for example, $H(z)\to 0$, the imaginary part of $H(x)$ is the Hilbert transform of the real part of $H(x)$: i.e., writing

$$h(x)=\operatorname{Re}H(x)$$

we have

$$\text{(11)}\qquad \operatorname{Im}H(x)=\lim_{y\to 0}\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x-t}{(x-t)^2+y^2}h(t)=-\!\!\!\!\!\!\int_{-\infty}^{\infty}\frac{h(t)}{\pi(x-t)}dt=\tilde{h}(x)\quad\text{a.e.}$$

Now consider the function analytic in the upper half-plane

$$\text{(12)}\qquad H(z;\gamma)=\frac{1}{(z-1)^{\gamma}(z+1)^{1-\gamma}},$$

where $\gamma=\alpha+i\beta$, $0<\alpha<1$, $-\infty<\beta<\infty$, and we take the branches so that

$$0\le\arg(z-1)\le\pi,\qquad 0\le\arg(z+1)\le\pi$$

and on the real axis

$$\arg(x+i0-1)=\begin{cases}0 & (x>1),\\ \pi & (x<1),\end{cases}$$

$$\arg(x+i0+1)=\begin{cases}0 & (x>-1),\\ \pi & (x<-1).\end{cases}$$

Then writing

$$H(z,\gamma)=e^{-\gamma\log(z-1)-(1-\gamma)\log(z+1)}\qquad(\gamma=\alpha+i\beta)$$

and

$$\phi(x)=\log\left|\frac{1+x}{1-x}\right|$$

we have

$$\text{(13)}\qquad H(x;\gamma)=\begin{cases}\dfrac{e^{i\beta\phi(x)}}{|1-x|^{\alpha}|1+x|^{1-\alpha}} & (x>1),\\[2ex] \dfrac{e^{-i\pi(\alpha+i\beta)}e^{i\beta\phi(x)}}{|1-x|^{\alpha}|1+x|^{1-\alpha}} & (-1<x<1),\\[2ex] -\dfrac{e^{i\beta\phi(x)}}{|1-x|^{\alpha}|1+x|^{1-\alpha}} & (x<-1),\end{cases}\qquad(\gamma=\alpha+i\beta).$$

Now defining

$$\text{(14)}\qquad H_1(z;\alpha,\beta)=iH(z;\alpha+i\beta)+iH(a;\alpha-i\beta),$$

$$\text{(15)}\qquad H_2(z;\alpha,\beta)=H(z;\alpha+i\beta)-H(z;\alpha-i\beta)$$

we see that on the real axis both H_1 and H_2 take pure imaginary values outside $[-1,1]$; i.e., the imaginary parts of these functions are Hilbert transforms of functions supported on $(-1,1)$. We have

(16)

$$h_1(x)=\operatorname{Re}H_1(x;\alpha,\beta)=\begin{cases}\dfrac{-e^{\beta\pi}\sin(\beta\phi-\alpha\pi)+e^{-\beta\pi}\sin(\beta\phi+\alpha\pi)}{(1-x)^{\alpha}(1+x)^{1-\alpha}} & (-1<x<1),\\ 0 & (|x|>1)\end{cases}$$

(17)

$$\tilde{h}_1(x)=\operatorname{Im}H_1(x;\alpha,\beta)=\begin{cases}\dfrac{e^{\beta\pi}\cos(\beta\phi-\alpha\pi)+e^{-\beta\pi}\cos(\beta\phi+\alpha\pi)}{(1-x)^{\alpha}(1+x)^{1-\alpha}} & (-1<x<1)\\ -(\operatorname{sgn}x)\dfrac{2\cos\beta\phi}{|1-x|^{\alpha}|1+x|^{1-\alpha}} & (|x|>1),\end{cases}$$

(18)

$$h_2(x)=\operatorname{Re}H_2(x;\alpha,\beta)=\begin{cases}\dfrac{e^{\beta\pi}\cos(\beta\phi-\alpha\pi)-e^{-\beta\pi}\cos(\beta\phi+\alpha\pi)}{(1-x)^{\alpha}(1+x)^{1-\alpha}} & (-1<x<1),\\ 0 & (|x|>1),\end{cases}$$

(19)

$$\tilde{h}_2(x)=\operatorname{Im}H_2(x;\alpha,\beta)=\begin{cases}\dfrac{e^{\beta\pi}\sin(\beta\phi-\alpha\pi)+e^{-\beta\pi}\sin(\beta\phi+\alpha\pi)}{(1-x)^{\alpha}(1+x)^{1-\alpha}} & (-1<x<1),\\ (\operatorname{sgn}x)\dfrac{2\sin\beta\phi}{|1-x|^{\alpha}|1+x|^{1-\alpha}} & (|x|>1).\end{cases}$$

(In (16)–(19), $\phi\equiv\phi(x)=\log|(1+x)/(1-x)|$, $0<\alpha<1$, $-\infty<\beta<\infty$.) Now a certain linear combination of h_1 and h_2 will give the result (10).

First we find, eliminating $\sin\beta\phi$ between h_1 and h_2, that

$$(20)\qquad g_1(x)=\cosh\beta\pi\,\sin\alpha\pi\,h_1(x)+\sinh\beta\pi\,\cos\alpha\pi\,h_2(x)$$

$$=\begin{cases}(\cosh 2\beta\pi-\cos 2\alpha\pi)\dfrac{\cos\beta\phi(x)}{(1-x)^{\alpha}(1+x)^{1-\alpha}} & (-1<x<1),\\ 0 & (|x|>1)\end{cases}$$

with the corresponding linear combinations of $\tilde{h}_1$ and $\tilde{h}_2$ giving

$$(21)\qquad \tilde{g}_1(x)=\cosh\beta\pi\,\sin\alpha\pi\,\tilde{h}_1(x)+\sinh\beta\pi\,\cos\alpha\pi\,\tilde{h}_2(x)$$

$$=\frac{\sinh 2\beta\pi\,\sinh\beta\phi(x)+\sin 2\alpha\pi\,\cos\beta\phi(x)}{(1-x)^{\alpha}(1+x)^{1-\alpha}}\qquad(-1<x<1).$$

Then, eliminating $\cos\beta\phi$ between h_1 and h_2, we find

$$(22)\quad g_2(x)=\cosh\beta\pi\,\sin\alpha\pi\,h_2(x)-\sinh\beta\pi\,\cos\alpha\pi\,h_1(x)$$
$$=\begin{cases}(\cosh2\beta\pi-\cos2\alpha\pi)\dfrac{\sin\beta\phi(x)}{(1-x)^{\alpha}(1+x)^{1-\alpha}} & (-1<x<1),\\ 0 & (|x|>1),\end{cases}$$

with the corresponding linear combinations of $\tilde{h}_2$ and $\tilde{h}_1$ giving

$$(23)\quad \tilde{g}_2(x)=\cosh\beta\pi\,\sin\alpha\pi\,\tilde{h}_2(x)-\sinh\beta\pi\,\cos\alpha\pi\,\tilde{h}_1(x)$$
$$=\frac{-\sinh2\beta\pi\,\cos\beta\phi(x)+\sin2\alpha\pi\,\sin\beta\phi(x)}{(1-x)^{\alpha}(1+x)^{1-\alpha}}\qquad(-1<x<1).$$

Then we have

$$(24)\quad g_1(x)+ig_2(x)=\begin{cases}(\cosh2\beta\pi-\cos2\alpha\pi)\dfrac{e^{i\beta\phi(x)}}{(1-x)^{\alpha}(1+x)^{1-\alpha}} & (-1<x<1),\\ 0 & (|x|>1),\end{cases}$$

$$(25)\quad \tilde{g}_1(x)+i\tilde{g}_2(x)=(\sin2\alpha\pi-i\sinh2\beta\pi)\frac{e^{i\beta\phi(x)}}{(1-x)^{\alpha}(1+x)^{1-\alpha}}\qquad(-1<x<1).$$

Writing

$$\frac{e^{i\beta\phi(x)}}{(1-x)^{\alpha}(1+x)^{1-\alpha}}=\frac{1}{(1-x)^{\gamma}(1+x)^{1-\gamma}}\qquad(-1<x<1,\gamma=\alpha+i\beta),$$

we have

(26)

$$\int_{-1}^{1}\frac{1}{(1-t)^{\gamma}(1+t)^{1-\gamma}}\frac{dt}{\pi(x-t)}=\mu\cdot\frac{1}{(1-x)^{\gamma}(1+x)^{1-\gamma}}\qquad(-1<x<1,0<\operatorname{Re}\gamma<1),$$

where $\gamma=\alpha+i\beta$, and

$$\mu=\frac{\sin2\alpha\pi-i\sinh2\beta\pi}{\cosh2\beta\pi-\cos2\alpha\pi}=\frac{1}{\tan\pi\gamma},$$

which is the desired result (10).

We note from the asymptotics of the Kummer function that $f(t)$ defined in (2), with $\gamma=\alpha+i\beta$, satisfies

$$|f(t)|=O\left(\frac{1}{|t|^{\alpha}}+\frac{1}{|t|^{1-\alpha}}\right)\qquad(t\to\pm\infty).$$

Thus for $0<\alpha<1$, the integral in (1) with $f(t)$ given by (2) is absolutely convergent.

REFERENCES

[1] M. G. KREIN AND P. JA. NUDEL'MAN, *On some new problems for Hardy class functions and continuous families of functions with double orthogonality*, Soviet Math. Dokl., 14 (1973), pp. 435–439.

[2] F. RIESZ AND B. SZ.-NAGY, *Functional Analysis*, Ungar, New York, 1955.

[3] H. J. LANDAU AND H. P. POLLAK, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*, III. *The dimension of the space of essentially-time-and band-limited signals*, Bell Syst. Tech. J., 41 (1962), pp. 1295–1336.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics
008

FOURIER INVERSION OF THE ATTENUATED X-RAY TRANSFORM*

ANDREW MARKOE[†]

Abstract. A variably attenuated x-ray transform is shown to be invertible via an integral formula for the inversion of the exponential x-ray transform.

The attenuation must be known and constant in a convex set containing the unknown emitter. However the attenuation can be otherwise arbitrary.

If μ denotes the attenuation constant of the exponential x-ray transform then the integral formula computes the Fourier transform of the emitter on all of R^n from the values of the Fourier transform on the set $A^\mu=\{\sigma+i\mu\omega\in C^n|\omega\in S^{n-1}, \sigma\perp\omega\}$. Of course F. Natterer [Numer. Math., 32 (1979), pp. 431–438] showed that the values of the Fourier transform of the emitter can be obtained from the Fourier transform of the exponential x-ray transform. In essence however the basic method is analytic continuation from the set A^μ.

A consequence of the integral formula is a uniqueness theorem for attenuated x-ray transforms of the type considered here: if the transforms of two objects agree at infinitely many directions, then the objects are the same.

1. Introduction and notation. The attenuated x-ray transform occurs in mathematical models of single photon emission tomography. This paper is concerned with inverting a variably attenuated x-ray transform by obtaining the Fourier transform of the unknown emitter from its attenuated x-ray transform. The main tool is an extension theorem in several complex variables since it was shown by F. Natterer [3] that the (constantly) attenuated x-ray transform can give the Fourier transform on a portion of complex Euclidean space. The emitter is assumed to exist in a convex region of constant attenuation, but otherwise there is no restriction on the attenuator except that it be measurable and bounded. The problem of inverting the attenuated x-ray transform for an arbitrary bounded and measurable attenuator is apparently still open.

The problem of inverting the constantly attenuated x-ray transform has also been solved by Piacentini et al. [1], Quinto [4] and Tretiak and Metz [5]. F. Natterer [3] has found an approximate inversion, but he also proved the important analogue of the projection theorem referred to above and the spirit of this paper most closely follows Natterer's. The paper along with the others cited assume that the attenuation is known. In practice this can be obtained by transmission tomography.

Let B^n be the unit ball of R^n and let S^{n-1} be the unit sphere. If $\omega\in S$ then $\omega^\perp$ denotes the hyperplane through the origin orthogonal to ω.

The attenuated x-ray transform is defined by

$$P^\mu f(s,\omega)=\int_{-\infty}^{\infty} f(s+t\omega)\exp\left(-\int_t^\infty \mu(s+\tau\omega)\,d\tau\right)dt,$$

where $f\in L^2(B^n)$, $\omega\in S^{n-1}$, $s\in\omega^\perp$ and μ is the attenuation.

If μ is constant then P^μ is called the constantly attenuated x-ray transform.

The exponential x-ray transform is defined by

$$Q^\mu f(x,\omega)=\int_{-\infty}^{\infty} f(s+t\omega)\exp(\mu t)\,dt$$

* Received by the editors December 4, 1981, and in revised form March 2, 1983.

[†] Department of Mathematics and Physics, Rider College, Lawrenceville, New Jersey 08648.

(cf. Tretiak and Metz [5]). It is useful since if μ is constant then the relation

$$P^{\mu}f(s,\omega)=\exp\Big(-\mu\big(1-|s|^2\big)^{1/2}\Big)Q^{\mu}f(s,\omega)$$

holds. Furthermore it will be shown that the inversion of the variably attenuated x-ray transform considered here can be reduced to the inversion of the exponential transform.

The series of papers by Budinger and Gullberg ([2] for example) are a good source of results and references on the practical side of attenuated x-ray transforms.

2. An extension problem in several complex variables. In this section it is shown how to analytically continue an entire function whose values are known on

$$A^{\mu}=\{\sigma+i\mu\omega\,|\,\omega\in S^{n-1},\ \sigma\in\omega^{\perp}\}$$

from A^{μ} to R^n. This will lead to an inversion of the attenuated x-ray transform.

DEFINITION 2.1. Let $t\colon R^n\to R$ be defined by

$$t(\sigma)=\left(\sigma_1^2+\sigma_2^2+\mu^2\right)^{1/2}. \tag{2.1}$$

Now define the parametrized family of plane curves Γ_σ by

$$\theta\to\Gamma_\sigma(\theta)=t(\sigma)\sin\theta+i\mu\cos\theta. \tag{2.2}$$

Then define $Z\colon C\times R^n\to C^n$ by

$$\begin{aligned} &Z_1(\gamma,\sigma)=\gamma,\\ &Z_2(\gamma,\sigma)=\left[\sigma_1^2+\sigma_2^2-\gamma^2\right]^{1/2},\\ &Z_j(\gamma,\sigma)=\sigma_j,\qquad j=3,\cdots,n. \end{aligned} \tag{2.3}$$

The branch of the square root defining Z_2 is arbitrary except in two cases. If the argument is positive then the positive branch is taken; if $\gamma=\Gamma_\sigma(\theta)$ then $Z_2(\gamma,\sigma)$ is taken to be

$$Z_2(\gamma,\sigma)=-t(\sigma)\cos\theta+i\mu\sin\theta. \tag{2.4}$$

It is easily shown that in this case $Z_2(\gamma,\sigma)^2=[\sigma_1^2+\sigma_2^2-\gamma^2]^{1/2}$.

Finally if $u\in C^n$, then u_- denotes the reflection of u in the second coordinate. The notation also applies to functions; $F_-(z)=[F(z)]_-$

THEOREM 2.1. *Let F be an entire holomorphic function on C^n and follow the notation of Definition* 2.1. *Then*

$$Z(\Gamma_\sigma,\sigma)\subset A^{\mu} \tag{2.5}$$

for all $\sigma\in R^n$.

Furthermore $F|_{R^n}$ can be recovered from values of F on A^{μ} by the integral formula

$$F(\sigma)=\frac{i}{4\pi}\int_0^{2\pi}\frac{[t(\sigma)\cos\theta-i\mu\sin\theta]\,\Xi(Z(\Gamma_\sigma(\theta),\sigma))+\sigma_2\Omega(Z(\Gamma_\sigma(\theta),\sigma))}{t(\sigma)\sin\theta+i\mu\cos\theta-\sigma_1}\,d\theta \tag{2.6}$$

where $\Xi=F+F_-$ is twice the z_2-even part of F and $\Omega=F-F_-$ is twice the z_2-odd part of F.

Proof. Assume that $\sigma \in R^n$ is fixed. Then define

$$\omega(\theta)=(\cos\theta,\sin\theta,0,0,\cdots,0)\in S^{n-1},$$
$$s(\theta)=(t(\sigma)\sin\theta,-t(\sigma)\cos\theta,\sigma_3,\cdots,\sigma_n)\in R^n.$$

It is then clear from Definition 2.1 (especially (2.4)) that

$$Z((\Gamma_\sigma,\sigma))(\theta)=s(\theta)+i\mu\omega(\theta)\in A^\mu,$$

thus proving (2.5).

It is also clear from the definition of Z that for fixed σ, $Z(\gamma,\sigma)$ is not continuous, much less holomorphic in γ. But if G is holomorphic on C^n and even in z_2, then $G(Z_\sigma)$ is indeed holomorphic. Here we have let $Z_\sigma(\gamma)=Z(\gamma,\sigma)$. In such a case the Cauchy integral formula holds:

$$G(Z_\sigma(\gamma_0))=\frac{i}{2\pi}\int_{\Gamma_\sigma}\frac{G(Z_\sigma(\gamma))}{\gamma-\gamma_0}d\gamma \tag{2.7}$$

for γ_0 interior to Γ_σ. Note that Γ_σ is oriented clockwise.

For an arbitrary entire F on C^n we let F_1 be the z_2-odd part of F: $F_1=\frac{1}{2}(F-F_-)$. It is easy to see that $F_1=z_2G_1$ for an entire G_1 which is z_2-even. Now let $F_2=G_2$ be the z_2-even part of F.

Observe that if $\gamma_0=\sigma_1$, then $Z_\sigma(\gamma_0)=\sigma$. Thus

$$F(\sigma)=F(Z_\sigma(\gamma_0))=(F_1+F_2)(Z_\sigma(\gamma_0))$$
$$=((z_2G_1)(Z_\sigma(\gamma_0))+G_2(Z_\sigma(\gamma_0))).$$

Since both G_1 and G_2 are z_2-even, (2.7) holds. Also $z_2(Z_\sigma(\gamma_0))=z_2(\sigma)=\sigma_2$. So

$$F(\sigma)=\frac{\sigma_2 i}{2\pi}\int_{\Gamma_\sigma}\frac{G_1(Z_\sigma(\gamma))}{\gamma-\gamma_0}d\gamma+\frac{i}{2\pi}\int_{\Gamma_\sigma}\frac{G_2(Z_\sigma(\gamma))}{\gamma-\gamma_0}d\gamma. \tag{2.8}$$

The integral formula (2.6) now appears after using the parametrization (2.2) of Γ_σ in the formula (2.7), thus completing the proof.

Remarks. Although Theorem 2.1 only states that the values of F on R^n can be recovered from its values on A^μ, actually more is true: A^μ is a set of uniqueness for holomorphic functions in C^n. This fact actually follows directly from Theorem 2.1 as was pointed out by one of the referees. For it is enough that an entire function vanishing on A^μ should vanish on C^n also. But Theorem 2.1 shows that an F vanishing on A^μ vanishes on R^n and hence also on the complexification C^n of R^n, by the Cauchy–Riemann equations.

The second referee produced a stronger result: Let Ω be an infinite subset of S^{n-1}. If F is holomorphic on C^n and $F(\sigma+i\mu\omega)=0$ whenever $\omega\in\Omega$ and $\sigma\in\omega^\perp$, then $F=0$ on C^n.

Proof. Let ω_0 be a limit point of Ω. Let $\omega_C^\perp=\{\sigma_1+i\sigma_2|\sigma_1,\sigma_2\in\omega^\perp\}$ be the complexification of $\omega^\perp$. As above, F vanishes on $\omega_C^\perp$ if it vanishes on $\omega^\perp$. By continuity then F vanishes on $\omega_{0,C}^\perp$. Choose coordinates so that $\omega_0=e_n$, the direction of the x_n-axis. Then $F(z+i\mu e_n)$ is divisible by z_n. But now $F(z+i\mu e_n)/z_n$ vanishes for $z=\sigma+i\mu(\omega-e_n)$ when $\omega\in\Omega$. Again by continuity $F(z+i\mu e_n)/z_n$ vanishes on $\omega_{0,C}^\perp$. Thus $F(z+i\mu e_n)/z_n$ is divisible by z_n. By continuing in this manner, $F(z+i\mu e_n)$ vanishes on $e_n^\perp$ and is divisible by z_n^k for any k. Hence $F=0$. *Note*: In this proof F denotes $F(z+i\mu\omega)$ for $z\in\omega_c^\perp$.

3. Fourier inversion of the attenuated Radon transform. F. Natterer [3] observed that the attenuated Radon transform determines the Fourier transform on A^{μ}. This is the analogue for attenuated transforms of the Helgason–Ludwig slice-projection theorem.

The *Fourier transform* of an L^1-function f on R^n is defined by

$$\hat{f}(\xi)=(2\pi)^{-n/2}\int_{R^n}\exp(-i\xi\cdot x)f(x)\,dx, \tag{3.1}$$

The next theorem is a precise statement of Natterer's result in n-dimensions.

THEOREM (Natterer [3]). *If* $\omega\in S^{n-1}$ *and* $\sigma\in\omega^{\perp}$, *then*

$$\hat{f}(\sigma+i\mu\omega)=(2\pi)^{-1/2}(Q^{\mu}f)\hat{\ }(\sigma,\omega). \tag{3.2}$$

Recall that the transform Q^{μ} was defined in the introduction ((1.2)).

Proof. (omitted, but follows directly by making the change of variable $x=\sigma+i\mu\omega$ in the definition of the Fourier transform).

Natterer used this result [3] to approximately invert the attenuated x-ray transform on R^2 by integral equation techniques.

An exact inversion of Q^{μ} is then accomplished by using the projections $Q^{\mu}f$ in (3.2) to find $\hat{f}$ on A^{μ}. Since f is a compactly supported L^2-function, $\hat{f}$ is entire so (2.6) can be used to find $\hat{f}$ on R^n. Finally f can be recovered by standard Fourier inversion.

Now assume that $\operatorname{supp} f\subset D\subset\operatorname{supp}\mu$, that D is convex and that $\mu|_D$ is constant, with α the constant value.

Let $T=\{(s,\omega)\,|\,s\in\omega^{\perp}\}$.

LEMMA. *There are bounded measurable functions* ε^{μ}: $T\to R$ *such that*

$$\varepsilon^{\mu}(s,\omega)\exp\left[-\int_t^{\infty}\mu(s+\tau\omega)\,d\tau\right]=\exp(\alpha t)$$

for all t *such that* $s+t\omega\in D$.

Proof. Let $L(s,\omega)$ denote the line through s in the direction ω. For every $(s,\omega)\in T$ such that $L(s,\omega)\cap D\neq\varnothing$, choose $d(s,\omega)$ such that $s+d(s,\omega)\omega\in D$. Then let $d=d(s,\omega)$ and define

$$\varepsilon^{\mu}(s,\omega)=\begin{cases}\exp\left[\int_d^{\infty}\mu(s+\tau\omega)\,d\tau+\alpha d(s,\omega)\right] & \text{if } L(s,\omega)\cap D\neq\varnothing,\\ 1 & \text{otherwise.}\end{cases}$$

Now

$$\varepsilon^{\mu}(s,\omega)\exp\left[-\int_t^{\infty}\mu(s+\tau\omega)\,d\tau\right]=\exp\left[\int_d^{t}\mu(s+\tau\omega)\,d\tau+\alpha d(s\omega)\right].$$

But if $s+\tau\omega\in D$, then convexity implies that the line segment $\{zs+\tau\omega\,|\,\tau$ between t and $d(s,\omega)\}$ is contained in D. Since $\mu=\alpha$ on D we get

$$\exp\left[\int_d^{\tau}\mu(s+\tau\omega)\,d\tau+\alpha d(s,\omega)\right]=\exp[\alpha\tau-\alpha d+\alpha d]=\exp(\alpha\tau).$$

LEMMA. $\varepsilon^{\mu}P^{\mu}f=Q^{\alpha}f$.

Proof. By the previous lemma

$$\varepsilon^{\mu}(s,\omega)P^{\mu}f(s,\omega)=\int_{-\infty}^{\infty}f(s+\tau\omega)\varepsilon^{\mu}(s,\omega)\exp\left[-\int_{t}^{\infty}\mu(s+\tau\omega)\,d\tau\right]dt$$

$$=\int_{-\infty}^{\infty}f(s+\tau\omega)\exp(\alpha\tau)\,dt=Q^{\alpha}f(s,\omega).$$

By combining this result with Natterer's theorem we get

THEOREM.

$$\hat{f}(\sigma+i\mu\alpha)=(2\pi)^{-1/2}(\varepsilon^{\mu}P^{\mu}f)^{\wedge}(\sigma,\omega)\quad \textit{if}\ \sigma\in\omega^{\perp}.$$

Since $P^{\mu}f$ is assumed to be known and since ε^{μ} can be calculated, f can be recovered by Fourier inversion.

Acknowledgments. I want to thank E. T. Quinto for carefully reading this paper and for offering several useful suggestions. I would also like to thank the referees for their suggestions and for their extensions of my results. In particular each referee gave an elegant proof of the fact that A^{μ} is a set of uniqueness for entire functions and one referee gave the generalization of Theorem 2.1 proved in the remark to that theorem.

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] G. T. GULLBERG AND T. F. BUDINGER, *The use of filtering methods to compensate for constant attenuation in single photon emission computed tomography*, IEEE Trans. Biomed. Engrg., BME-28 (1981), pp. 142–157.

[3] F. NATTERER, *On the inversion of the attenuated Radon transform*, Numer. Math., 32 (1979), pp. 431–438.

[4] E. T. QUINTO, *The invertibility of the rotation invariant Radon transforms*, J. Math. Anal. Appl., 91 (1983), pp. 510–522.

[5] O. J. TRETIAK AND C. METZ, *The exponential Radon transform*, SIAM J. Appl. Math., 39 (1980), pp. 341–354.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics

009

DECOMPOSITION OF HARDY FUNCTIONS INTO SQUARE INTEGRABLE WAVELETS OF CONSTANT SHAPE*

A. GROSSMANN† AND J. MORLET‡

Abstract. An arbitrary square integrable real-valued function (or, equivalently, the associated Hardy function) can be conveniently analyzed into a suitable family of square integrable wavelets of constant shape, (i.e. obtained by shifts and dilations from any one of them.) The resulting integral transform is isometric and self-reciprocal if the wavelets satisfy an "admissibility condition" given here. Explicit expressions are obtained in the case of a particular analyzing family that plays a role analogous to that of coherent states (Gabor wavelets) in the usual L_2-theory. They are written in terms of a modified Γ-function that is introduced and studied. From the point of view of group theory, this paper is concerned with square integrable coefficients of an irreducible representation of the nonunimodular $ax+b$-group.

1. Introduction.

1.1. It is well known that an arbitrary complex-valued square integrable function $\psi(t)$ admits a representation by Gaussians, shifted in direct and Fourier transformed space. If $g(t)=2^{-1/2}\pi^{-3/4}e^{-t^2/2}$ and t_0, ω_0 are arbitrary real, consider

$$g^{(t_0,\omega_0)}(t)=e^{-i\omega_0 t_0/2}e^{i\omega_0 t}g(t-t_0) \tag{1.1}$$

and form the inner product

$$\Psi(t_0,\omega_0)=\int \bar{g}^{(t_0,\omega_0)}(t)\psi(t)\,dt. \tag{1.2}$$

Then

$$\iint|\Psi(t_0,\omega_0)|^2 dt_0\,d\omega_0=\int|\psi(t)|^2 dt. \tag{1.3}$$

The function $\psi(t)$ can be recovered from the function $\Psi(t_0,\omega_0)$ through

$$\psi(t)=\iint g^{(t_0,\omega_0)}(t)\Psi(t_0,\omega_0)\,dt_0\,d\omega_0. \tag{1.4}$$

The above statements remain true if the Gaussian g is replaced by an arbitrary square integrable function. The advantages of the Gaussian are (i) maximal concentration in direct and Fourier transformed space and (ii) the possibility of a simple intrinsic characterization of the space of functions $\Psi(t_0,\omega_0)$.

This representation of functions has been used in quantum mechanics, quantum optics and signal theory. (See e.g. [1], [4], [5], [6].)

1.2. Consider now the case where the object of interest is not a complex-valued function $\psi(t)$, but a square integrable real-valued function $s(t)$, say the wiggle of a seismograph. It has been known for a long time that it is very useful to consider $s(t)$ as the real part of a complex-valued square integrable function $h(t)$ which has the special property that its Fourier transform vanishes on a half-line (say $\tilde{h}(\omega)=0$ for $\omega<0$). The

*Received by the editors September 21, 1982.

†Centre de Physique Théorique, Section II, Centre National de la Recherche Scientifique, Marseille, France.

‡Elf Aquitaine Company, O.R.I.C. Lab., 370 bis Av. Napoléon Bonaparte, 92500 Rueil-Malmaison, France.

space of such functions $h(t)$ is denoted by $\mathbf{H}^2$ and called the Hardy space on the line. It is a closed subspace of the space L_2 $(\mathbb{R},dt)$ of all square integrable functions. The functions s and h are in a natural one-to-one correspondence, and special properties of the function $h(t)$ (in particular its phase) make it a valuable tool.

1.3. This paper is concerned with the decomposition of functions $h\in\mathbf{H}^2$ into square integrable "elementary wavelets", and with the corresponding reconstruction problem. One can of course analyze the function $h(t)$ by applying to it the general results described in §1.1, applicable to any function in L_2. This is indeed what is done traditionally (see e.g. the famous paper [4]). It is however clear that, when we follow this procedure, we are not taking advantage of the special features of the function $h(t)$ which led us to introduce it in the first place; we are analyzing a function that belongs to the subspace $\mathbf{H}^2\subset L_2$ in terms of wavelets that do not belong to this subspace (the Fourier transform of a Gaussian does not vanish on a half-line). It will not help (at least in principle) to replace the Gaussian by an elementary wavelet that belongs to $\mathbf{H}^2$, since we have to consider all of its shifts in Fourier transformed space, and these are sure to bring it out of $\mathbf{H}^2$.

1.4. In several papers devoted to the study of seismic traces [7], [8], one of us has suggested analyzing them in terms of wavelets of *fixed shape*, and has produced strong numerical evidence for the soundness of such analysis. The aim of the present paper is to give mathematical underpinnings for this procedure, which also avoids the objections that were raised in §1.3. The main idea is to analyze functions in terms of wavelets obtained by shifts (only in direct space, not in Fourier transformed space) and *dilations* from a suitable basic wavelet.

1.5. The group G_2 of shifts and dilations (which is the only two-parameter Lie group and thus the "smallest" noncommutative Lie group), acts on $\mathbf{H}^2$ through a natural irreducible unitary representation $U(\gamma)$ $(\gamma\in G_2)$. If we fix a function $g\in\mathbf{H}^2$ ("the analyzing wavelet"), we obtain a correspondence between an arbitrary $h\in\mathbf{H}^2$, and the matrix element $m_h^{(g)}(\gamma)=(U(\gamma)g,h)$ considered as a function on the group G_2. The main question, both from a conceptual and practical point of view, is whether the correspondence $h\to m_h^{(g)}$ has a well-behaved inverse, allowing a "stable" reconstruction of h from $m_h^{(g)}$. Stated somewhat differently, the question is whether, for a suitable invariant measure $d\gamma$ on G_2, one has

$$\int|m_h^{(g)}(\gamma)|^2d\gamma=\int|h(t)|^2dt \tag{1.5}$$

in analogy to (1.3).

1.6. It turns out that the answer depends on the choice of the analyzing wavelet g. For (1.5) to hold, the wavelet g, in addition to being in $\mathbf{H}^2$, has to satisfy an "admissibility" condition.

The main general result, proved in §3, can be stated without reference to group theory:

Let $h(t)$ (the function to be analyzed) satisfy

$$\int|h(t)|^2dt<\infty \tag{i}$$

and

$$\text{(ii)} \qquad \tilde{h}(\omega)=0 \quad \text{for } \omega\leq 0.$$

(Conditions (i) and (ii) say that $h\in \mathbf{H}^2$.)

Let $g(t)$ (the analyzing wavelet) satisfy (i), (ii), and also the "admissibility condition"

(iii) $\int du\, e^u \int_0^\infty |\tilde{g}(\omega)\tilde{g}(e^u\omega)|^2 d\omega < \infty$. Associate to h the function $(\mathcal{C}h)(u,v)$ of two variables, defined by

$$(1.6) \qquad (\mathcal{C}h)(u,v)=\frac{1}{\sqrt{c_g}}e^{u/2}\int \bar{g}(e^u t-v)h(t)\,dt$$

where

$$(1.7) \qquad c_g=\frac{2\pi}{\|g\|^2}\int du\, e^u \int |\tilde{g}(\omega)\tilde{g}(e^u\omega)|^2 d\omega = 2\pi\int_0^\infty \frac{|\tilde{g}(\omega)|^2}{\omega}d\omega$$

and

$$\|g\|^2=\int |g(t)|^2 dt.$$

Then

(a)

$$(1.8) \qquad \iint |(\mathcal{C}h)(u,v)|^2 du\,dv = \int |h(t)|^2 dt$$

and

(b) $h(t)$ can be recovered from $(\mathcal{C}h)(u,v)$ through

$$h(t)=\frac{1}{\sqrt{c_g}}\iint e^{u/2}g(e^u t-v)(\mathcal{C}h)(u,v)\,du\,dv.$$

If g is not admissible, then $c_g=\infty$, and the transformation (1.6) is not defined.

The need for an admissibility condition may seem surprizing. It stems from the fact that G_2, in contrast to all other "everyday" groups, is nonunimodular (i.e. has no right-and-left-invariant measure) (compare [2], [3]).

Section 4 is devoted to a more detailed study of the transformation $\mathcal{C}$ in the special case where g is a "particularly good" wavelet which plays a role analogous to that of the Gaussian in the conventional theory. We find it convenient to introduce a special function $\Gamma_\alpha(z)$ which may be of independent interest. In §4 we gather the results necessary for an intrinsic characterization of the range of $\mathcal{C}$, which will be given in a forthcoming paper.

This paper can also be viewed as the description of a natural quantum-mechanical representation for particles that "only know how to move in one direction".

This interpretation and further developments will also be found in forthcoming papers.

2. Notation and preliminaries.

2.1. The inner product of square integrable functions is written as

$$(f,g)=\int \bar{f}(t)g(t)\,dt$$

where $\bar{f}$ is the complex conjugate of f.

The Fourier transform of $f(t)$ is

$$\tilde{f}(\omega)=(2\pi)^{-1/2}\int e^{-i\omega t}f(t)\,dt$$

inverted by

$$f(t)=(2\pi)^{-1/2}\int e^{i\omega t}\tilde{f}(\omega)\,d\omega.$$

We also write

$$\tilde{f}=\mathcal{F}f, \qquad f=\mathcal{F}^{-1}\tilde{f}.$$

The shift operator T^v is defined by

$$(T^vf)(t)=f(t-v) \qquad (v\in\mathbb{R}).$$

The corresponding multiplication operator is E^v;

$$(E^v\tilde{f})(\omega)=e^{iv\omega}\tilde{f}(\omega).$$

The dilation operator Z^u is defined by

$$(Z^uf)(t)=e^{-u/2}f(e^{-u}t).$$

The relations

$$T^vZ^u=Z^uT^{v/\exp u}, \qquad E^vZ^u=Z^uE^{v\exp u}, \qquad Z^uT^v=T^{v\exp u}Z^u, \qquad Z^uE^v=E^{v/\exp u}Z^u \tag{2.1}$$

will be basic for all that follows. They correspond to

$$(T^vZ^uf)(t)=e^{-u/2}f(e^{-ut}-e^{-u}v),$$
$$(Z^uT^vf)(t)=e^{-u/2}f(e^{-u}t-v).$$

The commutation properties with $\mathcal{F}$ are

$$\mathcal{F}T^v=E^{-v}\mathcal{F}, \tag{2.2}$$

$$\mathcal{F}Z^u=Z^{-u}\mathcal{F}. \tag{2.3}$$

We have $T^{v_1}T^{v_2}=T^{v_1+v_2}$, and $Z^{u_1}Z^{u_2}=Z^{u_1+u_2}$. The operators $\mathcal{F}$, T^v, E^v, Z^u are all unitary in $L_2(\mathbb{R})$.

2.2. We say that a function $h\in L_2(\mathbb{R},dt)$ belongs to the Hardy space $\mathbf{H}^2\subset L_2$ if $\tilde{h}(\omega)=0$ for $\omega<0$.

$\mathbf{H}^2$ is a closed subspace of L_2.

A real-valued function cannot belong to $\mathbf{H}^2$.

If $h\in\mathbf{H}^2$ then $\bar{h}$ and $\check{h}$ (where $\check{h}(t)=h(-t)$) are orthogonal to all of $\mathbf{H}^2$. However, $h^*\in\mathbf{H}^2$, where $h^*(t)=\bar{h}(-t)$.

The Hilbert transform in $L_2(\mathbb{R},dt)$ can be defined as $H=-\mathfrak{F}^{-1}\varepsilon\mathfrak{F}$, where ε is the operator of multiplication by $\operatorname{sgn}\omega$ (the sign of ω). We have $H^2=-1$; H is unitary, anti-Hermitian and real (commutes with complex conjugation). If $h\in\mathbf{H}^2$, then $h=iHh$, giving $\operatorname{Im}h=H\operatorname{Re}h$ and $\operatorname{Re}h=-H\operatorname{Im}h$.

If $s(t)$ is any real-valued, square integrable function then

$$h_s=s+iHs$$

belongs to $\mathbf{H}^2$. We have

$$(h_s,g)=2(s,g) \tag{2.4}$$

for every $g\in\mathbf{H}^2$. Also

$$\|h_s\|^2=2\|s\|^2 \tag{2.5}$$

where $\|f\|^2=(f,f)$

2.3. Shifts and dilations in $\mathbf{H}^2$. If $h\in\mathbf{H}^2$, then $T^v h\in\mathbf{H}^2$ and $Z^u h\in\mathbf{H}^2$ for all u,v.

For our purposes it is crucial to remark that the family T^v, Z^u acts *irreducibly* in $\mathbf{H}^2$. That is: If V is a closed subspace of $\mathbf{H}^2$, containing at least one nonzero vector; if V is stable under all Z^u, T^v (which means $T^u Z^v h\in V$ whenever $h\in V$), then V is all of $\mathbf{H}^2$. For a proof see e.g. [9].

Another important, if obvious, remark is that shifts and dilations are "real", in the sense that

$$\operatorname{Re}(T^v h)=T^v(\operatorname{Re}h) \tag{2.6}$$

and

$$\operatorname{Re}(Z^u h)=Z^u(\operatorname{Re}h). \tag{2.7}$$

3. The $\mathcal{C}$-transform: arbitrary admissible wavelet.

3.1. Admissible analyzing wavelet. We shall say that a function g, not identically zero, is an *admissible analyzing wavelet*, if

(i) g belongs to $\mathbf{H}^2$

and

(ii) g satisfies the condition

$$\iint|(Z^{-u}T^v g,g)|^2 du\,dv<\infty. \tag{3.1}$$

By (2.2) and (2.3), the condition (3.1) can also be written as

$$\iint|(Z^u E^{-v}\tilde{g},\tilde{g})|^2 du\,dv=2\pi\int du\, e^u\int|\tilde{g}(w)\tilde{g}(e^u\omega)|^2 d\omega<\infty \tag{3.2}$$

$$=2\pi\|g\|^2\int_0^\infty\frac{|\tilde{g}(\omega)|^2}{\omega}d\omega.$$

Examples. 1) Let $0<a<b<\infty$. Define $g(t)$ through its Fourier transform: $\tilde{g}(\omega)=1$ if $a<\omega<b$, and 0 otherwise. Then $g(t)$ is an admissible analyzing wavelet, as can be shown by a simple calculation.

2) Let $\alpha>0$. Define $g_\alpha(t)$ through its Fourier transform $\tilde{g}_\alpha(\omega)=\exp(-(\alpha/2)\ln^2\omega)=\omega^{-\alpha\ln\omega/2}$ for $\omega>0$, and $g_\alpha(\omega)=0$ for $\omega<0$. Then $g_\alpha(t)$ is an admissible analyzing wavelet which will be studied in §4.

Remarks. 1) There exist functions in $\mathbf{H}^2$ that are not admissible analyzing wavelets; this is the case e.g. if the Fourier transform of g is defined by

$$\tilde{g}(\omega)=\begin{cases}\omega^{-1/2+\varepsilon} & (0<\omega\leq 1)\\ \omega^{-1/2-\varepsilon} & (1<\omega<\infty)\end{cases}\quad(\varepsilon>0)$$

and $\tilde{g}(\omega)=0$ for $\omega\leq 0$.

2) A Gaussian cannot be an admissible analyzing wavelet since it does not belong to $\mathbf{H}^2$. However, if ω_0 is positive and sufficiently large, the function $e^{i\omega_0 t}\exp(-t^2/2)$ is very close to an admissible analyzing wavelet.

3) From (3.2) we see: If $\tilde{g}(\omega)$ is the Fourier transform of an admissible wavelet then, for any real-valued $\varphi(\omega)$, the function $e^{i\varphi(\omega)}\tilde{g}(\omega)$ is also the Fourier transform of an admissible wavelet.

3.2. The number c_g. If g is an admissible analyzing wavelet, we denote by c_g the number

$$c_g=\frac{1}{\|g\|^2}\iint|(Z^{-u}T^v g,g)|^2 du\,dv.$$

By (2.2), c_g can also be written as

$$c_g=\frac{1}{\|g\|^2}\iint|(Z^u E^{-v}\tilde{g},\tilde{g})|^2 du\,dv$$

which gives

$$c_g=\frac{2\pi}{\|g\|^2}\int_0^\infty d\omega\int du\, e^u|\tilde{g}(\omega)\tilde{g}(e^u\omega)|^2=2\pi\int_0^\infty\frac{|\tilde{g}(\omega)|^2}{\omega}d\omega. \tag{3.3}$$

3.3. The $\mathcal{C}$-transform. Let g be a fixed admissible analyzing wavelet. For arbitrary real u,v, define

$$g^{(u,v)}=Z^{-u}T^v g. \tag{3.4}$$

For every $h\in\mathbf{H}^2$, define the function $\mathcal{C}h$ of variables u,v by

$$(\mathcal{C}h)(u,v)=\frac{1}{\sqrt{c_g}}(g^{(u,v)},h) \tag{3.5}$$

i.e.

$$(\mathcal{C}h)(u,v)=\frac{1}{\sqrt{c_g}}e^{u/2}\int\bar{g}(e^u t-v)h(t)\,dt. \tag{3.6}$$

In words: $(\mathcal{C}h)(u,v)$ is obtained by "testing" the function h with the help of dilated and shifted analyzing wavelet. The dilation parameter is u, and the shift parameter is v. The result of testing is multiplied by a normalization factor which depends on the choice of the admissible analyzing wavelet.

We call $\mathcal{C}h$ the $\mathcal{C}$-transform of h (with respect to g). By (2.4) we have, with $s(t)=\operatorname{Re}h(t)$,

$$(\mathcal{C}h)(u,v)=\frac{2}{\sqrt{c_g}}e^{u/2}\int\bar{g}(e^u t-v)s(t)\,dt.$$

Alternative ways of writing $(\mathcal{C}h)(u,v)$ are, by (2.2), (2.3),

$$(3.7)\qquad (\mathcal{C}h)(u,v)=\frac{1}{\sqrt{c_g}}(Z^{-u}T^{v}g,h)=\frac{1}{\sqrt{c_g}}(Z^{u}E^{-v}\tilde{g},\tilde{h})$$

$$=\frac{1}{\sqrt{c_g}}e^{-u/2}\int_0^\infty \bar{\tilde{g}}(e^{-u}\omega)\tilde{h}(\omega)\exp[ive^{-u}\omega]\,d\omega.$$

From (3.7) one sees that

$$|(\mathcal{C}h)(u,v)|\le\frac{1}{\sqrt{c_g}}\|h\|\,\|g\|$$

for all u,v..

The correspondence $h\to\mathcal{C}h$ is linear.

3.4. Isometry of $\mathcal{C}$. We claim: *For every $h\in\mathbf{H}^2$, the function $(\mathcal{C}h)(u,v)$ is square integrable, and*

$$(3.8)\qquad \iint|(\mathcal{C}h)(u,v)|^2\,du\,dv=\|h\|^2.$$

Proof. (i) The equality (3.8) holds for $h=g$, since, by (3.5),

$$\iint|(\mathcal{C}g)(u,v)|^2\,du\,dv=\frac{1}{c_g}\iint|(Z^{-u}T^{v}g,g)|^2\,du\,dv=\frac{c_g}{c_g}\|g\|^2.$$

(ii) Equation (3.8) also holds for every h of the form $h=Z^{-u_0}T^{v_0}g$. Indeed,

$$(\mathcal{C}h)(u,v)=\frac{1}{\sqrt{c_g}}(Z^{-u}T^{v}g,Z^{-u_0}T^{v_0}g)=\frac{1}{\sqrt{c_g}}(T^{-v_0}Z^{u_0-u}T^{v}g,g)$$

$$=\frac{1}{\sqrt{c_g}}\left(Z^{u_0-u}T^{v-v_0\exp(u-u_0)}g,g\right)$$

$$=(\mathcal{C}g)(u-u_0,v-v_0e^{u-u_0}),$$

by (2.1), (3.5) and (3.7).

Now

$$\iint|(\mathcal{C}g)(u-u_0,v-v_0e^{u-u_0})|^2\,du\,dv=\iint|(\mathcal{C}g)(u',v-v_0e^{u'})|^2\,du'\,dv$$

$$=\iint|(\mathcal{C}g)(u',v')|^2\,du'\,dv'=\|g\|^2$$

with $u'=u-u_0$, $v'=v-v_0e^{u-u_0}$.

(iii) By standard arguments, (3.8) is extended to all finite linear combinations of vectors of the form $Z^{u}T^{v}g$. By irreducibility (§2.3.), these vectors are dense in $\mathbf{H}^2$, and so (3.8) is extended by continuity to all of $\mathbf{H}^2$. This completes the proof.

By the polarization identity, one has

$$(3.9)\qquad (\mathcal{C}h_1,\mathcal{C}h_2)_{L_2(\mathbf{R}^2,du\,dv)}=(h_1,h_2)$$

for all $h_1\in\mathbf{H}^2$, $h_2\in\mathbf{H}^2$.

3.5. Inversion of $\mathcal{C}$. We now sketch a verification of the fact that $\mathcal{C}$, considered as an integral transform, is self-reciprocal. In other words: If $\mathcal{C}h$ is given by (3.6), then h can be recovered from $\mathcal{C}h$ through the formula

$$h(t)=\frac{1}{\sqrt{c_g}}\iint e^{u/2}g(e^u t-v)(\mathcal{C}h)(u,v)\,du\,dv. \tag{3.10}$$

Since the integral (3.10) cannot converge for every h and every t, the formula (3.10) has a sense that is familiar from the L^2-theory of Fourier transforms or from [1].

In order to obtain (3.10) we write, with a slight stretching of notations, and using (2.4), (3.9),

$$h(t_0)=(\delta_{t_0},h)_{L_2(\mathbb{R},dt)}=\frac{1}{2}(\delta_{t_0}^{(+)},h)_{\mathbf{H}^2}=\frac{1}{2}(\mathcal{C}\delta_{t_0}^{(+)},\mathcal{C}h)_{L_2(\mathbb{R}^2,du\,dv)} \tag{3.11}$$

where $\delta_{t_0}(t)=\delta(t-t_0)$ (Dirac measure) and $\delta^{(+)}(t)=\delta(t-t_0)+(i/\pi)P/(t-t_0)$ (principal part). The function $(\mathcal{C}\delta_{t_0}^{(+)})(u,v)$ can be found by using (2.4):

$$\begin{aligned}(\mathcal{C}\delta_{t_0}^{(+)})(u,v)&=\frac{1}{\sqrt{c_g}}e^{u/2}\int\bar{g}(e^u t-v)\delta_{t_0}^{(+)}(t)\,dt\\&=2\frac{1}{\sqrt{c_g}}e^{u/2}\int\bar{g}(e^u t-v)\delta(t-t_0)\,dt=2\frac{1}{\sqrt{c_g}}e^{u/2}\bar{g}(e^u t_0-v).\end{aligned} \tag{3.12}$$

Inserting (3.12) into (3.11) gives (3.10).

Remark on redundancy. Equation (3.10) is a way of recovering the function $h(t)$ (and $s(t)=\operatorname{Re}h(t)$) from the function $(\mathcal{C}h)(u,v)$. The function $h(t)$ can also be recovered from values of $(\mathcal{C}h)(u,v)$ on suitable subsets of the plane, e.g. from the function

$$(\mathcal{C}h)(0,v)=\frac{1}{\sqrt{c_g}}\int_0^\infty\bar{\tilde{g}}(\omega)\tilde{h}(\omega)e^{iv\omega}\,d\omega.$$

We see that $\tilde{h}(\omega)$ can be obtained from $(\mathcal{C}h)(0,v)$ through Fourier transformation and division by $\bar{\tilde{g}}(\omega)$. The last step, however, corresponds—at best—to an unbounded operator. This makes the recovery of $h(t)$ from $(\mathcal{C}h)(0,v)$ an impractical proposition in general, and shows the advantage of working with the isometric transformation (3.10) or with suitable discrete approximations to it.

Covariance of $\mathcal{C}$. By the construction of $\mathcal{C}$, we have: If $h_1=Z^{u_1}h$, then

$$(\mathcal{C}h_1)(u,v)=(\mathcal{C}h)(u+u_1,v).$$

If $h_2=T^{v_2}h$, then

$$(\mathcal{C}h_2)(u,v)=(\mathcal{C}h)(u,v-v_2e^u).$$

3.6. Reproducing equation. The range of $\mathcal{C}$ is not all of $L_2(\mathbb{R}^2,du\,dv)$. In this section we derive a condition that has to be satisfied by all functions of the form $\mathcal{C}h$, with $h\in\mathbf{H}^2$. More specific results are given in §4, for a particular analyzing wavelet.

Define a kernel $G(u,v;u',v')$ by

$$\begin{aligned}G(u,v;u',v')&=(g^{(u,v)},g^{(u',v')})=(Z^{-u}T^v g,Z^{-u'}T^{v'}g)\\&=(Z^uE^{-v}\tilde{g},Z^{u'}E^{-v'}\tilde{g})=(Z^{u_1}E^{-v_1}\tilde{g},\tilde{g})\end{aligned} \tag{3.13}$$

with

$$u_1 = u - u' \quad \text{and} \quad v_1 = v - v'e^{u-u'}. \tag{3.14}$$

Then a function $f(u,v)$ that belongs to the range of $\mathcal{C}$ must satisfy

$$f(u,v) = \frac{1}{c_g} \iint G(u,v;u',v') f(u',v')\,du'\,dv'. \tag{3.15}$$

Indeed, by the definition and isometry of $\mathcal{C}$,

$$(\mathcal{C}h)(u,v) = \frac{1}{\sqrt{c_g}}\left(g^{(u,v)},h\right) = \frac{1}{\sqrt{c_g}}\left(\mathcal{C}g^{(u,v)},\mathcal{C}h\right)$$

$$= \frac{1}{\sqrt{c_g}} \iint \left(\overline{\mathcal{C}g^{(u,v)}}\right)(u',v')(\mathcal{C}h)(u',v')\,du'\,dv'.$$

Now

$$\left(\mathcal{C}g^{(u,v)}\right)(u',v') = \frac{1}{\sqrt{c_g}}\left(g^{(u',v')},g^{(u,v)}\right)$$

giving (3.15).

3.7. Cycle-octave representations.

THEOREM. *Let $s(t)$ be any real-valued square integrable function, and g an admissible analyzing wavelet. Associate to s the function $S(u,\tau)$ defined by*

$$S(u,\tau) = \frac{2}{\sqrt{c_g}} \int \bar{g}(e^u t - e^{-u}\tau) s(t)\,dt. \tag{3.16}$$

Then $s(t)$ can be recovered from $S(u,\tau)$ through

$$s(t) = \operatorname{Re} h(t) \tag{3.17}$$

where

$$h(t) = \frac{1}{\sqrt{c_g}} \iint g(e^u t - e^{-u}\tau) S(u,\tau)\,du\,d\tau \tag{3.18}$$

One has

$$\iint |S(u,\tau)|^2\,du\,d\tau = 2\int s(t)^2\,dt. \tag{3.19}$$

The function $h(t)$ defined by (3.18) *belongs to* $\mathbf{H}^2$.

An approximate discrete version of (3.16), (3.18), was discovered by one of us [7].

The statements of this theorem are an immediate consequence of the results proved so far, if we introduce the variable

$$\tau = e^u v.$$

3.8. Group-theoretical comments. The objects that we study, namely

$$\sqrt{c_g}(\mathcal{C}h)(u,-v)=(g,T^vZ^uh),$$

are matrix elements (coefficients,in another terminology) of the irreducible representation, in $\mathbf{H}^2$, of the two-parameter group of shifts and dilations. We have shown that these coefficients, considered as functions on the group, are square integrable with respect to the right Haar measure $dx_R=du\,dv$, if the vector g is suitably chosen.

If the standard theory of square integrable representations were applicable here (see e.g. [2]), it would follow that all coefficients of this representation are square integrable i.e. that all wavelets are admissible. However, the standard theory holds only for unimodular groups (i.e. groups possessing a right- and left-invariant Haar measure), while the group here is the prime example of nonunimodularity. (The left-invariant Haar measure is $dx_L=e^{-u}du\,dv$). Our results fit into the general theory of square integrable representations of nonunimodular groups, developed by Duflo and Moore [3].

4. The $\mathcal{C}$-transform: wavelet g_α.

4.1. The function $\tilde{g}_\alpha$. Among all admissible wavelets there is one that plays—in the $\mathbf{H}^2$-theory that we are concerned with—the same privileged role that the Gaussian plays in L^2-theory. The Fourier transform of this wavelet is just the image of a Gaussian under a natural map.

Let $\alpha>0$. Consider the function $\tilde{g}_\alpha(\omega)$ defined by

$$\tilde{g}_\alpha(\omega)=\begin{cases}\exp\left(-\frac{\alpha}{2}\ln^2\omega\right) & \text{for } \omega>0,\\ 0 & \text{for } \omega\leq 0.\end{cases} \tag{4.1}$$

Notice that $\tilde{g}_\alpha(\omega)$ is infinitely differentiable everywhere, in particular at $\omega=0$. Furthermore, $\tilde{g}_\alpha(\omega)$ tends to zero at infinity faster than any inverse polynomial.

We shall first verify that g_α is admissible, by using the criterion (3.2): one has

$$\int_0^\infty|\tilde{g}_\alpha(\omega)\tilde{g}_\alpha(e^u\omega)|^2d\omega=\sqrt{\frac{\pi}{2\alpha}}\,e^{1/8\alpha}e^{-\alpha u^2/2-u/2}$$

and consequently

$$\int du\,e^u\int_0^\infty|\tilde{g}_\alpha(\omega)\tilde{g}_\alpha(e^u\omega)|^2d\omega=\frac{\pi}{\alpha}e^{1/(4\alpha)}$$

which shows admissibility, and also gives

$$c_g=2\pi^{3/2}\alpha^{-1/2}. \tag{4.2}$$

The basic functional equation satisfied by $\tilde{g}_\alpha(\omega)$ is

$$(Z^u\tilde{g}_\alpha)(\omega)=e^{-\alpha u^2/2-u/2}\omega^{\alpha u}\tilde{g}_\alpha(\omega). \tag{4.3}$$

It corresponds to the equation relating a shifted Gaussian to the Gaussian multiplied by an exponential.

4.2. Condition satisfied by functions in the range of $\mathcal{C}$. We can use (4.3) to derive conditions satisfied by all functions in the range of $\mathcal{C}$. Writing

$$(\mathcal{C}h)(u,v)=\frac{1}{\sqrt{c_g}}\left(Z^uE^{-v}\tilde{g}_\alpha,\tilde{h}\right)=\left(E^{-v\exp(-u)}Z^u\tilde{g}_\alpha,\tilde{h}\right)\frac{1}{\sqrt{c_g}}$$

and evaluating $Z^u\tilde{g}$ by (4.3), we obtain

$$(\mathcal{C}h)(u,v)=\frac{1}{\sqrt{c_g}}e^{-\alpha u^2/2-u/2}\int_0^\infty \exp(ive^{-u}\omega)\,\omega^{\alpha u}\bar{\tilde{g}}(\omega)\tilde{h}(\omega)\,d\omega.$$

Introducing variables

$$z=\alpha u-1 \quad\text{and}\quad q=ve^{-u}$$

and writing

$$\Psi(z,q)=(\mathcal{C}h)(u,v),$$

we see that

$$\Psi(z,q)=\frac{1}{\sqrt{c_g}}e^{-(z^2-z)/2\alpha}\int_0^\infty e^{iq\omega}\omega^{z-1}\bar{\tilde{g}}_\alpha(\omega)\tilde{h}(\omega)\,d\omega. \tag{4.4}$$

It follows from (4.4) that $\Psi(z,q)$ satisfies

$$\frac{\partial\Psi}{\partial q}=ie^{z/\alpha}\Psi(z+1,q),$$

and, more generally

$$\Psi(z+n,q)=(-i)^ne^{-(nz+(n-1)n)/2\alpha}\frac{\partial^n\Psi}{\partial q^n}.$$

4.3. The function $\Gamma_\alpha(z)$. The wavelet $g_\alpha(t)$ is given by the inverse Fourier transform

$$g_\alpha(t)=(2\pi)^{-1/2}\int_0^\infty e^{i\omega t}\tilde{g}_\alpha(\omega)\,d\omega,$$

which does not seem expressible in closed form through special functions known to us. In order to evaluate it and related quantities we have found it convenient to introduce the function $\Gamma_\alpha(z)$ which will now be discussed and which, we believe, is also intrinsically interesting.

If $\alpha>0$ and if $z=x+iy$ is arbitrary complex, define $\Gamma_\alpha(z)$ by

$$\Gamma_\alpha(z)=\int_0^\infty \omega^{z-1}e^{-\omega}\exp\left(-\frac{\alpha}{2}\ln^2\omega\right)d\omega. \tag{4.5}$$

This definition is modeled on the definition

$$\Gamma(z)=\int_0^\infty \omega^{z-1}e^{-\omega}\,d\omega \qquad (\operatorname{Re} z>0)$$

of Euler's gamma function. Because of the factor $\exp(-\alpha\ln^2\omega/2)$, the function $\Gamma_\alpha(z)$ is entire analytic in z, in contrast to $\Gamma(z)$. If the factor $\exp(-\alpha\ln^2\omega/2)$ were replaced by a step function, the resulting integral would be an incomplete gamma function.

The substitution $\omega = e^s$ brings $\Gamma_\alpha(z)$ to the form

$$(4.6)\qquad \Gamma_\alpha(z) = \int_{-\infty}^{\infty} e^{zs} \exp\left(-e^s - \frac{\alpha}{2} s^2\right) ds.$$

We may think of $\Gamma_\alpha(z)$ as being a hybrid between a Gaussian and the Γ-function. This is made precise e.g. through the following statement:

If $\operatorname{Re} z > 0$, then

$$(4.7)\qquad \Gamma_\alpha(z) = \frac{1}{\sqrt{2\pi\alpha}} \int_{-\infty}^{\infty} e^{-u^2/2\alpha} \Gamma(z - iu)\, du.$$

The function $\Gamma_\alpha(z)$ satisfies a functional equation that goes over into the classical $\Gamma(z) = (z-1)\Gamma(z-1)$ in the limit $\alpha \to 0$. Denote by $\Gamma_\alpha^{(n)}$ the nth derivative of Γ_α with respect to z. We have then

$$(4.8)\qquad \Gamma_\alpha(z) = (z-1)\Gamma_\alpha(z-1) - \alpha\Gamma_\alpha^{(1)}(z-1)$$

and, more generally,

(4.9)

$$\Gamma_\alpha^{(n)}(z) = (z-1)\Gamma_\alpha^{(n)}(z-1) - \alpha\Gamma_\alpha^{(n+1)}(z-1) + n\Gamma_\alpha^{(n-1)}(z-1) \qquad (n = 1, 2, \cdots).$$

The function $\Gamma_\alpha(z)$ has asymptotic expansions for large $|z|$. For instance, we shall write the analogue of the formula $\Gamma(z) \simeq (2\pi)^{1/2} e^{-z + (z - 1/2)\ln z}(1 + 1/12z)$:

Denote by z_1 the solution of

$$z_1 = z - \alpha \ln z_1,$$

which is positive when z is large and positive. Then

$$(4.10)\qquad \Gamma_\alpha(z) \simeq \sqrt{\frac{2\pi}{z_1 + \alpha}}\, e^{-z_1 + (z^2 - z_1^2)/2\alpha} \left[1 - \frac{z_1}{8(z_1 + \alpha)^2} + \frac{5z_1}{24(z_1 + \alpha)^3}\right].$$

The expression (4.10) is numerically quite accurate even for small values of $|z|$.

If for fixed y, x is let go to $-\infty$, one has

$$\Gamma_\alpha(x + iy) \sim \sqrt{\frac{2\pi}{\alpha}}\, e^{(x+iy)^2/2\alpha}.$$

We consider next the variation of $\Gamma_\alpha(z)$ with α. From (4.6) we see that $\Gamma_\alpha(z)$ satisfies

$$(4.11)\qquad \frac{\partial \Gamma_\alpha}{\partial \alpha} = -\frac{1}{2} \frac{\partial^2 \Gamma_\alpha}{\partial z^2}.$$

Notice the minus sign in (4.11). We obtain the usual heat equation if we consider $\Gamma_\alpha(x + iy)$ as function of y for fixed x:

$$(4.12)\qquad \frac{\partial \Gamma_\alpha(x+iy)}{\partial \alpha} = \frac{1}{2} \frac{\partial^2 \Gamma_\alpha(x+iy)}{\partial y^2}.$$

The function $\Gamma_\alpha(x + iy)$ is bounded by its values on the real axis: $|\Gamma_\alpha(x + iy)| < \Gamma_\alpha(x)$. Along parallels to the imaginary axis, it decreases faster than any inverse polynomial. Considered as a function of α, $\Gamma_\alpha(x)$ is monotonically decreasing. On the positive real axis, $\Gamma_\alpha(x)$ is bounded by Euler's Γ-function to which it tends as $\alpha\to 0$. For every x,y, we have $|\Gamma_\alpha(x+iy)|<\sqrt{2\pi/\alpha}\,e^{x^2/2\alpha}$. If we let α tend to $+\infty$ while keeping z fixed, then $\Gamma_\alpha(z)$ behaves as $\sqrt{2\pi/\alpha}\,e^{\alpha z^2/2}$.

Details and further results will be given elsewhere.

4.4. A class of integrals. With the help of the function $\Gamma_\alpha(z)$ one can evaluate integrals of the form

$$h_{\alpha,\beta}(q)=\int_0^\infty \omega^{\beta-1}\exp\left(-\frac{\alpha}{2}\ln^2\omega\right)e^{i\omega q}d\omega \tag{4.13}$$

which will be needed below. Here β is complex and arbitrary, $\alpha>0$, $q\neq 0$, and $\operatorname{Im}q\geq 0$.

It is convenient to introduce the variable

$$\kappa=\ln\left(\frac{1}{i}q\right)=\ln|q|+i\arg q-i\frac{\pi}{2}. \tag{4.14}$$

If q is real, then $\operatorname{Im}\kappa=-(\pi/2)\operatorname{sgn}q$.

The integral (4.13) is first transformed into

$$h_{\alpha,\beta}(q)=\int_{-\infty}^\infty \exp\left(-e^{\kappa+s}-\frac{\alpha}{2}s^2+\beta s\right)ds$$

by the substitution $\omega=e^s$. Then the substitution $s'=s+\kappa$ and a shift of the path of integration back to the real axis bring it to the form

$$h_{\alpha,\beta}(q)=\exp\left(-\frac{\alpha}{2}\kappa^2-\beta\kappa\right)\int \exp\left[-e^s+(\alpha\kappa+\beta)s-\frac{\alpha}{2}s^2\right]ds,$$

giving the result

$$h_{\alpha,\beta}(q)=\exp\left(-\frac{\alpha}{2}\kappa^2-\beta\kappa\right)\Gamma_\alpha(\alpha\kappa+\beta). \tag{4.15}$$

Remarks. The value of the integral (4.13) for $q=0$ can be computed directly. It is

$$h_{\alpha,\beta}(0)=\sqrt{\frac{2\pi}{\alpha}}\exp\left(\frac{\beta^2}{2\alpha}\right).$$

The function $h_{\alpha,\beta}(q)$ defined by (4.15) and (4.16) is infinitely differentiable on the real axis, and it decreases at infinity faster than any inverse polynomial. Furthermore, $h_{\alpha,\beta}$ belongs to $\mathbf{H}^2$.

As a function of $\beta, h_{\alpha,\beta}$ is entire analytic, and square integrable on every parallel to the imaginary axis. We have, from (4.13)

$$\frac{d^n}{dq^n}h_{\alpha,\beta}=i^n h_{\alpha,\beta+n}.$$

4.5. Explicit expressions. We can now write down expressions for various quantities of interest:

1) The wavelet $g_\alpha(t)$ is given by

$$g_\alpha(t)=(2\pi)^{-1/2}e^{-\alpha\theta^2/2-\theta}\Gamma_\alpha(\alpha\theta+1)\qquad\left(\theta=\ln\left(\frac{1}{i}t\right)\right).$$

2) The $\mathcal{C}$-transform of g_α is given by

$$(\mathcal{C}g_\alpha)(u,v)=2^{-1/2}\pi^{-3/4}\alpha^{1/4}e^{-u^2\alpha/4}e^{-\alpha\eta^2-\eta}\Gamma_{2\alpha}(2\alpha\eta+1)$$

where $\eta=\ln((1/i)ve^{-u/2})$.

3) The kernel of the integral equation satisfied by the functions in the range of $\mathcal{C}$ is (compare (3.15))

$$G(u,v;u',v')=e^{-\alpha u_1^2/4-\alpha\eta_1^2-\eta_1}\Gamma_{2\alpha}(2\alpha\eta_1+1)$$

where $u_1=u-u'$ and

$$\eta_1=\ln\left(\frac{1}{i}ve^{(u-u')/2}-\frac{1}{i}v'e^{-(u-u')/2}\right).$$

Acknowledgments. One of us (A. G.) would like to thank H. Bacry for stimulating conversations.

REFERENCES

[1] V. Bargmann, *On a Hilbert space of analytic functions and an associated integral transform, Part* I, Comm. Pure Appl. Math, 14 (1961), pp. 187–214; *Part* II, ibid. 20 (1967), pp. 1–101.

[2] J. Dixmier, *Les C*-algèbres et leurs représentations*, Gauthier-Villars, Paris, 1969.

[3] M. Duflo and C. C. Moore, *On the regular representation of a nonunimodular locally compact group*, J. Funct. Anal., 21 (1976), pp. 209–243.

[4] D. Gabor, *Theory of Communication*, J. Inst. Electr. Engin. (London) 93 (III), 1946, pp. 429–457.

[5] C. W. Helstrom, *An expansion of a signal in Gaussian elementary signals*, IEEE Trans. Infor. Theory, IT 12 (1966), pp. 81–82.

[6] J. R. Klauder and E. C. Sudarshan, *Fundamentals of Quantum Optics*, Benjamin, New York, 1968.

[7] J. Morlet, G. Arens, E. Fourgeau and D. Giard, *Wave propagation and sampling theory, Part* II, Geophys. 47 (1982), pp. 222–236.

[8] J. Morlet, *Sampling theory and wave propagation*, Proc. 51st Annual International Meeting of the Society of Exploration Geophysicists, Los Angeles, 1981.

[9] N. Ya. Vilenkin, *Special Functions and the Theory of Group Representations*, American Mathematical Society, Providence, RI, 1968.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics

010

STARLIKE AND PRESTARLIKE HYPERGEOMETRIC FUNCTIONS*

B. C. CARLSON† AND DOROTHY B. SHAFFER‡

Abstract. A linear operator is defined which acts on an analytic function in the open unit disk by forming its Hadamard product with an incomplete beta function. The operator is shown to be convenient in discussing starlike, convex, and prestarlike functions. It is applied to the study of certain classes of hypergeometric functions which constitute dense subsets in the classes of starlike functions of order α, convex functions of order α, and prestarlike functions of order α. Integral representations of the functions in these classes are obtained from the integral representation of the starlike functions of order α.

1. Introduction. Connections between the theory of univalent functions and the theory of special functions have not received much attention. Univalence has been investigated for a few hypergeometric series, in particular the Gauss hypergeometric function [7] [S6] [S7], the Bessel function $z^{1-\nu}J_\nu(z)$ [S5] [S2] [S8], the error function $\mathrm{Erf}(z)$ [S3], and $\exp(z^2)\mathrm{Erf}(z)$ [S4]. See also [1, pp. 95–96] and [6]. At least one family of univalent functions, a class of Schwarz–Christoffel maps, is known to be connected with multivariate hypergeometric functions [4] [2, §8.2]. However, there are more extensive connections that have not previously been explored, much less exploited. The purpose of this paper is to uncover some of these connections in the hope that they will prove useful in the theory of univalent functions. The work reported here was initially stimulated by observing that an integral occurring in a paper on univalent functions by R. R. Hall [5] is an integral representation of a multivariate hypergeometric function.

In §2 we introduce a convolution operator that allows very concise proofs of the results in this paper and should be convenient in many other contexts related to univalent functions. At the beginning of §3 we summarize some properties of a hypergeometric function of k variables called the R-function [2]. It is a symmetric and homogeneous variant of Gauss's hypergeometric function if $k=2$, Appell's F_1 if $k=3$, and Lauricella's F_D for general k. In terms of the R-function we define classes of analytic functions on the open unit disk and show that certain of these classes are dense in either the starlike or prestarlike functions of order α. Some classes are shown to consist of convex functions, and the question of what other classes are convex is suggested as an open problem. Section 4 uses the F_D notation in dealing with convex functions of order α. In §5 we give integral representations of functions in the closure of the hypergeometric classes and hence, in particular, representations of prestarlike or convex functions of order α.

2. Convolution with an incomplete beta function. Let A be the class of analytic functions $f(z)$ on the open unit disk, normalized by $f(0)=0$ and $f'(0)=1$. The class A is

*Received by the editors March 10, 1982, and in final revised form August 11, 1983. This work was begun at the University of Maryland, where the first author was a visitor at the Institute for Physical Science and Technology and the second author was a visitor in the Department of Mathematics.

†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.

‡Department of Mathematics, Fairfield University, Fairfield, Connecticut 06430.

closed under the Hadamard product or convolution:

$$f*g(z)=\sum_{n=0}^{\infty}\alpha_n\beta_n z^{n+1},\quad \text{where } f(z)=\sum_{n=0}^{\infty}\alpha_n z^{n+1},\quad g(z)=\sum_{n=0}^{\infty}\beta_n z^{n+1}.$$

In particular we consider convolution with the function $\varphi(a,c)$ defined by

$$\varphi(a,c;z)=\sum_{n=0}^{\infty}\frac{(a)_n}{(c)_n}z^{n+1},\qquad |z|<1,\quad c\neq 0,-1,-2,\cdots, \tag{2.1}$$

where $(a)_n=\Gamma(a+n)/\Gamma(a)$, i.e., $(a)_0=1$, $(a)_n=a(a+1)\cdots(a+n-1)$, $n\geq 1$. The function $\varphi(a,c)$ is an incomplete beta function, related to the Gauss hypergeometric function by $\varphi(a,c;\ z)=z\,{}_2F_1(1,a;\ c;\ z)$. It has an analytic continuation to the z-plane cut along the positive real line from 1 to ∞. Note that $\varphi(a,1;\ z)=z/(1-z)^a$ and $\varphi(2,1)$ is the Koebe function.

We define a linear operator on A by

$$L(a,c)f=\varphi(a,c)*f,\qquad f\in A. \tag{2.2}$$

If $a=0,-1,-2,\cdots$, then $L(a,c)f$ is a polynomial. If $a\neq 0,-1,-2,\cdots$, application of the root test shows that the infinite series for $L(a,c)f$ has the same radius of convergence as that for f because $\lim_{n\to\infty}|(a)_n/(c)_n|^{1/n}=1$. Hence $L(a,c)$ maps A into itself. The Ruscheweyh derivatives [9, (2.2)] of f are $L(n+1,1)f$, $n=0,1,2,\cdots$.

If $c>a>0$, $L(a,c)$ has the integral representation

$$L(a,c)f(z)=\int_0^1 u^{-1}f(uz)\,d\mu_{(a,c-a)}(u), \tag{2.3}$$

where μ is a beta distribution:

$$d\mu_{(a,c-a)}(u)=\frac{u^{a-1}(1-u)^{c-a-1}}{B(a,c-a)}du.$$

Equation (2.3) is readily verified by expanding f in a power series. More generally, if $c\neq 0,-1,-2,\cdots, L(a,c)$ has the integral representation

$$L(a,c)f(z)=\frac{1}{2\pi i}\int_\gamma u^{-1}f(u)\varphi(a,c;\ z/u)\,du, \tag{2.4}$$

where γ is a simple closed contour contained in the open unit disk and encircling the line segment $[0,z]$ in the positive direction. Let K be the closed disk with center 0 and radius $r<1$. Choose γ to be the circle with center 0 and radius $(1+r)/2$, and note that $|\varphi(a,c;\ z/u)|$ is bounded for $z\in K$ and $u\in\gamma$. If a sequence $\{f_n\}_{n=0}^{\infty}, f_n\in A$, converges to f uniformly on compact subsets of the open unit disk, and in particular on γ, then $L(a,c)f_n\to L(a,c)f$ uniformly on K by (2.4). Hence $L(a,c)$ is continuous on A in the topology of uniform convergence on compact subsets.

If $a\neq 0,-1,-2,\cdots$, then $L(a,c)$ has a continuous inverse $L(c,a)$ and is a 1-1 mapping of A onto itself. Clearly $L(a,a)$ is the unit operator and

$$L(a,c)=L(a,b)L(b,c)=L(b,c)L(a,b),\qquad b,c\neq 0,-1,-2,\cdots. \tag{2.5}$$

The convolution operator provides a convenient representation of differentiation and integration: if $g(z)=zf'(z)$, then $g=L(2,1)f$ and $f=L(1,2)g$.

The class of normalized starlike functions of order $\alpha \leq 1$ is

$$S^*(\alpha)=\{f\in A\colon \operatorname{Re}[zf'(z)/f(z)]\geq\alpha,\ |z|<1\}, \tag{2.6}$$

and the class $K(\alpha)$ of normalized convex functions of order α is

$$K(\alpha)=L(1,2)S^*(\alpha). \tag{2.7}$$

The class $K(0)$ is denoted by K. Two theorems due to Suffridge [12] take the following forms:

THEOREM A0. *If $\alpha\leq 1$ and $f,g\in S^*(\alpha)$, then $f*g\in L(2-2\alpha,1)S^*(\alpha)$.*

THEOREM B. *If $\alpha\leq\beta\leq 1$ and $\alpha<1$, then $L(2-2\beta,2-2\alpha)S^*(\alpha)\subset S^*(\beta)\subset S^*(\alpha)$.*

We define the convolution $M*N$ of two classes M and N to be the class of all convolutions $f*g$ where $f\in M$ and $g\in N$. Then Theorem A0 states that

$$S^*(\alpha)*S^*(\alpha)\subset L(2-2\alpha,1)S^*(\alpha).$$

The inclusion is actually an equality, for the right side is the class of all convolutions $\varphi(2-2\alpha,1)*f$ where $f\in S^*(\alpha)$. Since $\varphi(2-2\alpha,1)$ also is in $S^*(\alpha)$, each such convolution is in $S^*(\alpha)*S^*(\alpha)$. Thus we may replace Theorem A0 by Theorem A:

THEOREM A. $S^*(\alpha)*S^*(\alpha)=L(2-2\alpha,1)S^*(\alpha)$, $\alpha\leq 1$.

This result is simpler in terms of the class $Q(\alpha)$ of normalized prestarlike functions of order α introduced by Ruscheweyh [10]:

$$Q(\alpha)=L(1,2-2\alpha)S^*(\alpha),\qquad \alpha<1, \tag{2.8}$$
$$Q(1)=\left\{f\in A\colon \operatorname{Re}[f(z)/z]>\frac{1}{2},\ |z|<1\right\}.$$

Note that $Q(0)=K$ and $Q(\frac{1}{2})=S^*(\frac{1}{2})$. Ruscheweyh [10] proves $Q(\alpha)*Q(\alpha)\subset Q(\alpha)$, $\alpha\leq 1$, and the inclusion is actually an equality because $Q(\alpha)$ contains the identity $\varphi(1,1)$ for convolution.

THEOREM A1. $Q(\alpha)*Q(\alpha)=Q(\alpha)$, $\alpha\leq 1$.

The case $\alpha<1$ follows from Theorem A by operating with $L^2(1,2-2\alpha)$, but results for $Q(1)$ go beyond Theorems A and B.

By substituting $S^*(\alpha)=L(2-2\alpha,1)Q(\alpha)$, $\alpha\leq 1$, in Theorem B, we obtain

$$L(2-2\beta,1)Q(\alpha)\subset L(2-2\beta,1)Q(\beta)\subset L(2-2\alpha,1)Q(\alpha),\qquad \alpha\leq\beta\leq 1,$$

the case $\alpha=\beta=1$ being trivial. The first inclusion yields $Q(\alpha)\subset Q(\beta)$, $\alpha\leq\beta<1$, but Ruscheweyh [10] proves this for $\alpha\leq\beta\leq 1$. The second inclusion yields $L(2-2\beta, 2-2\alpha)Q(\beta)\subset Q(\alpha)$ if $\alpha\leq\beta\leq 1$ and $\alpha<1$.

THEOREM B1. *If $\alpha\leq\beta\leq 1$ and $\alpha<1$, then $L(2-2\beta,2-2\alpha)Q(\beta)\subset Q(\alpha)\subset Q(\beta)$.*

3. Classes of multivariate hypergeometric functions.

Let $u=(u_1,\cdots,u_k)$ be a k-tuple of nonnegative weights with $\sum_{i=1}^k u_i=1$. The set E of all such k-tuples is the standard simplex in $\mathbb{R}^{k-1}$. Let $x=(x_1,\cdots,x_k)$ be a fixed k-tuple of complex numbers with $\operatorname{Re} x_i>0$, $1\leq i\leq k$. Then the set of all convex combinations $u\cdot x=\sum_{i=1}^k u_i x_i$ forms a polygon (including its interior) in the open right half-plane with one of the x's at each vertex. (Some of the x's may lie in the interior of the polygon.) If t is a real number and $b=(b_1,\cdots,b_k)$ is a k-tuple of positive parameters, $k\geq 2$, we define

$$R_t(b,x)=\int(u\cdot x)^t d\mu_b(u), \tag{3.1}$$

where the integration is over E and the probability measure μ_b, $\mu_b(E)=1$, is called a Dirichlet measure or multivariate beta distribution,

$$d\mu_b(u)=\frac{\Gamma(b_1+\cdots+b_k)}{\Gamma(b_1)\cdots\Gamma(b_k)}\prod_{i=1}^{k}u_i^{b_i-1}du_1\cdots du_{k-1}. \tag{3.2}$$

Thus R_t is a weighted integral average of z^t over a polygon in the right half-plane. If $k=1$, we define $R_t(b,x)=x_1^t$.

A binomial expansion of the integrand yields the series representation

$$R_{-a}(b,1-x)=\sum_{n=0}^{\infty}\frac{(a)_n}{n!}R_n(b,x),\qquad |x_i|<1,\quad 1\le i\le k, \tag{3.3}$$

where $1-x=(1-x_1,\cdots,1-x_k)$. The function R_n, $n=0,1,2,\cdots$, is a homogeneous polynomial with coefficients that can be determined from (3.1). The explicit formula [2, (6.2-1)] for R_n implies the generating relation

$$\prod_{i=1}^{k}(1-sx_i)^{-b_i}=\sum_{n=0}^{\infty}s^n\frac{(c)_n}{n!}R_n(b,x),\qquad |sx_i|<1,\quad 1\le i\le k, \tag{3.4}$$

where $c=\sum_{i=1}^{k}b_i$. Comparison of (3.3) and (3.4) yields

$$R_{-c}(b,x)=\prod_{i=1}^{k}x_i^{-b_i},\qquad c=\sum_{i=1}^{k}b_i. \tag{3.5}$$

From (3.3) it can be shown [2, §6.3] that $R_t(b,x)$ has an analytic continuation to all complex values of $b_1,\cdots,b_k$ such that $\sum_{i=1}^{k}b_i\neq 0,-1,-2,\cdots$. In particular, if $b_i=0$, then b_i and x_i can simply be omitted [2, (6.3-3)], as in (3.5) for example. In [2, Thm. 5.9-2] it is shown that $R_t(b,x)$ is analytic in $x_1,\cdots,x_k$ if $\operatorname{Re}x_i>0$, $1\le i\le k$.

For every $a\ge 0$ and $c>0$ we define a class of functions analytic on the open unit disk in the z-plane:

$$\Sigma(a,c)=\Big\{zR_{-a}(cw_1,\cdots,cw_k;\ 1-z\zeta_1,\cdots,1-z\zeta_k):\ w_i\ge 0,\ \sum_{i=1}^{k}w_i=1,\\ |\zeta_i|\le 1,\ 1\le i\le k,\ k=1,2,3,\cdots\Big\}. \tag{3.6}$$

Since $R_0=1$, $\Sigma(0,c)=\{z\}$. The class $\Sigma_1(a,c)$ differs from $\Sigma(a,c)$ only by the restriction that $|\zeta_i|=1$, $1\le i\le k$. Thus $\Sigma_1(a,c)\subset\Sigma(a,c)\subset A$. It follows from (3.5) for each $c>0$ that

$$\Sigma(c,c)=\Big\{z\prod_{i=1}^{k}(1-z\zeta_i)^{-cw_i}:\ w_i\ge 0,\ \sum_{i=1}^{k}w_i=1,\ |\zeta_i|\le 1,\ 1\le i\le k,\ k=1,2,3,\cdots\Big\}. \tag{3.7}$$

It is easy to verify from (2.6) that

$$\Sigma(c,c)\subset S^*(1-\tfrac{1}{2}c),\qquad c>0. \tag{3.8}$$

Equation (3.3) implies

$$(3.9)\qquad zR_{-a}(cw_1,\cdots,cw_k;\,1-z\zeta_1,\cdots,1-z\zeta_k)\\ =\sum_{n=0}^{\infty}\frac{(a)_n}{n!}R_n(cw_1,\cdots,cw_k;\,\zeta_1,\cdots,\zeta_k)z^{n+1},$$

and hence, if $a\geq 0$ and α, $c>0$,

$$(3.10)\qquad \Sigma(a,c)=L(a,\alpha)\Sigma(\alpha,c)\quad\text{and}\quad \Sigma_1(a,c)=L(a,\alpha)\Sigma_1(\alpha,c).$$

THEOREM 1. *If $c\geq a\geq 0$ and $c>0$, then $\Sigma(a,c)\subset S^*(1-\frac{1}{2}a)$.*

Proof. By (3.10) and (3.8), $\Sigma(a,c)=L(a,c)\Sigma(c,c)\subset L(a,c)S^*(1-\frac{1}{2}c)$. Use of Theorem B completes the proof.

COROLLARY 1. *If $c\geq 2$ then $\Sigma(1,c)\subset K$.*

Proof. By (3.10) and Theorem 1, $\Sigma(1,c)=L(1,2)\Sigma(2,c)\subset L(1,2)S^*(0)=K$.

If a, $\alpha>0$, then $L(a,\alpha)$ is a continuous operator with a continuous inverse. Hence (3.10) implies, if a, α, $c>0$,

$$(3.11)\qquad \overline{\Sigma}(a,c)=L(a,\alpha)\overline{\Sigma}(\alpha,c)\quad\text{and}\quad \overline{\Sigma}_1(a,c)=L(a,\alpha)\overline{\Sigma}_1(\alpha,c),$$

where an overbar signifies closure in the topology of uniform convergence on compact sets. Also, (3.11) holds trivially if $a=0$.

THEOREM 2. *If $a\geq 0$ and $c>0$, then $\overline{\Sigma}_1(a,c)=\overline{\Sigma}(a,c)=L(a,c)S^*(1-\frac{1}{2}c)$.*

Proof. By (3.8), $\Sigma_1(c,c)\subset\Sigma(c,c)\subset S^*(1-\frac{1}{2}c)$. It is well known [6, (1.2)] that $\overline{\Sigma}_1(c,c)=S^*(1-\frac{1}{2}c)$. Therefore,

$$(3.12)\qquad \overline{\Sigma}_1(c,c)=\overline{\Sigma}(c,c)=S^*(1-\tfrac{1}{2}c).$$

Application of $L(a,c)$ completes the proof by (3.11).

Since $R_{-a}(cw,x)$ tends to $\Sigma w_i x_i^{-a}$ as $c\to 0$ [2, Ex. (6.3-4)] and to $(\Sigma w_i x_i)^{-a}$ as $c\to\infty$, it is natural to define, for every $a\geq 0$,

$$(3.13)\\ \Sigma(a,0)=\left\{z\Sigma_{i=1}^{k}w_i(1-z\zeta_i)^{-a}\colon\, w_i\geq 0,\ \Sigma_{i=1}^{k}w_i=1,\ |\zeta_i|\leq 1,\ 1\leq i\leq k,\ k=1,2,\cdots\right\},\\ \Sigma(a,\infty)=\left\{z(1-z\zeta_0)^{-a}\colon\, |\zeta_0|\leq 1\right\}.$$

It is known [9, (1.3)] that $\Sigma(a,\infty)=\bigcap_{c>0}\overline{\Sigma}(a,c)$ for $a=1$ and hence, with the help of (3.11), for every $a\geq 0$. A theorem of Brickman, MacGregor, and Wilken (see [11, Thm. 2.11]) implies that $\overline{\Sigma}_1(1,0)=Q(1)$. This allows c to be 0 in the following Corollary, obtained for $c>0$ by writing $L(a,c)=L(a,1)L(1,c)$ in Theorem 2.

COROLLARY 2. *$\overline{\Sigma}(a,c)=L(a,1)Q(1-\frac{1}{2}c)$ if $a,c\in[0,\infty)$. In particular, $\overline{\Sigma}(1,c)=Q(1-\frac{1}{2}c)$ and $\overline{\Sigma}(1,2)=K$.*

Hall [5] observed that $\overline{\Sigma}_1(1,2)=K$. The functions in $\Sigma_1(1,2)$ are Schwarz–Christoffel maps of the unit disk onto convex (not necessarily bounded) polygonal regions.

By using Corollary 2 the next two theorems are easily proved from Theorems A1 and B1:

THEOREM A2. *$\overline{\Sigma}(\alpha,c)*\overline{\Sigma}(a,c)=L(\alpha,1)\overline{\Sigma}(a,c)$ if α, c, $a\in[0,\infty)$. In particular, $Q(1-\frac{1}{2}c)*\overline{\Sigma}(a,c)=\overline{\Sigma}(a,c)$.*

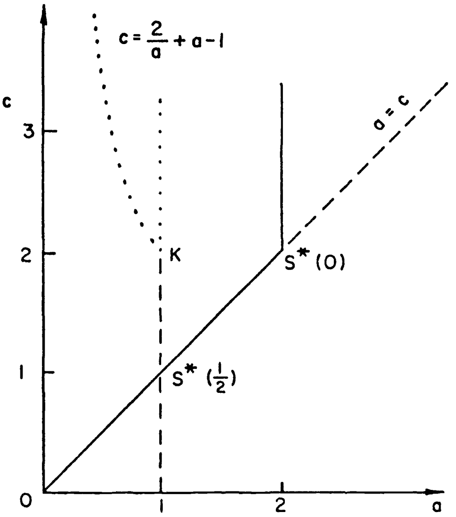


FIG. 1. *Classes of hypergeometric functions. Each point in the quadrant corresponds to a class $\bar{\Sigma}(a,c)$, and each class is contained in every class lying vertically below it. A point with $a=c$ is the class $S^*(1-\frac{1}{2}a)$ and is contained in each class with a larger value of $a=c$. All classes with fixed a and with $c \geq a$ are starlike of order $1-\frac{1}{2}a$ (and univalent if $a \leq 2$). A point with $a=1$ is the prestarlike class $Q(1-\frac{1}{2}c)$. The point with $a=1$ and $c=2$ is the class K of convex functions, and all classes vertically above it are convex. Each class with $a>1$ and each class below the curve $c=2/a+a-1$ contains a function that is not convex.*

THEOREM B2. *$L(c,e)\bar{\Sigma}(a,c) \subset \bar{\Sigma}(a,e) \subset \bar{\Sigma}(a,c)$ if $a,c \in [0,\infty)$, $e \in (0,\infty)$, and $c \leq e$.*

Since a starlike function of nonnegative order is univalent, every function in $\bar{\Sigma}(a,c)$ is univalent if $0 \leq a \leq 2$ and $c \geq a$. It is an open problem to determine what other classes are univalent. The class $\Sigma(a,c)$ contains the function $z(1-z)^{-a}$, which is not univalent if $a>2$. The following example proves that not every function in $\Sigma(a,c)$ is univalent if $c<\frac{1}{2}(a-1)(a+2)$. The coefficient of z^3 in the odd function $zR_{-a}(\frac{1}{2}c, \frac{1}{2}c; 1+z, 1-z)$ is $a(a+1)/2(c+1)$; this exceeds 1 if the inequality is satisfied, and hence the function is then not univalent by [8, Thm. 1.5].

It follows from [2, (6.2-24)] that the coefficient of z^{n+1} on the right side of (3.9) is bounded by $(a)_n/n!$. Hence the same coefficient bound holds for every function in $\bar{\Sigma}(a,c)$.

By Theorem B2 and Corollary 2, $\bar{\Sigma}(a,c) \subset K$ if $a=1$ and $c \geq 2$, but it is an open problem to determine other values of a and c for which this holds. The class $\Sigma(a,c)$ contains $z(1-z)^{-a}$, which is convex if $0 \leq a \leq 1$ and not convex if $a>1$. Since $\bar{\Sigma}(0,c)$ consists of the single convex function z, it suffices to consider the case $0<a \leq 1$. Some initial conjectures were disproved by plotting numerically the image of the unit circle for various examples. The elliptic integral $zR_{-1/2}(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}; 1-z, 1+iz, 1-iz)$ for $z=e^{i\theta}$, $0 \leq \theta < 2\pi$, shows a small deviation from convexity as $|\theta|$ approaches $\pi/2$ from below. The plot of the function

$$f(z)=zR_{-1/2}(1,1;\ 1+z, 1-z)=(1+z)^{1/2}-(1-z)^{1/2}$$

shows an even smaller deviation from convexity as θ approaches 0 or π. This can be verified analytically by writing $f(e^{i\theta})=u(\theta)+iv(\theta)$ and expanding in powers of θ. We find

$$\frac{dv}{du}=-1-\theta^{1/2}+O(\theta), \qquad \theta \to 0+.$$

The negative sign of the second term on the right side proves that f is not convex. The same procedure, starting from [2, Ex. 5.9-13], shows surprisingly that $zR_{-a}(1,1;\ 1+z,\ 1-z)$ is not convex if $0<a<1$, although it is convex if $a=0$ or $a=1$. The value of the function at $z=1$ is $2^{-a}/(1-a)$ and

$$\frac{dv}{du}=-\cot\left(\frac{a\pi}{2}\right)-\csc\left(\frac{a\pi}{2}\right)\left(\frac{\theta}{2}\right)^{a}+O(\theta),\qquad \theta\to 0+.$$

A similar but longer calculation starting from [2, (8.3-10)] shows that we can choose w_i so that $zR_{-a}(cw_1,cw_2;\ 1+z,1-z)$ is not convex, provided that $0<a<c<2/a+a-1$. In particular this proves that $\overline{\Sigma}(1,c)$ contains a nonconvex function if $1<c<2$. It leaves open the question whether $\overline{\Sigma}(a,c)\subset K$ if $0<a<1$ and $c\geq 2/a+a-1$.

4. Convexity of order α. The convex functions of order $1-\frac{1}{2}c$ comprise the class $K(1-\frac{1}{2}c)=L(1,2)S^*(1-\frac{1}{2}c)$, which is the closure of the class $L(1,2)\ \Sigma(c,c)$. It follows from (3.6) and (3.9) that every function in the latter class has the form

$$(4.1)\qquad \sum_{n=0}^{\infty}\frac{(c)_n}{(2)_n}R_n(cw_1,\cdots,cw_k;\ \zeta_1,\cdots,\zeta_k)z^{n+1}$$

for some value of k. By [2, (6.2-6)] and (3.3), this may be rewritten as

$$(4.2)\qquad \sum_{n=0}^{\infty}R_n(cw_1,\cdots,cw_k,2-c;\ \zeta_1,\cdots,\zeta_k,0)z^{n+1}$$
$$=zR_{-1}(cw_1,\cdots,cw_k,2-c;\ 1-z\zeta_1,\cdots,1-z\zeta_k,1).$$

If $c\leq 2$ (but not otherwise), this is a function in $\Sigma(1,2)$ with one of the ζ's equal to 0; the corresponding weight is $(2-c)/2$, the order of convexity. If $c=2$, that weight is 0; the last parameter and the last variable on the right side of (4.2) can then be omitted by [2, (6.3-3)], and we recover the general form of a function in $\Sigma(1,2)$.

The multiple power series F_D introduced by Lauricella (see [3]) is convenient in this context. We may write

$$(4.3)\qquad zF_D(a;\ cw_1,\cdots,cw_k;\ d;\ z\zeta_1,\cdots,z\zeta_k)$$
$$=zR_{-a}(cw_1,\cdots,cw_k,d-c;\ 1-z\zeta_1,\cdots,1-z\zeta_k,1)$$
$$=\sum_{n=0}^{\infty}\frac{(a)_n(c)_n}{(d)_n n!}R_n(cw_1,\cdots,cw_k;\ \zeta_1,\cdots,\zeta_k)z^{n+1},$$

where $d>0$ and $|z\zeta_i|<1$, $1\leq i\leq k$. This suggests defining, for every $a\geq 0$, $c\geq 0$, and $d>0$, a class of functions analytic on the open unit disk in the z-plane:

$$(4.4)\qquad T(a,c,d)=\left\{zF_D(a;\ cw_1,\cdots,cw_k;\ d;\ z\zeta_1,\cdots,z\zeta_k)\colon w_i\geq 0,\ \sum_{i=1}^{k}w_i=1,\right.$$
$$\left.|\zeta_i|\leq 1,\ 1\leq i\leq k,\ k=1,2,3,\cdots\right\}.$$

As in Theorem 2 the closure of this class is unchanged if the ζ's are restricted to lie on the unit circle. We define $T(a,0,0)=\Sigma(a,0)$ and $T(0,c,0)=\Sigma(c,c)$. Several relations follow at once from the series in (4.3) and (3.9). If $ac=0$ and $d>0$, then $T(a,c,d)=\{z\}$ by [2, (6.2-7)]. It is evident also that

$$(4.5)\qquad T(a,c,c)=\Sigma(a,c),$$

$$T(d,c,d)=\Sigma(c,c), \tag{4.6}$$

$$T(a,c,d)=L(a,\alpha)T(\alpha,c,d)=L(\delta,d)T(a,c,\delta), \qquad d,\alpha,\delta>0, \tag{4.7}$$

$$T(a,c,d)=L(a,d)\Sigma(c,c)=L(c,d)\Sigma(a,c), \qquad d>0. \tag{4.8}$$

THEOREM 4. *If $a,c\geq 0$ and $d>0$, then $\bar{T}(a,c,d)=L(a,d)S^*(1-\frac{1}{2}c)=L(c,d)\bar{\Sigma}(a,c)$. In particular, $\bar{T}(d,c,d)=S^*(1-\frac{1}{2}c)$, $\bar{T}(a,c,c)=\bar{\Sigma}(a,c)$, $\bar{T}(1,c,c)=Q(1-\frac{1}{2}c)$, and $\bar{T}(1,c,2)=K(1-\frac{1}{2}c)$.*

Proof. Since the case $ac=0$ is trivial, we assume $ac>0$. By continuity of $L(a,d)$, $L(c,d)$, and their inverses, (4.8) and (3.12) imply

$$\bar{T}(a,c,d)=L(a,d)\bar{\Sigma}(c,c)=L(a,d)S^*(1-\tfrac{1}{2}c)=L(c,d)\bar{\Sigma}(a,c).$$

The next two theorems follow easily from Theorems 4, A2, and B2.

THEOREM A3. *$\bar{T}(\alpha,c,\delta)*\bar{T}(a,c,d)=L(\alpha,1)L(c,\delta)\bar{T}(a,c,d)$, where $a,c,\alpha\in[0,\infty)$ and $d,\delta\in(0,\infty)$. In particular, $Q(1-\frac{1}{2}c)*\bar{T}(a,c,d)=\bar{T}(a,c,d)$.*

THEOREM B3. *$L(c,e)\bar{T}(a,e,d)\subset\bar{T}(a,c,d)\subset\bar{T}(a,e,d)$ if $a,c\in[0,\infty)$, $e,d\in(0,\infty)$, and $c\leq e$.*

COROLLARY 3. *$\bar{T}(a,c,d)\subset S^*(1-\frac{1}{2}a)$ if $a\leq d$ and $c\leq d$.*

Corollary 3 is proved by putting $e=d$ in the second inclusion in Theorem B3 and using (4.5) and Theorem 1. It follows that every function in $\bar{T}(a,c,d)$ is univalent if $a\leq 2$ and $a,c\leq d$. By [2, (6.2-24)] the coefficient of z^{n+1} on the right side of (4.3) is bounded by $(a)_n(c)_n/n!(d)_n$. Hence the same coefficient bound holds for every function in $\bar{T}(a,c,d)$.

5. Integral representations. It is well known [6, (1.2)], [11, Thm. 2.13] that $g\in S^*(1-\frac{1}{2}c)$ if and only if there exists a probability measure w on the unit circle T such that

$$g(z)=z\exp\left[-c\int_T \ln(1-z\zeta)\,dw(\zeta)\right]. \tag{5.1}$$

Since $\bar{T}(a,c,d)=L(a,d)S^*(1-\frac{1}{2}c)$, we can obtain integral representations of functions in $\bar{T}(a,c,d)$ by using (2.3) or (2.4). If $d>a>0$, then $f\in\bar{T}(a,c,d)$ if and only if there exists a w such that

$$f(z)=z\int_0^1\exp\left[-c\int_T \ln(1-uz\zeta)\,dw(\zeta)\right]d\mu_{(a,\,d-a)}(u). \tag{5.2}$$

If w consists of point masses $w_1,\cdots,w_k$ with unit sum, then $f\in T(a,c,d)$ and we recover the representation of F_D by a single integral. The right side of (5.2) is especially simple if $f\in K(1-\frac{1}{2}c)=\bar{T}(1,c,2)$, since $d\mu_{(1,1)}(u)=du$.

If $a,c\geq 0$ and $d>0$, then $f\in\bar{T}(a,c,d)$ if and only if there exists a w such that

$$f(z)=\frac{1}{2\pi i}\int_\gamma\exp\left[-c\int_T \ln(1-u\zeta)\,dw(\zeta)\right]\varphi(a,d;\,z/u)\,du, \tag{5.3}$$

where γ is the same as in (2.4). In the case $f\in K(1-\frac{1}{2}c)$ we note that $\varphi(1,2;\,z/u)=-\ln(1-z/u)$.

Acknowledgments. We thank Professor Stefan Ruscheweyh for discussion and some key remarks, Professor Glenn Luecke for helpful conversations, and Mr. John L. Gustafson for numerical computations.

REFERENCES

[1] R. A. Askey, *Orthogonal Polynomials and Special Functions*, CBMS Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[2] B. C. Carlson, *Special Functions of Applied Mathematics*, Academic Press, New York, 1977.

[3] ______, *Lauricella's hypergeometric function* F_D, J. Math. Anal. Appl., 7 (1963), pp. 452–470.

[4] J. Haeusler, *Eine Darstellung des unvollständigen Schwarz–Christoffelschen Abbildungsintegrals*, Z. Angew. Math. Mech., 46 (1966), p. 551.

[5] R. R. Hall, *On a conjecture of Clunie and Sheil-Small*, Bull. London Math. Soc., 12 (1980), pp. 25–28.

[6] J. L. Lewis, *Applications of a convolution theorem to Jacobi polynomials*, this Journal, 10 (1979), pp. 1110–1120.

[7] E. P. Merkes and W. T. Scott, *Starlike hypergeometric functions*, Proc. Amer. Math. Soc., 12 (1961), pp. 885–888.

[8] Chr. Pommerenke, *Univalent Functions*, Vandenhoeck and Ruprecht, Göttingen, 1975.

[9] St. Ruscheweyh, *New criteria for univalent functions*, Proc. Amer. Math. Soc., 49 (1975), pp. 109–115.

[10] ______, *Linear operators between classes of prestarlike functions*, Comment. Math. Helvetici, 52 (1977), pp. 497–509.

[11] G. Schober, *Univalent Functions—Selected Topics*, Lecture Notes in Mathematics 478, Springer-Verlag, New York, 1975.

[12] T. J. Suffridge, *Starlike functions as limits of polynomials*, in Advances in Complex Function Theory, W. E. Kirwan and L. Zalcman, eds., Lecture Notes in Mathematics 505, Springer-Verlag, New York, 1976, pp. 164–203.

Supplementary references on univalence of special functions that are not directly connected with the contents of the present paper:

[S1] S. D. Bernardi, *Bibliography of Schlicht Functions*, Mariner Publishing Co., Tampa, FL, 1982.

[S2] R. K. Brown, *Univalent solutions of* $w''+pw=0$, Canad. J. Math., 14 (1962), pp. 69–78.

[S3] E. Kreyszig and John Todd, *The radius of univalence of the error function*, Numer. Math., 1 (1959), pp. 78–89.

[S4] ______, *On the radius of univalence of the function* $\exp(z^2)\int_0^z \exp(-t^2)\,dt$, Pacific J. Math., 9 (1959), pp. 123–127.

[S5] ______, *The radius of univalence of Bessel functions*, Illinois J. Math., 4 (1960), pp. 143–149.

[S6] P. Todorov, *Ueber die Moeglichkeiten einer schlichten konformen Abbildung des Einheitskreises durch die hypergeometrische Funktion von Gauss*, Acad. Roy. Belg. Bull. Cl. Sci. (5), 53 (1967), pp. 432–441.

[S7] ______, *On certain univalent conformal mappings that are realized by Gauss's hypergeometric function*, (Bulgarian. Russian and French summaries), Plovdiv Univ. Naučn. Trud., 12 (1974), no. 1, pp. 59–64.

[S8] H. S. Wilf, *The radius of univalence of certain entire functions*, Illinois J. Math., 6 (1962), pp. 242–244.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics
011

THE QD-ALGORITHM FOR PADÉ-APPROXIMANTS IN OPERATOR THEORY*

ANNIE A. M. CUYT†

Abstract. It is well known that the quotient-difference algorithm can be used to construct univariate Padé-approximants. In this paper we see that the Padé-approximants for nonlinear operators $F: X \to Y$ where X is a Banach space and Y a commutative Banach algebra, introduced by the author, can also be obtained by means of the QD-algorithm and can consequently be obtained as convergents of a continued fraction, if the scalar QD-algorithm is reformulated as in §1. The definition of abstract Padé-approximants will be repeated in §2, while the operator QD-algorithm will be treated in §3.

1. The scalar QD-scheme. Let us consider a nonlinear real-valued function f of one real variable x, analytic at the origin

$$f(x)=\sum_{k=0}^{\infty} c_k x^k \quad \text{with } c_k=\frac{1}{k!}f^{(k)}(0).$$

We will present the QD-algorithm in a slightly different way than usual, but the two approaches are equivalent. The advantage of this approach is that it can be generalized to the case where F is a nonlinear operator from a Banach space X to a commutative Banach algebra Y.

Let the series f be normal, i.e.,

$$\begin{vmatrix} c_n x^n & c_{n+1}x^{n+1} & \cdots & c_{n+k-1}x^{n+k-1} \\ c_{n+1}x^{n+1} & & & \\ \vdots & & & \vdots \\ c_{n+k-1}x^{n+k-1} & & & c_{n+2k-2}x^{n+2k-2} \end{vmatrix} \neq 0$$

for $n=0,1,2,\cdots$ and $k=1,2,\cdots$. This determinant is a monomial of degree $k(n+k-1)$ in the variable x. Demanding that this monomial be nontrivial is equivalent to demanding that this determinant evaluated at $x=1$ be nonzero.

For a normal series we can construct a double entry table of numbers $q_k^{(n)}$ and $e_k^{(n)}$ defined as follows:

$$\begin{aligned} & e_0^{(n)}=0, && n=0,1,\cdots, \\ & q_1^{(n)}=\frac{c_{n+1}x^{n+1}}{c_n x^n}, && n=0,1,\cdots, \\ & e_k^{(n)}=q_k^{(n+1)}+e_{k-1}^{(n+1)}-q_k^{(n)}, && n=0,1,2,\cdots, \quad k=1,2,\cdots, \\ & q_{k+1}^{(n)}=\frac{q_k^{(n+1)}e_k^{(n+1)}}{e_k^{(n)}}, && n=0,1,2,\cdots, \quad k=1,2,\cdots. \end{aligned}$$

From this QD-algorithm we can obtain Padé-approximants to the function f as follows.

*Received by the editors November 24, 1982, and in revised form March 1, 1983.

†Aspirant M.F.W.O. Department of Mathematics, Universitaire Instelling Antwerpen, Universiteitsplein 1, B-2610 Wilrijk, Belgium.

The (l,m) Padé-approximant (numerator of degree l and denominator of degree m) for $l \geq m$ is equal to the $(2m)$th convergent K_{2m} of the continued fraction

$$c_0+c_1x+\cdots+c_{l-m}x^{l-m}+\frac{c_{l-m+1}x^{l-m+1}|}{|1}-\frac{q_1^{(l-m+1)}|}{|1}-\frac{e_1^{(l-m+1)}|}{|1}-\frac{q_2^{(l-m+1)}|}{|1}-\frac{e_2^{(l-m+1)}|}{|1}-\cdots$$

if $K_0=\sum_{k=0}^{l-m}c_kx^k$, and to the $(2m+1)$th convergent K_{2m+1} of the continued fraction

$$c_0+c_1x+\cdots+c_{l-m-1}x^{l-m-1}+\frac{c_{l-m}x^{l-m}|}{|1}-\frac{q_1^{(l-m)}|}{|1}-\frac{e_1^{(l-m)}|}{|1}-\frac{q_2^{(l-m)}|}{|1}-\frac{e_2^{(l-m)}|}{|1}-\cdots$$

if $K_0=\sum_{k=0}^{l-m-1}c_kx^k$[1].

The terms $q_k^{(n)}$ and $e_k^{(n)}$ each contain a factor x now because of the definition of $q_1^{(n)}$. Since the series f is normal the Padé-approximants are also normal [1].

2. Abstract Padé-approximants for operators. We briefly repeat the definition of Padé-approximant in operator theory and a determinantal formula for the calculation. More details can be found in [3]. Let X be a Banach space and Y a commutative algebra (0 denotes the unit for addition and I the unit for multiplication). Let $F: X \to Y$ be analytic in the open ball $B(0,r)$ with centre $0 \in X$ and radius $r>0$ [5, p. 113]

$$F(x)=\sum_{k=0}^{\infty}\frac{1}{k!}F^{(k)}(0)x^k \quad \text{for } \|x\|<r,$$

where $F^{(k)}(0)$ is the kth Fréchet-derivative of F in 0 and thus a symmetric k-linear bounded operator, and where $(1/0!)F^{(0)}(0)x^0=F(0)$.

DEFINITION 2.1. $F(x)=O(x^k)$ $(k \in \mathbb{N})$ if nonnegative constants $r<1$ and K exist such that $\|F(x)\| \leq K\|x\|^k$ for $\|x\|<r$.

Write $D(F)=\{x \in X \mid F(x)$ is regular in Y, i.e., there exists $y \in Y: F(x)\cdot y=I\}$.

DEFINITION 2.2. An *abstract polynomial* is a nonlinear operator $P: X \to Y$ with $P(x)=A_nx^n+A_{n-1}x^{n-1}+\cdots+A_0$, where A_i is a symmetric i-linear bounded operator $(i=0,\cdots,n)$ [5, p. 194].

When we have two abstract polynomials P and Q, we can construct an abstract rational operator $Q^{-1}\cdot P$ where $Q^{-1}(x)$ is the inverse element of $Q(x)$ for the multiplication in the Banach algebra Y. Of course division by $Q(x)$ can only be performed when x is in $D(Q)$.

DEFINITION 2.3. The pair of abstract polynomials $(P(x),Q(x))=(\sum_{i=0}^{l}A_{lm+i}x^{lm+i}, \sum_{j=0}^{m}B_{lm+j}x^{lm+j})$ such that the abstract power series $(F\cdot Q-P)(x)=O(x^{lm+l+m+1})$ is called a *solution of the Padé-approximation problem of order* (l,m).

The shift of degrees by $l\cdot m$ provides us with many nice properties [2], [3] and will also provide us with an abstract QD-scheme.

Let us denote by $Q_\triangle^{-1}\cdot P_\triangle$ a reduced form of the abstract rational operator $Q^{-1}\cdot P$; in other words $P=P_\triangle\cdot T$ and $Q=Q_\triangle\cdot T$ and we have cancelled this abstract polynomial T in both numerator and denominator. Different solutions of the Padé-approximation problem and different reduced forms are equivalent (denoted by $\simeq$); i.e., they satisfy the relation

$$(P,Q)\simeq(R,S) \Leftrightarrow P(x)\cdot S(x)=Q(x)\cdot R(x) \quad \forall x \in X.$$

DEFINITION 2.4. The *abstract Padé-approximant of order* (l,m) *for* F is the equivalence class containing all the pairs (P,Q) satisfying Definition 2.3 and all the pairs $(P_\triangle, Q_\triangle)$ which are the numerator and denominator of a reduced form of $Q^{-1}\cdot P$.

Let us write $C_k=(1/k!)F^{(k)}(0)$. We call the series F normal if there exists x in X such that

$$H_k(C_n)=\begin{vmatrix} C_nx^n & C_{n+1}x^{n+1} & \cdots & C_{n+k-1}x^{n+k-1} \\ \vdots & & & \vdots \\ C_{n+k-1}x^{n+k-1} & & & C_{n+2k-2}x^{n+2k-2} \end{vmatrix}$$

is regular in Y for $n=0,1,2,\cdots$ and $k=1,2,\cdots$. When the series

$$C_0+\sum_{k=1}^{\infty}\left(C_kx^k-C_{k-1}x^{k-1}\right)$$

is normal, a representation of $P(x)$ and $Q(x)$ satisfying Definition 2.3 is given by

$$P(x)=\begin{vmatrix} F_l(x) & F_{l-1}(x) & \cdots & F_{l-m}(x) \\ C_{l+1}x^{l+1} & & & C_{l-m+1}x^{l-m+1} \\ \vdots & & & \vdots \\ C_{l+m}x^{l+m} & & & C_lx^l \end{vmatrix},$$

$$Q(x)=\begin{vmatrix} I & I & \cdots & I \\ C_{l+1}x^{l+1} & \cdots & & C_{l-m+1}x^{l-m+1} \\ \vdots & & & \vdots \\ C_{l+m}x^{l+m} & \cdots & & C_lx^l \end{vmatrix},$$

where $F_l(x)=\sum_{k=0}^{l}C_kx^k$ [3, p. 25].

From now on we shall denote these determinants by $P_{[l,m]}(x)$ and $Q_{[l,m]}(x)$, respectively. The pair $(P_{[l,m]},Q_{[l,m]})$ can be considered as a representative of the abstract Padé-approximant of order (l,m) for F. If we introduce the notation

$$\Delta C_kx^k=C_{k+1}x^{k+1}-C_kx^k$$

then normality of the series $C_0+\sum_{k=1}^{\infty}(C_kx^k-C_{k-1}x^{k-1})$ is equivalent to $H_k(\Delta C_n)$ being regular in Y for some x in X and for all $n=0,1,2,\cdots$ and $k=1,2,\cdots$. So normality of the series $C_0+\sum_{k=0}^{\infty}\Delta C_kx^k$ implies regularity of $Q_{[l,m]}(x)=H_m(\Delta C_{l-m+1})$ for some x and thus existence of $Q_{[l,m]}^{-1}\cdot P_{[l,m]}$.

3. The abstract QD-scheme. For a normal series F we can define the abstract QD-scheme as follows:

$$\begin{aligned} &E_0^{(n)}=0, && n=0,1,\cdots,\\ &Q_1^{(n)}=\left(C_{n+1}x^{n+1}\right)\cdot\left(C_nx^n\right)^{-1}, && n=0,1,\cdots,\\ &E_k^{(n)}=Q_k^{(n+1)}+E_{k-1}^{(n+1)}-Q_k^{(n)}, && n=0,1,\cdots,\quad k=1,2,\cdots,\\ &Q_{k+1}^{(n)}=Q_k^{(n+1)}\cdot E_k^{(n+1)}\cdot\left(E_k^{(n)}\right)^{-1}, && n=0,1,\cdots,\quad k=1,2,\cdots. \end{aligned}$$

The existence of all the $E_k^{(n)}$ and $Q_k^{(n)}$ is proved as in [4, pp. 610–611]. Let us construct the following continued fractions in the Banach algebra Y:

$$(1)\qquad \sum_{k=0}^{l-m} C_k x^k + \cfrac{C_{l-m+1}x^{l-m+1}}{I-\cfrac{Q_1^{(l-m+1)}}{I-\cfrac{E_1^{(l-m+1)}}{I-\cfrac{Q_2^{(l-m+1)}}{I-\cfrac{E_2^{(l-m+1)}}{I-\dots}}}}}$$

and

$$(2)\qquad \sum_{k=0}^{l-m-1} C_k x^k + \cfrac{C_{l-m}x^{l-m}}{I-\cfrac{Q_1^{(l-m)}}{I-\cfrac{E_1^{(l-m)}}{I-\cfrac{Q_2^{(l-m)}}{I-\cfrac{E_2^{(l-m)}}{I-\dots}}}}}$$

where division means multiplication by the inverse element for multiplication in Y.

We shall now prove that these continued fractions are of the same form as in the univariate case where only a factor x remains in $q_k^{(n)}$ and $e_k^{(n)}$ after division of their numerator by denominator and we shall also prove that the convergents of these continued fractions yield our abstract Padé-approximants.

THEOREM 1. *If we write $Q_k^{(n)}=N_{q,k,n}/D_{q,k,n}$ and $E_k^{(n)}=N_{e,k,n}/D_{e,k,n}$ then $\partial N_{q,k,n}=\partial D_{q,k,n}+1$ and $\partial N_{e,k,n}=\partial D_{e,k,n}+1$, where ∂ indicates the degree of the abstract polynomial.*

Proof. The proof is by induction. For $k=1$ we have

$$N_{q,k,n}=C_{n+1}x^{n+1},\qquad D_{q,k,n}=C_n x^n,$$

$$N_{e,k,n}=\left(C_{n+1}x^{n+1}\right)^2-C_nx^n\cdot C_{n+2}x^{n+2},\qquad D_{e,k,n}=C_nx^n\cdot C_{n+1}x^{n+1},$$

so that

$$\partial N_{q,k,n}=n+1=\partial D_{q,k,n}+1,\qquad \partial N_{e,k,n}=2n+2=\partial D_{e,k,n}+1.$$

Suppose the theorem holds for $Q_1^{(n)},\cdots,Q_k^{(n)}$, $E_1^{(n)},\cdots,E_k^{(n)}$; we shall prove it then for $Q_{k+1}^{(n)}$ and $E_{k+1}^{(n)}$.

Since $Q_{k+1}^{(n)}=Q_k^{(n+1)}\cdot E_k^{(n+1)}\cdot(E_k^{(n)})^{-1}$, we have

$$Q_{k+1}^{(n)}=\frac{N_{q,k,n+1}\cdot N_{e,k,n+1}\cdot D_{e,k,n}}{N_{e,k,n}\cdot D_{q,k,n+1}\cdot D_{e,k,n+1}}=\frac{N_{q,k+1,n}}{D_{q,k+1,n}}.$$

Thus $\partial N_{q,k+1,n}=\partial N_{q,k,n+1}+\partial N_{e,k,n+1}+\partial D_{e,k,n}=\partial D_{q,k+1,n}+1$. For $E_{k+1,n}$ the proof is analogous.

Consider the following descending staircase:

$$\begin{array}{lll}
P_{[l-m,0]}(x)\cdot Q^{-1}_{[l-m,0]}(x) & & \\
P_{[l-m+1,0]}(x)\cdot Q^{-1}_{[l-m+1,0]}(x) & P_{[l-m+1,1]}(x)\cdot Q^{-1}_{[l-m+1,1]}(x) & \\
 & P_{[l-m+2,1]}(x)\cdot Q^{-1}_{[l-m+2,1]}(x) & \cdots \\
 & & \vdots
\end{array}$$

THEOREM 2. *$P_{[l,m]}(x)\cdot Q^{-1}_{[l,m]}(x)$ is the $(2m)$th convergent of the continued fraction* (1).

Proof. Let on the above staircase $P_{[l-m+i,j]}(x)\cdot Q^{-1}_{[l-m+i,j]}(x)$ be denoted by K_{i+j}, $i+j=0,1,\cdots$.

Regularity of the $H_k(C_n)$ and the $H_k(\Delta C_n)$ implies that [3, pp. 38–39]

$$\begin{aligned}
&K_{2i+1}-K_{2i}=(-1)^i H_{i+1}(C_{l-m+i+1})H_i(C_{l-m+i+1})H_i^{-1}(\Delta C_{l-m+i})H_i^{-1}(\Delta C_{l-m+i+1}),\\
&K_{2i}-K_{2i-1}=(-1)^{i-1} H_i(C_{l-m+i+1})H_i(C_{l-m+i})H_i^{-1}(\Delta C_{l-m+i})H_{i-1}^{-1}(\Delta C_{l-m+i}),\\
&K_{i+j}-K_{i+j-2}=(-1)^{j-1}[H_j(C_{l-m+i})]^2 H_j^{-1}(\Delta C_{l-m+i})H_{j-1}^{-1}(\Delta C_{l-m+i+1})
\end{aligned}$$

are regular.

So it is possible to construct the continued fraction

$$(3)\qquad \cfrac{\sum_{k=0}^{l-m} c_k x^k + c_{l-m+1}x^{l-m+1}}{I+\cfrac{\dfrac{K_1-K_2}{K_2-K_0}}{I+\sum_{n=3}^{\infty}\left|\dfrac{\dfrac{(K_{n-1}-K_n)(K_{n-2}-K_{n-3})}{(K_n-K_{n-2})(K_{n-1}-K_{n-3})}}{I}\right.}}$$

with convergents $K_0, K_1, K_2, \cdots$, where division again means multiplication by the inverse element for multiplication defined in Y. It is easy to verify that

$$\frac{K_1-K_2}{K_2-K_0}=Q_1^{(l-m+1)}\quad\text{and}\quad\frac{(K_2-K_3)(K_1-K_0)}{(K_3-K_1)(K_2-K_0)}=E_1^{(l-m+1)},$$

using the representation of $P_{[l-m,0]}(x)$, $Q_{[l-m,0]}(x)$, $P_{[l-m+1,0]}(x)$, $Q_{[l-m+1,0]}(x),\cdots$ given in the previous section.

Let us denote

$$\frac{(K_{n-1}-K_n)(K_{n-2}-K_{n-3})}{(K_n-K_{n-2})(K_{n-1}-K_{n-3})}$$

by $A_{n/2}^{(l-m+1)}$ if n is even and by $B_{(n-1)/2}^{(l-m+1)}$ if n is odd. We write also $A_1^{(l-m+1)}=Q_1^{(l-m+1)}$.

If we write down the continued fraction that is the even contraction of (3) (i.e., a continued fraction having as convergents K_{2n} for $n=0,1,2,\cdots$), we get

$$(4)\qquad \sum_{k=0}^{l-m} C_k x^k + \cfrac{C_{l-m+1}x^{l-m+1}}{I - A_1^{(l-m+1)} - \cfrac{A_1^{(l-m+1)}B_1^{(l-m+1)}}{I - B_1^{(l-m+1)} - A_2^{(l-m+1)} - \cdots}}$$

If we write down the continued fraction that is the odd contraction of (3) with $l-m$ replaced by $l-m-1$ (i.e, a continued fraction having as convergents the $P_{[l-m,0]}(x)\cdot Q_{[l-m,0]}^{-1}(x)$, $P_{[l-m+1,1]}(x)\cdot Q_{[l-m+1,1]}^{-1}(x),\cdots$ on the descending staircase (6)), we get

$$(5)\qquad \sum_{k=0}^{l-m-1} C_k x^k + \cfrac{C_{l-m}x^{l-m}A_1^{(l-m)}}{I - A_1^{(l-m)} - B_1^{(l-m)} - \cfrac{B_1^{(l-m)}A_2^{(l-m)}}{I - A_2^{(l-m)} - B_2^{(l-m)} - \cdots}}$$

Because (4) and (5) have the same convergents, we have

$$A_k^{(l-m+1)}B_k^{(l-m+1)} = B_k^{(l-m)}A_{k+1}^{(l-m)},\quad B_{k-1}^{(l-m+1)} + A_k^{(l-m+1)} = B_k^{(l-m)} + A_k^{(l-m)},$$

$$k=1,2,\cdots,$$

if we put $B_0^{(l-m+1)}=0$. So

$$A_k^{(l-m+1)} = Q_k^{(l-m+1)},\quad B_k^{(l-m+1)} = E_k^{(l-m+1)},\qquad k=1,2,\cdots.$$

This completes the proof.

Analogously we can formulate and prove the next theorem.

THEOREM 3. *$P_{[l,m]}(x)\cdot Q_{[l,m]}^{-1}(x)$ is the $(2m+1)$th convergent of the continued fraction* (2).

This can easily be seen by writing down the continued fraction (3) with $l-m$ replaced by $l-m-1$; the convergents of this continued fraction are the abstract Padé-approximants on the following descending staircase:

$$(6)\qquad \begin{array}{lll} P_{[l-m-1,0]}(x)\cdot Q_{[l-m-1,0]}^{-1}(x) & & \\ P_{[l-m,0]}(x)\cdot Q_{[l-m,0]}^{-1}(x) & P_{[l-m,1]}(x)\cdot Q_{[l-m,1]}^{-1}(x) & \\ & P_{[l-m+1,1]}(x)\cdot Q_{[l-m+1,1]}^{-1}(x) & \cdots \\ & & \vdots \end{array}$$

We illustrate Theorems 2 and 3 by means of a simple example. Consider

$$F: C'([0,T]) \to C([0,T]): x(t) \to e^{x(t)}\frac{dx}{dt} - (1+d).$$

The unit in the Banach algebra $C([0,T])$ is the constant function $x(t)=1$, so we shall write $I=1$.

The Taylor series development of F around $x(t)=0$, is

$$F(x)=-(1+d)+\frac{dx}{dt}\sum_{k=0}^{\infty}\frac{1}{k!}x^k(t).$$

Let us calculate, for instance, the $(l,2)$ abstract Padé-approximants for $l\geq 0$. If we use the determinantal representation of $P_{[l,2]}(x)$ and $Q_{[l,2]}(x)$ we find that

$$P_{[l,2]}(x)=\left(\frac{dx}{dt}\right)^2\frac{x^{2l-2}(t)}{l!(l-1)!}\left[\sum_{k=0}^{l}C_kx^k-\frac{2x(t)}{l+1}\sum_{k=0}^{l-1}C_kx^k+\frac{x^2(t)}{l(l+1)}\sum_{k=0}^{l-2}C_kx^k\right],$$

$$Q_{[l,2]}(x)=\left(\frac{dx}{dt}\right)^2\frac{x^{2l-2}(t)}{l!(l-1)!}\left[1-\frac{2x(t)}{l+1}+\frac{x^2(t)}{l(l+1)}\right].$$

Now $P_{[l,2]}(x)\cdot Q_{[l,2]}^{-1}(x)$ is the 4th convergent of the continued fraction (1). We calculate the necessary elements in the QD-table

$$Q_1^{(n)}=\frac{C_{n+1}x^{n+1}}{C_nx^n}=\frac{x(t)}{n},$$

$$E_1^{(n)}=Q_1^{(n+1)}-Q_1^{(n)}=\frac{-x(t)}{n(n+1)},$$

$$Q_2^{(n)}=\frac{Q_1^{(n+1)}\cdot E_1^{(n+1)}}{E_1^{(n)}}=x(t)\frac{n}{(n+1)(n+2)}.$$

Note that in $Q_1^{(n)}$ the quotient of an $(n+1)$-linear operator by an n-linear operator is a linear operator, which is not true in general but which simplifies the calculations a lot.

It is easy to check that

$$P_{[l,2]}(x)\cdot Q_{[l,2]}^{-1}(x)=\sum_{k=0}^{l-2}C_kx^k+\cfrac{C_{l-1}x^{l-1}}{1-\cfrac{Q_1^{(l-1)}}{1-\cfrac{E_1^{(l-1)}}{1-\cfrac{Q_2^{(l-1)}}{1}}}}$$

where the division is here a division of continuous functions. Analogously we can see that $P_{[l,2]}(x)\cdot Q_{[l,2]}^{-1}(x)$ is also the 5th convergent of the continued fraction (2).

REFERENCES

[1] C. Brezinski, *Padé-type Approximation and General Orthogonal Polynomials*, ISNM 50, Birkhäuser Verlag, Basel, 1980.

[2] Annie A. M. Cuyt, *The ε-algorithm and Padé-approximants in operator theory*, this Journal, 14 (1983), pp. 1009–1014.

[3] ———, *Abstract Padé-approximants for operators: theory and applications*, Ph. D. thesis, University of Antwerp, 1982.

[4] P. Henrici, *Applied and Computational Complex Analysis* I, John Wiley, New York, 1974.

[5] Louis B. Rall, *Computational Solution of Nonlinear Operator Equations*, Krieger, Huntington, New York, 1979.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics
0012

RESTRICTIONS OF NORMAL OPERATORS, PADÉ APPROXIMATION AND AUTOREGRESSIVE TIME SERIES*

GEORGE CYBENKO†

Abstract. This work studies restrictions of normal operators on a Hilbert space to so-called Krylov subspaces with special attention to selfadjoint and unitary operators. It is shown that the characteristic polynomials of these restrictions are orthogonal polynomials and furthermore are intimately related to denominators of Padé approximations to certain moment generating functions. These relations are seen to unify certain aspects of Lanczos methods for eigenvalue approximations of selfadjoint operators and autoregressive modeling of time series.

1. Introduction. In recent years, there has been a growing suspicion that similar ideas underlie Rayleigh-Ritz-Krylov-Lanczos eigenvalue approximation methods [19], [20], [25], [26] (see [24] for a modern treatment), Padé approximation [1], [15], [16], and autoregressive time series modeling [3] and linear prediction [23]. Perhaps one of the most compelling clues to such a connection has been at the algorithmic level, where numerous recursions were noted to be quite similar, many involving orthogonal polynomials. The sense in which autoregressive modeling is an eigenvalue approximation method has never been made explicit and will be precisely described by the end of this paper.

Whereas connections between the Lanczos method [20], polynomials orthogonal on the real line and Padé approximation are not too difficult to extract, the connections between linear prediction and Padé approximation have not been fully realized, perhaps because the relation between Padé approximation and polynomials orthogonal over the unit circle has not been succinctly identified, although Gragg has shown how Fourier-Padé approximation and such orthogonal polynomials interact [15] (see [6] also).

In this paper, we derive exact relationships between certain restrictions of normal operators on a Hilbert space and polynomials orthogonal over the spectrum of the operator. The case for a selfadjoint operator leads to orthogonal polynomials on the real line and Padé approximation to a certain moment generating function. The case for a unitary operator leads to orthogonal polynomials on the unit circle and corresponding Padé approximations of generating functions for moments on the unit circle. In time series analysis and linear predictive signal processing, the unitary operator is the bi-infinite shift operator on $l_2(\mathbb{Z})$, the Hilbert space of bi-infinite square summable sequences and the restriction of the shift to a finite dimensional subspace of $l_2(\mathbb{Z})$ is an object of primary interest.

What is particularly striking about the difference between Rayleigh-Ritz-Krylov-Lanczos methods and autoregressive modeling of time series is that in the former case the operator is the object of primary interest in that its spectrum is sought, the particular subspace is essentially arbitrary, and the approximation gives information about the spectrum of the operator, while in the latter case, the operator is completely trivial, in the sense that its spectrum is completely known, the subspace is determined by the time series and its shifted version, and the approximation then gives information

* Received by the editors June 28, 1982, and in revised form January 28, 1983. This work was supported in part by the National Science Foundation under grant MCS-8003364.

† Associate Director, Applied Mathematics, National Science Foundation, Washington, DC 20550. On leave from Department of Mathematics, Tufts University, Medford, Massachusetts 02155.

about the spectral content of the particular sequence with respect to the shift operator. These statements will be made more precise after the main results in the general case are derived.

Numerous authors have observed and used relationships between some of the subjects under discussion in this paper. For instance, Toeplitz matrices arise naturally in the computation of Padé approximants. Furthermore, moment matrices for measures on the unit circle are Toeplitz so that recursions for orthogonal polynomials on the unit circle are related to recursions for denominators of Padé approximants. These connections were implicitly or explicitly exploited in [4], [5], [6], [17], [22]. On the other hand, the relationship between moments of real measures, Padé approximants and rational functions were important ingredients of the works [2], [18], [28]. Further background with excellent bibliographic material can be found in [4], [16].

Section 2 studies the restrictions of a normal operator to so-called "Krylov" subspaces and establishes a precise relationship between the characteristic polynomial of the restriction and orthogonal polynomials over the spectrum of the operator. Section 3 looks at the special cases of selfadjoint and unitary operators while §4 develops the case for time series analysis, namely for autoregressive modeling and maximum entropy spectral estimation.

2. Restrictions of normal operators to Krylov subspaces. Let A be a bounded normal operator on a Hilbert space H and let $\mathbf{b}\in H$ be an arbitrary nonzero element. Consider the subspaces for $p\geq 0$

$$K_p=\operatorname{span}\left[\mathbf{b},A\mathbf{b},A^2\mathbf{b},\cdots,A^{p-1}\mathbf{b}\right].$$

These subspaces have been used extensively for approximating the spectrum $\sigma(A)$ of A and are commonly called Krylov subspaces. Rayleigh used the subspace K_1 [25] to approximate the spectrum of a selfadjoint operator, Krylov used K_p for $H=\mathbb{C}^p$ to find the characteristic polynomial of a $p\times p$ matrix A [19], while more recently, Lanczos used K_p to approximate eigenvalues of linear differential and integral operators [20]. Lanczos's method is currently one of the most efficient methods for computing eigenvalues of large, sparse symmetric matrices [24]. In the case of time series $H=l_2(\mathbb{Z})$ and A is the shift, so that K_p is the space generated by a finite "history" of the series, in the sense that, componentwise, elements of K_p involve information about the series drawn from a finite time frame.

In general, K_p is not invariant under A, but clearly, if $AK_p\subset K_p$ for some p, then $AK_{p'}\subset K_p$ for all $p'\geq p$. We shall let $p^*>0$ be defined by

$$p^*=\text{minimal } p \text{ for which } AK_p\subset K_p \text{ if such a } p \text{ exists,}$$
$$p^*=+\infty \text{ otherwise.}$$

For each K_p, let π_p denote the orthogonal projection of H onto K_p so that

$$A_p=\pi_p A|_{K_p}$$

is the restriction of A to K_p followed by orthogonal projection onto K_p. Clearly, A_p is a linear operator on the finite dimensional space K_p. Let us assume that $p\leq p^*$ so that K_p is in fact p-dimensional. Now let

$$d_p(t)=\det(tI_p-A_p)$$

be the characteristic polynomial of A_p acting on K_p. Here I_p is the identity on K_p. Clearly there is nothing to be gained by the study of cases $p>p^*$ since then $K_p=K_{p^*}$,

$\pi_p=\pi_{p^*}$, $A_p=A_{p^*}$ and $d_p(t)=d_{p^*}(t)$. Note also that

$$t^p d_p\left(\frac{1}{t}\right)=\det(I_p-tA_p).$$

Our goal in this section is to identify the polynomials $d_p(t)$, $p=0,1,2,\cdots,p^*$, with the monic orthogonal polynomials determined by a finite measure on the spectrum of $A, \sigma(A)$.

We recall the spectral resolution of a normal bounded operator A [27], [31]; namely there is a resolution of the identity P_λ for which

$$A=\int_{\sigma(A)} \lambda dP_\lambda .$$

Now the measure on $\mathbf{C}$ defined by

$$\int_S d\mu(\lambda)=\int_{S\cap\sigma(A)} d\langle P_\lambda \mathbf{b},\mathbf{b}\rangle=\int_{S\cap\sigma(A)} d\langle P_\lambda \mathbf{b},P_\lambda \mathbf{b}\rangle$$

is clearly finite and positive. In particular, we have

$$\langle A^i\mathbf{b},A^j\mathbf{b}\rangle=\int_{\sigma(A)} \bar{\lambda}^i\lambda^j d\mu(\lambda)$$

where $\langle\cdot,\cdot\rangle$ is the inner product on H. Let $\mathcal{P}$ be the space of complex polynomials of degree no greater than p^*, so that for $u,v\in\mathcal{P}$

$$(u,v)=\int_{\sigma(A)} \overline{u(\lambda)}v(\lambda)\,d\mu(\lambda) \tag{1}$$

defines an inner product on $\mathcal{P}$ and so there is a unique sequence of monic orthogonal polynomials $q_0,q_1,q_2,\cdots,q_{p^*}$ so that

$$\text{degree}(q_i)=i, \qquad (q_i,q_j)\begin{cases} =0 & \text{if } i\neq j,\\ \neq 0 & \text{if } i=j.\end{cases}$$

Our first result is then

THEOREM 1. *For* $p\le p^*$,

$$d_p(\lambda)=\det(\lambda I_p-A_p)=q_p(\lambda).$$

Proof. Consider the matrix representation of A_p on K_p with respect to the basis $\mathbf{b},A\mathbf{b},A^2\mathbf{b},\cdots,A^{p-1}\mathbf{b}$, which is of the form

$$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -a_0\\ 1 & 0 & 0 & \cdots & 0 & -a_1\\ 0 & 1 & 0 & \cdots & 0 & -a_2\\ & \ddots & \ddots & \ddots & \vdots & \vdots\\ 0 & & \ddots & 1 & 0 & -a_{p-2}\\ & & & 0 & 1 & -a_{p-1}\end{bmatrix}$$

where $a_0,a_1,\cdots,a_{p-1}$ are determined by the condition that

$$\left\|A^p\mathbf{b}+a_{p-1}A^{p-1}\mathbf{b}+\cdots a_0\mathbf{b}\right\|^2$$

is minimized over all $a_{p-1},\cdots,a_0 \in \mathbf{C}$. This follows from the fact that $A_p(A^{p-1}\mathbf{b})$is the orthogonal projection of $A^p\mathbf{b}=AA^{p-1}\mathbf{b}$ on K_p. Putting $d_p(t)=\det(tI_p-A_p)$, it is clear that

$$d_p(t)=t^p+a_{p-1}t^{p-1}+\cdots+a_1t+a_0,$$

and we have that $\|d_p(A)\mathbf{b}\|^2$ is minimal over all monic polynomials of degree p. However,

$$\|d_p(A)\mathbf{b}\|^2=\left\langle d_p(A)\mathbf{b},d_p(A)\mathbf{b}\right\rangle=\int_{\sigma(A)}\overline{d_p(\lambda)}d_p(\lambda)\,d\mu(\lambda),$$

and it is a classical fact that the monic polynomial solving this minimization is precisely the monic orthogonal polynomial $q_p(\lambda)$ of degree p. Hence $d_p(\lambda)$ is precisely that polynomial. □

COROLLARY 1. *The eigenvalues of A_p lie in the convex hull of the spectrum of A.*

Proof. A classical result of Fejer's [10], [13] states that the roots of polynomials orthogonal with respect to a measure on $\mathbf{C}$ like (1) lie inside the convex hull of the support of the measure. □

It should be noted that the normality of the operator A is used only to show that the inner product is actually a classical inner product induced by an integral in the complex plane. Contained in the proof of Theorem 1 is the fact that the minimizing polynomial of any operator A defined by the condition that

$$\left\|\left(A^p+a_{p-1}A^{p-1}+\cdots+a_1A+a_0I\right)\mathbf{b}\right\|$$

is minimal over all $a_{p-1},\cdots,a_0$ also gives the characteristic polynomial of A_p and is the monic pth orthogonal polynomial with respect to the inner product $(u,v)=\langle u(A)b,v(A)b\rangle$. Furthermore, in the general setting, one has that if λ is an eigenvalue of A_p, then $\pi_pA\pi_p\mathbf{c}=\lambda\pi_p\mathbf{c}$ for some $\mathbf{c}$ in H, whence $\langle\pi_p\mathbf{c}, A\pi_p\mathbf{c}\rangle=\lambda$ and so λ is in the field of values of A. For a normal A we can then conclude that λ then belongs to the closed convex hull of the spectrum of A, since the field of values for a normal operator is precisely the closure of the convex hull of the spectrum of A.

This result simultaneously explains the fact that Lanczos polynomials have roots inside the interval containing the spectrum of a selfadjoint operator and that predictor polynomials have roots inside the unit circle, since as already mentioned the autoregressive case involves a unitary operator whose spectrum is the unit circle.

This is furthermore a generalization of the fact that Rayleigh quotients lie inside the convex hull of the spectrum of A since this is the case $p=1$. The Hausdorff-Toeplitz theorem [9] states that

$$\{\langle\mathbf{b},A\mathbf{b}\rangle/\langle\mathbf{b},\mathbf{b}\rangle,\ \mathbf{b}\in H\}$$

is convex. A natural conjecture is that the roots of the above polynomials, as $\mathbf{b}$ varies, form a convex set in $\mathbf{C}^p$:

Conjecture.

$$\left\{(t_1,\cdots,t_p)\in\mathbf{C}^p \,\middle|\, \prod_{i=1}^{p}(t-t_i)=d_p(t)\text{ some }\mathbf{b}\in H\right\}$$

is convex.

As shown above, the finite dimensional sections of A on Krylov subspaces are intimately related to polynomials orthogonal over the spectrum of A. The two cases where a significant orthogonal polynomial theory over a set has evolved are for the real line, corresponding to selfadjoint operators, and for the unit circle in $\mathbb{C}$, corresponding to unitary operators.

The next section develops the relationships between $\det(\lambda I_p - A_p)$ and Padé approximations of "moment generating functions" of the form

$$\int_{-\infty}^{\infty} \frac{1}{(1-\lambda w)} d\mu(w)$$

in the selfadjoint case and

$$\int_{|w|=1} \frac{w^{-p}}{(1-\lambda w)} d\mu(w)$$

in the unitary case.

Before proceeding however, it is worth pointing out how Theorem 1 explains the occurrence of orthogonal polynomials both in Rayleigh-Ritz-Krylov-Lanczos eigenvalue approximations from the subspaces K_p and in time series analysis and signal processing. In the case of Lanczos eigenvalue methods for symmetric operators, A is a selfadjoint operator and $\mathbf{b}$ is essentially arbitrary. The information in $d_p(t)$, namely its zeros, is used to model A's spectrum. The situation for time series analysis, and autoregressive modeling in particular, is similar in form but quite different in spirit. Here $H = l_2(\mathbb{Z})$, the space of bi-infinite square summable sequences, $A = Z$, is the shift operator with $\mathbf{b}$ the time series. In that case, $d_p(t)$ is the pth order autoregressive scheme that best describes $\mathbf{b}$ in the least squares sense. The spectrum of A in this case is completely trivial, namely the unit circle, and the spectral information in K_p and $d_p(t)$ then give information on the "periodic" properties contained in $\mathbf{b}$ in the following sense. If $e^{i\theta} \in \sigma(\mathbb{Z})$ then $Z - e^{i\theta} I$ has only the zero vector in its null space (that is, has no proper eigenvectors) but for any $\varepsilon > 0$ there exists vectors $\mathbf{u}$ so that

$$\|\mathbf{u}\|^2 = 1 \quad \text{and} \quad \|(Z - e^{i\theta} I)\mathbf{u}\| < \varepsilon.$$

A completely heuristic argument might go as follows: For an eigenvalue λ of Z_p on K_p, there is a value of the spectrum, $e^{i\theta}$, which is close to λ and an almost periodic $\mathbf{u}$ so that $\mathbf{u}$ would have a large component in the eigenvector of Z_p in K_p corresponding to λ. It is well known that in the symmetric case, there need not be very strong relationships between approximate eigenvectors (in the finite dimensional case) and the exact eigenvalues [24], and so this is indeed completely heuristic.

In the next section, we therefore show that the moment generating function for Z_p is a Padé approximant of the moment generating function for Z with respect to $\mathbf{b}$. More precisely, this approximant should be called a variant of Fourier-Laurent-Padé approximation as will be made precise shortly.

In any case, although the occurence of orthogonal polynomials in both Lanczos eigenvalue approximations and autoregressive time series modeling has typically been treated as coincidental, the results of this section have shown that they are part of the same underlying principle.

3. Selfadjoint and unitary operators. Recall that the (n,m) Padé approximant, $n\geq 0$, $m\geq 0$, denoted by $r_{n,m}(t)$, to a function $f(t)$ analytic in a neighborhood of $t=0$ is determined by the condition that

> $r_{n,m}$ is a rational function which is a quotient of a numerator polynomial of degree n and a denominator polynomial of degree m, that is,
>
> $$r_{n,m}=\frac{p_{n,m}}{q_{n,m}},\quad \text{degree}(p_{n,m})=n,\quad \text{degree}(q_{n,m})=m$$

and

$$f(t)q_{n,m}(t)-p_{n,m}(t)=O(t^{n+m+1})\quad \text{as } t\to 0.$$

See [1], [16] for details about Padé approximants and their theory.

First we let A be selfadjoint, so that $\sigma(A)\subset\mathbb{R}$. Define for $|t|<1/\|A\|$

$$f(t)=\int_{\mathbb{R}}\frac{1}{(1-tw)}d\mu(w)$$

where μ is defined as in the previous section. Now

$$f(t)=\int_{\mathbb{R}}\sum_{j=0}^{\infty}(tw)^j d\mu(w)=\sum_{j=0}^{\infty}\int_{\mathbb{R}}t^jw^jd\mu(w)=\sum_{j=0}^{\infty}t^j\langle \mathbf{b},A^j\mathbf{b}\rangle,$$

which justifies calling f the moment generator of A with respect to $\mathbf{b}$.

Our basic result is then an identification of the relationship between the characteristic polynomial of A_p, the $(p-1,p)$ Padé approximant to $f(t)$, and the moment generating function of A_p with respect to b. Our first step is to show that A_p is also selfadjoint.

LEMMA 1. *If A is selfadjoint, then so is A_p (with respect to $\langle\cdot,\cdot\rangle$).*

Proof. Since π_p is an orthogonal projection, it is selfadjoint. Now for $\mathbf{c},\mathbf{d}\in K_p$, we have

$$\langle \mathbf{c},A_p\mathbf{d}\rangle=\langle \mathbf{c},\pi_pA\pi_p\mathbf{d}\rangle=\langle \pi_p\mathbf{c},A\pi_p\mathbf{d}\rangle=\langle A\pi_p\mathbf{c},\pi_p\mathbf{d}\rangle=\langle A_p\mathbf{c},\mathbf{d}\rangle. \qquad \square$$

It follows that A_p has a similar spectral resolution and that there is a positive measure on $\mathbb{R}$ so that

$$\langle \mathbf{b},A_p^j\mathbf{b}\rangle=\int_{\mathbb{R}}t^jd\mu_p(t)\quad \text{for all } j\geq 0.$$

We thus define $f_p(t)$ by

$$f_p(t)=\int_{\mathbb{R}}\frac{1}{1-tw}d\mu_p(w)$$

and by Corollary 1, $f_p(t)$ is also analytic for $|t|<1/\|A\|$. In fact, $f_p(t)$ is clearly rational.

THEOREM 2. *With the notation of above, the following are true*:

(i) *$f_p(t)$ is the $(p-1,p)$ Padé approximant to $f(t)$.*

(ii) *If $f_p(t)=u(t)/v(t)$*, degree$(u)=p-1$, degree$(v)=p$ *and v is normalized to have constant coefficient* 1, *then*

$$\det(I_p-tA_p)=t^pd_p\left(\frac{1}{t}\right)=v(t).$$

(iii) *The poles of $f_p(t)$ are located at the reciprocals of the eigenvalues of A_p and the residue at $1/\lambda_j$ is precisely $-|\langle \mathbf{b}, \mathbf{e}_j\rangle|^2/\lambda_j$ where $\mathbf{e}_j$ is the eigenvector of A_p for eigenvalue λ_j and $\langle \mathbf{e}_j, \mathbf{e}_j\rangle = 1$.*

Proof. To show (i), it is sufficient to establish that

$$\langle \mathbf{b}, A^j\mathbf{b}\rangle = \left\langle \mathbf{b}, A_p^j\mathbf{b}\right\rangle \quad \text{for } j=0,1,\cdots,2p-1.$$

Now for $j<p$, $A^j\mathbf{b}=A_p^j\mathbf{b}$ while for $j=p$ we have that $(A^p-A_p^p)\mathbf{b}$ is orthogonal to K_p and so

$$\left\langle A^i\mathbf{b}, A^p\mathbf{b}\right\rangle = \left\langle A_p^i\mathbf{b}, A_p^p\mathbf{b}\right\rangle \quad \text{for } i=0,1,\cdots,p-1.$$

Since both A and A_p are selfadjoint, we in fact have that

$$\langle \mathbf{b}, A^{p+i}\mathbf{b}\rangle = \left\langle \mathbf{b}, A_p^{p+i}\mathbf{b}\right\rangle \quad \text{for } i=0,\cdots,p-1.$$

This establishes (i).

To see (ii), note that μ_p has support contained in the spectrum of A_p and hence the support of μ_p consists of precisely p points since we previously assumed that $p\le p^*$. This means that $f_p(t)$ has poles located at the reciprocals of the eigenvalues of A_p. On the other hand, the roots of $d_p(t)$ are the eigenvalues and so $t^p d_p(1/t)$ has roots at the reciprocals of the eigenvalues and must therefore be, up to a constant multiple, the same as the denominator of $f_p(t)$ expressed as a rational function.

Finally, (iii) follows from the observation that $f_p(t)$ has the form

$$f_p(t) = \sum_{j=1}^{p} \frac{\langle \mathbf{b}, \mathbf{e}_j\rangle^2}{(1-t\lambda_j)} = \sum_{j=1}^{p} \frac{\langle \mathbf{b}, \mathbf{e}_j\rangle^2/\lambda_j}{(1/\lambda_j - t)},$$

establishing the final claim. □

Throughout the remainder of this section, we shall assume that A is unitary.

Although the relationship between orthogonal polynomials on the real line and denominators of certain Padé approximations are well known, albeit difficult to trace historically, the relationship between orthogonal polynomials on the unit circle and denominators of Padé approximations has not been identified. This is perhaps due to the fact that polynomials orthogonal on the unit circle, although thoroughly studied [14], [29], have not been studied by a wide audience. Our first result is to make this relationship evident.

To this end, let

$$g_p(t) = \int_{|w|=1} \frac{w^{-p}}{(1-tw)} d\mu(w)$$

where μ is a positive measure on $|w|=1$. We shall suppose that μ is supported by more than p points over $|w|=1$. For $|t|<1$, we have

$$g_p(t) = \int_{|w|=1} w^{-p} \sum_{j=0}^{\infty} (wt)^j d\mu(w)$$

$$= \sum_{j=0}^{\infty} t^j \int_{|w|=1} w^{j-p} d\mu(w).$$

THEOREM 3. *Let $c_p(t)$ be the denominator of the $(p-1,p)$ Padé approximation to $g_p(t)$ at 0. Then $\bar{c}_p(t)$ is precisely the pth orthogonal polynomial with respect to μ, normalized so that $c_p(t)=t^p+$ lower order terms.*

Proof. We first show that $c_p(t)$ can indeed be normalized so that its leading term in t^p has a nonzero coefficient. Looking at the desired relationship and letting c_p denote the vector of coefficients of $c_p(t)$,

$$g_p(t)=\frac{u_{p-1}(t)}{c_p(t)}+O(t^{2p}),$$

we require that

$$T_{p+1}c_p=\begin{bmatrix} \mu_0 & \mu_{-1} & \cdots & \mu_{-p} \\ \mu_1 & & & \vdots \\ \vdots & & & \\ & & & \mu_{-1} \\ \mu_p & \cdots & \mu_1 & \mu_0 \end{bmatrix} c_p=\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ X \end{bmatrix}$$

where X is some possibly nonzero entry and $\mu_j=\int w^j d\mu(w)$. Now if the last entry of c_p were in fact zero, then $c_p(t)$ would in fact be of degree $p-1$ or less and so T_p, the $p\times p$ leading submatrix of T_{p+1}, would be singular. Now if c_{p-1} is the vector consisting of the first p entries of c_p then we would have

$$c^H_{p-1}T_pc_{p-1}=0=\int_{|t|=1}|c_{p-1}(t)|^2 d\mu(t).$$

But $c_{p-1}(t)$ has only $p-1$ zeros and so μ must be supported on $p-1$ or fewer points. This contradicts our earlier assumption that μ is supported on at least p points.

Finally, since $T_{p+1}c_p$ has p leading zeros, we have for any $a(t)$ of degree $p-1$ or less,

$$a^H\bar{T}_{p+1}\bar{c}_p=0=\int_{|t|=1}\overline{a(t)}\bar{c}_p(t)\,d\mu(t)$$

so $\bar{c}_p(t)$ is indeed the pth monic orthogonal polynomial with respect to μ. Here we have used the fact that T_{p+1} is Hermitian, and of course Toeplitz. □

This result also follows from the determinant representation for orthogonal polynomials on the circle [29] and the determinant representation for Padé approximants [1], [16].

We know from §2 that the characteristic polynomial of A_p is precisely the pth orthogonal polynomial with respect to the inner product

$$(u,v)=\int_{|t|=1}\overline{u(t)}v(t)\,d\mu(t),$$

and now we have seen that the denominator of the $(p-1,p)$ Padé approximant to $g_p(t)$ is the conjugate of the characteristic polynomial of A_p. It remains to be seen how the analogous function for A_p is related to the various quantities introduced above. To this end, we have the following results:

THEOREM 4. *Let*

$$g_{A_p}(t)=\left\langle A_p^p\mathbf{b},\mathbf{b}\right\rangle+\sum_{j=0}^{\infty}\left\langle A_p^{p-1}\mathbf{b},A_p^j\mathbf{b}\right\rangle t^{j+1}.$$

Then $g_{A_p}(t)$ is a rational function of degree (p,p) and is precisely the (p,p) Padé approximant to $g_p(t)$ defined above. Furthermore, the denominator polynomial of $g_{A_p}(t)$, say $v(t)$, satisfies

$$t^pv\left(\frac{1}{t}\right)=\det\left(tI-A_p\right)=\bar{c}_p(t),$$

provided that v is normalized to have constant coefficient $+1$.

That is, v is the reflection of the characteristic polynomial of A_p and the reflection of the conjuguate of the denominator polynomial of the $(p-1,p)$ Padé approximant to $g_p(t)$.

Proof. First we note that

$$\left\langle A^p\mathbf{b},\mathbf{b}\right\rangle=\left\langle A_p^p\mathbf{b},\mathbf{b}\right\rangle=\left\langle\mathbf{b},A^{-p}\mathbf{b}\right\rangle$$

and

$$\left\langle\mathbf{b},A^{j-p+1}\mathbf{b}\right\rangle=\left\langle A_p^{p-1}\mathbf{b},A_p^j\mathbf{b}\right\rangle=\left\langle A^{p-1}\mathbf{b},A^j\mathbf{b}\right\rangle\quad\text{for } j=0,1,\cdots,2p-1.$$

Thus $g_{A_p}(t)$ and $g_p(t)$ have the same MacLaurin series coefficients up to and including the term for t^{2p}. Hence it suffices to show that $g_{A_p}(t)$ is a rational function with numerator and denominator both polynomials of degree p, in which case it will be established that $g_{A_p}(t)$ is precisely the (p,p) Padé approximant to $g_p(t)$.

To this end, let $d_p(t)$ be the characteristic polynomial of A_p acting on K_p, and let

$$g_{A_p}(t)=\sum_{j=0}^{\infty}c_jt^j.$$

By the Cayley-Hamilton theorem, we know that $d_p(A_p)=0$, so that if $d_p(t)=t^p+a_{p-1}t^{p-1}+\cdots+a_1t+a_0$ then

$$c_n+a_{p-1}c_{n-1}+\cdots+a_1c_{n-p+1}+a_0c_{n-p}=\left\langle A_p^{p-1}b,d_p\left(A_p\right)A_p^{n-p}b\right\rangle=0$$

for $n>p$. Note that the relationship does not necessarily hold for $n=p$ because c_0 is anomalous in a sense. Hence the sequence $c_1,c_2,\cdots$ satisfies a pth order difference equation and must therefore have the form

$$c_j=\sum_m\sum_{k=0}^{k_m-1}j^k\lambda_m^jg_{m,k}$$

where λ_m is the mth root of $d_p(t)$ occuring with multiplicity k_m. It is straightforward to verify then that

$$\sum_{j=1}^{\infty}c_jt^j=t\frac{u(t)}{\Pi_m\left(1-\lambda_mt\right)^{k_m}}=\frac{tu(t)}{t^pd_p(1/t)}=\frac{w(t)}{t^pd_p(1/t)}$$

where $u(t)$ is a polynomial of degree $p-1$. Hence

$$g_{A_p}(t)=c_0+\frac{tu(t)}{t^pd_p(1/t)}=\frac{w(t)}{t^pd_p(1/t)}$$

and so the theorem is proved. □

It is important to note that unlike the selfadjoint case, the unitary case does not guarantee that A_p is itself unitary, nor even normal. It is for this reason that the theorem is more complicated in this case, and that a moment generating formula for A_p cannot be written in terms of a spectral measure directly. Although A_p is not unitary it is close to unitary in the following sense.

LEMMA 2. *There is a rank one selfadjoint operator E so that if A_p^* denotes the adjoint of A_p then*

$$A_pA_p^*+E=I_p$$

where I_p is the identity on K_p.

Proof. Clearly, since π_p is an orthogonal projection operator and hence selfadjoint, we have that

$$A_p^*=(\pi_pA\pi_p)^*=\pi_pA^{-1}\pi_p$$

on K_p and so $A_pA_p^*$ is the identity on the subspace generated by $A\mathbf{b},A^2\mathbf{b},\cdots,A^{p-}\mathbf{b}$ and so there exists a rank one operator E so that $A_pA_p^*+E=I_p$ and the selfadjointness of E follows from taking adjoints. □

We actually have the following characterization of the situation when A_p is unitary also.

THEOREM 5. *A_p is unitary if and only if $d_p(t)$ has all of its roots on the unit circle $|t|=1$, and in that case K_p is an invariant subspace for A and so $p=p^*$.*

Proof. Let $d_p(t)=a_0+tq(t)+t^p$ where $q(t)$ is a polynomial of degree $p-2$. Hence from the definitions and previous results,

$$A_p^p\mathbf{b}=-a_0\mathbf{b}-Aq(A)\mathbf{b}.$$

Since A is unitary, it follows that

$$A_p^*\mathbf{b}=-A^{-1}\bar{q}(A^{-1})A^{p-1}\mathbf{b}-\bar{a}_0A^{p-1}\mathbf{b}$$

whence

$$A_pA_p^*\mathbf{b}=-\bar{q}(A^{-1})A^{p-1}\mathbf{b}+\bar{a}_0(a_0\mathbf{b}+Aq(A)\mathbf{b}).$$

For $A_pA^*{}_p=I$, it is then necessary for $|a_0|^2=1$ and

$$-A^{p-2}\bar{q}(A^{-1})A\mathbf{b}+\bar{a}_0q(A)A\mathbf{b}=0.$$

It follows that

$$-t^{p-2}\bar{q}\left(\frac{1}{t}\right)+\bar{a}_0q(t)=0$$

and so

$$-\bar{d}_p\left(\frac{1}{t}\right)t^p+\bar{a}_0d_p(t)=0.$$

Thus if t is a root of $d_p(t)$ then so is $1/\bar{t}$, but all the roots of $d_p(t)$ lie inside the unit circle or on it, so that t and $1/\bar{t}$ are actually on the unit circle whence $t=1/\bar{t}$. Hence we have established that if A_p is unitary, then all of its eigenvalues lie on the unit circle.

Conversely, should $d_p(t)$ have all of its roots on the unit circle then it is obvious that the identity

$$-\bar{d}_p\left(\frac{1}{t}\right)t^p+\bar{a}_0d_p(t)=0$$

must hold and so A_p is indeed unitary by working the above argument backwards.

It remains to be seen that K_p is an invariant subspace for A if the roots of $d_p(t)$ happen to lie on the unit circle. Now since we are assuming that A_p is unitary, it follows that

$$\left\|A_pA^{p-1}\mathbf{b}\right\|^2=\left\|\pi_pA^p\mathbf{b}\right\|^2=\left\|A^p\mathbf{b}\right\|^2,$$

and since π_p is the orthogonal projection onto K_p the only way for this equality to hold is for $A^p\mathbf{b}\in K_p$ so that K_p is invariant under A. □

4. Linear prediction, autoregressive modeling and maximum entropy spectral estimation. In this section, the previous results are interpreted in terms of the various constructs that occur in linear prediction, autoregressive modeling of time series and maximum entropy spectral estimation. The reader unfamiliar with these areas is referred to [23] for an introduction to linear prediction and its applications, to [3] for autoregressive modeling of time series, and to [7] for maximum entropy spectral estimation. Although all of these topics are traditionally studied in the context of a stochastic system, our discussion involves no statistical ingredients. Thus, this presentation can be viewed as being analogous to the relationship between multiple linear regression and linear least-squares theory—linear regression consists of a purely geometric component, namely linear least-squares which is free of statistical hypotheses, plus a statistical component, which interprets the least-squares solution. Our presentation in this section is independent of statistical assumptions, dealing solely with the algebraic and analytic parts of linear prediction, autoregressive modeling, and maximum entropy spectral estimation.

First of all, we should note that all three of the above named topics are essentially the same, the choice of name being dictated by subtleties of the information being sought and the discipline of the author.

The ingredients are as follows:

$H=l_2(\mathbb{Z})$, the Hilbert space of bi-infinite sequences,

$$\langle x,y\rangle=\sum_j\bar{x}_jy_j\quad\text{for } \mathbf{x},\mathbf{y}\in l_2(\mathbb{Z});$$

A is the bi-infinite shift operator, namely

$$(A\mathbf{x})_j=x_{j-1},\quad\text{so that } A \text{ is clearly unitary.}$$

It is well known that the spectrum of A is precisely the unit circle, $|w|=1$, in the complex plane. We shall begin with some simple observations.

LEMMA 3. *Let* $\mathbf{b}\in l_2(\mathbb{Z})$, $\mathbf{b}=(b_j)$. *Then*

$$\int_{|w|=1}w^kd\langle P_w\mathbf{b},\mathbf{b}\rangle=\sum_jb_jb_{j+k}=R_k$$

is the so-called kth autocorrelation of $\mathbf{b}$.

Proof. Trivially,

$$\int_{|w|=1} w^k d\langle P_w \mathbf{b},\mathbf{b}\rangle = \langle \mathbf{b}, A^k \mathbf{b}\rangle = R_k,$$

as claimed. □

Given $\mathbf{b}\neq 0$, $\mathbf{b}\in l_2(\mathbb{Z})$ and a positive integer p, we now define the predictor coefficients $a_0, a_1, \cdots, a_{p-1}$ to be the values minimizing the expression

$$\sum_j |b_j + a_{p-1}b_{j-1} + a_2 b_{j-2} + \cdots + a_0 b_{j-p}|^2.$$

The corresponding predictor polynomial is $q_p(t) = t^p + t^{p-1}a_{p-1} + \cdots + a_0$. The name "predictor" comes from the fact that the minimizing coefficients are the coefficients which best predict the next future value of b_j given the previous p values $b_{j-1}, \cdots, b_{j-p}$.

THEOREM 6. *The predictor polynomial is the reflection of the denominator of the* (p,p) *Padé approximant to the function*

$$g(t) = R_p + R_{p-1}t + R_{p-2}t^2 + \cdots + R_0 t^p + R_1 t^{p+1} + \cdots$$

is the characteristic polynomial of A_p, *and is the conjugate of the denominator of the* $(p-1,p)$ *Padé approximant to* $g(t)$. *Of course, the pth predictor polynomial is also precisely the pth monic orthogonal polynomial with respect to the polynomial inner product*

$$(u,v) = \int_{|w|=1} \overline{u(w)}v(w)\langle dP_w \mathbf{b},\mathbf{b}\rangle.$$

Proof. From the proof of Theorem 1, $q_p(t)$ is the characteristic polynomial of A_p and is the pth monic orthogonal polynomial with respect to the above inner product on polynomials. The claims about being denominators of various Padé approximations follows from Theorem 4 and Lemma 3. □

It is quite clear that for $\mathbf{b}\neq 0$, none of the subspaces K_p can be invariant under A (since this would mean that the coordinates of $\mathbf{b}$ satisfy a finite difference equation which is impossible for a nonzero element of $l_2(\mathbb{Z})$). Thus, by Theorem 5, the characteristic polynomial cannot have all of its roots on the unit circle. The question remains whether it is possible for some, but not all, of A_p's eigenvalues to lie on the unit circle. The following result settles the issue.

THEOREM 7. *All roots of* $d_p(t)$ *lie strictly inside the unit circle.*

Proof. Suppose that λ is a root of $d_p(t)$ with $|\lambda| = 1$. We first show that the measure $d\langle P_w \mathbf{b},\mathbf{b}\rangle = d\mu(w)$ is then concentrated on a discrete set of the unit circle. Since λ is a root, we can write

$$d_p(t) = (t-\lambda) = (t-\lambda)q(t), \qquad \text{degree}(q) = p-1.$$

Thus

$$0 = (q, d_p) = \int_{|w|=1} \overline{q(w)}d_p(w)d\langle P_w \mathbf{b},\mathbf{b}\rangle = \int_{|w|=1} (w-\lambda)|q(t)|^2 d\mu(w)$$

since $d_p(t)$ is the pth orthogonal polynomial with respect to the above polynomial inner product and q is of degree $p-1$. Solving for λ we get

$$\lambda = \frac{\int_{|w|=1} w|q(w)|^2 d\mu(w)}{\int_{|w|=1} |q(w)|^2 d\mu(w)}.$$

Now μ is a positive measure, so this expresses λ as a convex combination of the points on the unit circle. But the unit circle is strictly convex, so this can only happen if the measure $|q(w)|^2 d\mu(w)$ is concentrated only at λ. Hence, the measure $d\mu(w)$ has support contained in the set consisting of λ and the roots of $q(w)$ on the unit circle. This set consists of p or fewer points. It cannot be p, however, since then $d_p(t)$ would have all p roots on the unit circle, which is impossible by Theorem 5 and the previous comments. Now if the set consists of $p'<p$ points, then there is a monic polynomial $r(w)$ of degree p' with

$$0 = \int_{|w|=1} \overline{r(w)} r(w)\, d\mu(w) = \int_{|w|=1} |r(w)|^2 d\mu(w) = \langle r, r\rangle.$$

By Theorem 1, this polynomial must be the p' monic orthogonal polynomial with respect to the inner product induced by $\mu(w)$, and so by Theorem 5, $K_{p'}$ is an invariant subspace for A. This is also impossible, so that none of the roots of $d_p(w)$ are on the unit circle as claimed. □

Although this result was stated for the shift operator, it remains true for arbitrary unitary operators A provided that K_p is not an invariant subspace for A. Using the notation of the previous section, we thus have proved: If $p<p^*$, and A is unitary, then the roots of $d_p(t)$ lie strictly inside the unit circle.

We now turn our attention to the computational aspects of finding the characteristic polynomials $d_p(t)$ for $p=0,1,2,\cdots$. As we have seen, $d_p(t)$ is precisely the pth monic orthogonal polynomial with respect to a certain measure over the spectrum of A. In the selfadjoint case, the polynomials are orthogonal over the real line, and so a three-term recursion can be used to compute them. In the selfadjoint case, we thus have

$$d_{p+1}(t) = (t-\alpha_p) d_p(t) - \beta_p d_{p-1}(t)$$

where

$$\alpha_p = \frac{\langle td_p, d_p\rangle}{\langle d_p, d_p\rangle}, \quad \beta_p = \frac{\langle td_{p-1}, d_{p-1}\rangle}{\langle d_{p-1}, d_{p-1}\rangle}.$$

Clearly all that is needed for this recursion are the values of $d_p(A)\mathbf{b} = \mathbf{b}_p$ which gives rise to a recursion of the form

$$\mathbf{b}_{p+1} = (A-\alpha_p I)\mathbf{b}_p - \beta_p \mathbf{b}_{p-1}, \quad \mathbf{b}_0 = \mathbf{b}, \quad \mathbf{b}_{-1} = \mathbf{0}.$$

This is precisely the Lanczos algorithm for tridiagonalizing a symmetric matrix [20], [24].

In the unitary case of linear prediction and autoregressive time series modeling, the spectrum of the operator is the unit circle and so the recursion relations for polynomial orthogonal over the circle are used. The recursion is named after Szegö who discovered them [29]. For real $l_2(\mathbb{Z})$, let $\hat{d}_p(t)$ be the reflection of $d_p(t)$; that is,

$$\hat{d}_p(t)=t^p d_p\left(\frac{1}{t}\right).$$

Then the Szegö recursions are $d_0=\hat{d}_0=1$,

$$d_{p+1}(t)=td_p(t)+k_p\hat{d}_p(t), \qquad \hat{d}_{p+1}(t)=tk_pd_p(t)+\hat{d}_p(t),$$

where

$$k_p=-\frac{\left\langle td_p,\hat{d}_p\right\rangle}{\left\langle \hat{d}_p,\hat{d}_p\right\rangle}.$$

Once again, all that is needed are the vectors $d_p(A)\mathbf{b}=\mathbf{b}_p$, so the recursion resembles

$$\mathbf{b}_{p+1}=A\mathbf{b}_p+k_p\hat{\mathbf{b}}_p, \quad \hat{\mathbf{b}}_{p+1}=k_pA\mathbf{b}_p+\hat{\mathbf{b}}_p, \quad k_p=-\frac{\left\langle A\mathbf{b}_p,\hat{\mathbf{b}}_p\right\rangle}{\left\langle \hat{\mathbf{b}}_p,\hat{\mathbf{b}}_p\right\rangle}.$$

This recursion is described in greater detail in [8], while the use of the recurrence relations for polynomials orthogonal over the unit circle is implicitly used in [12], [21], [30] to find the polynomial $d_p(t)$ directly from the matrix of moments $M=(m_{jk})$ where $m_{jk}=\int w^{j-k}d\mu(w)$. Note that the case for selfadjoint operators leads to moment matrices which are Hankel, while the unitary case arising in time series analysis leads to Toeplitz matrices. Thus the theory developed above serves to unify the computational aspects of Lanczos methods and linear prediction by showing that they are both manifestations of the same basic construction. The relations with Padé approximations, although not fully exploited yet in the selfadjoint case but which are known, appear to be new for the unitary case.

5. Summary. In this paper, certain relationships between Rayleigh-Ritz-Krylov-Lanczos methods, Padé approximation, and autoregressive time series modeling have been made explicit. This serves to unify the constructions arising in these areas and explains the similarity exhibited by numerous algorithms used in these areas. In cases where H is a finite dimensional Hilbert space, the moment generating functions used in this paper are in fact rational functions themselves, so that using results from the theory of Padé approximation, it is possible to describe explicitly the relationships between the poles of the moment generating function and the poles of the approximations. This direction is currently being pursued and should be forthcoming in a separate paper.

Although a heuristic description of the sense in which roots of predictor polynomials approximate eigenvalues of the shift matrix (in the finite case, the cyclical shift matrix has eigenvalues which are the roots of unity) it is still far from being precise. Work in this direction is also being pursued.

Acknowledgments. The author wishes to thank G. W. Stewart for suggesting ideas that led to this work. Both Y. Saad and P. Delsarte noted that Theorem 1 has a weaker analogue for general matrices. Valuable references and comments were supplied by P. Delsarte, Bill Gragg and the anonymous referee.

REFERENCES

[1] G. A. Baker, Jr., *Essentials of Padé Approximants*, Academic Press, New York, 1975.

[2] F. L. Bauer and A. S. Householder, *Moments and characteristic roots*, Numer. Math., 2 (1960), pp. 42–53.

[3] G. E. Box and G. M. Jenkins, *Time Series Forecasting and Control*, Holden-Day, San Francisco, 1970.

[4] R. P. Brent, F. G. Gustavson and D. Y. Y. Yun, *Fast solution of Toeplitz systems of equations and computation of Padé approximants*, J. Algorithms, 1 (1980), pp. 259–295.

[5] A. Bultheel, *On a special Laurent-Hermite interpolation problem*, in Numerische Methoden der Approximations-theorie, Birkhauser, Basel, 1982, pp. 63–79.

[6] A. Bultheel and P. Dewilde, *On the relation between Padé approximation algorithms and Levinson/Schur recursive methods*, in Signal Processing, Theories and Applications, M. Kunt and F. deCoulon, eds., North-Holland, Amsterdam, 1980, pp. 517–523.

[7] J. P. Burg, *Maximum entropy spectral analysis*, Proc. 37th Meeting of the Society of Exploratory Geophysicists, 1967.

[8] G. Cybenko, *A general orthogonalization method with applications to time series analysis and signal processing*, Math. Comp., 40 (1983), pp. 323–336.

[9] C. Davis, *The Toeplitz-Hausdorff theorem explained*, Canad. Math. Bull., 14 (1971), pp. 245–246.

[10] P. J. Davis, *Interpolation and Approximation*, Blaisdell, New York, 1963.

[11] R. E. Dubroff, *The effective autocorrelation function of maximum entropy spectra*, Proc. Inst. Electr. Engrs., 63 (1975), pp. 1622–1623.

[12] J. Durbin, *The fitting of time-series models*, Rev. Internat. Inst. Statist., 28 (1960), pp. 233–243.

[13] L. Fejér, *Über die Lage der Nullstellen von Polynomen, die aus Minimumforderungen gewisser Art entspringen*, Math. Ann., 85 (1922), pp. 41–48.

[14] L. Y. Geronimus, *Orthogonal Polynomials*, Consultants Bureau, New York, 1961.

[15] W. B. Gragg, *Laurent, Fourier and Chebyshev-Padé tables*, in Padé and Rational Approximation, Academic Press, New York, 1977.

[16] ———, *The Padé table and its relation to certain algorithms of numerical analysis*, SIAM Rev., 14 (1972), pp. 1–62.

[17] P. R. Graves-Morris, *The numerical calculation of Padé approximants*, in Padé Approximation and Its Applications, L. Wuytack, ed., Lecture Notes in Mathematics, 765, Springer-Verlag, Berlin, Heidelberg, New York, 1979, pp. 231–245.

[18] A. S. Householder, *Moments and characteristic roots*, II, Numer. Math., 11 (1968), pp. 126–128.

[19] A. N. Krylov, *On the numerical solution of equations which in technical questions are determined by the frequency of small vibrations of material systems*, Izv. Akad. Nauk SSSR, Otd. Mat. Estest., 48 (1931), pp. 491–539. (In Russian.)

[20] C. Lanczos, *An iteration method for the solution of the eigenvalue problem for linear differential and integral operators*, J. Res. Nat. Bur. Standards Sect. B, 45 (1950), pp. 225–280.

[21] N. Levinson, *The Wiener* RMS (*root mean square*) *error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.

[22] R. J. McEliece and J. B. Shearer, *A property of Euclid's algorithm and an application to Padé approximation*, SIAM J. Appl. Math., 34 (1978), pp. 611–617.

[23] J. Makhoul, *Linear prediction: A tutorial review*, Proc. Inst. Electr. Engrs., 63 (1975), pp. 561–580.

[24] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[25] Lord Rayleigh (J. W. Strutt), *On the calculation of the frequency of vibration of a system in its gravest mode, with an example from hydrodynamics*, Philos. Mag., 47 (1899), pp. 556–572.

[26] W. Ritz, *Über eine neue Method zur Lösung Gewisser Variationsprobleme der Mathematischen Physik*, J. Reine Angew. Math., 135 (1909), pp. 1–61.

[27] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1973.

[28] H. Rutishauser, *Der Quotienten-Differenzen-Algorithms*, Mitt. Inst. Angew. Math., Birkhauser Verlag, Basel, 1957.

[29] G. Szegö, *Orthogonal Polynomials*, AMS Colloquium Publications, XXIII, American Mathematical Society, Providence, RI, 1959.

[30] W. F. Trench, *An algorithm for the inversion of finite Toeplitz matrices*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 515–522.

[31] K. Yosida, *Functional Analysis*, Springer, New York, 1971.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics
013

PRODUCT FORMULAS OF WATSON, BAILEY AND BATEMAN TYPES AND POSITIVITY OF THE POISSON KERNEL FOR q-RACAH POLYNOMIALS*

GEORGE GASPER† AND MIZAN RAHMAN‡

Abstract. A new method is introduced for proving certain important formulas due to Watson, Bailey and Bateman for products of ${}_2F_1$ hypergeometric series, and it is used to extend these formulas to products of ${}_4\phi_3$ basic hypergeometric series. The ${}_4\phi_3$ analogue of Watson's product formula is used to give conditions under which the Poisson kernels for q-Racah polynomials, q-Hahn polynomials and little q-Jacobi polynomials are positive. A transformation formula for a certain ${}_4\phi_3$ series and expansion formulas for basic hypergeometric series are also derived.

1. Introduction. Among the most useful product formulas for hypergeometric functions are Watson's formula [22], expressing the product of two terminating hypergeometric functions in terms of an F_4 Appell function

$$(1.1)\quad {}_2F_1\left[\begin{matrix}-n,n+a\\ c\end{matrix};z\right]{}_2F_1\left[\begin{matrix}-n,n+a\\ c\end{matrix};Z\right]$$
$$=\frac{(-1)^n(1+a-c)_n}{(c)_n}F_4[-n,n+a;c,1+a-c;zZ,(1-z)(1-Z)],$$

Bailey's formula [3], [4, p. 81]

$$(1.2)\quad {}_2F_1\left[\begin{matrix}a,b\\ c\end{matrix};z\right]{}_2F_1\left[\begin{matrix}a,b\\ 1+a+b-c\end{matrix};Z\right]=F_4[a,b;c,1+a+b-c;z(1-Z),Z(1-z)],$$

and Bateman's formula [5, p. 392]

$$(1.3)\quad {}_2F_1\left[\begin{matrix}-n,n+a+b+1\\ a+1\end{matrix};z\right]{}_2F_1\left[\begin{matrix}-n,n+a+b+1\\ a+1\end{matrix};Z\right]$$
$$=\frac{(-1)^n(b+1)_n}{(a+1)_n}\sum_{k=0}^{n}\frac{(-n)_k(n+a+b+1)_k}{k!(b+1)_k}(1-z-Z)^k$$
$$\cdot\,{}_2F_1\left[\begin{matrix}-k,k+a+b+1\\ a+1\end{matrix};\frac{zZ}{1-z-Z}\right],$$

where $(a)_n=a(a+1)\cdots(a+n-1)$ for $n\geq 1$ and $(a)_0=1$. Bailey [3] showed that (1.1) follows easily from (1.2) and used (1.1) to derive an F_4 representation for the Poisson kernel for the Jacobi polynomials

$$(1.4)\qquad P_n^{(\alpha,\beta)}(x)=\frac{(\alpha+1)_n}{n!}\,{}_2F_1\left[\begin{matrix}-n,n+\alpha+\beta+1\\ \alpha+1\end{matrix};\frac{1-x}{2}\right],$$

which was the first known formula to give the positivity of this kernel for $\alpha,\beta>-1$. Watson [23, pp. 372, 413] used (1.1) to derive an integral representation for a certain

*Received by the editors May 26, 1982, and in revised form March 9, 1983.

†Department of Mathematics, Northwestern University, Evanston, Illinois 60201. The research of this author was supported in part by the National Science Foundation under grant MCS-8002507 A01.

‡Department of Mathematics, Carleton University, Ottawa, Ontario,Canada K1S 5B6. The research of this author was supported in part by the Natural Science and Engineering Research Council under grant A6197.

sum of a triple product of hypergeometric functions, which was the main tool used by Gasper [6], [7] to derive an integral representation for the product of Jacobi polynomials and a convolution structure with a positive kernel for Jacobi expansions. Bateman's formula (1.3) was used by Koornwinder in [11] to give a relatively simple proof of his integral representation for the product of Jacobi polynomials. In view of these and other applications, one is naturally led to look for generalizations of these product formulas which are applicable to other orthogonal polynomials. It was in order to prove the positivity of a discrete Poisson kernel and a kernel in a projection formula for Hahn polynomials

$$(1.5)\qquad Q_n(x;\alpha,\beta,M)={}_3F_2\left[\begin{matrix}-n,n+\alpha+\beta+1,-x\\ \alpha+1,-M\end{matrix};1\right]$$

that Gasper [8], [9] showed how (1.1) and (1.2) could be extended to ${}_3F_2$ series. Extensions of (1.3) to ${}_3F_2$ series and of (1.1) to ${}_4F_3$ series were given by Rahman [13], [14], [15], and extensions of (1.1) and (1.2) to ${}_2\phi_1$ and ${}_3\phi_2$ basic hypergeometric series were recently given by Verma and Jain [21].

Here, as usual, an ${}_{r+1}\phi_r$ basic hypergeometric series in a base q is defined by

$$(1.6)\qquad {}_{r+1}\phi_r\left[\begin{matrix}a_1,\cdots,a_{r+1}\\ b_1,\cdots,b_r\end{matrix};z\right]={}_{r+1}\phi_r\left[\begin{matrix}a_1,\cdots,a_{r+1}\\ b_1,\cdots,b_r\end{matrix};q,z\right]$$
$$=\sum_{n=0}^{\infty}\frac{(a_1;q)_n\cdots(a_{r+1};q)_n}{(b_1;q)_n\cdots(b_r;q)_n}\frac{z^n}{(q;q)_n},$$

whenever it converges, where

$$(1.7)\qquad (a;q)_n=\begin{cases}1, & n=0,\\ (1-a)(1-aq)\cdots(1-aq^{n-1}), & n=1,2,\cdots.\end{cases}$$

The ${}_{r+1}\phi_r$ series in (1.6) is said to be balanced if $b_1b_2\cdots b_r=qa_1a_2\cdots a_{r+1}$ and $z=q$, well-poised if $qa_1=b_1a_2=b_2a_3=\cdots=b_ra_{r+1}$, and to be very well-poised if it is well-poised and $a_2=q\sqrt{a_1}=-a_3$, $b_1=\sqrt{a_1}=-b_2$ where the same value of the square root is used throughout. Since from now on we will be only dealing with basic hypergeometric series, to simplify the notation we will write $(a)_n$ in place of $(a;q)_n$. The Pochhammer symbols used in (1.1), (1.3) and (1.4), which indicate the usual shifted factorials, are not to be confused with the basic $(a)_n$'s used throughout the rest of the paper.

We shall say that a formula is of Watson, Bailey or Bateman type if it has formula (1.1), (1.2) or (1.3), respectively, as a special or limiting case. In this paper we introduce a new method for proving (1.1) and (1.2) and use it to derive very general extensions of them to products of ${}_4\phi_3$ series which are applicable to the q-Racah polynomials [2]

$$(1.8)\qquad W_n(x;q)\equiv W_n(x;a,b,c,M;q)$$
$$={}_4\phi_3\left[\begin{matrix}q^{-n},abq^{n+1},q^{-x},cq^{x-M}\\ aq,q^{-M},bcq\end{matrix};q\right],$$

$n=0,1,\cdots,M$. Note that this ${}_4\phi_3$ series is balanced and that $W_n(x;a,b,c,M;q)$ is a polynomial of degree n in the variable $\mu(x)=q^{-x}+cq^{x-M}$. Askey and Wilson [2] showed (in a slightly different notation) that these polynomials satisfy the orthogonality relation

$$(1.9)\qquad \sum_{x=0}^{M} \rho(x;q)W_m(x;q)W_n(x;q)=\frac{\delta_{m,n}}{h_n(q)},$$

where

$$(1.10)\qquad \rho(x;q)\equiv\rho(x;a,b,c,M;q)$$
$$=\frac{(cq^{-M})_x(1-cq^{2x-M})(aq)_x(bcq)_x(q^{-M})_x(abq)^{-x}}{(q)_x(1-cq^{-M})(ca^{-1}q^{-M})_x(b^{-1}q^{-M})_x(cq)_x},$$

$$(1.11)\qquad h_n(q)\equiv h_n(a,b,c,M;q)$$
$$=\frac{(abq)_n(1-abq^{2n+1})(aq)_n(bcq)_n(q^{-M})_n(bq)_M(ac^{-1}q)_M(c^{-1}q^M)^n}{(q)_n(1-abq)(bq)_n(ac^{-1}q)_n(abq^{M+2})_n(abq^2)_M(c^{-1})_M},$$

and it is assumed that a,b,c are real numbers such that $\rho(x;q)$ is nonnegative for $x=0,1,\cdots,M$.

In addition to the Racah polynomials [12], [2]

$$(1.12)\qquad W_n(x;\alpha,\beta,\gamma,M)=\lim_{q\to 1} W_n(x;q^\alpha,q^\beta,q^\gamma,M;q)$$
$$={}_4F_3\left[\begin{matrix}-n,n+\alpha+\beta+1,-x,x+\gamma-M\\ \alpha+1,-M,\beta+\gamma+1\end{matrix};1\right],$$

the q-Racah polynomials also contain as limit cases the classical orthogonal polynomials (Jacobi, Laguerre and Hermite polynomials), their discrete analogues (Hahn, Meixner, Krawtchouk and Charlier polynomials) and their q-analogues.

Our extension of (1.1) (and of (1.2)) to the product of ${}_4\phi_3$ series, which is more general than that recently given in Rahman [17], is derived in §2 and then used in §3 to give conditions under which the Poisson kernel

$$(1.13)\qquad \sum_{n=0}^{M} t^n h_n(q)W_n(x;q)W_n(y;q),\qquad 0\le t<1,$$

and the so-called discrete Poisson kernel

$$(1.14)\qquad \sum_{n=0}^{z} \frac{(q^{-z})_n}{(q^{-M})_n} h_n(q)W_n(x;q)W_n(y;q),\qquad z=0,1,\cdots,M,$$

are nonnegative for $x,y=0,1,\cdots,M$. Additional bilinear sums are considered in §4, and extensions of Bateman's formula (1.3) are derived in §5. This work has also led to a new transformation for ${}_4\phi_3$ series derived in §3 and to the expansion formulas derived in §5.

2. Product formulas of Watson and Bailey types. Our method originates with the observation that since

$$(q^{-x})_j(aq^x)_j=\prod_{k=0}^{j-1}\left(1+aq^{2k}-q^k(q^{-x}+aq^x)\right)$$

is a polynomial of (exact) degree j in powers of the variable $q^{-x}+aq^{x}$, there must exist an expression of the form

$$(2.1)\qquad (q^{-x})_j(aq^x)_j(q^{-x})_k(aq^x)_k=\sum_{m=0}^{j+k} A_m(j,k,a;q)(q^{-x})_m(aq^x)_m,$$

and hence, replacing x by a nonnegative integer n,

$$(2.2)\qquad (q^{-n})_j(aq^n)_j(q^{-n})_k(aq^n)_k=\sum_{m=0}^{j+k} A_m(j,k,a;q)(q^{-n})_m(aq^n)_m.$$

To compute the coefficients it suffices to observe from the formula [19, p. 247]

$$(2.3)\qquad {}_3\phi_2\left[\begin{matrix} q^{-n},a,b \\ c,abc^{-1}q^{1-n} \end{matrix};q\right]=\frac{(c/a)_n(c/b)_n}{(c)_n(c/ab)_n}$$

that

$${}_3\phi_2\left[\begin{matrix} q^{-j},q^{k-x},aq^{k+x} \\ aq^k,q^{1+k-j} \end{matrix};q\right]=\frac{(q^{-x})_j(aq^x)_j}{(q^{-k})_j(aq^k)_j}$$

for $k\geq j$, and hence

$$\begin{aligned}
&(q^{-x})_j(aq^x)_j(q^{-x})_k(aq^x)_k\\
&\quad=(q^{-k})_j(aq^k)_j(q^{-x})_k(aq^x)_k\sum_{m=0}^{j}\frac{(q^{-j})_m(q^{k-x})_m(aq^{k+x})_m}{(q)_m(aq^k)_m(q^{1+k-j})_m}q^m\\
&\quad=(q^{-k})_j(a)_{j+k}\sum_{m=k}^{j+k}\frac{(q^{-j})_{m-k}(q^{-x})_m(aq^x)_mq^{m-k}}{(q)_{m-k}(a)_m(q^{1+k-j})_{m-k}}\\
&\quad=(q)_j(q)_k(a)_{j+k}q^{\binom{j}{2}+\binom{k}{2}}\sum_{m=k}^{j+k}\frac{(-1)^{j+k+m}(q^{-x})_m(aq^x)_mq^{\binom{m+1}{2}-m(j+k)}}{(q)_{m-j}(q)_{m-k}(q)_{j+k-m}(a)_m},
\end{aligned}$$

where we used the standard identities in [19, App. II] and the notation $\binom{j}{2}=j(j-1)/2$. Thus both (2.1) and (2.2) hold for $j,k=0,1,\cdots$ with

$$(2.4)\qquad A_m(j,k,a;q)=\frac{(q)_j(q)_k(a)_{j+k}(-1)^{j+k+m}}{(q)_{m-j}(q)_{m-k}(q)_{j+k-m}(a)_m}q^{\binom{j}{2}+\binom{k}{2}+\binom{m+1}{2}-m(j+k)}.$$

Observe that since $(q^{-n})_m=0$ for $m>n$ and $1/(a)_{m-j}=0$ for $j>m$, the terms in the sum in (2.2) are zero unless $\max(j,k)\leq m\leq\min(j+k,n)$. Hence, using (2.2) and (2.4), we find for $n=0,1,\cdots$ that

$$\begin{aligned}
(2.5)\quad &{}_4\phi_3\left[\begin{matrix} q^{-n},aq^n,b,c \\ d,e,f \end{matrix};q\right]{}_4\phi_3\left[\begin{matrix} q^{-n},aq^n,g,h \\ u,v,w, \end{matrix};q\right]\\
&\quad=\sum_{j,k,m}\frac{(-1)^{j+k+m}(q^{-n})_m(aq^n)_m(a)_{j+k}(b)_j(c)_j(g)_k(h)_k}{(q)_{m-j}(q)_{m-k}(q)_{j+k-m}(a)_m(d)_j(e)_j(f)_j(u)_k(v)_k(w)_k}\\
&\qquad\cdot q^{\binom{j}{2}+\binom{k}{2}+\binom{m+1}{2}+(1-m)(j+k)},
\end{aligned}$$

where the triple sum is taken over all (finitely many) nonzero terms, i.e. over all (nonnegative) integer values of j,k,m such that $\max(j,k)\le m\le \min(j+k,n)$. Now let $j=m-r$ to see that

$$(2.6)\qquad \sum_j \frac{(-1)^j(a)_{j+k}(b)_j(c)_j q^{\binom{j}{2}+(1-m)j}}{(q)_{m-j}(q)_{j+k-m}(d)_j(e)_j(f)_j}$$

$$=\sum_r \frac{(-1)^{m-r}(a)_{m+k-r}(b)_{m-r}(c)_{m-r}q^{\binom{m-r}{2}+(1-m)(m-r)}}{(q)_r(q)_{k-r}(d)_{m-r}(e)_{m-r}(f)_{m-r}}$$

$$=\frac{(-1)^m(a)_{m+k}(b)_m(c)_m q^{-\binom{m}{2}}}{(q)_k(d)_m(e)_m(f)_m}$$

$$\cdot {}_4\phi_3\left[\begin{matrix} q^{-k},d^{-1}q^{1-m},e^{-1}q^{1-m},f^{-1}q^{1-m} \\ a^{-1}q^{1-m-k},b^{-1}q^{1-m},c^{-1}q^{1-m}\end{matrix};\frac{def}{abc}\right].$$

If the first ${}_4\phi_3$ in (2.5) is balanced, i.e. if $abcq=def$, then the ${}_4\phi_3$ in (2.6) is also balanced and we can apply Watson's transformation formula [19, (3.4.1.5)] to find that the ${}_4\phi_3$ in (2.6) equals

$$\frac{(ae^{-1}q)_k(af^{-1}q)_k}{(ae^{-1}f^{-1}q^{2-m})_k(aq^m)_k}$$

$$\cdot {}_8\phi_7\left[\begin{matrix} db^{-1}c^{-1}q^{-m},q\sqrt{\ },-q\sqrt{\ },db^{-1},dc^{-1},e^{-1}q^{1-m},f^{-1}q^{1-m},q^{-k} \\ \sqrt{\ },-\sqrt{\ },c^{-1}q^{1-m},b^{-1}q^{1-m},af^{-1}q,ae^{-1}q,ae^{-1}f^{-1}q^{2-m+k}\end{matrix};ad^{-1}q^{k+1}\right],$$

where, as elsewhere, the square root denotes one of the square roots of the first numerator parameter a_1 in the ${}_8\phi_7$. Using this and (2.6) in (2.5) with the above ${}_8\phi_7$ having r as its summation index, setting $k=r+j$ and $m=r+s$, changing the order of summation and simplifying, we obtain the rather general product formula

$$(2.7)\qquad {}_4\phi_3\left[\begin{matrix} q^{-n},aq^n,b,c \\ d,e,\dfrac{abcq}{de}\end{matrix};q\right]{}_4\phi_3\left[\begin{matrix} q^{-n},aq^n,g,h \\ u,v,w\end{matrix};q\right]$$

$$=\sum_{r=0}^{n}\sum_{s=0}^{n-r}\frac{(q^{-n})_{r+s}(aq^n)_{r+s}(b)_s(c)_s(g)_r(h)_r(d/b)_r}{(q)_r(q)_s(d)_{r+s}(e)_s(abcq/de)_s(u)_r(v)_r(w)_r}$$

$$\cdot\frac{(d/c)_r}{(db^{-1}c^{-1}q^{1-s})_r}\frac{1-db^{-1}c^{-1}q^{r-s}}{1-db^{-1}c^{-1}q^{-s}}q^{r+s-rs}$$

$$\cdot {}_5\phi_4\left[\begin{matrix} q^{-s},gq^r,hq^r,ae^{-1}q^{r+1},deb^{-1}c^{-1}q^r \\ uq^r,vq^r,wq^r,db^{-1}c^{-1}q^{1+r-s}\end{matrix};q\right].$$

If we also assume that the second ${}_4\phi_3$ in (2.7) is balanced, i.e. if $aghq=uvw$, then the above ${}_5\phi_4$ is balanced and it can be summed by (2.3) whenever it reduces to a ${}_3\phi_2$

series. For instance, setting $v=aq/e$ and $w=de/bc$ in (2.7) yields the product formula

$$
(2.8)\qquad {}_4\Phi_3\left[\begin{matrix} q^{-n},aq^n,b,c \\ d,e,\dfrac{abcq}{de} \end{matrix};q\right] {}_4\Phi_3\left[\begin{matrix} q^{-n},aq^n,g,h \\ \dfrac{bcgh}{d},\dfrac{aq}{e},\dfrac{de}{bc} \end{matrix};q\right]
$$

$$
=\sum_{r=0}^{n}\sum_{s=0}^{n-r}\frac{(q^{-n})_{r+s}(aq^n)_{r+s}(g)_r(h)_r(d/b)_r(d/c)_r(b)_s(c)_s}{(q)_r(q)_s(d)_{r+s}(bcgh/d)_{r+s}(aq/e)_r(dq/bc)_r(de/bc)_r}
$$

$$
\cdot\frac{(bcg/d)_s(bch/d)_s}{(e)_s(bc/d)_s(abcq/de)_s}\cdot\frac{1-db^{-1}c^{-1}q^{r-s}}{1-db^{-1}c^{-1}q^{-s}}q^{r+s}.
$$

If $b=q^{-x}$ and $g=q^{-y}$ with $x,y=0,1,\cdots$, then (2.7) and (2.8) also hold even if n is not a nonnegative integer, and so by relabeling the parameters we obtain the following generalizations of Bailey's product formula (1.2) (and of its discrete analogue in Gasper [9,(1)]).

$$
(2.9)\qquad {}_4\Phi_3\left[\begin{matrix} a,b,q^{-x},cq^x \\ d,e,\dfrac{abcq}{de} \end{matrix};q\right] {}_4\Phi_3\left[\begin{matrix} a,b,q^{-y},hq^y \\ u,v,w \end{matrix};q\right]
$$

$$
=\sum_{r=0}^{y}\sum_{s=0}^{x}\frac{(a)_{r+s}(b)_{r+s}(q^{-x})_s(cq^x)_s(q^{-y})_r(hq^y)_r}{(q)_r(q)_s(d)_{r+s}(e)_s(abcq/de)_s(u)_r(v)_r}
$$

$$
\cdot\frac{(dq^x)_r(dc^{-1}q^{-x})_r}{(w)_r(dc^{-1}q^{1-s})_r}\cdot\frac{1-dc^{-1}q^{r-s}}{1-dc^{-1}q^{-s}}q^{r+s-rs}
$$

$$
\cdot{}_5\Phi_4\left[\begin{matrix} q^{-s},q^{r-y}hq^{r+y},abe^{-1}q^{r+1},dec^{-1}q^r \\ uq^r,vq^r,wq^r,dc^{-1}q^{1+r-s} \end{matrix};q\right],
$$

$$
(2.10)\qquad {}_4\Phi_3\left[\begin{matrix} a,b,q^{-x},cq^x \\ d,e,\dfrac{abcq}{de} \end{matrix};q\right] {}_4\Phi_3\left[\begin{matrix} a,b,q^{-y},hq^y \\ \dfrac{ch}{d},\dfrac{abq}{e},\dfrac{de}{c} \end{matrix};q\right]
$$

$$
=\sum_{r=0}^{y}\sum_{s=0}^{x}\frac{(a)_{r+s}(b)_{r+s}(q^{-y})_r(hq^y)_r(dq^x)_r(dc^{-1}q^{-x})_r}{(q)_r(q)_s(d)_{r+s}(ch/d)_{r+s}(abq/e)_r(dq/c)_r}
$$

$$
\cdot\frac{(q^{-x})_s(cq^x)_s(cd^{-1}q^{-y})_s(chd^{-1}q^y)_s}{(de/c)_r(e)_s(c/d)_s(abcq/de)_s}\cdot\frac{1-dc^{-1}q^{r-s}}{1-dc^{-1}q^{-s}}q^{r+s},
$$

where $x,y=0,1,\cdots$. Formula (3.9) in Verma and Jain [21] is a limit case of (2.10).

Additional product formulas can be derived by applying Sears' transformation formula [18]

$$
(2.11)\qquad {}_4\Phi_3\left[\begin{matrix} q^{-n},a,b,c \\ d,e,f \end{matrix};q\right]=\frac{(de/bc)_n(df/bc)_n}{(f)_n(e)_n}\left(\frac{bc}{d}\right)^n {}_4\Phi_3\left[\begin{matrix} q^{-n},a,\dfrac{d}{b},\dfrac{d}{c} \\ d,\dfrac{de}{bc},\dfrac{df}{bc} \end{matrix};q\right],
$$

where $abcq^{1-n}=def$, to the balanced ${}_4\phi_3$ series in (2.7)–(2.10). In particular, application of (2.11) to the first ${}_4\phi_3$ in (2.7) and the substitutions $b\to d/b$, $c\to d/c$, $e\to aq/e$ yields the following generalization of Watson's formula (1.1) (and of the product formulas [8,(2.2)], [21,(3.1)], [21,(3.2)], [15,(1.9)], [17,(4.10)])

(2.12)

$$
\begin{aligned}
&{}_4\Phi_3\left[\begin{matrix} q^{-n},aq^n,b,c \\ d,e,\dfrac{abcq}{de} \end{matrix};q\right] {}_4\Phi_3\left[\begin{matrix} q^{-n},aq^n,g,h \\ u,v,w \end{matrix};q\right] \\
&\quad=\frac{(aq/e)_n(de/bc)_n}{(e)_n(abcq/de)_n}\left(\frac{bc}{d}\right)^n \sum_{r=0}^{n}\sum_{s=0}^{n-r} \frac{(q^{-n})_{r+s}(aq^n)_{r+s}(d/b)_s(d/c)_s}{(q)_r(q)_s(d)_{r+s}(aq/e)_s(de/bc)_s} \\
&\qquad\cdot\frac{(g)_r(h)_r(b)_r(c)_r}{(u)_r(v)_r(w)_r(bcd^{-1}q^{1-s})_r}\cdot\frac{1-bcd^{-1}q^{r-s}}{1-bcd^{-1}q^{-s}}q^{r+s-rs} \\
&\qquad\cdot{}_5\Phi_4\left[\begin{matrix} q^{-s},gq^r,hq^r,eq^r,abcd^{-1}e^{-1}q^{r+1} \\ uq^r,vq^r,wq^r,bcd^{-1}q^{1+r-s} \end{matrix};q\right].
\end{aligned}
$$

Similarly, (2.8) gives the double-sum Watson type formula

(2.13)

$$
\begin{aligned}
&{}_4\Phi_3\left[\begin{matrix} q^{-n},aq^n,b,c \\ d,e,\dfrac{abcq}{de} \end{matrix};q\right] {}_4\Phi_3\left[\begin{matrix} q^{-n},aq^n,g,h \\ e,\dfrac{dgh}{bc},\dfrac{abcq}{de} \end{matrix};q\right] \\
&\quad=\frac{(aq/e)_n(de/bc)_n}{(e)_n(abcq/de)_n}\left(\frac{bc}{d}\right)^n \sum_{r=0}^{n}\sum_{s=0}^{n-r} \frac{(q^{-n})_{r+s}(aq^n)_{r+s}(g)_r(h)_r(b)_r}{(q)_r(q)_s(d)_{r+s}(dgh/bc)_{r+s}(e)_r} \\
&\qquad\cdot\frac{(c)_r(d/b)_s(d/c)_s(dg/bc)_s(dh/bc)_s}{(bcq/d)_r(abcq/de)_r(aq/e)_s(d/bc)_s(de/bc)_s}\cdot\frac{1-bcd^{-1}q^{r-s}}{1-bcd^{-1}q^{-s}}q^{r+s},
\end{aligned}
$$

which is equivalent to formula (4.10) in Rahman [17] and yields the product formulas for q-Racah and q-Wilson polynomials in [17, §5].

An extension of formulas (16) and (17) in [9] to a product of ${}_4\phi_3$ series can be derived from (2.9) as follows. Let $v=abq/e$ and $w=eh/u$ so that the ${}_5\phi_4$ series in (2.9) reduces to a balanced ${}_4\phi_3$ series which, by Watson's transformation formula [19,(3.4.1.5)], equals

$$
\begin{aligned}
&\frac{(dc^{-1}h^{-1}q^{1-y-s})_s(dc^{-1}q^{1+y-s})_s}{(dc^{-1}h^{-1}q^{1-r-s})_s(dc^{-1}q^{1+r-s})_s} \\
&\qquad\cdot{}_8\Phi_7\left[\begin{matrix} chd^{-1}q^{r-1},q\sqrt{\ },-q\sqrt{\ },\dfrac{ch}{du},\dfrac{cu}{de},q^{r-y},hq^{r+y},q^{-s} \\ \sqrt{\ },-\sqrt{\ },uq^r,heu^{-1}q^r,chd^{-1}q^y,cd^{-1}q^{-y},chd^{-1}q^{r+s} \end{matrix};eq^s\right].
\end{aligned}
$$

Then, using this in (2.9) with j as the index of summation in the above ${}_8\phi_7$, replacing s by $s+j$, and changing the order of summation, we obtain the desired expansion formula

(2.14)

$$
{}_4\Phi_3\left[\begin{matrix} a,b,q^{-x},cq^{x} \\ d,e,\dfrac{abcq}{de} \end{matrix};q\right]{}_4\Phi_4\left[\begin{matrix} a,b,q^{-y},hq^{y} \\ u,\dfrac{abq}{e},\dfrac{he}{u} \end{matrix};q\right]
$$

$$
=\sum_{j=0}^{\min(x,y)}\frac{(a)_j(b)_j(ch/du)_j(cu/de)_j(ch/d)_j(q^{-x})_j(cq^{x})_j}{(q)_j(d)_j(e)_j(u)_j(eh/u)_j(c/d)_j(abcq/de)_j}
$$

$$
\cdot\frac{(q^{-y})_j(hq^{y})_j}{(ch/d)_{2j}}(-eq)^j q^{\binom{j}{2}}\sum_{r=0}^{y-j}\sum_{s=0}^{x-j}\frac{(aq^j)_{r+s}(bq^j)_{r+s}(dq^x)_r}{(q)_r(q)_s(dq^j)_{r+s}}
$$

$$
\cdot\frac{(dc^{-1}q^{-x})_r(q^{j-y})_r(hq^{j+y})_r(chd^{-1}q^j)_r(q^{j-x})_s(cq^{j+x})_s}{(chd^{-1}q^{2j})_{r+s}(uq^j)_r(abq/e)_r(dq/c)_r(ehu^{-1}q^j)_r(eq^j)_s}
$$

$$
\cdot\frac{(chd^{-1}q^{j+y})_s(cd^{-1}q^{j-y})_s(1-dc^{-1}q^{r-j-s})(1-chd^{-1}q^{r-1+2j})}{(cd^{-1}q^j)_s(abcd^{-1}e^{-1}q^{1+j})_s(1-dc^{-1}q^{-j-s})(1-chd^{-1}q^{r-1+j})}q^{r+s},
$$

$x,y=0,1,2,\cdots$, which reduces to a double sum whenever $u=de/c$ or $u=ch/d$ (giving (2.10)).

These formulas yield such a large collection of product formulas for the q-Racah polynomials that we shall not present them here. In addition, because of the (essential) duality of the q-Racah polynomials in n and x, duals of these formulas follow, e.g., by replacing n by x in (2.7), (2.8), (2.12), (2.13) and x and y by n and m in (2.9) and (2.10), and choosing the other parameters appropriately.

3. Generalized Poisson kernels. In this section we shall start out by considering the following generalization of the Poisson kernel (1.13) for q-Racah polynomials

(3.1)

$$
P_z(x,y)=P_z(x,y;a,b,c,\alpha,\gamma,K,M,N;q)
$$

$$
=\sum_{n=0}^{z}\frac{(q^{-z})_n}{(q^{-K})_n}h_n(a,b,c,M;q)\,W_n(x;a,b,c,M;q)\,W_n\left(y;\alpha,\frac{ab}{\alpha},\gamma,N;q\right),
$$

where $z=0,1,\cdots,\ \min(K,M)$ and $M\leq N$. If $\alpha=a$, $\gamma=c$ and $N=M$, then (3.1) reduces to the discrete Poisson kernel (1.14) when $K=M$ and it has the Poisson kernel (1.13) as a limit case. From the Watson type product formula (2.12) it follows that

(3.2)

$$
W_n(x;a,b,c,M;q)W_n\left(y;\alpha,\frac{ab}{\alpha},\gamma,N;q\right)
$$

$$
=\frac{(bq)_n(aq/c)_n}{(aq)_n(bcq)_n}c^n\sum_{r=0}^{n}\sum_{s=0}^{n-r}\frac{(q^{-n})_{r+s}(abq^{n+1})_{r+s}(q^{-x})_r(cq^{x-M})_r}{(q)_r(q)_s(q^{-M})_{r+s}(\alpha q)_r(q^{-N})_r}
$$

$$
\cdot\frac{(q^{-y})_r(\gamma q^{y-N})_r(q^{x-M})_s(c^{-1}q^{-x})_s}{(ab\gamma q/\alpha)_r(bq)_s(aq/c)_s(cq^{1-s})_r}\cdot\frac{1-cq^{r-s}}{1-cq^{-s}}A_{r,s}q^{r+s-rs},
$$

with

$$(3.3)\qquad A_{r,s}={}_5\phi_4\left[\begin{matrix}q^{-s},q^{r-y},\gamma q^{r+y-N},aq^{r+1},bcq^{r+1}\\ \alpha q^{r+1},q^{r-N},ab\gamma\alpha^{-1}q^{r+1},cq^{1+r-s}\end{matrix};q\right].$$

By using (3.2) in (3.1) and changing the order of summation we find thaı

(3.4)

$$P_z(x,y)=\frac{(bq)_M(aq/c)_M}{(abq^2)_M(c^{-1})_M}\sum_{r=0}^{z}\sum_{s=0}^{z-r}\frac{(q^{-z})_{r+s}(abq^2)_{2r+2s}}{(q)_r(q)_s(q^{-K})_{r+s}(abq^{M+2})_{r+s}}$$

$$\cdot\frac{(q^{-x})_r(cq^{x-M})_r(q^{-y})_r(\gamma q^{y-N})_r(q^{x-M})_s(c^{-1}q^{-x})_s}{(\alpha q)_r(q^{-N})_r(ab\gamma q/\alpha)_r(bq)_s(aq/c)_s(cq^{1-s})_r}\cdot\frac{1-cq^{r-s}}{1-cq^{-s}}$$

$$\cdot(-1)^{r+s}A_{r,s}B_{r,s}q^{M(r+s)-rs-\binom{r+s}{2}},$$

with

$$(3.5)\qquad B_{r,s}={}_5\phi_4\left[\begin{matrix}abq^{2r+2s+1},q\sqrt{\ },-q\sqrt{\ },q^{r+s-M},q^{r+s-z}\\ \sqrt{\ },-\sqrt{\ },abq^{M+2+r+s},q^{r+s-K}\end{matrix};q^{M-r-s}\right].$$

(A similar formula holds with the term $1-abq^{2n+1}$ in $h_n(a,b,c,M;q)$ replaced by 1.)

We will now point out some special cases in which $P_z(x,y)$ is nonnegative.

Suppose $K=M$. Then the ${}_5\phi_4$ in (3.5) reduces to a nearly poised ${}_4\phi_3$ series, which by means of the transformation formula

$$(3.6)\qquad {}_4\phi_3\left[\begin{matrix}a,q\sqrt{a},-q\sqrt{a},b\\ \sqrt{a},-\sqrt{a},w\end{matrix};\frac{w}{aq}\right]=\frac{(wb/aq)_\infty(w/b)_\infty}{(w/aq)_\infty(w)_\infty}\,{}_2\phi_1\left[\begin{matrix}b,bq\\ \frac{wb}{a}\end{matrix};\frac{w}{b}\right],$$

where $(a)_\infty=(a;q)_\infty=(1-a)(1-aq)(1-aq^2)\cdots$, is equal to

$$(3.7)\qquad \frac{(q^{M-z})_{z-r-s}}{(abq^{M+2+r+s})_{z-r-s}}\,{}_2\phi_1\left[\begin{matrix}q^{r+s-z},q^{1+r+s-z}\\ q^{1+M-z}\end{matrix};abq^{M+2+z}\right],$$

and hence is clearly nonnegative when $z=0,1,\cdots,M, r+s\le z$, $0\le q<1$, $0\le abq^{M+2}<1$.

The special case of (3.6) in which b is a negative integer power of q was first proved by George Andrews at the 1980 Summer Meeting of the American Mathematical Society shortly after George Gasper told him that in view of this work such a transformation should exist. Here we shall give a proof of (3.6) which is much shorter than Andrews' proof of the terminating case. It suffices to observe that, since $1-aq^{2k}=1-q^k+q^k(1-aq^k)$ and by [1, p. 576, I2]

$$(3.8)\qquad {}_2\phi_1\left[\begin{matrix}a,b\\ c\end{matrix};t\right]=\frac{(bt)_\infty(c/b)_\infty}{(t)_\infty(c)_\infty}\,{}_2\phi_1\left[\begin{matrix}b,\frac{abt}{c}\\ bt\end{matrix};\frac{c}{b}\right],$$

the left-hand side of (3.6) equals

$$
{}_2\phi_1\left[\begin{matrix} aq,b \\ w \end{matrix};\frac{w}{a}\right]+\frac{w(1-b)}{aq(1-w)}\,{}_2\phi_1\left[\begin{matrix} aq,bq \\ wq \end{matrix};\frac{w}{aq}\right]
$$

$$
=\frac{(wb/a)_\infty(w/b)_\infty}{(w/a)_\infty(w)_\infty}\,{}_2\phi_1\left[\begin{matrix} b,bq \\ \frac{wb}{a} \end{matrix};\frac{w}{b}\right]+\frac{w(1-b)(wb/a)_\infty(w/b)_\infty}{aq(w/aq)_\infty(w)_\infty}\,{}_2\phi_1\left[\begin{matrix} b,bq \\ \frac{wb}{a} \end{matrix};\frac{w}{b}\right]
$$

$$
=\frac{(wb/aq)_\infty(w/b)_\infty}{(w/aq)_\infty(w)_\infty}\,{}_2\phi_1\left[\begin{matrix} b,bq \\ \frac{wb}{a} \end{matrix};\frac{w}{b}\right].
$$

Returning to the case $K=M$ of (3.4), we also need to consider the sign of $A_{r,s}$. If $\alpha=a$ then the ${}_5\phi_4$ series for $A_{r,s}$ in (3.3) reduces to a ${}_4\phi_3$ series which, by (2.11), equals

$$
\tag{3.9}
\frac{(q^{1+N-y-s})_s(b^{-1}\gamma^{-1}q^{-y-s})_s}{(q^{r-N})_s(b\gamma q^{r+1})_s}(b\gamma q^{r+s+y-N})^s
$$

$$
\cdot{}_4\phi_3\left[\begin{matrix} q^{-s},q^{r-y},b^{-1}q^{-s},c\gamma^{-1}q^{1+N-y-s} \\ cq^{1+r-s},q^{1+N-y-s},b^{-1}\gamma^{-1}q^{-y-s} \end{matrix};q\right].
$$

From (3.4), (3.7) and (3.9) it follows that

$$
\tag{3.10}
P_z(x,y;a,b,c,a,\gamma,M,M,N;q)\geq 0
$$

for $x=0,1,\cdots,M$, $y=0,1,\cdots,N$, $z=0,1,\cdots,M$ when $0<q<1$, $0<aq<1$, $0\leq bq<1$, $0<c<aq^M$ and $cq\leq\gamma<q^{N-1}\leq q^{M-1}$. Hence the discrete Poisson kernel (1.14) is nonnegative for $x,y,z=0,1,\cdots,M$ when $0<q<1$, $0<aq<1$, $0\leq bq<1$ and $0<c<aq^M$. Also notice that the balanced ${}_4\phi_3$ series in (3.9) reduces to a summable ${}_3\phi_2$ series when $\gamma=c$.

If in (3.1) we write the sum with M as the upper limit of summation, replace $(q^{-z})_n$ by $(tq^{-K})_n$ and let $K\to\infty$, it follows from (3.4) that

(3.11)

$$
L_t(x,y;a,b,c,\alpha,\gamma,M,N;q)
$$

$$
\equiv\sum_{n=0}^{M}t^nh_n(a,b,c,M;q)W_n(x;a,b,c,M;q)W_n\left(y;\alpha,\frac{ab}{\alpha},\gamma,N;q\right)
$$

$$
=\frac{(bq)_M(aq/c)_M}{(abq^2)_M(c^{-1})_M}\sum_{r=0}^{x}\sum_{s=0}^{M-x}\frac{(-t)^{r+s}(abq^2)_{2r+2s}(q^{-x})_r(cq^{x-M})_r}{(q)_r(q)_s(abq^{M+2})_{r+s}(\alpha q)_r(q^{-N})_r}
$$

$$
\cdot\frac{(q^{-y})_r(\gamma q^{y-N})_r(q^{x-M})_s(c^{-1}q^{-x})_s}{(ab\gamma q/\alpha)_r(bq)_s(aq/c)_s(cq^{1-s})_r}\frac{1-cq^{r-s}}{1-cq^{-s}}A_{r,s}C_{r,s}q^{M(r+s)-rs-\binom{r+s}{2}},
$$

for $x=0,1,\cdots,M$ with $A_{r,s}$ defined in (3.3) and

$$(3.12)\qquad C_{r,s}={}_4\phi_3\left[\begin{matrix}abq^{2r+2s+1},q\sqrt{\ },-q\sqrt{\ },q^{r+s-M}\\ \sqrt{\ },-\sqrt{\ },abq^{M+2+r+s}\end{matrix};tq^{M-r-s}\right].$$

In our work on the nonnegativity of the Poisson kernel for the continuous q-ultraspherical polynomials [10] we derived the transformation formula

$$(3.13)\qquad {}_4\phi_3\left[\begin{matrix}a,q\sqrt{a},-q\sqrt{a},b^{-1}\\ \sqrt{a},-\sqrt{a},abq\end{matrix};tb\right]=\frac{(t)_\infty(aq)_\infty}{(tb)_\infty(abq)_\infty}\,{}_2\phi_1\left[\begin{matrix}b,tb\\ tq\end{matrix};aq\right],$$

which turns out to be exactly what we need here to show that

$$(3.14)\quad C_{r,s}=(t)_{M-r-s}(abq^{2r+2s+2})_{M-r-s}\cdot{}_2\phi_1\left[\begin{matrix}q^{M-r-s},tq^{M-r-s}\\ tq\end{matrix};abq^{2r+2s+2}\right],$$

from which it is obvious that $C_{r,s}\geq 0$ for $0\leq t<1$, $r+s\leq M$ when $0\leq abq^2<1$. Combining this with our previous observation that $A_{r,s}$ equals the expression in (3.9) when $\alpha=a$, it follows from (3.11) that

$$(3.15)\qquad L_t(x,y;a,b,c,a,\gamma,M,N;q)>0$$

for $x=0,1,\cdots,M$, $y=0,1,\cdots,N$, $0\leq t<1$ when $0<q<1$, $0<aq<1$, $0\leq bq<1$, $0<c<aq^M$ and $cq\leq\gamma<q^{N-1}\leq q^{M-1}$. In particular, the Poisson kernel (1.13) is positive for $x,y=0,1,\cdots,M$, $0\leq t<1$ when $0<q<1$, $0<aq<1$, $0\leq bq<1$ and $0<c<aq^M$. The depth of this result can be seen from the observation that the Poisson kernel equals zero when $t=1$ and $x\neq y$.

Since $W_n(x;a,b,c,M;q)$ tends to the q-Hahn polynomial

$$(3.16)\qquad Q_n(x;a,b,M;q)={}_3\phi_2\left[\begin{matrix}q^{-n},abq^{n+1},q^{-x}\\ aq,q^{-M}\end{matrix};q\right]$$

as $c\to 0$, (3.4) and (3.11) lead to similar formulas for the q-Hahn polynomials and, in particular, our observations above lead to

(3.17)

$$\sum_{n=0}^{z}\frac{(abq)_n(1-abq^{2n+1})(aq)_n(q^{-z})_n}{(q)_n(1-abq)(bq)_n(abq^{M+2})_n}\left(\frac{-1}{a}\right)^n q^{Mn-n-\binom{n}{2}}$$
$$\cdot Q_n(x;a,b,M;q)Q_n(y;a,b,;q)$$
$$=\sum_{r=0}^{z}\sum_{s=0}^{z-r}\frac{(q^{-z})_{r+s}(abq^2)_{2r+2s}(q^{-x})_r(q^{-y})_r(q^{x-M})_s(q^{y-N})_s}{(q)_r(q)_s(q^{-M})_{r+s}(q^{-N})_{r+s}(abq^{M+2})_{r+s}(aq)_r(bq)_s}$$
$$\cdot\frac{(q^{M-z})_{z-r-s}}{(abq^{M+2+r+s})_{z-r-s}}(-1)^{r+s}a^{-s}q^{M(r+s)-(x+y+1)s-\binom{r+s}{2}}$$
$$\cdot{}_2\phi_1\left[\begin{matrix}q^{r+s-z},q^{1+r+s-z}\\ q^{1+M-z}\end{matrix};abq^{M+2+z}\right]\geq 0,$$

(3.18)

$$\sum_{n=0}^{M} \frac{(abq)_n(1-abq^{2n+1})(aq)_n(q^{-M})_n}{(q)_n(1-abq)(bq)_n(abq^{M+2})_n}\left(\frac{-t}{a}\right)^n q^{Mn-n-\binom{n}{2}}$$

$$\cdot Q_n(x;a,b,M;q)Q_n(y;a,b,N;q)$$

$$=\sum_{r=0}^{x}\sum_{s=0}^{M-x}\frac{(abq^2)_{2r+2s}(q^{-x})_r(q^{-y})_r(q^{x-M})_s(q^{y-N})_s}{(q)_r(q)_s(q^{-N})_{r+s}(abq^{M+2})_{r+s}(aq)_r(bq)_r}$$

$$\cdot(t)_{M-r-s}(abq^{2r+2s+2})_{M-r-s}(-t)^{r+s}a^{-s}q^{M(r+s)-(x+y+1)s-\binom{r+s}{2}}$$

$$\cdot{}_2\phi_1\left[\begin{matrix} q^{M-r-s}, tq^{M-r-s} \\ tq \end{matrix}; abq^{2r+2s+2}\right]>0,$$

which hold for $x=0,1,\cdots,M$, $y=0,1,\cdots,N$, $z=0,1,\cdots,M$, $0\le t<1$ when $0<q<1$, $0<aq<1$, $0\le bq<1$ and $M\le N$.

In addition, since the q-Hahn polynomial $Q_n(M-x;a,b,M;q)$ tends to the little q-Jacobi polynomial

$$P_n(q^x;a,b;q)={}_2\phi_1\left[\begin{matrix} q^{-n}, abq^{n+1} \\ aq \end{matrix}; q^{x+1}\right] \tag{3.19}$$

as $M\rightarrow\infty$, it follows from (3.18) that

$$\sum_{n=0}^{\infty}\frac{(abq)_n(1-abq^{2n+1})(aq)_n}{(q)_n(1-abq)(bq)_n}(aq)^{-n}t^nP_n(q^x;a,b;q)P_n(q^y;a,b;q) \tag{3.20}$$

$$=(t)_\infty(abq^2)_\infty\sum_{r=0}^{\infty}\sum_{s=0}^{\min(x,y)}\frac{(q^{-x})_s(q^{-y})_sa^{-s}}{(q)_r(q)_s(aq)_r(bq)_s}q^{(x+y)(r+s)-2rs-s^2}$$

$$\cdot t^{r+s}\sum_{j=0}^{\infty}\frac{(abq^{2r+2s+2})^j}{(q)_j(tq)_j},$$

which gives the positivity of the Poisson kernel for the little q-Jacobi polynomials for $x,y=0,1,\cdots$, $0\le t<1$ when $0<q<1$, $0<aq<1$ and $0<bq<1$. A formal power series formula for the left side of (3.20) with t^n replaced by $(abq^2t)^nq^{n(n-1)/2}$ is given in Stanton [20].

4. Additional bilinear sums. Analogous to a bilinear generating function considered for the Racah polynomials in Rahman [15, Thm. 2], we shall here consider the sum

$$K_z(x,y)=K_z(x,y;a,b,c,\alpha,\gamma,M,N;q) \tag{4.1}$$

$$=\sum_{n=0}^{z}\frac{(aq)_n(abq)_n(bcq)_n(-c)^{-n}}{(q)_n(bq)_n(aq/c)_n(abq)_{2n}}\lambda_n(z)q^{\binom{n}{2}}$$

$$\cdot W_n(x;a,b,c,M;q)W_n\left(y;\alpha,\frac{ab}{\alpha},\gamma,N;q\right),$$

where $z=0,1,\cdots,M$, $M\le N$, and

$$\lambda_n(z)=\sum_{k=0}^{z-n}\frac{(q^{-z})_{n+k}\mu_{n+k}}{(q)_k(abq^{2n+2})_k} \tag{4.2}$$

for $n=0,1,\cdots,z$ and an arbitrary (fixed) sequence $\{\mu_k\}$ of constants.

Using (3.2) in (4.1), setting $n=r+s+j$, $k=m-s-j$, and observing that the sum over j is a multiple of the very well-poised series

$$
{}_4\phi_3\left[\begin{matrix} abq^{2r+2s+1}, q\sqrt{\ }, -q\sqrt{\ }, q^{s-m} \\ \sqrt{\ }, -\sqrt{\ }, abq^{2+2r+s+m} \end{matrix} ; q^{m-s}\right],
$$

(which equals 1 if $m=s$ and 0 if $m>s$ by a q-analogue of Dixon's theorem [19, (3.3.1.5)]) we find that

$$
(4.3)\quad K_z(x,y)=\sum_{r=0}^{z}\sum_{s=0}^{z-r}\frac{(q^{-z})_{r+s}\mu_{r+s}(q^{-x})_r(cq^{x-M})_r}{(q)_r(q)_s(q^{-M})_{r+s}(\alpha q)_r(q^{-N})_r}
\cdot\frac{(q^{-y})_r(\gamma q^{y-N})_r(q^{x-M})_s(c^{-1}q^{-x})_s}{(ab\gamma q/\alpha)_r(bq)_s(aq/c)_s(cq^{1-s})_r}\,\frac{1-cq^{r-s}}{1-cq^{-s}}A_{r,s}q^{-rs},
$$

with $A_{r,s}$ as defined in (3.3).

Formula (4.3) has many interesting special cases, some of which we shall now consider.

Case 1. $\lambda_n(z)$ *proportional to a balanced* ${}_3\phi_2$. Let

$$
(4.4)\qquad \mu_k=\frac{(d)_k(e)_k}{(q^{-z-1}de/ab)_k}q^k,
$$

where d and e are arbitrary parameters. Then $\lambda_n(z)$ is a multiple of a balanced ${}_3\phi_2$ series which can be summed via (2.3) to give

$$
(4.5)\quad \lambda_n(z)=\frac{(abq^2/d)_z(abq^2/e)_z}{(abq^2)_z(abq^2/de)_z}
\cdot\frac{(q^{-z})_n(d)_n(e)_n(abq^2)_{2n}}{(abq^{z+2})_n(abq^2/d)_n(abq^2/e)_n}\left(-\frac{ab}{de}\right)^n q^{nz+2n-\binom{n}{2}}.
$$

Hence it follows from (4.3) that

(4.6)

$$
\sum_{r=0}^{z}\sum_{s=0}^{z-r}\frac{(q^{-z})_{r+s}(d)_{r+s}(e)_{r+s}(q^{-x})_r(cq^{x-M})_r(q^{-y})_r(\gamma q^{y-N})_r}{(q)_r(q)_s(q^{-z-1}de/ab)_{r+s}(q^{-M})_{r+s}(\alpha q)_r(q^{-N})_r(ab\gamma q/\alpha)_r}
\cdot\frac{(q^{x-M})_s(c^{-1}q^{-x})_s}{(bq)_s(aq/c)_s(cq^{1-s})_r}\,\frac{1-cq^{r-s}}{1-cq^{-s}}A_{r,s}q^{r+s-rs}
$$

$$
=\frac{(abq^2/d)_z(abq^2/e)_z}{(abq^2)_z(abq^2/de)_z}\sum_{n=0}^{z}\frac{(q^{-z})_n(d)_n(e)_n(abq^2)_{2n}(aq)_n(abq)_n}{(q)_n(abq^{z+2})_n(abq^2/d)_n(abq^2/e)_n(bq)_n}
\cdot\frac{(bcq)_n}{(aq/c)_n(abq)_{2n}}\left(\frac{ab}{cde}q^{z+2}\right)^n
\cdot W_n(x;a,b,c,M;q)W_n\left(y;\alpha,\frac{ab}{\alpha},\gamma,N;q\right).
$$

From the left-hand side of (4.6) and our observations concerning (3.9), it is clear that if $\alpha=a$, $x=0,1,\cdots,M$, $y=0,1,\cdots,N$, $z=0,1,\cdots,M$, $0<q<1$, $0<aq<1$, $0\leq bq<1$, $0<c<aq^M$ and $cq\leq\gamma<q^{N-1}\leq q^{M-1}$, then the kernel on the right-hand side of (4.6) is nonnegative if either

(i) $|d|<1, |e|<1, abq^2<de$, or

(ii) both d and e are nonnegative integer powers of q, or

(iii) $d,e\geq q^{-z}$.

Case 2. $\lambda_n(z)$ *proportional to a balanced* ${}_4\phi_3$. Let

$$\mu_k=\frac{(d)_k(e)_k(q^{-M})_k}{(f)_k(q^{-M-z-1}de/abf)_k}q^k. \tag{4.7}$$

Then, for $n=0,1,\cdots,z$,

$$\begin{aligned}\lambda_n(z)&=\frac{(q^{-z})_n(q^{-M})_n(d)_n(e)_n}{(f)_n(q^{-M-z-1}de/abf)_n}q^n\\&\quad\cdot{}_4\phi_3\left[\begin{matrix}dq^n,eq^n,q^{n-M},q^{n-z}\\fq^n,\dfrac{q^{n-M-z-1}de}{abf},abq^{2n+2}\end{matrix};q\right]\\&=\frac{(abfq^2/de)_z(f^{-1}q^{1-M-z})_z}{(f)_z(q^{-M-z-1}de/abf)_z}\cdot\frac{(q^{-z})_n(q^{-M})_n(d)_n(e)_n}{(abfq^2/de)_n(f^{-1}q^{1-M-z})_n}\left(\frac{de}{abq^2}\right)^{z-n}q^n\\&\quad\cdot{}_4\phi_3\left[\begin{matrix}\dfrac{q^{n+2}ab}{d},\dfrac{q^{n+2}ab}{e},q^{n-M},q^{n-z}\\f^{-1}q^{n-M-z+1},\dfrac{q^{n+2}abf}{de},abq^{2n+2}\end{matrix};q\right]\end{aligned}\tag{4.8}$$

by (2.11). These ${}_4\phi_3$ series can be summed only in very special cases. If

$$d=q^{3/2}\sqrt{ab},\quad e=q\sqrt{ab},\quad f=q^{(1-M-z)/2}, \tag{4.9}$$

then

$$\begin{aligned}\lambda_n(z)&=\frac{(q^{-(M+z)/2})_z(q^{-z})_n(q^{-M})_n(q^{3/2}\sqrt{ab})_n(q\sqrt{ab})_n}{(q^{(2-M-z)/2})_z(q^{-(M+z)/2})_n(q^{(1-M-z)/2})_n}\\&\quad\cdot{}_4\phi_3\left[\begin{matrix}q^{n+3/2}\sqrt{ab},q^{n+1}\sqrt{ab},q^{n-M},q^{n-z}\\q^{n+(1-M-z)/2},q^{n-(M+z)/2},abq^{2n+2}\end{matrix};q\right].\end{aligned}\tag{4.10}$$

Since this ${}_4\phi_3$ series can be summed by the formula

$$\begin{aligned}&{}_4\phi_3\left[\begin{matrix}q^{a/2},q^{(a+1)/2},q^{a+1-w},q^{-m}\\q^{a+1},q^{(1+a-w-m)/2},q^{(2+a-w-m)/2}\end{matrix};q\right]\\&=\frac{(q^{w/2};q^{1/2})_m(-q^{1/2};q^{1/2})_m}{(q^{(w-a)/2};q^{1/2})_m(-q^{(a+1)/2};q^{1/2})_m},\qquad m=0,1,\cdots,\end{aligned}\tag{4.11}$$

which is a special case $b=0$ of [16,(1.8)], it follows from (4.3), (4.7), (4.10) and (4.11) that we have the rather strange looking formula

(4.12)

$$\sum_{r=0}^{z}\sum_{s=0}^{z-r}\frac{(q^{-z})_{r+s}\left(q\sqrt{ab}\,;q^{1/2}\right)_{2r+2s}(q^{-x})_r(cq^{x-M})_r}{(q)_r(q)_s(q^{(1-M-z)/2};q^{1/2})_{2r+2s}(\alpha q)_r(q^{-N})_r}$$
$$\cdot\frac{(q^{-y})_r(\gamma q^{y-N})_r(q^{x-M})_s(c^{-1}q^{-x})_s}{(ab\gamma q/\alpha)_r(bq)_s(aq/c)_s(cq^{1-s})_r}\frac{1-cq^{r-s}}{1-cq^{-s}}A_{r,s}q^{-rs}$$
$$=\frac{\left(q^{-(M+z)/2}\right)_z}{\left(q^{(2-M-z)/2}\right)_z}\sum_{n=0}^{z}\frac{(q^{-z})_n(-q^{1/2};q^{1/2})_{z-n}\left(q^{(n+2-M)/2}\sqrt{ab}\,;q^{1/2}\right)_{z-n}}{(q)_n(q^{-(M+z)/2};q^{1/2})_{2n}\left(q^{(1+m-n)/2};q^{1/2}\right)_{z-n}}$$
$$\cdot\frac{\left(q\sqrt{ab}\,;q^{1/2}\right)_{2n}(q^{-M})_n(abq)_n(bcq)_n}{\left(-q^{n+1}\sqrt{ab}\,;q^{1/2}\right)_{z-n}(bq)_n(abq)_{2n}(aq/c)_n}\left(-\frac{q^{1/2}}{c}\right)^n q^{z/2}q^{\binom{n}{2}}$$
$$\cdot W_n(x;a,b,c,M;q)W_n\left(y;\alpha,\frac{ab}{\alpha},\gamma,N;q\right),$$

where the same square root of ab is used throughout. As in our consideration of the sign of the kernel in (4.6), from the left-hand side of (4.12) it follows that if $\alpha=a$, $x=0,1,\cdots,M$, $y=0,1,\cdots,N$, $z=0,1,\cdots,M$, $0<q<1$, $0<aq<1$, $0\le bq<1$, $0<c<aq^M$ and $cq\le\gamma<q^{N-1}\le q^{M-1}$, then the kernel on the right-hand side of (4.12) is nonnegative.

If (4.9) is replaced by

$$d=q\sqrt{ab}\,,\quad e=q^{1/2}\sqrt{ab}\,,\quad f=-q^{(M+z)/2},\tag{4.13}$$

then $\lambda_n(z)$ becomes a multiple of the series

$$ {}_4\phi_3\left[\begin{matrix}q^{n+1}\sqrt{ab}\,,q^{n+3/2}\sqrt{ab}\,,q^{n-M},q^{n-z}\\ q^{n+(1-M-z)/2},q^{n+(2-M-z)/2},abq^{2n+2}\end{matrix};q\right].\tag{4.14}$$

This series can be summed by the formula [16,(4.8) with $b=0$]

$$ {}_4\phi_3\left[\begin{matrix}q^{(a+1)/2},q^{(a+2)/2},q^{1+a-w},q^{-m}\\ q^{a+1},q^{(2+a-w-m)/2},q^{(3+a-w-m)/2}\end{matrix};q\right]\tag{4.15}$$
$$=\frac{(-q^{1/2};q^{1/2})_m(q^{w/2};q^{1/2})_m(q^{(1+q-w-m)/2})_m(q^{(w-a-m)/2})_m}{(-q^{(a+1)/2};q^{1/2})_m(q^{(w-a)/2};q^{1/2})_m(q^{(w+a-m-1)/2})_m(q^{(2+a-w-m)/2})_m},$$

to yield a formula similar to (4.12), which we shall omit.

5. Product formulas of Bateman type. In order to derive our Bateman type product formulas we first set $s=k-r$ in (2.13) and change the order of summation to obtain the expansion

(5.1)

$$
{}_4\phi_3\left[\begin{matrix} q^{-n}, aq^n, b, c \\ d, e, \dfrac{abcq}{de} \end{matrix} ; q\right] {}_4\phi_3\left[\begin{matrix} q^{-n}, aq^n, g, h \\ e, \dfrac{dgh}{bc}, \dfrac{abcq}{de} \end{matrix} ; q\right]
$$

$$
= \frac{(aq/e)_n(de/bc)_n}{(e)_n(abcq/de)_n}\left(\frac{bc}{d}\right)^n \sum_{k=0}^{n} \frac{(q^{-n})_k(aq^n)_k(d/b)_k(d/c)_k(dg/bc)_k}{(q)_k(d)_k(aq/e)_k(d/bc)_k(de/bc)_k}
$$

$$
\cdot \frac{(dh/bc)_k}{(dgh/bc)_k}\, q^k
$$

$$
\cdot {}_{10}\phi_9\left[\begin{matrix} \dfrac{q^{-k}bc}{d}, q\sqrt{\ }, -q\sqrt{\ }, b, c, g, h, \dfrac{q^{-k}e}{a}, \dfrac{q^{1-k}bc}{de}, q^{-k} \\ \sqrt{\ }, -\sqrt{\ }, \dfrac{q^{1-k}c}{d}, \dfrac{q^{1-k}b}{d}, \dfrac{q^{1-k}bc}{dg}, \dfrac{q^{1-k}bc}{dh}, \dfrac{abcq}{de}, e, \dfrac{bcq}{d} \end{matrix} ; \frac{abcq^2}{d^2gh}\right].
$$

If $abcq^2/d^2gh$ were equal to q, then this ${}_{10}\phi_9$ could be transformed via Jackson's transformation formula [19,(3.4.2.4)]

(5.2)

$$
{}_{10}\phi_9\left[\begin{matrix} a, q\sqrt{a}, -q\sqrt{a}, c, d, e, f, g, h, q^{-m} \\ \sqrt{a}, -\sqrt{a}, \dfrac{aq}{c}, \dfrac{aq}{d}, \dfrac{aq}{e}, \dfrac{aq}{f}, \dfrac{aq}{g}, \dfrac{aq}{h}, aq^{m+1} \end{matrix} ; q\right]
$$

$$
= \frac{(aq)_m(aq/fg)_m(aq/fh)_m(aq/gh)_m}{(aq/f)_m(aq/g)_m(aq/h)_m(aq/fgh)_m}
$$

$$
\cdot {}_{10}\phi_9\left[\begin{matrix} \dfrac{a^2q}{cde}, q\sqrt{\ }, -q\sqrt{\ }, \dfrac{aq}{de}, \dfrac{aq}{ce}, \dfrac{aq}{cd}, f, g, h, q^{-m} \\ \sqrt{\ }, -\sqrt{\ }, \dfrac{aq}{c}, \dfrac{aq}{d}, \dfrac{aq}{e}, \dfrac{a^2q^2}{cdef}, \dfrac{a^2q^2}{cdeg}, \dfrac{a^2q^2}{cdeh}, \dfrac{a^2q^{m+2}}{cde} \end{matrix} ; q\right],
$$

where $a^3q^{m+2} = cdefgh$ and m is a nonnegative integer. However, in order to avoid the restriction $abcq^2/d^2gh = q$ and derive our Bateman type formula we first need to use the following expansion formula (which will be proved at the end of this section) in an appropriate way:

(5.3)

$$
{}_{10}\phi_9\left[\begin{matrix} a, q\sqrt{a}, -q\sqrt{a}, c, d, e, f, g, h, q^{-m} \\ \sqrt{a}, -\sqrt{a}, \dfrac{aq}{c}, \dfrac{aq}{d}, \dfrac{aq}{e}, \dfrac{aq}{f}, \dfrac{aq}{g}, \dfrac{aq}{h}, aq^{m+1} \end{matrix} ; zq\right]
$$

$$
= \frac{(q^{1-m}/h)_m(aq/hz)_m}{(q^{1-m}/hz)_m(aq/h)_m} \sum_{j=0}^{m} \frac{(q^{-m}/hz)_j\left(q\sqrt{\ }\right)_j\left(-q\sqrt{\ }\right)_j(z^{-1})_j(q^{-m}/a)_j}{(q)_j\left(\sqrt{\ }\right)_j\left(-\sqrt{\ }\right)_j(q^{1-m}/h)_j(aq/hz)_j}
$$

$$
\cdot \frac{(q^{-m})_j}{(q/hz)_j}\left(\frac{aq^{m+1}}{h}\right)^j
$$

$$\cdot {}_{10}\phi_9\left[\begin{matrix} a, q\sqrt{a}, -q\sqrt{a}, c, d, e, f, g, hzq^{-j}, q^{j-m} \\ \sqrt{a}, -\sqrt{a}, \frac{aq}{c}, \frac{aq}{d}, \frac{aq}{e}, \frac{aq}{f}, \frac{aq}{g}, \frac{q^{j+1}a}{hz}, aq^{m-j+1} \end{matrix}; q\right].$$

where $a^3q^{m+2} = cdefghz$.

Assume, temporarily, that $b = q^{-x}$, where x is a nonnegative integer, and replace $a, c, d, e, f, g, h, q^{-m}$ in (5.3) by $q^{-k-x}c/d$, $g, h, q^{-k}e/a$, $q^{1-k-x}c/de$, q^{-k}, c, q^{-x}, respectively, to obtain that the ${}_{10}\phi_9$ in (5.1) equals

(5.4)

$$\frac{(c^{-1}q^{1-x})_x(q^{-k}dgh/ac)_x}{(d^2gh/ac^2)_x(d^{-1}q^{1-k-x})_x}\sum_{j=0}^{x}\frac{(d^2gh/ac^2q)_j(q\sqrt{\ })_j(-q\sqrt{\ })_j}{(q)_j(\sqrt{\ })_j(-\sqrt{\ })_j}$$

$$\cdot\frac{(q^{x-1}d^2gh/ac)_j(q^kd/c)_j(q^{-x})_j(d^{-1}q^{1-k})^j}{(c^{-1}q^{1-x})_j(q^{-k}dgh/ac)_j(q^xd^2gh/ac^2)_j}$$

$$\cdot {}_{10}\phi_9\left[\begin{matrix} \frac{q^{-k-x}c}{d}, q\sqrt{\ }, -q\sqrt{\ }, g, h, \frac{q^{-k}e}{a}, \frac{q^{1-k-x}c}{de}, q^{-k}, \frac{q^{1-j-x}ac^2}{d^2gh}, q^{j-x} \\ -\sqrt{\ }, -\sqrt{\ }, \frac{q^{1-k-x}c}{dg}, \frac{q^{1-k-x}c}{dh}, \frac{q^{1-x}ac}{de}, e, \frac{q^{1-x}c}{d}, \frac{q^{j-k}dgh}{ac}, \frac{q^{1-j-k}c}{d} \end{matrix}; q\right].$$

Now we can apply (5.2) to the above ${}_{10}\phi_9$ to find that it equals

(5.5)

$$\frac{(q^xd/c)_k(q^{1-x}ac/dgh)_k(acq/dh)_k(dg/c)_k(dh/ac)_j(q^{-k}dgh/ac)_j}{(q^xdg/c)_k(q^{1-x}ac/dh)_k(acq/dgh)_k(d/c)_k(q^{-k}dh/ac)_j(dgh/ac)_j}$$

$$\cdot\frac{(q^kdg/c)_j\,(d/c)_j}{(dg/c)_j\,(q^kd/c)_j}$$

$$\cdot {}_{10}\phi_9\left[\begin{matrix} \frac{q^{-x}ac}{dh}, q\sqrt{\ }, -q\sqrt{\ }, \frac{e}{h}, \frac{q^{1-x}ac}{deh}, aq^k, g, \frac{q^{1-j-x}ac^2}{d^2gh}, q^{j-x}, q^{-k} \\ \sqrt{\ }, -\sqrt{\ }, \frac{q^{1-x}ac}{de}, e, \frac{q^{1-k-x}c}{dh}, \frac{q^{1-x}ac}{dgh}, \frac{q^jdg}{c}, \frac{q^{1-j}ac}{dh}, \frac{q^{1+k-x}ac}{dh} \end{matrix}; q\right].$$

Writing the above ${}_{10}\phi_9$ series with m as the index of summation, substituting (5.5) into (5.4) and summing over j we find that

$$\sum_j \frac{(d^2gh/ac^2g)_j\left(q\sqrt{\ }\right)_j\left(-q\sqrt{\ }\right)_j(q^{x-1}d^2gh/ac)_j(q^kd/c)_j(q^{-x})_j(d^{-1}q^{1-k})^j}{(q)_j(\ )_j\left(-\sqrt{\ }\right)_j(c^{-1}q^{1-x})_j(q^{-k}dgh/ac)_j(q^xd^2gh/ac^2)_j}$$

$$\cdot\frac{(dh/ac)_j(q^{-k}dgh/ac)_j(q^kdg/c)_j(d/c)_j(q^{1-j-x}ac^2/d^2gh)_m(q^{j-x})_m}{(q^{-k}dh/ac)_j(dgh/ac)_j(dg/c)_j(q^kd/c)_j(q^jdg/c)_m(q^{1-j}ac/dh)_m}$$

$$=\frac{(q^{-x})_m(q^{1-x}ac^2/d^2gh)_m}{(acq/dh)_m(dg/c)_m}$$

$$\cdot{}_8\phi_7\left[\begin{matrix}\dfrac{d^2gh}{ac^2q},q\sqrt{\ },-q\sqrt{\ },\dfrac{q^{x-1}d^2gh}{ac},q^{m-x},\dfrac{q^{-m}dh}{ac},\dfrac{d}{c},\dfrac{q^kdg}{c}\\ \sqrt{\ },-\sqrt{\ },c^{-1}q^{1-x},\dfrac{q^{x-m}d^2gh}{ac^2},\dfrac{q^mdg}{c},\dfrac{dgh}{ac},\dfrac{q^{-k}dh}{ac}\end{matrix};d^{-1}q^{1-k}\right]$$

$$=\frac{(d^2gh/ac^2)_x(d^{-1}q^{1-x})_x(q^{-x})_m(c)_m(q^{1-x}ac/dgh)_m}{(dgh/ac)_x(c^{-1}q^{1-x})_x(acq/dh)_m(dg/c)_m(d)_m}$$

$$\cdot{}_4\phi_3\left[\begin{matrix}q^{m-k},q^{m-x},\dfrac{q^{x-1}d^2gh}{ac},\dfrac{d}{c}\\ dq^m,\dfrac{q^mdg}{c},\dfrac{q^{-k}dh}{ac}\end{matrix};q\right]$$

by [19,(3.4.1.5)]. Finally, using the above observations in (5.1) and simplifying, we obtain the following generalization of Bateman's formula (1.3):

(5.6)

$${}_4\phi_3\left[\begin{matrix}q^{-n},aq^n,b,c\\ d,e,\dfrac{abcq}{de}\end{matrix};q\right]{}_4\phi_3\left[\begin{matrix}q^{-n},aq^n,g,h\\ e,\dfrac{dgh}{bc},\dfrac{abcq}{de}\end{matrix};q\right]$$

$$=\frac{(aq/e)_n(de/bc)_n}{(e)_n(abcq/de)_n}\left(\frac{bc}{d}\right)^n\sum_{k=0}^n\frac{(q^{-n})_k(aq^n)_k(dh/bc)_k(acq/dh)_k}{(q)_k(aq/e)_k(de/bc)_k(dgh/bc)_k}$$

$$\cdot\frac{(dg/c)_k}{(abcq/dh)_k}q^k\sum_{m=0}^k\frac{(abc/dh)_m\left(q\sqrt{\ }\right)_m\left(-q\sqrt{\ }\right)_m(q^{-k})_m(aq^k)_m}{(q)_m\left(\sqrt{\ }\right)_m\left(-\sqrt{\ }\right)_m(q^{k+1}abc/dh)_m(q^{1-k}bc/dh)_m}$$

$$\cdot\frac{(b)_m(e/h)_m(abcq/deh)_m(g)_m(c)_m}{(acq/dh)_m(abcq/de)_m(e)_m(d)_m(dg/c)_m}q^m$$

$$\cdot{}_4\phi_3\left[\begin{matrix}q^{m-k},bq^m,\dfrac{d^2gh}{abcq},\dfrac{d}{c}\\ dq^m,\dfrac{q^mdg}{c},\dfrac{q^{-k}dh}{ac}\end{matrix};q\right],$$

where, by analytic continuation, b can now be an arbitrary parameter.

It follows from (5.6) that the q-Racah polynomials have a Bateman type formula of the form

(5.7)

$$W_n(x;a,b,c,N;q)W_n(y;a,b,c',N;q)$$

$$=\frac{(bq)_n(abq^{N+2})_n}{(aq)_n(q^{-N})_n}(bq^{N+1})^{-n}\sum_{k=0}^{n}\frac{(q^{-n})_k(abq^{n+1})_k(bc'q^{y+1})_k}{(q)_k(bq)_k(abq^{N+2})_k}$$

$$\cdot\frac{(q^{1+x-y}a/c')_k(bq^{1+N-x-y})_k}{(bc'q)_k(q^{1-y}a/c')_k}q^k\sum_{m=0}^{k}\frac{(q^{-y}a/c')_m(q\sqrt{\ })_m(-q\sqrt{\ })_m}{(q)_m(\sqrt{\ })_m(-\sqrt{\ })_m}$$

$$\cdot\frac{(q^{-k})_m(abq^{k+1})_m(q^{-x})_m(q^{1+N-y}a/c')_m(q^{-y})_m(q^{-y}/c')_m(cq^{x-N})_m}{(q^{1+k-y}a/c')_m(q^{-y-k}/bc')_m(q^{1+x-y}a/c')_m(q^{-N})_m(bcq)_m(aq)_m(bq^{1+N-x-y})_m}q^m$$

$$\cdot{}_4\phi_3\left[\begin{matrix}q^{m-k},q^{m-x},\dfrac{bcc'}{a},bq^{1+N-x}\\ bcq^{m+1},bq^{1+N-x-y+m},\dfrac{q^{y-x-k}c'}{a}\end{matrix};q\right].$$

A similar formula for the product $W_n(x;a,b,c,N;q)W_n(y;a,b,c,N';q)$ also follows from (5.6).

By letting c and c' tend to zero in (5.7) we find that the q-Hahn polynomials have a Bateman type formula of the form

(5.8)

$${}_3\phi_2\left[\begin{matrix}q^{-n},abq^{n+1},q^{-x}\\ aq,q^{-N}\end{matrix};q\right]{}_3\phi_2\left[\begin{matrix}q^{-n},abq^{n+1},q^{-y}\\ aq,q^{-N}\end{matrix};q\right]$$

$$=\frac{(bq)_n(abq^{N+2})_n}{(aq)_n(q^{-N})_n}(bq^{N+1})^{-n}\sum_{k=0}^{n}\frac{(q^{-n})_k(abq^{n+1})_k(bq^{1+N-x-y})_k}{(q)_k(bq)_k(abq^{N+2})_k}q^{(x+1)k}$$

$$\cdot\sum_{m=0}^{k}\frac{(q^{-k})_m(abq^{k+1})_m(q^{-x})_m(q^{-y})_m}{(q)_m(aq)_m(q^{-N})_m(bq^{1+N-x-y})_m}(bq^{N+2-x})^m$$

$$\cdot{}_3\phi_2\left[\begin{matrix}q^{m-k},q^{m-x},bq^{1+N-x}\\ bq^{1+N+m-x-y},0\end{matrix};q\right].$$

A Bateman type formula for Hahn polynomials follows from (5.8) by setting $a=q^\alpha$, $b=q^\beta$ and letting $q\to 1-$. In addition, by replacing q^{-x} and q^{-y} in (5.8) by xq^{-N} and yq^{-N}, respectively, and letting $N\to\infty$ we find that the little q-Jacobi polynomials have a Bateman type formula

$$(5.9)\qquad {}_2\phi_1\left[\begin{matrix} q^{-n},abq^{n+1} \\ aq \end{matrix};xq\right]{}_2\phi_1\left[\begin{matrix} q^{-n},abq^{n+1} \\ aq \end{matrix};yq\right]$$

$$=\frac{(bq)_n}{(aq)_n}(-b)^{-n}q^{-n(n+1)/2}\sum_{k=0}^{n}\frac{(q^{-n})_k(abq^{n+1})_k}{(q)_k(bq)_k}(-byq^2)^k q^{\binom{k}{2}}$$

$$\cdot\sum_{m=0}^{k}\frac{(q^{-k})_m(abq^{k+1})_m}{(q)_m(aq)_m}(xq)^m{}_2\phi_1\left[\begin{matrix} q^{m-k},bxq, \\ 0, \end{matrix};b^{-1}y^{-1}\right].$$

To complete our proof of (5.6), it remains to prove (5.3). However, since (5.3) is a special case of more general expansion formulas, with a view to future applications we shall first derive these more general formulas. Observe that from Jackson's formula [19,(3.3.1.1)]

$$(5.10)\qquad {}_8\phi_7\left[\begin{matrix} q,q\sqrt{a},-q\sqrt{a},b,c,d,\dfrac{q^{1+m-k}a^2}{bcd},q^{k-m} \\ \sqrt{a},-\sqrt{a},\dfrac{aq}{b},\dfrac{aq}{c},\dfrac{aq}{d},\dfrac{q^{k-m}bcd}{a},aq^{1+m-k} \end{matrix};q\right]$$

$$=\frac{(aq)_{m-k}(q^{k-m}bc/a)_{m-k}(q^{k-m}cd/a)_{m-k}(aq/bd)_{m-k}}{(aq/d)_{m-k}(q^{k-m}bcd/a)_{m-k}(aq/b)_{m-k}(q^{k-m}c/a)_{m-k}}$$

$$=\frac{(q^{-m}b/a)_k(q^{-m}c/a)_k(q^{-m}d/a)_k(q^{-m}bcd/a)_k}{(q^{-m}cd/a)_k(q^{-m}bd/a)_k(q^{-m}bc/a)_m(q^{-m}cd/a)_m}A_m^{-1},$$

with

$$A_m=\frac{(aq/b)_m(aq/d)_m(q^{-m}c/a)_m(q^{-m}bcd/a)_m}{(aq)_m(aq/bd)_m(q^{-m}bc/a)_m(q^{-m}cd/a)_m}.$$

Hence

(5.11)

$${}_{r+5}\phi_{r+4}\left[\begin{matrix} a_1,\cdots,a_{r+1},\dfrac{q^{-m}b}{a},\dfrac{q^{-m}c}{a},\dfrac{q^{-m}d}{a},q^{-m} \\ b_1,\cdots,b_r,\dfrac{q^{-m}cd}{a},\dfrac{q^{-m}bd}{a},\dfrac{q^{-m}bc}{a},\dfrac{q^{-m}bcd}{a^2} \end{matrix};t\right]$$

$$=A_m\sum_{k=0}^{m}\frac{(a_1)_k\cdots(a_{r+1})_k(q^{-m}/a)_k(q^{-m})_k}{(q)_k(b_1)_k\cdots(b_r)_k(q^{-m}bcd/a)_k(q^{-m}bcd/a^2)_k}t^k$$

$$\cdot\sum_{j=0}^{m-k}\frac{(a)_j(q\sqrt{a})_j(-q\sqrt{a})_j(b)_j(c)_j(d)_j(q^{1+m-k}a^2/bcd)_j(q^{k-m})_j}{(q)_j(\sqrt{a})_j(-\sqrt{a})_j(aq/b)_j(aq/c)_j(aq/d)_j(q^{k-m}bcd/a)_j(aq^{1+m-k})_j}q^j$$

$$=A_m \sum_{j=0}^{m} \frac{(a)_j\left(q\sqrt{a}\right)_j\left(-q\sqrt{a}\right)_j(b)_j(c)_j(d)_j\left(q^{m+1}a^2/bcd\right)_j\left(q^{-m}\right)_j}{(q)_j\left(\sqrt{a}\right)_j\left(-\sqrt{a}\right)_j(aq/b)_j(aq/c)_j(aq/d)_j\left(q^{-m}bcd/a\right)_j\left(aq^{m+1}\right)_j} q^j$$

$$\cdot {}_{r+3}\phi_{r+2}\left[\begin{matrix} a_1,\cdots,a_{r+1},a^{-1}q^{-j-m},q^{j-m} \\ b_1,\cdots,b_r,\dfrac{q^{j-m}bcd}{a},\dfrac{q^{-j-m}bcd}{a^2} \end{matrix};t\right]$$

by a change in order of summation. Setting $c=e/d$ and letting $d\to\infty$ in (5.11) gives

(5.12)

$${}_{r+3}\phi_{r+2}\left[\begin{matrix} a_1,\cdots,a_{r+1},\dfrac{q^{-m}b}{a},q^{-m} \\ b_1,\cdots,b_r,\dfrac{q^{-m}e}{a},\dfrac{q^{-m}be}{a^2} \end{matrix};\frac{t}{b}\right]$$

$$=\frac{(aq/b)_m(q^{-m}be/a)_m}{(aq)_m(q^{-m}e/a)_m} \sum_{j=0}^{m} \frac{(a)_j\left(q\sqrt{a}\right)_j\left(-q\sqrt{a}\right)_j\left(-q\sqrt{a}\right)_j(b)_j\left(q^{m+1}a^2/be\right)_j}{(q)_j\left(\sqrt{a}\right)_j\left(-\sqrt{a}\right)_j(aq/b)_j(q^{-m}be/a)_j}$$

$$\cdot \frac{(q^{-m})_j}{\left(aq^{m+1}\right)_j}\left(\frac{e}{a}\right)^j {}_{r+3}\phi_{r+2}\left[\begin{matrix} a_1,\cdots,a_{r+1},a^{-1}q^{-j-m},q^{j-m} \\ b_1,\cdots,b_r,\dfrac{q^{j-m}be}{a},\dfrac{q^{-j-m}be}{a^2} \end{matrix};t\right].$$

Formula (5.3) now follows from the case $r=7$ of (5.12) by replacing $t,a,b,e,a_1,\cdots,a_8$, $b_1,\cdots,b_7$, respectively, by $q,q^{-m}/hz$, z^{-1}, aq/h^2z, $a,q\sqrt{a}$, $-q\sqrt{a}$, $c,d,e,f,g,\sqrt{a}$, $-\sqrt{a}$, aq/c, aq/d, aq/e, aq/f, aq/g.

REFERENCES

[1] G. Andrews, *q-identities of Auluck, Carlitz and Rogers*, Duke Math. J., 33 (1966), pp. 575–582.

[2] R. Askey and J. Wilson, *A set of orthogonal polynomials that generalize the Racah coefficients or 6−j symbols*, this Journal, 10 (1979), pp. 1008–1016.

[3] W. N. Bailey, *A reducible case of the fourth type of Appell's hypergeometric functions of two variables*, Quart. J. Math., Oxford, 4 (1933), pp. 305–308.

[4] ______, *Generalized Hypergeometric Series*, Stechert-Hafner Service Agency, New York and London, 1964.

[5] H. Bateman, *Partial Differential Equations of Mathematical Physics*, Cambridge Univ. Press, Cambridge, England, 1932.

[6] G. Gasper, *Positivity and the convolution structure for Jacobi series*, Ann. of Math., 93 (1971), pp. 112–118.

[7] ______, *Banach algebras for Jacobi series and positivity of a kernel*, Ann. of Math., 95 (1972), pp. 261–280.

[8] ______, *Nonnegativity of a discrete Poisson kernel for the Hahn polynomials*, J. Math. Anal. Appl., 42 (1973), pp. 438–451.

[9] ______, *Products of terminating ${}_3F_2$ series*, Pacific J. Math., 56 (1975), pp. 87–95.

[10] G. Gasper and Mizan Rahman, *Positivity of the Poisson kernel for the continuous q-ultraspherical polynomials*, this Journal, 14 (1983), pp. 409–420.

[11] T. Koornwinder, *Jacobi polynomials*, II. *An analytic proof of the product formula*, this Journal, 5 (1974), pp. 125–137.

[12] G. Racah, *Theory of complex spectra* II., Phys. Rev. 62 (1942), pp. 438–462.

[13] MIZAN RAHMAN, *A five-parameter family of positive kernels from Jacobi polynomials*, this Journal, 7 (1976), pp. 386–413.

[14] ______, *Some positive kernels and bilinear sums for Hahn polynomials*, this Journal, 7 (1976), pp. 414–435.

[15] ______, *A product formula and a nonnegative Poisson kernel for Racah-Wilson polynomials*, Canad. J. Math., 32 (1980), pp. 1501–1517.

[16] MIZAN RAHMAN AND B. NASSRALLAH, *On the q-analogues of some transformations of nearly-poised hypergeometric series*, Trans. Amer. Math. Soc., 268, (1981), pp. 211–229.

[17] MIZAN RAHMAN, *Reproducing kernels and bilinear sums for q-Racah and q-Wilson polynomials*, Trans. Amer. Math. Soc., 273 (1982), pp. 483–508.

[18] D. B. SEARS, *On the transformation theory of basic hypergeometric functions*, Proc. London Math. Soc. (2), 53 (1951), pp. 158–180.

[19] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge Univ. Press, Cambridge, 1966.

[20] D. STANTON, *A short proof of a generating function for Jacobi polynomials*, Proc. Amer. Math. Soc., 80 (1980), pp. 398–400.

[21] A. VERMA AND V. K. JAIN, *Some transformations of basic hypergeometric functions, part* I., this Journal, 12 (1981), pp. 943–956.

[22] G. N. WATSON, *The product of two hypergeometric functions*, Proc. London Math. Soc. (2), 20 (1922), pp. 189–195.

[23] ______, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge Univ. Press, Cambridge, 1962.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics
014

GENERALIZED CHEBYSHEVIAN SPLINES*

G. NÜRNBERGER†, L. L. SCHUMAKER‡, M. SOMMER† AND H. STRAUSS†

Abstract. In this paper we study a space of splines in which the pieces are drawn from a linear space spanned by an ECT-system U. The splines here generalize the usual Chebyshevian splines in that the pieces in the various intervals are restricted to come from varying subspaces of U. For our class of generalized splines we discuss zeros, determinants associated with certain Lagrange and Hermite interpolation problems, and properties of certain local support basis splines.

1. Introduction. Over the past twenty years, the concept of a polynomial spline has been generalized in many ways, resulting in an extensive hierarchy of splines—see e.g. [5] and references therein. Chebyshevian splines, which lie in the middle of this hierarchy, are of particular interest because they have almost all of the nice features of the classical polynomial splines.

The purpose of this paper is to study a class of generalized splines which retains most of the features of the Chebyshevian splines, including interpolation properties (which are studied here) as well as approximation properties (which are studied in [3]). The spline spaces of interest are introduced later in this section.

Section 2 of this paper is devoted to zero properties of our splines, including a version of the Budan–Fourier theorem. In §3 we examine Lagrange and Hermite interpolation, and show that interpolation is possible precisely when a certain interlacing property holds. In §4 we use our results on Hermite interpolation to construct a basis of local-support B-splines.

Section 5 of the paper contains a variety of examples to illustrate the material. Our examples also show that any further generalization of the spline spaces will result in the loss of some of the key properties. We conclude the paper with remarks and references.

We devote the remainder of this section to definitions. For convenience, we follow the notation of [5]. Given positive functions $w_i \in C^{m-i}[a,b]$, $i=1,2,\cdots,m$, we define

$$(1.1)\qquad \begin{aligned} u_1(x)&=w_1(x),\\ u_2(x)&=w_1(x)\int_a^x w_2(s_2)\,ds_2,\\ &\cdots\\ u_m(x)&=w_1(x)\int_a^x w_2(s_2)\int_a^{s_2}\cdots\int_a^{s_{m-1}} w_m(s_m)\,ds_m\cdots ds_2. \end{aligned}$$

These functions form a canonical extended complete Chebyshev-(ECT-) system on $[a,b]$. We write $\mathcal{U}=\operatorname{span}\{u_i\}_1^m$.

Suppose now that $\mathfrak{N}=(n_0,\cdots,n_k)$ is a vector of integers with $0\le n_i\le m$, $i=0,\cdots,k$. We write

$$(1.2)\qquad \mathcal{U}_i=\operatorname{span}\{u_j\}_{j=1}^{n_i},\qquad i=0,\cdots,k.$$

The pieces of our splines will be drawn from these spaces. We now need to introduce a partition of the interval $[a,b]$. Let $a=x_0<x_1<\cdots<x_{k+1}=b$. Then $\Delta=\{x_i\}_1^k$ divides the interval $[a,b]$ into subintervals $I_i=[x_i,x_{i+1})$, $i=0,\cdots,k-1$ and $I_k=[x_k,x_{k+1}]$.

*Received by the editors October 14, 1982.

† Institut für Angewandte Mathematik, Universität Erlangen, Erlangen, Germany.

‡ Center for Approximation Theory, Texas A&M University, College Station, Texas 77843.

Finally, we need a way to describe the smoothness of the splines. We do this with a vector $\mathcal{R}=(r_1,\cdots,r_k)$ of integers satisfying $0\leq r_i\leq n_i$, $i=1,\cdots,k$.

DEFINITION 1.1 (generalized Chebyshevian splines). Given $\mathcal{U},\mathcal{N},\mathcal{R}$, and Δ as above, we define

$$(1.3)\quad \mathcal{S}(\mathcal{U};\mathcal{N};\mathcal{R};\Delta)=\big\{s\colon s|_{I_i}\in\mathcal{U}_i,\ i=0,\cdots,k \text{ and}$$
$$D_-^{j-1}s(x_i)=D_+^{j-1}s(x_i), j=1,\cdots,r_i \text{ and } i=1,\cdots,k\big\}.$$

When there is no chance of confusion, we shall usually shorten the notation for the space (1.3) to $\mathcal{S}$. In addition, we shall use the abbreviations T-spline for Chebyshevian spline and gT-spline for generalized Chebyshevian spline. The space of gT-splines defined here generalizes the classical T-spline spaces in that here the structure of the splines is allowed to vary from interval to interval (although all pieces are drawn from subspaces of one fixed ECT-space $\mathcal{U}$). The usual T-splines correspond to the case where $m=n_0=\cdots=n_k$.

THEOREM 1.2. *The space* $\mathcal{S}$ *defined in* (1.3) *is a linear space of dimension*

$$(1.4)\qquad n=n_0+\sum_{i=1}^{k}(n_i-r_i).$$

Proof. We observe that if $s\in\mathcal{S}$, then

$$s(x)=\begin{cases}\sum_{j=1}^{n_i}c_{ij}u_j(x) & \text{in } I_i \text{ if } n_i>0,\\ 0 & \text{in } I_i \text{ if } n_i=0,\end{cases}$$

for $i=0,1,\cdots,k$. Now writing down all the smoothness conditions leads to a system of $r_1+\cdots+r_k$ equations in the $n_0+\cdots+n_k$ unknowns. Since it is easily seen that this system is of full rank, the result follows. □

The proof of Theorem 1.2 is very much like the proof for ordinary T-splines. The result does not, however, follow from the usual dimensionality theorems (cf. [5, Thm. 11.4]) since here we have allowed $r_i>n_{i-1}$ in general.

The usual approach to studying a space of splines once the dimension is identified is to then define a one-sided basis for the space. While this is possible here (cf. [5, 11.2]), the resulting one-sided splines are rather complicated due to the fact that the nature of the spline varies from interval to interval. A basis for $\mathcal{S}$ will be given in §4.

2. Zero properties. In this section we show that any spline s in the space $\mathcal{S}$ defined in (1.3) can have at most $n-1$ zeros in $[a,b]$, counting multiplicities in a very strong way, where $n=\dim(\mathcal{S})$. In addition, we shall establish a Budan–Fourier-type theorem for gT-splines. While the results here are very similar in nature to well-known results for T-splines, the proofs require some modification to make them work in the gT-spline setting.

We begin by defining our scheme for counting zeros. Suppose that $s\in\mathcal{S}$. Since s restricted to a subinterval I_i of the partition belongs to the ECT-space $\mathcal{U}_i\subseteq\mathcal{U}$, it follows that s must either vanish identically throughout the interval I_i, or it can vanish only at a finite number (at most n_i-1) of isolated points in this interval. Taking our cue from [4], [5], we are led to the following definitions.

DEFINITION 2.1 (isolated zero). Suppose that

$$s(t-)=D_-s(t)=\cdots=D_-^{l-1}s(t)=0\neq D_-^{l}s(t),$$
$$s(t+)=D_+s(t)=\cdots=D_+^{r-1}s(t)=0\neq D_+^{r}s(t),$$

and that s does not vanish identically on any interval containing t. Let $\alpha=\max(l,r)$. Then we say that s has an *isolated zero at t of multiplicity*

$$z=\begin{cases}\alpha+1 & \text{if } \alpha \text{ is even and } s \text{ changes sign at } t,\\ \alpha+1 & \text{if } \alpha \text{ is odd and } s \text{ does not change sign at } t,\\ \alpha & \text{otherwise.}\end{cases}$$

DEFINITION 2.2 (left end interval zero). Suppose that $s(x)=0$ for all $a<x<x_p$ while $s(y)\neq 0$ for some $x_p<y<x_{p+1}$. Then we say that $[a,x_p)$ is an *interval zero of s of multiplicity*

$$z=n_0+\sum_{i=1}^{p-1}(n_i-r_i).$$

DEFINITION 2.3 (right end interval zero). Suppose that $s(x)=0$ for all $x_q<x<b$ while $s(y)\neq 0$ for some $x_{q-1}<y<x_q$. Then we say that $(x_q,b]$ is an *interval zero of s of multiplicity*

$$z=n_q+\sum_{i=q+1}^{k}(n_i-r_i).$$

DEFINITION 2.4 (interior interval zero). Suppose that $s(x)=0$ for all $x_p<x<x_q$ and does not vanish identically on any larger interval containing (x_p,x_q). Let $\alpha=n_p+\sum_{i=p+1}^{q-1}(n_i-r_i)$. Then we say that (x_p,x_q) is an *interval zero of s of multiplicity*

$$z=\begin{cases}\alpha+1 & \text{if } \alpha \text{ is even and } s \text{ changes sign,}\\ \alpha+1 & \text{if } \alpha \text{ is odd and } s \text{ does not change sign,}\\ \alpha & \text{otherwise.}\end{cases}$$

Given a spline $s\in\mathcal{S}$, we use the notation $Z^{\mathcal{S}}(s)$ to stand for the number of zeros of s in the interval $[a,b]$, counting multiplicities as in Definitions 2.1–2.4. The key tool in establishing bounds on $Z^{\mathcal{S}}(s)$ is an appropriate version of Rolle's theorem which we now present. Before stating it, we need some additional notation.

Associated with the canonical ECT-system (1.1), we introduce the differential operators

$$D_0f=f,\qquad D_if=D(f/w_i),\quad i=1,\cdots,m \tag{2.1}$$

and

$$L_i=D_iD_{i-1}\cdots D_0,\qquad i=0,\cdots,m. \tag{2.2}$$

At times we will write these operators with a superscript of + or − to indicate right or left derivatives, respectively.

It follows directly from the definitions that if $s\in\mathcal{S}(\mathcal{U};\mathcal{N};\mathcal{R};\Delta)$, then $L_1^+s\in\mathcal{S}(\mathcal{U}';\mathcal{N}';\mathcal{R}';\Delta)$, where

$$\mathcal{N}'=(n_0',\cdots,n_k'),\qquad n_i'=\max(n_i-1,0),\qquad i=0,\cdots,k, \tag{2.3}$$

$$\mathcal{R}'=(r_1',\cdots,r_k'),\qquad r_i'=\max(r_i-1,0),\qquad i=1,\cdots,k, \tag{2.4}$$

and where $\mathcal{U}'=\operatorname{span}\{u_i'\}_{i=1}^{m-1}$ is the first reduced space defined by

$$\begin{aligned}&u_1'(x)=w_2(x),\\ &\cdots\\ &u_{m-1}'(x)=w_2(x)\int_a^x w_3(s_3)\int_a^{s_3}\cdots\int_a^{s_{m-1}} w_m(s_m)\,ds_m\cdots ds_3.\end{aligned}\tag{2.5}$$

THEOREM 2.5 (Rolle's theorem for gT-splines). *For any* $s\in\mathcal{S}(\mathcal{U};\mathcal{N};\mathcal{R};\Delta)\cap C[a,b]$,

$$Z^{\mathcal{D}\mathcal{S}}(L_1^+s)\geq Z^{\mathcal{S}}(s)-1,\tag{2.6}$$

where $\mathcal{D}\mathcal{S}=\mathcal{S}(\mathcal{U}';\mathcal{N}';\mathcal{R}';\Delta)$.

Proof. The proof follows along the same lines as for T-splines; see [5, Thm. 9.29].

We are now ready for the main result of this section.

THEOREM 2.6. *For any nontrivial* $s\in\mathcal{S}$, $Z^{\mathcal{S}}(s)\leq n-1$, *where n is the dimension of* $\mathcal{S}$.

Proof. We consider first the case where $m=1$. Let $\{i_1<\cdots<i_p\}=\{j\colon 0\leq j\leq k$ and $n_j-r_j>0\}$, where we set $r_0=0$. Then it is easy to see that the spline $s\in\mathcal{S}$ defined by

$$s(x)=\left\{(-1)^j w_1(x)\text{ in }\left[x_{i_j},x_{i_{j+1}}\right),\ j=1,\cdots,p\right\}$$

has a maximal number of zeros $(p-1)$ among splines in $\mathcal{S}$. Since

$$n=\sum_{i=0}^{k}(n_i-r_i)=\sum_{j=1}^{p}(n_{i_j}-r_{i_j})=p,$$

we have established the result for $m=1$.

We now proceed by induction on m. Suppose that the theorem is correct for gT-spline spaces of order $m-1$, and suppose that s is a spline of order m. We define

$$\{\nu_1<\cdots<\nu_q\}=\{1\leq i\leq k\colon s\text{ has a jump at }x_{\nu_i}\}.$$

Let $J_i=[x_{\nu_i},x_{\nu_{i+1}}]$, where for convenience, we set $\nu_0=0$, $\nu_{q+1}=k+1$. Finally, set

$$s_i(x)=\begin{cases}s(x), & x_{\nu_i}\leq x<x_{\nu_{i+1}},\\ \lim\limits_{t\uparrow x}s(t), & x=x_{\nu_{i+1}}\end{cases}$$

for $i=0,1,\cdots,q$. We examine the zeros of s_i on J_i.

Case 1 (s_i vanishes identically on J_i). In this case J_i counts as an interval zero of multiplicity z_i, where

$$z_i=n_{\nu_i}+\sum_{j=\nu_i+1}^{\nu_{i+1}-1}(n_j-r_j).$$

Case 2 (s_i does not vanish identically on J_i). In this case s_i has at most z_i-1 zeros on J_i. Indeed, if it had more, then since it is continuous on this interval, we could apply Rolle's theorem 2.5 to deduce that $L_1^+s_i$ has z_i-1 zeros on J_i. But since $L_1^+s_i$ belongs to a (z_i-1)-dimensional space of splines on J_i, this would be a contradiction of the inductive hypothesis.

It remains to count the total number of zeros of s. Let α = number of intervals J_i where Case 2 applies. Then we have

$$Z^{\mathbb{S}}(s) \leq \sum_{i=0}^{q} Z_{J_i}(s_i) + (\alpha - 1) \leq \sum_{i=0}^{q} z_i - \alpha + \alpha - 1 = \sum_{i=0}^{q} z_i - 1,$$

since the number of zeros of s is at most equal to the total number of zeros of its pieces s_i in the intervals J_i, plus a possible extra 1 for each interval or knot where a sign change occurs. But there can be at most $\alpha - 1$ sign changes. □

It follows immediately from Theorem 2.6 that $\mathbb{S}$ is a weak-Chebyshev space. The technique of proof used in Theorem 2.6 can be used to establish a refined bound on $Z^{\mathbb{S}}(s)$ in which we take account of the behavior at the endpoints.

THEOREM 2.7 (Budan–Fourier theorem). *For any nontrivial spline $s \in \mathbb{S}$,*

$$Z^{\mathbb{S}}_{(a,b)}(s) \leq n - 1 - A(s,a) - B(s,b), \tag{2.7}$$

where for a general t,

$$\begin{aligned} A(s,t) &= S^{+}\left[s(t+), -L_1^{+}s(t), \cdots, (-1)^{\alpha-1}L_{\alpha-1}^{+}s(t)\right], \\ B(s,t) &= S^{+}\left[s(t-), L_1^{-}s(t), \cdots, L_{\beta-1}^{-}s(t)\right]. \end{aligned} \tag{2.8}$$

Here α and β are the exact orders of s on the intervals $[a, x_1)$ and $[x_k, b]$, respectively, and S^{+} counts weak sign changes (cf. [5]).

Proof. For $m = 1$, the result reduces to Theorem 2.6. We now proceed by induction on the order m of the spline. Assume the theorem is correct for splines of order $m-1$, and suppose that s is a spline of order m. If s is continuous on $[a,b]$, then we apply the same argument as in the polynomial- or T-spline case (cf. [5, Thms. 4.58 and 9.32]) to obtain the result. (Here it is necessary to use Rolle's theorem for ECT-systems—see [5, Thm. 9.11].)

If s is not continuous, then we break the interval $[a,b]$ into pieces in the same way as in the proof of Theorem 2.6. In particular, let $\nu_0 < \cdots < \nu_{q+1}$ be the integers defined there. Then the restriction of s to each interval $J_i = [x_{\nu_i}, x_{\nu_{i+1}}]$ is continuous. Let J_i and s_i be as in the proof of Theorem 2.6, $i = 0, \cdots, q$. We examine the zeros of s_i on J_i. In the interval J_0 we now use the Budan–Fourier theorem which we have already established. This gives us a bound of $z_0 - 1 - A(s,a)$. In the last interval we again use our result for continuous splines to give a bound of $z_q - 1 - B(s,b)$. Adding the zeros in the various intervals together and taking account of what happens at the knots $x_{\nu_1}, \cdots, x_{\nu_q}$, we arrive at (2.7). □

3. Interpolation. In this section we discuss interpolation using the n dimensional space $\mathbb{S}$ defined in (1.3). We begin by formulating three typical interpolation problems, each one more general than its predecessor. Throughout we assume $\{B_j\}_1^n$ is some basis for $\mathbb{S}$.

Problem 3.1 (Lagrange interpolation). Let $a \leq t_1 < \cdots < t_n \leq b$, and suppose that $v_1, \cdots, v_n$ are given real numbers. Find s such that

$$s(t_i) = v_i, \qquad i = 1, 2, \cdots, n. \tag{3.1}$$

Discussion. It is well known that this problem has a unique solution for every choice of v_i's provided that the determinant

$$D\begin{pmatrix} t_1,\cdots,t_n \\ B_1,\cdots,B_n \end{pmatrix} = \det\left(B_j(t_i)\right)_{i,j=1}^n \tag{3.2}$$

is nonzero.

Problem 3.2 (Hermite interpolation). Let $a \le t_1 \le \cdots \le t_n \le b$, and suppose that $v_1,\cdots,v_n$ are given real numbers. Define

$$d_i = \max\{\nu\colon t_i = \cdots = t_{i-\nu}\}, \qquad i=1,\cdots,n. \tag{3.3}$$

Find $s \in \mathbb{S}$ such that

$$L_{d_i}^+ s(t_i) = v_i, \qquad i=1,2,\cdots,n. \tag{3.4}$$

Discussion. Clearly we do not want to specify more interpolation conditions in any one subinterval of the partition than the dimension of the ECT-space from which that piece of the spline is drawn. Thus, in posing the Hermite interpolation problem, it is natural to make the assumption

$$\text{if } x_{j-1} \le t_i < x_j, \text{ then } d_i < n_{j-1}, \tag{3.5}$$

for all $i=1,\cdots,n$. The unique solvability of the Hermite interpolation problem is equivalent to the nonvanishing of the following determinant:

$$D\begin{pmatrix} t_1,\cdots,t_n \\ B_1,\cdots,B_n \end{pmatrix} = \det\left(L_{d_i}^+ B_j(t_i)\right)_{i,j=1}^n. \tag{3.6}$$

Problem 3.3 (extended Hermite interpolation). Let $a \le t_1 \le \cdots \le t_n \le b$, and for $i=1,\cdots,n$ define

$$\rho_i = \begin{cases} r_j & \text{if } t_i = x_j \text{ for some } 1 \le j \le k, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

Suppose $\theta_1,\cdots,\theta_n$ is a sequence of signs, and define

$$d_i = \begin{cases} \max\{\nu\colon t_i = \cdots = t_{i-\nu} \text{ with } \theta_i = \cdots = \theta_{i-\nu}\} & \text{if } \theta_i = +, \\ \rho_i + \max\{\nu\colon t_i = \cdots = t_{i+\nu} \text{ and } \theta_i = \cdots = \theta_{i+\nu}\} & \text{if } \theta_i = - \end{cases} \tag{3.8}$$

for $i=1,\cdots,n$. Then given real numbers $v_1,\cdots,v_n$, we seek $s \in \mathbb{S}$ such that

$$L_{d_i}^{\theta_i} s(t_i) = v_i, \qquad i=1,\cdots,n. \tag{3.9}$$

Discussion. This problem generalizes the Hermite interpolation Problem 3.2. (It reduces to it if we take all θ_i's to be $+$ except for those associated with t_i's which are equal to b.) The idea here is that if t_i falls at a knot x_j, then we can specify as many right derivatives as we want (up to $n_j - 1$), and that in addition, we can also specify left derivatives of order r_j and up at the same point. In order for the problem to make sense, we impose the following restrictions:

$$\text{if } t_i \notin \Delta \cup \{x_{k+1}\}, \text{ then } \theta_i = +, \tag{3.10}$$

$$\text{if } t_i = x_{k+1}, \text{ then } \theta_i = -, \tag{3.11}$$

$$\text{if } x_{j-1} < t_i < x_j, \text{ then } d_i < n_{j-1}, \tag{3.12}$$

$$\text{if } t_i=x_j, \text{ then } d_i<\begin{cases} n_j & \text{if } \theta_i=+, \\ n_{j-1} & \text{if } \theta_i=-, \end{cases} \tag{3.13}$$

$$\text{if } \theta_i=+ \text{ and } \theta_{i+1}=-, \text{ then } t_i<t_{i+1}, \tag{3.14}$$

$$\text{if } \theta_i=- \text{ with } t_i<x_{k+1}, \text{ then there exists } j>i \text{ with } \theta_j=+ \text{ and } t_i=t_j, \tag{3.15}$$

$$\text{if } \theta_i=- \text{ and } \theta_{i+1}=+, \text{ then for some } j, \tag{3.16}$$

$$t_i=\cdots=t_{i+r_j}=x_j \text{ and } \theta_{i+1}=\cdots=\theta_{i+r_j}=+.$$

Condition (3.10) requires that we specify only right derivatives at points which are not knots. Condition (3.11) requires that we specify only left derivatives at x_{k+1}. Conditions (3.12)–(3.13) are to insure that the number of interpolation conditions forced on a piece of the spline does not exceed the dimension of the space to which that piece belongs. Condition (3.14) makes sure that the t's are in a natural lexicographical order. Finally, conditions (3.15)–(3.16) insure that a full set of right derivatives is specified at a point before any left ones are.

It is now clear that the extended Hermite interpolation problem has a unique solution for an arbitrary set of data values $v_1,\cdots,v_n$ if and only if the determinant

$$D\begin{pmatrix} t_1,\cdots,t_n \\ \theta_1,\cdots,\theta_n \\ B_1,\cdots,B_n \end{pmatrix}=\det\left[\theta_i^{d_i}L_{d_i}^{\theta_i}B_j(t_i)\right]_{i,j=1}^{n} \tag{3.17}$$

is nonzero. (We have introduced the powers of θ_i's in the definition of this determinant in order to make it have a certain sign—see Theorem 4.5 below.)

We are now ready for the main result of this section, in which we give certain interlacing conditions which are equivalent to the nonvanishing of the determinant in (3.17). We shall specialize this result later to Hermite and Lagrange interpolation.

THEOREM 3.4. *Suppose* $\mathbb{S}$ *is the spline space defined in* (1.3), *and that* $B_1,\cdots,B_n$ *is any basis for it. Let* $a\le t_1\le\cdots\le t_n\le b$ *and suppose* $\theta_1,\cdots,\theta_n$ *is a corresponding sequence of signs such that* (3.10)–(3.16) *are satisfied. Then*

$$D\begin{pmatrix} t_1,\cdots,t_n \\ \theta_1,\cdots,\theta_n \\ B_1,\cdots,B_n \end{pmatrix}\neq 0 \tag{3.18}$$

if and only if for each $i=1,\cdots,k$ *the following conditions are satisfied*:

$$t_{n-n_{i,k+1}}\le x_i\le t_{n_{0i}+1}, \tag{3.19}$$

$$\textit{if } x_i=t_{n_{0i}+1}, \textit{ then } t_{n_{0i}+1}=\cdots=t_{n_{0i}+1-r_i} \textit{ and } \theta_{n_{0i}+1}=\cdots=\theta_{n_{0i}+1-r_i}=+, \tag{3.20}$$

$$\textit{if } t_{n-n_{i,k+1}}=x_i, \textit{ then } \theta_{n-n_{i,k+1}}=-. \tag{3.21}$$

Here

$$n_{\nu\mu}=\dim\mathbb{S}|_{[x_\nu,x_\mu)} \quad \textit{all } 0\le\nu<\mu\le k+1. \tag{3.22}$$

Proof. By the discussion of the extended Hermite interpolation Problem 3.3, if (3.18) holds, then we can find a spline $s\in\mathbb{S}$ interpolating arbitrary data. Now if

(3.19)–(3.21) fails, this would mean that we could interpolate $n_{0i}+1$ pieces of data at points in $[a,x_i]$ by a spline which comes from a n_{0i}-dimensional spline space, or we could interpolate $n_{i,k+1}+1$ pieces of data at points in $[x_i,b]$ by a spline which comes from a $n_{i,k+1}$-dimensional spline space. In either case we have a contradiction, and we have proved that (3.18) implies (3.19)–(3.21).

We turn now to the converse. Suppose that (3.19)–(3.21) hold, but that (3.18) does not. Then there exists some nontrivial spline $s\in\mathbb{S}$ with

$$L_{d_i}^{\theta_i}s(t_i)=0, \qquad i=1,\cdots,n.$$

We now show that this leads to a contradiction of our results on the zeros of gT-splines.

Let $J=[x_i,x_j]$ be a largest subinterval of the partition such that s does not vanish on any subinterval of J and such that there is no t_ν in (x_i,x_j) with $\theta_\nu=-$. It is easy to see that

$$n_{ij}=\dim\mathbb{S}|_J=n_i+\sum_{\nu=i+1}^{j-1}(n_\nu-r_\nu)=n-n_{0i}-n_{j,k+1}+r_i+r_j$$

where we recall our convention that $r_0=r_{k+1}=0$. We now show that

$$Z^{\mathbb{S}}(\tilde{s})\geq n_{ij} \quad \text{where } \tilde{s}=s|_J. \tag{3.23}$$

To count the zeros of $\tilde{s}$, we need to know exactly how many t's are equal to x_i and x_j. Suppose $0\leq l,r$ are integers such that

$$\begin{aligned}x_i=t_{n_{0i}+1}=\cdots=t_{n_{0i}+l}&<t_{n_{0i}+l+1}\leq\cdots\\ \leq t_{n-n_{j,k+1}-r}&<t_{n-n_{j,k+1}-r+1}=\cdots=t_{n-n_{j,k+1}}=x_j.\end{aligned}$$

Clearly $\tilde{s}$ has $n-n_{0i}-n_{j,k+1}-r-l$ zeros in (x_i,x_j). It remains to count the multiplicities at the ends.

First, we claim that $\tilde{s}$ has an (r_i+l)-tuple zero at x_i. This is clear if $i=0$ as $r_0=0$. Suppose now that $i>0$. Then either $\tilde{s}$ vanishes identically to the left of x_i, or some left derivatives are specified at x_i. In either case $\tilde{s}$ has at least an r_i-tuple zero at x_i and our claim is correct if $l=0$. It remains to check the case where $l>0$. In this case (3.20) implies that

$$\begin{aligned}t_{n_{0i}+1-r_i}&=\cdots=t_{n_{0i}}=x_i \quad\text{with}\\ \theta_{n_{0i}+1-r_i}&=\cdots=\theta_{n_{0i}}=+\end{aligned}$$

and since $t_{n_{0i}+1},\cdots,t_{n_{0i}+l}$ also fall at x_i with positive θ's, we conclude that $\tilde{s}$ has an (r_i+l)-tuple zero at x_i as asserted.

The analysis at the right end point x_j is similar. We claim that $\tilde{s}$ has an (r_j+r)-tuple zero at x_j. If $j=k+1$ this is clear since we define $r_{k+1}=0$. Suppose now that $j<k+1$. Then either $\tilde{s}$ vanishes identically to the right of x_j, or some left derivatives are specified at x_j. In either case $\tilde{s}$ has at least an r_j-tuple zero at x_j and our claim is correct if $r=0$. It remains to check the case where $r>0$. In this case we have $t_{n-n_{j,k+1}}=x_j$, and thus by (3.21), $\theta_{n-n_{j,k+1}}=-$. But then by (3.15)–(3.16), there are at least an additional r_j t's with positive θ's which fall at the point x_j. It follows that $\tilde{s}$ has an r_j+r-tuple zero at x_j as asserted.

We now add up the zeros of $\tilde{s}$ in J. We have r_i+l at x_i, r_j+r at x_j, and a total of $n-n_{0i}-n_{j,k+1}-r-l$ in (x_i,x_j). This adds up to n_{ij}, and we have established (3.23). Since n_{ij} is the dimension of the space from which $\tilde{s}$ came, we have a contradiction of Theorem 2.6. This completes the proof of the theorem. □

In the following two corollaries we specialize this result to the cases of Hermite and Lagrange interpolation.

COROLLARY 3.5. *Suppose* $\mathbb{S}$ *is the spline space defined in* (1.3), *and that* $B_1,\cdots,B_n$ *is a basis for it. Let* $a\le t_1\le\cdots\le t_n\le b$ *be a set of points defining a Hermite interpolation problem as in Problem* 3.2 *and satisfying* (3.5). *Then*

$$D\begin{pmatrix} t_1,\cdots,t_n \\ B_1,\cdots,B_n \end{pmatrix}\neq 0 \tag{3.24}$$

if and only if for each $i=1,\cdots,k$ *the following conditions hold*:

$$t_{n-n_{i,k+1}}<x_i\le t_{n_{0i}+1}, \tag{3.25}$$

$$\textit{if } t_{n_{0i}+1}=x_i, \textit{ then } t_{n_{0i}+1}=\cdots=t_{n_{0i}+1-r_i}. \tag{3.26}$$

COROLLARY 3.6. *Suppose that* $\mathbb{S}$ *is the spline space defined in* (1.3) *and that* $B_1,\cdots,B_n$ *is a basis for it. Let* $a\le t_1<\cdots<t_n\le b$. *Then*

$$D\begin{pmatrix} t_1,\cdots,t_n \\ B_1,\cdots,B_n \end{pmatrix}\neq 0 \tag{3.27}$$

if and only if

$$t_{n-n_{i,k+1}}<x_i<t_{n_{0i}+1}, \qquad i=1,\cdots,k, \tag{3.28}$$

where equality is allowed on the right when $r_i=0$.

Corollary 3.6 asserts that if $r_i>0$, $i=1,\cdots,k$ so that $\mathbb{S}\subseteq C[a,b]$, then $\mathbb{S}$ satisfies the interlacing property of [2]. It then follows from [2, Thm. 2.5] that $\mathbb{S}$ has the properties listed in the following corollary.

COROLLARY 3.7. *Suppose* $\mathbb{S}$ *is a gT-spline space as in* (1.3) *with* $r_i>0$, $i=1,\cdots,k$. *Then*

(3.29) $\mathbb{S}$ *is a weak-Chebyshev space*,

(3.30) $|\mathrm{bd}\,Z(s)|\le n_{ij}^0$ *for all* $s\in\mathbb{S}_{ij}^0$, $0\le i<j\le k+1$,

(3.31) $\dim\left(\mathbb{S}_{0p}^0\cap\mathbb{S}_{q,k+1}^0\right)=\max\{n-n_{0p}-n_{q,k+1},0\}, \qquad 1\le p<q\le k+1.$

Here

$$\mathbb{S}_{ij}^0=\{s\in\mathbb{S}: s(x)=0 \textit{ for all } x\in[x_i,x_j]\}, \tag{3.32}$$

$$n_{ij}^0=\dim\mathbb{S}_{ij}^0 \tag{3.33}$$

and

$Z(s)=\{x\in[a,b]: s(x)=0\}$,

$\mathrm{bd}\,Z(s)=$ *boundary of the set* $Z(s)$ *in* $[a,b]$,

$|\mathrm{bd}\,Z(s)|=$ *number of points in* bd $Z(s)$.

4. *B*-splines. In this section we construct a basis for the space $\mathcal{S}$ of gT-splines. Our basis will consist of analogues of the classical B-splines. We begin with some notation and a lemma which will be useful in our construction. Given $0 \le p \le k$ and $1 \le j \le n_p - r_p$, let

$$q(p,j) = \min\{\nu\colon \nu > p \text{ and } n_{p\nu} - r_p - j \ge r_\nu\}. \tag{4.1}$$

Such a q always exists in view of our convention that $r_0 = r_{k+1} = 0$.

LEMMA 4.1. *For any $0 \le p \le k$ and $1 \le j \le n_p - r_p$, the integer $q(p,j)$ satisfies*

$$r_q \le n_{pq} - r_p - j < n_{q-1}. \tag{4.2}$$

Proof. The left-hand inequality follows from the definition of q. The definition also implies $n_{p,q-1} - r_p - j < r_{q-1}$. Now using the elementary relationship $n_{pq} = n_{p,q-1} + (n_{q-1} - r_{q-1})$ leads to (4.2). □

Our next theorem states the existence of certain splines in the space $\mathcal{S}$ with special properties. In view of these properties, we call them *B-splines.*

THEOREM 4.2. *Fix $0 \le p \le k$ and $1 \le j \le n_p - r_p$, and let $q = q(p,j)$ be defined as in (4.1). Then there exists a spline $M_{pj} \in \mathcal{S}$ with*

$$M_{pj}(x) > 0 \quad \textit{for all } x_p < x < x_q, \tag{4.3}$$

$$M_{pj}(x) = 0 \quad \textit{for all } x \in [x_0, x_p) \cup (x_q, x_{k+1}], \tag{4.4}$$

$$\int_{x_p}^{x_q} M_{pj}(x)\,dx = 1, \tag{4.5}$$

$$L_{\nu-1}^{+} M_{pj}(x_p) = 0, \qquad \nu = 1, \cdots, r_p + j - 1, \tag{4.6}$$

$$L_{r_p+j-1}^{+} M_{pj}(x_p) > 0, \tag{4.7}$$

$$L_{\nu-1}^{-} M_{pj}(x_q) = 0, \qquad \nu = 1, \cdots, n_{pq} - r_p - j. \tag{4.8}$$

Proof. Consider the spline space $\tilde{\mathcal{S}} = \mathcal{S}(\tilde{\mathfrak{U}}; \tilde{\mathfrak{M}}; \tilde{\mathfrak{R}}; \tilde{\Delta})$ on $[x_p, x_q]$ corresponding to $\tilde{\Delta} = \{x_{p+1}, \cdots, x_{q-1}\}$, $\tilde{\mathfrak{M}} = (n_p, \cdots, n_{q-1})$, and $\tilde{\mathfrak{R}} = (r_{p+1}, \cdots, r_{q-1})$. Clearly $\tilde{\mathcal{S}}$ is of dimension n_{pq}. Now consider finding $\varphi \in \tilde{\mathcal{S}}$ satisfying the Hermite interpolation problem

$$\begin{aligned} &L_{\nu-1}^{+} \varphi(x_p) = 0, \qquad \nu = 1, \cdots, r_p + j - 1, \\ &L_{r_p+j-1}^{+} \varphi(x_p) = 1, \\ &L_{\nu-1}^{-} \varphi(x_q) = 0, \qquad \nu = 1, \cdots, n_{pq} - r_p - j. \end{aligned}$$

Lemma 4.1 assures that this problem makes sense, and it is easily checked that the interlacing conditions of Corollary 3.5 are satisfied, and hence there exists a unique $\varphi \in \tilde{\mathcal{S}}$ with these properties. By Theorem 2.6, φ can have at most $n_{pq} - 1$ zeros in $[x_p, x_q]$. Since it has precisely these many zeros (counting multiplicities) at the points x_p and x_q, it cannot have any other zeros in (x_p, x_q). Since the first nonzero derivative of φ is positive at x_p, it follows that φ is positive throughout (x_p, x_q). Now we define

$$M_{pq}(x) = \begin{cases} \varphi(x) \Big/ \displaystyle\int_{x_p}^{x_q} \varphi(t)\,dt, & x_p \le x < x_q, \\ 0 & \text{otherwise.} \end{cases}$$

It is easily checked that $M_{pq} \in \mathcal{S}$ and satisfies (4.4)–(4.8). □

It is clear from the definition of the B-spline M_{pj} that its support starts at the point x_p. How far its support extends to the right depends on the size of q—in some cases we may have to choose $q=k+1$ in which case the support set will extend all the way to b. (See Example 5.1.) The following theorem shows that each B-spline M_{pj} has a minimal support set.

THEOREM 4.3. *Fix $0\leq p\leq k$ and $1\leq j\leq n_p-r_p$, and let M_{pj} be the B-spline defined in Theorem* 4.2. *Then there does not exist any spline $s\in\mathbb{S}$ and integer $\tilde{q}<q$ with*

$$s(x)=0 \quad \text{for all } x\in[x_0,x_p)\cup(x_{\tilde{q}},x_{k+1}], \tag{4.9}$$

$$s(x)>0 \quad \text{for all } x_p<x<x_{\tilde{q}}, \tag{4.10}$$

$$L_{\nu-1}^{+}s(x_p)=0, \qquad \nu=1,\cdots,r_p+j-1. \tag{4.11}$$

Proof. Indeed, we can show that there does not exist any nontrivial spline $s\in\mathbb{S}$ satisfying (4.9), (4.11) and not vanishing on subintervals of $[x_p,x_{\tilde{q}}]$. If s were such a spline, then it would have an (r_p+j-1)-tuple zero at x_p and a $r_{\tilde{q}}$ -tuple zero at $x_{\tilde{q}}$. By the definition of $q, n_{p\tilde{q}}-r_p-j<r_{\tilde{q}}$, and we conclude that s has at least

$$r_p+j-1+r_{\tilde{q}}>n_{p\tilde{q}}-1$$

zeros on $[x_p,x_{\tilde{q}}]$. But since s restricted to this interval is a member of a spline space of dimension $n_{p\tilde{q}}$, this contradicts Theorem 2.6. □

We now show that the B-splines constructed in Theorem 4.2 form a basis for the gT-spline space $\mathbb{S}$.

THEOREM 4.4. *The B-splines $\{M_{pj}\}_{j=1,p=0}^{n_p-r_p,k}$ form a basis for $\mathbb{S}$.*

Proof. Suppose that for some set of coefficients

$$\sum_{p=0}^{k}\sum_{j=1}^{n_p-r_p} c_{pj}M_{pj}(x)=0 \quad \text{for all } x_0<x<x_{k+1}.$$

Then on the interval $[x_0,x_1)$ we have $c_{01}M_{01}+\cdots+c_{0,n_0}M_{0,n_0}=0$. But then properties (4.6)–(4.7) show that $c_{01}=\cdots=c_{0,n_0}=0$. Now we can look in the interval $[x_1,x_2)$ to eliminate the next set of c's (namely $c_{11},\cdots,c_{1,n_1-r_1}$). This process can be continued to show that all of the c's are 0, thus establishing the linear independence of the M's. As the M's are clearly in $\mathbb{S}$, the theorem is established. □

In §3 we showed that certain determinants formed with any basis for the spline space $\mathbb{S}$ are nonzero under appropriate interlacing conditions. We now apply those results to obtain some important total positivity properties of the B-splines.

THEOREM 4.5. *Let $B_1,\cdots,B_n$ be the B-spline basis for $\mathbb{S}$ given in Theorem* 4.4. *Suppose that $a\leq t_1\leq\cdots\leq t_n\leq b$ is a set of points and $\theta_1,\cdots,\theta_n$ is a sequence of signs such that conditions* (3.10)–(3.16) *are satisfied. Then*

$$D\begin{pmatrix} t_1,\cdots,t_n \\ \theta_1,\cdots,\theta_n \\ B_1,\cdots,B_n \end{pmatrix}\geq 0 \tag{4.12}$$

and strict positivity holds if and only if (3.19)–(3.21) *hold.*

Proof. We already know that D is nonzero precisely under the conditions (3.19)–(3.21). The fact that D has one sign (and that that sign is $+$) is established by the same continuity and perturbation argument which works in the polynomial spline case (cf. [5, Thm. 4.72]).

Theorem 4.5 can now be used to obtain a total positivity result.

THEOREM 4.6. *For any integers* $1 \le \nu_1 < \cdots < \nu_p \le n$, *any points* $a \le t_1 \le \cdots \le t_p \le b$, *and any sequence of signs* $\theta_1, \cdots, \theta_p$,

$$D\begin{pmatrix} t_1, \cdots, t_p \\ \theta_1, \cdots, \theta_p \\ B_{\nu_1}, \cdots, B_{\nu_p} \end{pmatrix} \ge 0 \tag{4.13}$$

and strict positivity holds if and only if for $i = 1, \cdots, p$

$$t_i \in \sigma\left(L^{\theta_i}_{d_i} B_{\nu_i}\right) = \left\{x\colon B_{\nu_i}(x) \neq 0\right\} \cup \left\{x\colon L^{\theta_i}_{d_i} B_{\nu_i}(x) \neq 0\right\}. \tag{4.14}$$

Proof. The proof uses the same algebraic argument as for polynomial splines—cf. [5, Thm. 4.73].

Theorem 4.6 asserts that the B-splines $\{B_j\}_1^n$ forming a basis for the gT-space $\mathcal{S}$ form an order-complete-weak-Chebyshev-(OCWT-) system—see [5, p. 41]. It then follows immediately (cf. [5, Thm. 2.42]) that gT-splines have the following *variation diminishing property*:

$$S^-\left(\sum_{j=1}^{n} c_j B_j\right) \le S^-(c_1, \cdots, c_n) \tag{4.15}$$

for every nontrivial B-spline expansion. Here S^- counts strong sign changes (cf. [5]).

5. Examples. In this section we give two examples of gT-spline spaces to illustrate the above material. In addition, we give two further examples to show that some natural related generalized spline spaces do not retain all of the features of gT-splines.

Example 5.1. Let $[a,b] = [0,3]$, $\mathcal{U} = \operatorname{span}\{1, x, x^2\}$, $\Delta = \{1,2\}$, $\mathcal{M} = (3,1,2)$ and $\mathcal{R} = (1,1)$. Let $\mathcal{S} = \mathcal{S}(\mathcal{U}; \mathcal{M}; \mathcal{R}; \Delta)$.

Discussion. It is easily checked that $\dim \mathcal{S} = 4$. Each spline in $\mathcal{S}$ belongs to $C[0,3]$ and consists of quadratic, constant, and linear polynomials, respectively, on the three subintervals defined by the partition. The B-spline basis for this space is given by

$$M_{01}(x) = \begin{cases} 3(1-x)^2, & 0 \le x < 1, \\ 0 & \text{otherwise}, \end{cases}$$

$$M_{02}(x) = \begin{cases} 6x(1-x), & 0 \le x < 1, \\ 0 & \text{otherwise}, \end{cases}$$

$$M_{03}(x) = \frac{6}{11}\begin{cases} x^2 & 0 \le x < 1, \\ 1, & 1 \le x < 2, \\ 3 - x, & 2 \le x \le 3, \end{cases}$$

$$M_{21}(x) = 2(x-2)_+.$$

Example 5.2. Let $[a,b] = [0,3]$, $\mathcal{U} = \operatorname{span}\{1, x, x^2\}$, $\Delta = (1,2)$, $\mathcal{M} = (3,0,2)$, and $\mathcal{R} = (0,1)$.

Discussion. It is easily checked that $\dim \mathcal{S} = 4$. Here the splines in $\mathcal{S}$ must vanish identically on $[1,2]$ and be continuous at 2. They may have a jump discontinuity at the knot located at 1. In this case the B-spline basis for $\mathcal{S}$ is given by the B-splines M_{01}, M_{02}, and M_{21} coupled with

$$M_{03}(x) = \begin{cases} 3x^2, & 0 \le x < 1, \\ 0 & \text{otherwise}. \end{cases}$$

A natural way to generalize the space of gT-splines considered in this paper is to allow the spaces $\mathcal{U}_1,\cdots,\mathcal{U}_k$ from which the pieces of the splines are drawn to be arbitrary ECT-spaces (rather than all subspaces of one fixed ECT-space). The following example shows that, in general, we will not be able to perform interpolation with such generalized splines.

Example 5.3. Let $[a,b]=[-15,5]$, and let $\Delta=\{1\}$. Suppose $\mathcal{U}_0=\operatorname{span}\{1,x,x^2\}$ and $\mathcal{U}_1=\operatorname{span}\{x,x^2,x^3\}$. Let

$$\mathcal{S}=\left\{s\in C^1[-15,5]: s|_{[x_i,x_{i+1})}\in\mathcal{U}_i i=0,1\right\}.$$

Discussion. $\mathcal{U}_0$ is an ECT-space on $\mathbb{R}$ while $\mathcal{U}_1$ is an ECT-space on $[1,\infty)$. It is easily checked that $\dim\mathcal{S}=4$. Thus it would be natural to consider interpolation at four points. Then the analogue of the interlacing condition (3.28) for Lagrange interpolation would be $t_1<x_1<t_4$. But the points $t_1=-14$, $t_2=-9$, $t_3=3$, and $t_4=4$ satisfy this interlacing condition, while Lagrange interpolation at these four points is not uniquely defined. Indeed, the spline

$$s(x)=\begin{cases}(x+9)(x+14), & 0\le x<1,\\ 25x(x-3)(x-4), & 1\le x\le 2\end{cases}$$

interpolates 0 values at these four points.

Our last example shows that if we weaken our assumption on the space $\mathcal{S}$ from which the pieces of our gT-splines are drawn, then the Hermite interpolation property may no longer hold.

Example 5.4. Let $[a,b]=[0,2\pi]$ and $\Delta=\{\pi\}$. Let $\mathcal{U}=\operatorname{span}\{1,\cos(x),\sin(x)\}$, $\mathcal{M}=(3,3)$ and $\mathcal{R}=(2)$. Consider the spline space $\mathcal{S}=\mathcal{S}(\mathcal{U};\mathcal{M};\mathcal{R};\Delta)$.

Discussion. It is well known that $\mathcal{U}$ is an ET-space on any subinterval of $[a,b]$ which does not contain both end points—thus in particular, it is an ET-space on both $[0,\pi]$ and $[\pi,2\pi]$. The smoothness conditions assure that each spline $s\in\mathcal{S}$ belongs to $C^1[0,2\pi]$. It is easily checked that $\dim\mathcal{S}=4$. Now consider Hermite interpolation at the four points $t_1=t_2=0$ and $t_3=t_4=2\pi$. These points satisfy the interlacing conditions (3.25)–(3.26) for Hermite interpolation. But Hermite interpolation at these four points is not uniquely defined since the spline $s(x)=1-\cos(x)$ interpolates 0 at these points.

Although Hermite interpolation does not work for this spline space, it is interesting to note that Lagrange interpolation is always possible at any set of points $t_1<\cdots<t_4$ which satisfy the interlacing conditions (3.28). This fact follows from [2, Thm. 2.5] once we check its hypotheses. These hypotheses are precisely conditions (3.29)–(3.31).

We begin with (3.29). We need to show that there is no spline $s\in\mathcal{S}$ with 4 sign changes. Suppose s were such a spline. Then s' would have at least 3 zeros in $(0,2\pi)$, and thus at least 2 zeros in one of the intervals $(0,\pi]$ or $[\pi,2\pi)$. In either case this is a contradiction since $\mathcal{U}'=\operatorname{span}\{\cos(x),-\sin(x)\}$ is a Chebyshev space on both $(0,\pi]$ and $[\pi,2\pi)$.

Condition (3.30) is easy to check since the space $\mathcal{S}_{01}^0$ is spanned by the spline

$$s_1(x)=\begin{cases}0, & 0\le x<\pi,\\ 1+\cos(x), & \pi\le x\le 2\pi.\end{cases}$$

while $\mathcal{S}_{12}^0$ is spanned by

$$s_2(x)=\begin{cases}1+\cos(x), & 0\le x<\pi,\\ 0, & \pi\le x\le 2\pi.\end{cases}$$

Finally, condition (3.31) is trivial.

6. Remarks. 1). For some historical notes on Chebyshevian splines, see [5, p. 417]. Some notes on generalized splines can be found on [5, p. 482].

2). Throughout this paper we have assumed that the basic space $\mathcal{U}$ is spanned by an ECT-system. All of the results given here are also valid in the more general case where $\mathcal{U}$ is spanned by a *canonical complete Chebyshev* (CCT-) system as introduced in [4]. A CCT-system is a set of functions defined on $[a,b]$ such that u_1 is positive and

$$u_2(x)=u_1(x)\int_a^x d\sigma_2(s_2),$$
$$\cdots$$
$$u_m(x)=u_1(x)\int_a^x\cdots\int_a^{s_{m-1}} d\sigma_m(s_m)\cdots d\sigma_2(s_2),$$

where $\sigma_2,\cdots,\sigma_m$ are bounded, right continuous, monotone increasing functions on $[a,b]$. In this case we must replace the differential operators $D_0,\cdots,D_m$ by the following operators:

$$D_0f(x)=\frac{f(x)}{u_1(x)},$$
$$D_jf(x)=\lim_{\delta\downarrow 0}\frac{f(x+\delta)-f(x)}{\sigma_{j+1}(x+\delta)-\sigma_{j+1}(x)},\qquad j=1,\cdots,m-1.$$

See also [5, p. 407].

3). We have posed the Hermite and extended Hermite interpolation problems in terms of the operators L_j (cf. (3.4) and (3.9)). It follows immediately from the form of the L's that it is equivalent to work with ordinary derivatives; i.e., the specification of $f(t), L_1f(t),\cdots,L_rf(t)$ is equivalent to the specification of $f(t), Df(t),\cdots,D^rf(t)$.

4). While interlacing conditions are well known in the theory of splines, the general kind of interlacing which appears in (3.19) is relatively novel. This kind of condition was introduced in [7] in connection with the study of generalized spline spaces consisting of pieces of Chebyshev spaces tied together continuously.

5). For a discussion of the history of extended Hermite interpolation problems, see the historical notes for [5, §4.8].

6). A complete characterization of all generalized spline spaces for which Lagrange interpolation is possible if and only if the interpolation points interlace the knots of the spline appropriately is given in [2].

7). The space of gT-splines (along with certain other spaces of generalized splines having a certain interlacing property) enjoys a variety of nice approximation properties. For a treatment of approximation by generalized splines, see [3].

8). Usually B-splines are introduced by some kind of divided difference process. Our approach here via Hermite interpolation is entirely different. The divided difference approach leads to a number of very nice results for polynomial and Chebyshevian splines which are missing here. These include a partition of unity, a Marsden identity, and recursions for computing the B-splines. (See [6] for more on recursions for generalized B-splines.)

9). Many of the results presented here have analogues for certain spaces of periodic and discrete generalized Chebyshevian splines.

REFERENCES

[1] S. KARLIN AND Z. ZIEGLER, *Chebyshevian spline functions*, SIAM J. Numer. Anal., 3 (1966), pp. 514–543.

[2] G. NÜRNBERGER, L. L. SCHUMAKER, M. SOMMER AND H. STRAUSS, *Interpolation by generalized splines*, CAT #22, 1982; Numer. Math., to appear.

[3] ———, *Approximation by generalized splines*, CAT #24, 1982.

[4] L. L. SCHUMAKER, *On Tchebycheffian spline functions*, J. Approx. Theory, 18 (1976), pp. 278–303.

[5] ———, *Spline Functions: Basic Theory*, Wiley-Interscience, New York, 1981.

[6] ———, *On recursions for generalized splines*, J. Approx. Theory, 36 (1982), pp. 16–31.

[7] M. SOMMER, *Characterization of continuous selections of the metric projection for generalized spline functions*, this Journal, 11 (1980), pp. 23–40.

© 1984 Society for Industrial and Applied Mathematics

FORMULAS FOR ELEMENTARY SPHERICAL FUNCTIONS AND GENERALIZED JACOBI POLYNOMIALS*

LARS VRETARE†

Abstract. Elementary spherical functions on symmetric spaces can be considered as orthogonal polynomials in several variables. This paper deals with the weight function

$$w(x_1,\cdots,x_l)=\Pi(1-x_i)^{\alpha}(1+x_i)^{\beta}\Pi(x_i-x_j)^{2\gamma+1},\qquad -1\le x_i\le 1.$$

Recurrence relations for polynomials corresponding to different spaces are derived and generalized to other values of the parameters α,β and γ.

1. Introduction. Elementary spherical functions or more generally intertwining functions on real and complex Grassmann manifolds were considered as generalized Jacobi polynomials, $p_N^{\alpha,\beta,\gamma}$, by James and Constantine [6]. See also [5]. In the case of two variables, these polynomials have been studied for general values of the parameters α,β and γ as well. This was done by Koornwinder [7] together with Sprinkhuizen-Kuyper [10] and [9]. See also [8].

In the present paper we consider generalized Jacobi polynomials in l variables. We establish the correspondence between these polynomials and intertwining functions on the spaces $K_m\backslash U/K_l$, where $U=SO(n)$, $SU(n)$ or $Sp(n)$ and $K_l=SO(n-l)\times SO(l)$, $S(U_{n-l}\times U_l)$ and $Sp(n-l)\times Sp(l)$ respectively. By combining the polynomial and the group-theoretical aspects we obtain formulas for $p_N^{\alpha,\beta,\gamma}$ which are generalizations of the well-known formulas for ordinary Jacobi polynomials

$$(x-1)p_n^{(\alpha+1,\beta)}(x)=\frac{2(n+1)}{2n+\alpha+\beta+2}p_{n+1}^{(\alpha,\beta)}(x)-\frac{2(n+\alpha+1)}{2n+\alpha+\beta+2}p_n^{(\alpha,\beta)}(x)$$

and

$$p_n^{(\alpha,\beta)}(x)=\frac{n+\alpha+\beta+1}{2n+\alpha+\beta+1}p_n^{(\alpha+1,\beta)}(x)-\frac{n+\beta}{2n+\alpha+\beta+1}p_{n-1}^{(\alpha+1,\beta)}(x).$$

The importance of such formulas is obvious. They express intertwining functions on one space in terms of intertwining functions on another space. For example it is possible to express elementary spherical functions on $SO(n)/SO(n-l)$ in terms of those on $SO(n+2)/SO(n+2-l)$.

We are also led to simple expressions for $p_N^{\alpha,\beta,\pm 1/2}$ in terms of ordinary Jacoby polynomials. In this way we get an explicit formula for the intertwining functions on

$$S(U_{n-m}\times U_m)\backslash SU(n)/S(U_{n-l}\times U_l)$$

which improves the result for elementary spherical functions obtained by Berezin and Karpelevič [1].

In the case of two and three variables where explicit calculations are possible we derive some more formulas of the same type.

The plan of the paper is as follows. After the necessary preliminaries in §2, concerning elementary spherical functions and intertwining functions, a more detailed investigation follows in §3, in particular of the orthogonality relations. The polynomials

*Received by the editors March 30, 1982.

† Department of Mathematics, Lund Institute of Technology, Lund, Sweden.

$p_N^{\alpha,\beta,\gamma}$ are introduced in §4, where we also establish their interpretation as intertwining functions. A list of possible values of α,β and γ is given in Theorem 4.2. Section 5 contains the statement and proof of the formulas and finally in §§6, 7 and 8 we consider special cases in two and three variables.

2. Notation and definition of intertwining functions. Let U/K be a compact symmetric space of type AIII, BDI or CII, i.e., U/K is one of the following spaces:

$$SU(n)/S(U_{n-l}\times U_l),\quad SO(n)/SO(n-l)\times SO(l),\quad Sp(n)/Sp(n-l)\times Sp(l).$$

Let $\mathfrak{u}$ and $\mathfrak{k}$ be the Lie algebras of U and K, and let

$$\mathfrak{u}=\mathfrak{k}+i\mathfrak{p}$$

be a decomposition of $\mathfrak{u}$ into eigenspaces of an involutive automorphism of $\mathfrak{u}$. Choose a maximal abelian subspace $\mathfrak{h}_\mathfrak{p}$ of $\mathfrak{p}$. Extend $i\mathfrak{h}_\mathfrak{p}$ to a maximal abelian subalgebra $\mathfrak{h}$ of $\mathfrak{u}$ and put

$$\mathfrak{h}_\mathfrak{k}=\mathfrak{h}\cap\mathfrak{k}.$$

When considering two symmetric spaces corresponding to the same group U we use the rank as an index to avoid confusion. For example the maximal compact subgroups of U will be denoted by K_l and K_m respectively. Throughout this paper we also assume that $l\le m$.

Let Λ be the highest weight of an irreducible representation, T_Λ, of U on a finite dimensional vector space V with scalar product $(\cdot,\cdot)$. If there is a unit vector $e\in V$ which is invariant under all $T_\Lambda(k)$, $k\in K$, the representation is said to be of *class one* with respect to K. In this case the function

$$\psi_\Lambda\colon u\to(e,T_\Lambda(u)e)$$

is constant on two-sided cosets $K\backslash U/K$. ψ_Λ will be called an elementary spherical function. More generally, if T_Λ is of class one with respect to K_l and K_m, the function

$$\psi_\Lambda^{l,m}\colon u\to(e_m,T_\Lambda(u)e_l)$$

is constant on $K_m\backslash U/K_l$. $\psi_\Lambda^{l,m}$ will be called an intertwining function.

The weights as well as the roots will be considered as real valued linear forms on $h_{\mathfrak{p}_l}$ we also assume that coordinates in $h_{\mathfrak{p}_l}$, denoted by $\theta_1,\cdots,\theta_l$, have been chosen in such a way that the roots are

$$\pm\theta_i\pm\theta_j,\quad \pm\theta_i,\quad \pm2\theta_i.$$

Among the roots we have the simple roots

$$\alpha_i=\theta_i-\theta_{i-1},\qquad i=1,\cdots,l-1$$

and

$$\alpha_l=\theta_l$$

also called the fundamental roots. Each weight of the form

$$\sum_{i=1}^{l} m_i\theta_i$$

where m_i are nonnegative integers, is the highest weight of some irreducible representation of U. In particular,

$$\mu_k = 2\sum_{i=1}^{k} \theta_i, \qquad k=1,\cdots,l$$

are called the fundamental weights for class one representations, which refers to the fact that the highest weights of class one representations with respect to K_l are those of the form

$$\Lambda = \sum_{k=1}^{l} n_k\mu_k, \qquad n_k \text{ nonnegative integers.}$$

Remark 2.1. Since the multiplicity of the roots $\pm 2\theta_i$ is zero in the case BDI, the correct definition of μ_l should be

$$\mu_l = \sum_{i=1}^{l} \theta_i.$$

Our definition of μ_l means that we do not consider all class one representations of

$$SO(n)/SO(n-l)\times SO(l)$$

but only those of

$$O(n)/O(n-l)\times O(l).$$

Let us identify $h_{\mathfrak{p}_l}$ and its dual by means of the Killing form $\langle\cdot,\cdot\rangle$. In this way $\theta_1,\cdots,\theta_l$ can be considered as an orthogonal basis for $h_{\mathfrak{p}_l}$. We also introduce a partial ordering in $h_{\mathfrak{p}_l}$ by setting

$$h_1 \preccurlyeq h_2 \quad \text{if } \langle h_2 - h_1, \mu_k\rangle \geq 0 \text{ for all } k.$$

The restriction to $\exp i\mathfrak{h}_{\mathfrak{p}_l}$ of the elementary spherical function $\psi_\Lambda^{l,l}$ is a trigonometric polynomial

$$\psi_\Lambda^{l,l}(\exp h) = \sum_{\lambda \preccurlyeq \Lambda} c(\lambda,\Lambda)e^{-\lambda(h)}, \qquad h\in i\mathfrak{h}_{\mathfrak{p}_l}.$$

It is invariant under the Weyl group W_l, i.e.

$$\psi_\Lambda^{l,l}(\exp Sh) = \psi_\Lambda^{l,l}(\exp h)$$

for all $S\in W_l$. W_l is the group generated by all permutations of $\theta_1,\cdots,\theta_l$ and arbitrary sign changes.

3. Further properties of intertwining functions. The following two lemmas are fundamental for further investigations of intertwining functions.

LEMMA 3.1. *There exists an automorphism of U*

$$u \to x^{-1}ux, \qquad x,u\in U,$$

such that

$$x^{-1}\exp i\mathfrak{h}_l\, x = \exp i\mathfrak{h}_m, \quad x^{-1}\exp i\mathfrak{h}_{\mathfrak{p}_l} x \subset \exp i\mathfrak{h}_{\mathfrak{p}_m}, \quad x^{-1}\exp i\mathfrak{h}_{\mathfrak{k}_l} x \supset \exp i\mathfrak{h}_{\mathfrak{k}_m}.$$

LEMMA 3.2. *To each $S \in W_l$ there is a $k \in K_l \cap K_m$ such that*

$$\exp Sh = k^{-1} \exp h k.$$

for all $h \in i\mathfrak{h}_{\mathfrak{p}_l}$.

Proof. The two lemmas can be verified directly in each of the three cases under consideration. Because of similarity we only treat C II (cf. [3, p. 351]).

Let I_n denote the unit matrix of order n and put

$$J = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}.$$

A matrix u belongs to the group $U = Sp(n)$ if and only if u is a unitary matrix of order $2n$, satisfying $Ju = \bar{u}J$. This latter condition means that u has the form

$$u = \begin{pmatrix} A & B \\ -\bar{B} & \bar{A} \end{pmatrix} \quad \text{or} \quad u = (u_1, \cdots, u_n, -J\bar{u}_1, \cdots, -J\bar{u}_n),$$

where $u_1, \cdots, u_n$ are column vectors of length $2n$. A maximal compact subgroup K_l, where $1 \leq l \leq n/2$, is given by $Sp(n-l) \times Sp(l)$ embedded into $Sp(n)$ according to the rule

$$\begin{pmatrix} A & B \\ -\bar{B} & \bar{A} \end{pmatrix}, \begin{pmatrix} C & D \\ -\bar{D} & \bar{C} \end{pmatrix} \rightarrow \begin{pmatrix} A & 0 & B & 0 \\ 0 & C & 0 & D \\ -\bar{B} & 0 & \bar{A} & 0 \\ 0 & -\bar{D} & 0 & \bar{C} \end{pmatrix}.$$

The Lie algebra $\mathfrak{u} = \mathfrak{sp}(n)$ is the set of all skew Hermitian matrices for which $Ju = \bar{u}J$ and $i\mathfrak{h}_{\mathfrak{p}_l}$ is given by the matrices of the form

$$\begin{pmatrix} 0 & \theta & 0 & 0 \\ \theta' & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta \\ 0 & 0 & -\theta' & 0 \end{pmatrix}$$

where the order of θ is $n - l \times l$ and

$$\theta = \begin{pmatrix} \theta_1 & & \\ & \ddots & \\ & & \theta_l \\ & & \\ & & \end{pmatrix}.$$

To verify Lemma 3.1 in this case let $\tilde{x}$ be obtained from I_n by permutation of suitable columns and put

$$x = \begin{pmatrix} \tilde{x} & 0 \\ 0 & \tilde{x} \end{pmatrix}.$$

Then $x \in U$ and since the mapping

$$y \rightarrow x^{-1} y x$$

permutes rows as well as columns, it is clear that

$$x^{-1} i\mathfrak{h}_{\mathfrak{p}_l} x \subset i\mathfrak{h}_{\mathfrak{p}_m}$$

for a suitable choice of x. Here the multiplication means ordinary matrix multiplication. If we now define $\mathfrak{h}_m$ by

$$\mathfrak{h}_m = x^{-1}\mathfrak{h}_l x,$$

we automatically obtain

$$x^{-1}i\mathfrak{h}_{\mathfrak{k}_l}x \supset i\mathfrak{h}_{\mathfrak{k}_m}$$

from the fact that $\mathfrak{h}_{\mathfrak{k}}$ is the orthogonal complement of $i\mathfrak{h}_{\mathfrak{p}}$ in $\mathfrak{h}$. This shows that Lemma 3.1 holds with x chosen as above.

For the verification of Lemma 3.2 we note that the Weyl group W_l is generated by transpositions and change of signs of $\theta_1, \cdots, \theta_l$. The corresponding $\tilde{k}$ and k can be chosen in a similar way as $\tilde{x}$ and x and it is easily checked that $k \in K_l \cap K_m$.

Consider two equivalent representations of U belonging to the same highest weight,

$$u \to T(u) \quad \text{and} \quad u \to T(x^{-1}ux).$$

On one hand the highest weight can be considered as a linear functional ν on $\mathfrak{h}_l$, corresponding to a weight vector f.

$$T(\exp h)f = e^{\nu(h)}f, \qquad h \in \mathfrak{h}_l.$$

On the other hand it can be considered as a linear functional μ on $\mathfrak{h}_m$ corresponding to the weight vector $T(x^{-1})f$. Since

$$T(x^{-1}\exp hx)T(x^{-1})f = e^{\nu(h)}T(x^{-1})f, \qquad h \in \mathfrak{h}_l$$

the relation between ν and μ must be

$$\mu(x^{-1}hx) = \nu(h), \qquad h \in \mathfrak{h}_l$$

In view of Lemma 3.1 we now see that if the restriction of ν to $\mathfrak{h}_{\mathfrak{k}_l}$ is zero, then the restriction of μ to $\mathfrak{h}_{\mathfrak{k}_m}$ is also zero. Using [2, Lemma 2] which states that if ν is zero on $\mathfrak{h}_{\mathfrak{k}_l}$ then 2ν is of class one with respect to K_l, we obtain

LEMMA 3.3. *Let ν be the highest weight of an irreducible representation of U and assume that the restriction of ν to $\mathfrak{h}_{\mathfrak{k}_l}$ is zero. Then there is an irreducible representation T_Λ, of class one with respect to K_l and K_m, having highest weight $\Lambda = 2\nu$.*

COROLLARY 3.4. *To each weight of the form*

$$\Lambda = \sum_{k=1}^{l} n_k \mu_k, \qquad n_k \text{ nonnegative integers}$$

there is an intertwining function $\psi_\Lambda^{l,m}$. Moreover, in the cases AIII *and* CII *there are no more.*

As an immediate consequence of Lemma 3.2 we have

COROLLARY 3.5. *The restriction of $\psi_\Lambda^{l,m}$ to* $\exp i\mathfrak{h}_{\mathfrak{p}_l}$ *is W_l-invariant. $\psi_\Lambda^{l,m}(Sh) = \psi_\Lambda^{l,m}(h)$, $S \in W_l$, $h \in \exp i\mathfrak{h}_{\mathfrak{p}_l}$.*

By use of the expansion of elementary spherical functions in [2, Lemma 2] we also get some information about the other weights appearing in the expansion of $\psi_\Lambda^{l,m}(\exp h)$, $h \in i\mathfrak{h}_{\mathfrak{p}_l}$. From

$$\psi_\Lambda^{l,l}(\exp h) = (e_l, T_\Lambda(\exp h)e_l) = \sum_{j=0}^{q} |c_j|^2 e^{-2\nu_j(h)}, \qquad c_0 \neq 0,$$

and

$$\psi_\Lambda^{m,m}(\exp h)=(e_m,T_\Lambda(\exp h)e_m)=\sum_{j=0}^{q}|d_j|^2e^{-2\nu_j(h)},\qquad d_0\neq 0,$$

we conclude that

$$\psi_\Lambda^{l,m}(\exp h)=(e_l,T_\Lambda(\exp h)e_m)=\sum_{j=0}^{q}c_j\bar{d}_je^{-2\nu_j(h)},\qquad c_0\bar{d}_0\neq 0.$$

Here ν_j denote the other weights of the representation with highest weight ν, and $\nu_0=\nu$. Note that $2\nu_j\preccurlyeq\Lambda$. The properties of $\psi_\Lambda^{l,m}$ obtained so far can now be summarized:

THEOREM 3.6. *The restriction to* $\exp i\mathfrak{h}_{\mathfrak{p}_l}$ *of the intertwining function* $\psi_\Lambda^{l,m}$ *is a* W_l *invariant trigonometric polynomial*

$$\psi_\Lambda^{l,m}(\exp h)=\sum_{\lambda\preccurlyeq\Lambda}c(\lambda,\Lambda)e^{-\lambda(h)},$$

where the sum ranges over weights of the form

$$\lambda=\sum_{k=1}^{l}n_k\mu_k,\qquad n_k\ \textit{integers}.$$

Belonging to nonequivalent irreducible representations of U, the intertwining functions must be orthogonal with respect to the invariant measure on $K_m\backslash U/K_l$. This measure has been computed by James [4] in the cases AIII and BDI. We now treat the remaining case CII. The method is due to James but we use a somewhat different formulation. Our first goal is to obtain a decomposition of a matrix $u\in Sp(n)$

$$u=k_m\exp h\,k_l,$$

where $k_m\in K_m$, $k_l\in K_l$ and $h\in i\mathfrak{h}_{\mathfrak{p}_l}$. This will be broken up into several lemmas.

LEMMA 3.7. *A skew Hermitian matrix A of order $2k$, satisfying $JA=\bar{A}J$ can be diagonalized by a matrix in $Sp(k)$. More precisely, there is a $S\in Sp(k)$ such that*

$$A=S^*\begin{pmatrix}\Lambda & 0\\ 0 & \Lambda\end{pmatrix}S,\qquad \Lambda=\operatorname{diag}(\lambda_1,\cdots,\lambda_k)$$

Proof. In the space of column vectors of length $2k$, equipped with the scalar product u^*v, any vector u is orthogonal to $-J\bar{u}$. Furthermore if E_λ denotes the eigenspace of A corresponding to the real eigenvalue λ then, by the assumption $JA=\bar{A}J$, it is clear that $u\in E_\lambda$ implies that $-J\bar{u}\in E_\lambda$ and vice versa. To build up a basis for E_λ assume that there are orthogonal column vectors

$$u_1,\cdots,u_{j-1},\ -J\bar{u}_1,\cdots,-J\bar{u}_{j-1}$$

spanning a proper subspace D of E_λ. Choose $u_j\in D^\perp$. Then $-J\bar{u}_j$ also belongs to $D^\perp$. In this way all the columns of S^* can be chosen,

$$S^*=(u_1,\cdots,u_k,-J\bar{u}_1,\cdots,-J\bar{u}_k).$$

Hence S^*, and then also S, belong to $Sp(k)$.

LEMMA 3.8. *Let X and Y be matrices with equal number of rows and assume that X as well as Y have orthogonal column vectors. Then all eigenvalues of X^*YY^*X lie in the interval* $0\le\lambda\le 1$.

Proof. Denote the columns of X by $x_1,\cdots,x_q$ and the orthogonal projection of x_i on the space spanned by the columns of Y by Px_i i.e.

$$YY^*X=(Px_1,\cdots,Px_q).$$

Then since

$$|Px_i|\le|x_i|$$

the diagonal elements in the matrix,

$$(YY^*X)^*YY^*X,$$

are numbers between 0 and 1. Now let S be a unitary matrix diagonalizing X^*YY^*X,

$$SX^*YY^*XS^*=\operatorname{diag}(\lambda_1,\cdots,\lambda_q).$$

Applying the above argument with X replaced by XS^* we find that the eigenvalues $\lambda_1,\cdots,\lambda_q$ considered as diagonal elements in

$$(YY^*XS^*)^*YY^*XS^*$$

lie in the interval $0\le\lambda\le1$.

Definition 3.9. The critical angles between the matrices X and Y are defined by

$$\cos^2\theta_i=\lambda_i,\qquad 0\le\theta_i\le\frac{\pi}{2}$$

where λ_i denote the eigenvalues of X^*YY^*X.

Here we find it convenient to introduce the following notation. A matrix Y is said to belong to $Sp(n,m)$ if Y is a $2n\times 2m$ matrix, $Y^*Y=I_{2m}$ and $J_nY=\bar{Y}J_m$. We also say that Y spans a vector space or Y belongs to a vector space if its columns do. Finally we write

$$f(\theta)=\begin{bmatrix} f(\theta_1) & & & & & \\ & \ddots & & & & \\ & & f(\theta_l) & & & \\ & & & f(\theta_1) & & \\ & & & & \ddots & \\ & & & & & f(\theta_l) \end{bmatrix}$$

Now let $Y\in Sp(n,m)$ be a fixed matrix spanning a vector space M and let $X\in Sp(n,l)$ span L. Choose $S\in Sp(l)$ according to Lemma 3.7 and 3.8 such that

$$X^*YY^*X=S^*\cos^2\theta S.$$

We also assume that

$$0<\theta_1<\cdots<\theta_l<\frac{\pi}{2}.$$

Under this assumption the matrix Y^*XS^* $(\cos\theta)^{-1}$ belongs to $Sp(m,l)$ i.e. Y^*X can be expressed as a product

$$Y^*X=T\cos\theta S,\quad T\in Sp(m,l),\quad S\in Sp(l).$$

Define α and $\beta \in Sp(n,l)$ by

$$\alpha = YT$$

and

$$\beta = (XS^* - YT\cos\theta)(\sin\theta)^{-1}.$$

Then $\alpha \in M$, $\beta \in M^\perp$ and

$$XS^* = \alpha\cos\theta + \beta\sin\theta$$

is an orthogonal decomposition of XS^*. Extend

$$\alpha = (\alpha_1, \cdots, \alpha_l, -J\bar{\alpha}_1, \cdots, -J\bar{\alpha}_l)$$

to a basis for M, denoted by $\tilde{\alpha}$, such that

$$\tilde{\alpha} = (\alpha_1, \cdots, \alpha_m, -J\bar{\alpha}_1, \cdots, -J\bar{\alpha}_m)$$

belongs to $Sp(n,m)$. Similarly, extend β to a basis for $M^\perp$ such that $\tilde{\beta} \in Sp(n,n-m)$ and consider the matrices A, B and C defined by

$$A = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & I_{n-2l} & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix},$$

$$B = (\beta_1, \cdots, \beta_{n-m}, \alpha_{l+1}, \cdots, \alpha_m, \alpha_1, \cdots, \alpha_l)$$

and

$$C = (B \quad -J\bar{B})\begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix}.$$

LEMMA 3.10. *Given $X \in Sp(n,l)$ and $Y \in Sp(n,m)$, let C be determined from X and Y as above. Then there is a $D \in Sp(n-l) \times Sp(l)$ such that CD is any prescribed matrix in $Sp(n)$ containing X as the columns with numbers $n-l+1, \cdots, n$ and $2n-l+1, \cdots, 2n$.*

Proof. Being a product of two elements in $Sp(n)$, C also belongs to $Sp(n)$. Denote its columns by $c_1, \cdots, c_{2n}$ and put

$$C' = (c_1, \cdots, c_{n-l}, c_{n+1}, \cdots, c_{2n-l})$$

and

$$C'' = (c_{n-l+1}, \cdots, c_n, c_{2n-l+1}, \cdots, c_{2n}).$$

The mapping

$$P\colon C \to (C', C'')$$

maps $Sp(n)$ into $Sp(n, m-l) \times Sp(n,l)$ in such a way that if $D = (D_1, D_2) \in Sp(n-l) \times Sp(l)$ then

$$P(CD) = (C'D_1, C''D_2).$$

Thus, to prove the lemma we have to choose D_1 and D_2 such that $C''D_2 = X$ and $C'D_1$ is any prescribed matrix in $Sp(n, n-l)$, orthogonal to X. This is clearly possible since $C'' = XS^*$ spans L and C' spans $L^\perp$.

COROLLARY 3.11. *For all $u \in Sp(n)$ there is a decomposition*

$$u = k_m \exp h\, k_l$$

where $k_m \in Sp(n-m) \times Sp(m)$, $k_l \in Sp(n-l) \times Sp(l)$ and $h \in i\mathfrak{h}_{\mathfrak{p}_l}$.

Proof. Given

$$u = (u_1, \cdots, u_{2n})$$

put

$$X = (u_{n-l+1}, \cdots, u_n, u_{2n-l+1}, \cdots, u_{2n}), \qquad Y = \begin{pmatrix} R & 0 \\ 0 & R \end{pmatrix}$$

where the order of R is $n \times m$ and

$$R = \begin{pmatrix} 0 \\ I_m \end{pmatrix}.$$

The desired decomposition follows from Lemma 3.10 with

$$k_m = (B, -J\bar{B}), \qquad k_l = D$$

and

$$\exp h = \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix}.$$

The particular choice of Y implies that $k_m \in Sp(n-m) \times Sp(m)$.

The derived decomposition shows that an intertwining function as well as any function on $K_m \backslash U / K_l$ is completely determined by its values on $\exp i\mathfrak{h}_{\mathfrak{p}_l}$. Furthermore, this decomposition makes it possible to compute the invariant measure on $K_m \backslash U / K_l$ in terms of the critical angles.

THEOREM 3.12. *The intertwining functions are orthogonal with respect to the measure*

$$\prod_{i=1}^{l} (\cos\theta_i)^{4(m-l)+3} (\sin\theta_i)^{4(n-l-m)+3} \prod_{1 \le i < j \le l} (\cos^2\theta_i - \cos^2\theta_j)^4.$$

Proof. All details can be found in James [4], where the case $SO(n)$ is treated. Since the proof can be carried over to our case without problems, we only give a brief sketch.

Decompose $u \in Sp(n)$ according to Corollary 3.11 and keep the notation from Lemma 3.10. The invariant measure on U/K_l is

$$\prod c_i^* dc_j$$

where c_i and c_j ranges over the columns of C' and C'' respectively. Denote the first $n-l$ columns of C' by $b_1, \cdots, b_{n-l}$ and the first l columns of C'' by $a_1, \cdots, a_l$. Then the remaining columns of C' and C'' are $-J\bar{b}_1, \cdots, -J\bar{b}_{n-l}$ and $-J\bar{a}_1, \cdots, -J\bar{a}_l$ respectively. The measure written in terms of b_j and a_i becomes

$$\prod_{\substack{1 \le j \le n-l \\ 1 \le i \le l}} b_j^* da_i \overline{b_j^* da_i} b_j' J da_i \overline{b_j' J da_i}.$$

Here

$$a_i = \alpha_i \cos\theta_i + \beta_i \sin\theta_i,$$
$$da_i = (-\alpha_i \sin\theta_i + \beta_i \cos\theta_i)\, d\theta_i + d\alpha_i \cos\theta_i + d\beta_i \sin\theta_i,$$

and

$$\begin{array}{ll} b_j = -\alpha_j \sin\theta_j + \beta_j \cos\theta_j & \text{if } 1 \leq j \leq l, \\ b_j = \beta_j & \text{if } l+1 \leq j \leq n-m, \\ b_j = \alpha_{j-n+m+l} & \text{if } n-m+1 \leq j \leq n-l. \end{array}$$

Multiplication according to the rules for differential forms yields

$$\begin{array}{ll} b_j^* da_i b_i^* da_j = \alpha_j^* d\alpha_i \beta_j^* d\beta_i \left(\cos^2\theta_i - \cos^2\theta_j\right) & \text{if } i \neq j \leq l, \\ b_j' J da_i b_i' J da_j = \alpha_j' J d\alpha_i \beta_j' J d\beta_i \left(\cos^2\theta_i - \cos^2\theta_j\right) & \text{if } i \neq j \leq l, \\ b_i^* da_i \overline{b_i^* da_i} = 2(\beta_i^* d\beta_i - \alpha_i^* d\alpha_i) \cos\theta_i \sin\theta_i d\theta_i, & \\ b_i' J da_i = (\beta_i' J d\beta_i - \alpha_i' J d\alpha_i) \cos\theta_i \sin\theta_i, & \\ b_j^* da_i = \beta_j^* d\beta_i \sin\theta_1 & \text{if } l+1 \leq j \leq n-m, \\ b_j' J da_i = \beta_j' J d\beta_i \sin\theta_i & \text{if } l+1 \leq j \leq n-m, \\ b_j^* da_i = \alpha_{j-n+m+l}^* d\alpha_i \cos\theta_i & \text{if } n-m+1 \leq j \leq n-l, \\ b_j' J da_i = \alpha_{j-n+m+l}' J d\alpha_i \cos\theta_i & \text{if } n-m+1 \leq j \leq n-l. \end{array}$$

Here we have used the facts that

$$\tilde{\alpha}^* d\tilde{\beta} = \tilde{\alpha}' J d\tilde{\beta} = 0$$

and that $\tilde{\beta}^* d\tilde{\beta}$ and $\tilde{\alpha}^* d\tilde{\alpha}$ are skew Hermitian. Inserting into the expression for the invariant measure we obtain the desired angular part.

4. The polynomials $p_N^{\alpha,\beta,\gamma}$. From now on weights and roots will be considered as elements in Z^l. In particular,

$$\begin{array}{ll} \alpha_i = (0, \cdots, 1, -1, 0, \cdots, 0), & i = 1, \cdots, l-1, \\ \alpha_l = (0, \cdots, 1) & \end{array}$$

and

$$\mu_i = (2, \cdots, 2, 0, \cdots, 0), \qquad i = 1, \cdots, l.$$

We denote by r the square of the length of the "unit" vectors $(1, \cdots, 0) \cdots (0, \cdots, 1)$.

$$r = \langle \alpha_l, \alpha_l \rangle.$$

As a total ordering of Z^l we use lexicographic ordering with respect to $\alpha_1, \cdots, \alpha_l$. This will be denoted by $<$ which should be carefully distinguished from the partial ordering $\prec$ defined in §2 by $M \prec N$ if $M \neq N$ and

$$\sum_{i=1}^{k} m_i \leq \sum_{i=1}^{k} n_i \quad \text{for } k = 1, \cdots, l.$$

Here $M = (m_1, \cdots, m_l)$ and $N = (n_1, \cdots, n_l)$. For example $(3,3,0) < (4,1,1)$ but $(3,3,0) \prec (4,1,1)$ is false. A polynomial $p_N(x)$ in l variables $x_1, \cdots, x_l$ is said to have degree N and leading term

$$c_N x^N = c_N x_1^{n_1} \cdots x_l^{n_l}$$

if

$$p_N(x)=\sum_{M\leq N} c_M x^M.$$

If p_N is a symmetric polynomial of degree N then

$$n_1\geq n_2\geq\cdots\geq n_l\geq 0.$$

For $\alpha>-1$, $\beta>-1$ and $\gamma\geq -1/2$ put

$$w(x)=w^{\alpha,\beta,\gamma}(x)=\prod_{i=1}^{l}(1-x_i)^{\alpha}(1+x_i)^{\beta}\prod_{i<j}(x_i-x_j)^{2\gamma+1}, \qquad x\in\Omega$$

where Ω is the region

$$\Omega:\ -1\leq x_l\leq x_{l-1}\leq\cdots\leq x_1\leq 1.$$

Note that the integral of w over Ω is essentially Selberg's integral.

DEFINITION 4.1. The polynomials $p_N=p_N^{\alpha,\beta,\gamma}$ are defined by

1) $p_0=1$,

2) p_N is a symmetric polynomial with leading term x^N,

3) $\int_\Omega p_N^{\alpha,\beta,\gamma}(x)q(x)w^{\alpha,\beta,\gamma}(x)\,dx=0$ if q is a symmetric polynomial and degree $q<N$.

THEOREM 4.2. *For the values of α, β and γ given in Table* 1 *a suitable normalization of the polynomials $p_N^{\alpha,\beta,\gamma}$ can be interpreted as intertwining functions* $(a-c)$, *or as elementary spherical functions* $(a-h)$. *Moreover except for the cases a and h all such functions are obtained as polynomials of this type.*

TABLE 1

space	α	β	γ
a) $K_m\backslash U/K_l$, BD I	$(n-m-l-1)/2$	$(m-l-1)/2$	0
b) $K_m\backslash U/K_l$, A III	$n-m-l$	$m-l$	$1/2$
c) $K_m\backslash U/K_l$, C II	$2(n-m-l)+1$	$2(m-l)+1$	$3/2$
d) $Sp(l)/U(l)$	0	0	0
e) $SO(4l)/U(2l)$	0	0	$3/2$
f) $SO(4l+2)/U(2l+1)$	2	0	$3/2$
g) $Sp(l)$	$1/2$	$1/2$	$1/2$
h) $SO(2l+1)$	$1/2$	$-1/2$	$1/2$

Proof. We have seen in Theorem 3.6 that the intertwining functions are W-invariant trigonometric polynomials

$$\psi_\Lambda^{l,m}=\sum_{\lambda\preccurlyeq\Lambda} c(\lambda,\Lambda)e^{-\lambda(h)},$$

where

$$\lambda=\sum_{k=1}^{l} n_k\mu_k, \qquad n_k \text{ integers.}$$

By induction over Λ it is easy to see that $\psi_\Lambda^{l,m}$ can be written as an algebraic polynomial of degree $\Lambda/2$ in the variables

$$x_i=\cos 2\theta_i, \qquad i=1,\cdots,l.$$

In each particular case, orthogonality relations hold for the trigonometric polynomials with respect to the weight function

$$w'=\prod_{i=1}^{l}(\sin\theta_i)^{2\alpha+1}(\cos\theta_i)^{2\beta+1}\prod_{i<j}(\cos 2\theta_i-\cos 2\theta_j)^{2\gamma+1}$$

defined in the region

$$\Omega':\ 0\leq\theta_1\leq\theta_2\leq\cdots\leq\theta_l\leq\frac{\pi}{2}.$$

Transformation of the variables shows that the algebraic polynomials fulfill Definition 4.1 except for the normalization condition. Hence they coincide with $p_N^{\alpha,\beta,\gamma}$.

Consider now the differential operator corresponding to the Laplace–Beltrami operator on a symmetric space.

$$D=D^{\alpha,\beta,\gamma}=4r\sum_{i=1}^{l}(1-x_i^2)\frac{\partial^2}{\partial x_i^2}+\left[w^{\alpha,\beta,\gamma}(x)\right]^{-1}\frac{\partial}{\partial x_i}\left[w^{\alpha,\beta,\gamma}(x)(1-x_i^2)\right]\frac{\partial}{\partial x_i}.$$

THEOREM 4.3. *Let f and g be polynomials. Then there holds*

$$\int_\Omega (Df)gw\,dx=\int_\Omega f(Dg)w\,dx.$$

Proof. If we make the change of variables

$$x_i=\cos\theta_i,\qquad i=1,\cdots,l$$

the left-hand side becomes

$$\int_{\Omega'}(D'f)gw'\,d\theta,$$

where

$$D'=r\sum_{i=1}^{l}\frac{1}{w'}\frac{\partial}{\partial\theta_i}w'\frac{\partial}{\partial\theta_i}.$$

An application of the Gauss formula yields

$$\int_{\Omega'}(D'f)gw'\,d\theta-\int_{\Omega'}f(D'g)w'\,d\theta$$

$$=\sum_{i=1}^{l}\int_{\partial\Omega'}\left(\frac{\partial f}{\partial\theta_i}g-\frac{\partial g}{\partial\theta_i}f\right)w'\,d\theta_1\cdots d\theta_{i-1}d\theta_{i+1}\cdots d\theta_l,$$

which is equal to zero if $\alpha,\beta,\gamma>-1/2$ since w' vanishes on $\partial\Omega'$. For other values of α,β and γ the theorem holds by analytic continuation.

COROLLARY 4.4. *$p_N^{\alpha,\beta,\gamma}$ is an eigenfunction of $D^{\alpha,\beta,\gamma}$ with eigenvalue $-|2N+\rho|^2+|\rho|^2$, where*

$$\rho=(\alpha+\beta+1)(1,\cdots,1)+(2\gamma+1)(l-1,\cdots,1,0).$$

Proof. Put

$$\pi=\prod_{i<j}(x_i-x_j)$$

and write the operator D in the form

$$D=4r\sum_{i=1}^{l}\left(1-x_i^2\right)\frac{\partial^2}{\partial x_i^2}+[(1+\beta)(1-x_i)-(1+\alpha)(1+x_i)]\frac{\partial}{\partial x_i}$$
$$+(2\gamma+1)\left(1-x_i^2\right)\frac{1}{\pi}\frac{\partial\pi}{\partial x_i}\frac{\partial}{\partial x_i}.$$

It is clear that Dp_N is a symmetric polynomial and that

$$Dp_N=\left(4r\sum_{i=1}^{l}-n_i(n_i-1)-(1+\beta)n_i-(1+\alpha)n_i-(2\gamma+1)n_i(l-i)\right)x^N$$
$$+\text{lower terms.}$$

Moreover if $M<N$

$$\int_\Omega (Dp_N)p_M w\,dx=\int_\Omega p_N(Dp_M)w\,dx=0.$$

These conditions determine Dp_N uniquely according to Definition 4.1.

Let $p_n^{\alpha,\beta}(x)$ be the ordinary Jacobi polynomial of degree n, normalized such that the leading term is x^n.

THEOREM 4.5. *For $\gamma=-1/2$ there holds*

$$p_N^{\alpha,\beta,-1/2}(x)=\sum p_{i_1}^{\alpha,\beta}(x)\cdots p_{i_l}^{\alpha,\beta}(x),$$

where the sum ranges over the different permutations of $n_1,\cdots,n_l$.

Proof. The right-hand side is a symmetric polynomial with leading term x^N. All other terms are of the form

$$x^M=x_1^{m_1}\cdots x_l^{m_l},$$

with $m_j\le i_j$. Then

$$\sum_{j=1}^{k} m_j\le\sum_{j=1}^{k} i_j\le\sum_{j=1}^{k} n_j,$$

where the last inequality is a consequence of the fact that $n_1\ge n_2\ge\cdots\ge n_l\ge 0$. This implies that $M\prec N$, hence also $M<N$. Orthogonality relations follow from the corresponding ones for Jacobi polynomials by noting that for symmetric functions f there holds

$$\int_\Omega fw\,dx=C\int_{|x_i|\le 1} f|w|\,dx$$

THEOREM 4.6. *For $\gamma=1/2$ there holds*

$$p_N^{\alpha,\beta,1/2}(x)=\frac{A^{\alpha,\beta}(n_1+l-1,n_2+l-2,\cdots,n_l)}{A^{\alpha,\beta}(l-1,l-2,\cdots,0)}$$

where

$$A(M)=A^{\alpha,\beta}(m_1,\cdots,m_l)=\begin{vmatrix} p_{m_1}^{\alpha,\beta}(x_1) & \cdots & p_{m_1}^{\alpha,\beta}(x_l) \\ \vdots & & \\ p_{m_l}^{\alpha,\beta}(x_1) & \cdots & p_{m_l}^{\alpha,\beta}(x_l) \end{vmatrix}.$$

Proof. If $m_1>m_2>\cdots>m_l>0$, then $A(M)$ is an antisymmetric polynomial of degree M. Its leading term is x^M. In particular

$$A(l-1,\cdots,0)=\pi=\prod_{i<j}(x_i-x_j).$$

It follows that the quotient is a symmetric polynomial with leading term x^N. In view of the equality

$$w^{\alpha,\beta,1/2}=\pi^2 w^{\alpha,\beta,-1/2}$$

the orthogonality relations are obtained as in Theorem 4.5.

As can be seen from the last two theorems a monomial $c_M x^M$ appears in $p_N^{\alpha,\beta,\pm 1/2}$ with nonzero coefficient c_M only if $M\preccurlyeq N$. This is also true for $\gamma=0$ and $\gamma=3/2$. For the proof we first need a lemma.

LEMMA 4.7. *If* $2\gamma+1$ *is an integer then the coefficients* c_M *in the expansion*

$$p_N^{\alpha,\beta,\gamma}(x)=\sum_{M\leq N} c_M x^M$$

are rational functions of α *and* β.

Proof. We use induction over N. Suppose that the coefficients of p_M are rational if $M<N$ and write

$$p_N(x)=Sx^N-\sum_{M<N}\frac{(Sx^N,p_M)_\gamma}{(p_M,p_M)_\gamma}p_M(x),$$

where

$$(f,g)_\gamma=\int_\Omega f(x)g(x)w^{\alpha,\beta,\gamma}(x)\,dx$$

and Sx^N denotes the symmetrization of x^N

$$l!\cdot Sx^N=\sum_{T\in W} x^{TN}$$

By the induction hypothesis it is sufficient to prove the rationality for monomials

$$\frac{(x^K,1)_\gamma}{(x^J,1)_\gamma}.$$

Since $2\gamma+1$ is integral it is no restriction to assume that $\gamma=-1/2$. Now

$$\frac{(x^K,1)_{-1/2}}{(1,1)_{-1/2}}$$

is easily computed in terms of Γ-functions and it is found to be rational.

COROLLARY 4.8. *If $\gamma=-1/2, 0, 1/2$ or $3/2$ then c_M can be nonzero only if $M\preccurlyeq N$, i.e.*

$$p_N^{\alpha,\beta,\gamma}(x)=\sum_{M\preccurlyeq N} c_M x^M.$$

Proof. For $\gamma=\pm 1/2$ this is clear from the explicit expressions of p_N given in Theorems 4.5 and 4.6. Moreover according to the group-theoretical interpretation in Theorem 4.2 the corollary also holds true for the values of α,β and γ listed there. Let γ be $3/2$ or 0, choose β from the list and consider a coefficient c_M which do not satisfy $M\preccurlyeq N$. There are infinitely many values of α in the list for which $c_M=0$. Being a rational function c_M has to vanish identically. Now keep α fixed. Then c_M vanishes for infinitely many values of β. Hence $c_M=0$ for all α and β.

DEFINITION 4.9. Put $N=(n_1,\cdots,n_l)$, $q_i=n_i+(\gamma+1/2)(l-i)$ for $i=1,\cdots,l$ and let Γ denote the gamma-function. We then define $c(N)=c(N,l,\alpha,\beta,\gamma)$ by

$$c(N)=\frac{\tilde{c}(N)}{\tilde{c}(0)},$$

where

$$\tilde{c}(N)=\prod_{1\le i<j\le l}\frac{\Gamma(q_i+q_j+\alpha+\beta+1)\Gamma(q_i-q_j)}{\Gamma(q_i+q_j+\alpha+\beta+\gamma+3/2)\Gamma(q_i-q_j+\gamma+1/2)}$$

$$\cdot\prod_{1\le i\le l}\frac{2^{-2q_i}\Gamma(2q_i+\alpha+\beta+1)}{\Gamma(q_i+\alpha+\beta+1)\Gamma(q_i+\alpha+1)}.$$

For the values of α,β and γ corresponding to elementary spherical functions $c(N)$ is Harish-Chandra's c-function evaluated at $-i(2N+\rho)$ where

$$\rho=(2\gamma+1)(l-1,l-2,\cdots,0)+(\alpha+\beta+1)(1,\cdots,1).$$

In this case the usual normalization is

$$\varphi_N^{\alpha,\beta,\gamma}(x)=c(N,l,\alpha,\beta,\gamma)2^{n_1+\cdots+n_l}p_N^{\alpha,\beta,\gamma}(x) \tag{1}$$

and

$$\left\|\varphi_N^{\alpha,\beta,\gamma}\right\|^2=\frac{\left(\varphi_N^{\alpha,\beta,\gamma},\varphi_N^{\alpha,\beta,\gamma}\right)_\gamma}{(1,1)_\gamma}=\frac{1}{d(N,l,\alpha,\beta,\gamma)} \tag{2}$$

where

$$d(N)=\frac{c(0)c(-\rho)}{c(N)c(-N-\rho)}$$

is the dimension of the representation with highest weight $2N$. The connection between the c-function and the leading term of φ_N is proved in [2, p. 291] while the formula for the dimension can be deduced recursively from a recurrence formula for φ_N. See [11, Lemma 4.6].

THEOREM 4.10. *If $\gamma=-1/2, 0, 1/2$ or $3/2$ the normalization* (1) *and* (2) *holds for all $\alpha,\beta>-1$.*

Proof. The case $\gamma=-1/2$ can be checked directly in view of Theorem 4.5. In the remaining cases it is sufficient to prove the theorem for intertwining functions. Full generality is then obtained by analytic continuation with respect to α and β as in the proof of Corollary 4.8. Thus we consider three triples (α,β,γ), (α',β,γ) and $(\alpha'',\beta'',\gamma)$ corresponding to the spaces $K_l\backslash U/K_l$, $K_m\backslash U/K_m$ and $K_m\backslash U/K_l$ respectively. To treat the three different values of γ at the same time note that the parameters can be written

$$\begin{aligned}
&\alpha=(\gamma+1/2)(n-2l)+\gamma-1/2,\\
&\beta=(\gamma-1/2),\\
&\alpha'=(\gamma+1/2)(n-2m)+(\gamma-1/2),\\
&\alpha''=(\gamma+1/2)(n-l-m)+(\gamma-1/2),\\
&\beta''=(\gamma+1/2)(m-l)+(\gamma-1/2).
\end{aligned}$$

Let us first define

$$\psi_N^{\alpha'',\beta'',\gamma}=\left[c(N,l,\alpha,\beta,\gamma)c(N',m,\alpha',\beta,\gamma)\right]^{1/2}2^{n_1+\cdots+n_l}p_N^{\alpha'',\beta'',\gamma},$$

where

$$N'=(n_1,\cdots,n_l,0,\cdots,0).$$

From the discussion preceding Theorem 3.6 we know that this is the correct normalization of the intertwining function $\psi_N^{\alpha'',\beta'',\gamma}$. Its quadratic norm is the same as for the elementary spherical function $\varphi_N^{\alpha,\beta,\gamma}$.

$$\left\|\psi_N^{\alpha'',\beta'',\gamma}\right\|^2=\frac{1}{d(N,l,\alpha,\beta,\gamma)}.$$

Next we put

$$\varphi_N^{\alpha'',\beta'',\gamma}=\left(\frac{d(N,l,\alpha,\beta,\gamma)}{d(N,l,\alpha'',\beta'',\gamma)}\right)^{1/2}\psi_N^{\alpha'',\beta'',\gamma}.$$

Then condition (2) is fulfilled and for the verification of (1) we have to show that

$$\frac{d(N,l,\alpha,\beta,\gamma)}{d(N,l,\alpha'',\beta'',\gamma)}c(N,l,\alpha,\beta,\gamma)\cdot c(N',m,\alpha',\beta,\gamma)=(c(N,l,\alpha'',\beta'',\gamma))^2$$

This can be done by lengthy but elementary calculations. Except for the normalization $d(0)=c(0)=1$ we have

$$\frac{d(N,l,\alpha,\beta,\gamma)}{d(N,l,\alpha'',\beta'',\gamma)}=\prod_{i=1}^{l}\frac{f_i(n-l)f_i(m)}{f_i(n-m)f_i(l)},$$

$$\frac{c(N',m,\alpha',\beta,\gamma)}{c(N,l,\alpha,\beta,\gamma)}=\prod_{i=1}^{l}\frac{f_i(l)f_i(n-l)}{f_i(m)f_i(n-m)},$$

$$\frac{c(N,l,\alpha'',\beta'',\gamma)}{c(N,l,\alpha,\beta,\gamma)}=\prod_{i=1}^{l}\frac{f_i(n-l)}{f_i(n-m)},$$

where

$$f_i(k)=\Gamma(n_i+(\gamma+1/2)(k-i+1)).$$

5. Formulas. Let α,β and γ take such values that the normalization φ is defined (cf. Theorem 4.10).

THEOREM 5.1. *In the following formulas changing one of the parameter values, the number of terms is independent of N.*

$$\varphi_N^{\alpha,\beta,\gamma}=\sum_{0\preccurlyeq\varepsilon\preccurlyeq\mu_l/2} a_1(N,\varepsilon)\varphi_{N-\varepsilon}^{\alpha+1,\beta,\gamma}, \tag{1}$$

$$\prod_{1\leq i\leq l}(x_i-1)\varphi_N^{\alpha+1,\beta,\gamma}=\sum_{0\preccurlyeq\varepsilon\preccurlyeq\mu_l/2} b_1(N,\varepsilon)\varphi_{N+\varepsilon}^{\alpha,\beta,\gamma}, \tag{2}$$

$$\varphi_N^{\alpha,\beta,\gamma}=\sum_{0\preccurlyeq\varepsilon\preccurlyeq\mu_l/2} a_2(N,\varepsilon)\varphi_{N-\varepsilon}^{\alpha,\beta+1,\gamma}, \tag{3}$$

$$\prod_{1\leq i\leq l}(x_i+1)\varphi_N^{\alpha,\beta+1,\gamma}=\sum_{0\preccurlyeq\varepsilon\preccurlyeq\mu_l/2} b_2(N,\varepsilon)\varphi_{N+\varepsilon}^{\alpha,\beta,\gamma}, \tag{4}$$

$$\varphi_N^{\alpha,\beta,\gamma}=\sum_{0\preccurlyeq\varepsilon\preccurlyeq 2\delta} a_3(N,\varepsilon)\varphi_{N-\varepsilon}^{\alpha,\beta,\gamma+1}, \tag{5}$$

$$\prod_{i<j}(x_i-x_j)^2\varphi_N^{\alpha,\beta,\gamma+1}=\sum_{0\preccurlyeq\varepsilon\preccurlyeq 2\delta} b_3(N,\varepsilon)\varphi_{N+\varepsilon}^{\alpha,\beta,\gamma}, \tag{6}$$

$$\delta=\frac{1}{2}\sum_{1\leq i\leq l}\mu_i.$$

Proof. To prove (1) and (2) consider the expansions

$$\varphi_N^{\alpha,\beta,\gamma}=\sum_{M\preccurlyeq N} A(N,M)\varphi_M^{\alpha+1,\beta,\gamma}$$

and

$$w^{1,0,-1/2}\varphi_N^{\alpha+1,\beta,\gamma}=\sum_{M\preccurlyeq N+\mu_l/2} B(N,M)\varphi_M^{\alpha,\beta,\gamma},$$

where

$$w^{1,0,-1/2}(x)=\prod_{1\leq i\leq l}(1-x_i).$$

Note that

$$w^{1,0,-1/2}(x)w^{\alpha,\beta,\gamma}(x)=w^{\alpha+1,\beta,\gamma}(x).$$

In view of the orthogonality relations we have on one hand

$$\int\varphi_N^{\alpha,\beta,\gamma}(x)\varphi_M^{\alpha+1,\beta,\gamma}(x)w^{\alpha+1,\beta,\gamma}(x)\,dx=A(N,M)\int\left|\varphi_M^{\alpha+1,\beta,\gamma}(x)\right|^2 w^{\alpha+1,\beta,\gamma}(x)\,dx.$$

On the other the left-hand side is equal to

$$\int\varphi_N^{\alpha,\beta,\gamma}(x)\varphi_M^{\alpha+1,\beta,\gamma}(x)w^{1,0,-1/2}(x)w^{\alpha,\beta,\gamma}(x)\,dx=B(M,N)\int\left|\varphi_N^{\alpha,\beta,\gamma}(x)\right|^2 w^{\alpha,\beta,\gamma}(x)\,dx.$$

We conclude that $A(N,M)$ is nonzero if and only if $B(M,N)$ is nonzero, and this is possible only if

$$M\preccurlyeq N \quad\text{and}\quad N\preccurlyeq M+\mu_l/2$$

Writing $M=N-\varepsilon$, this means that $0\leqslant\varepsilon\leqslant\mu_l/2$, and the proof of (1) and (2) is complete. Formulas (3)–(6) can be proved in the same way.

We now turn our attention to the group theoretical interpretation of $\varphi_N^{\alpha,\beta,\gamma}$ given in Theorem 4.2, trying to obtain explicit formulas. As we have seen, only the cases $\gamma=0$ and $\gamma=3/2$ offer difficulties. Let us consider the case $\gamma=3/2$. Formula (6) of Theorem 5.1 expresses $\varphi_N^{\alpha,\beta,3/2}$ in terms of $\varphi_M^{\alpha,\beta,1/2}$ for which we have an explicit expression in terms of Jacobi polynomials. Unfortunately formula (6) contains too many unknown coefficients but nevertheless some valuable information can be obtained. For $\alpha=2n-4l+1$, $\beta=1$ and $\gamma=3/2$, $\varphi_N^{\alpha,\beta,\gamma}(x)$ is an elementary spherical function which can be continued analytically with respect to N and x, to the noncompact analogue. A similar extension of $\varphi_N^{\alpha,\beta,1/2}$ involves Jacobi functions rather than Jacobi polynomials. Denote the extension by

$$\phi_\lambda^{\alpha,\beta,\gamma}=\varphi_{(i\lambda-\rho)/2}^{\alpha,\beta,\gamma}$$

where

$$\rho=\rho(\alpha,\beta,\gamma)=(2\gamma+1)(l-1,l-2,\cdots,0)+(\alpha+\beta+1)(1,\cdots,1).$$

Then

$$\varphi_N=\phi_{-i(2N+\rho)}$$

and

$$\phi_{S\lambda}=\phi_\lambda \quad \text{for all } S\in W.$$

LEMMA 5.2. *For all $\alpha>-1$, $\varphi_N^{\alpha,1,3/2}$ can be continued analytically to a W-invariant function $\phi_\lambda^{\alpha,1,3/2}$.*

Proof. By the methods in [11], including repeated application of the Laplace–Beltrami operator, it can be shown that $b_3(N,\varepsilon)$ in formula (6) is rational in N if $\alpha=2n-4l+1$. The natural extension of $b_3(N,\varepsilon)$ to complex N defines a function $B(\lambda,\varepsilon)$,

$$B(\lambda,\varepsilon)=b_3((i\lambda-\rho(\alpha,1,3/2))/2,\varepsilon).$$

This is of course also rational in α. Hence $B(\lambda,\varepsilon)$ can be defined for general values of α. Following again the methods of [11] we have for $\alpha=2n-4l+1$

$$(6') \qquad \pi^2\phi_\lambda^{\alpha,1,3/2}=\sum_{0\leqslant\varepsilon\leqslant2\delta} B(\lambda,\varepsilon)\phi_{\lambda-2i\varepsilon}^{\alpha,1,1/2}.$$

Because of the W-invariance of ϕ_λ, $B(\lambda,\varepsilon)$ must satisfy

$$B(S\lambda,\varepsilon)=B(\lambda,S^{-1}\varepsilon)$$

By passing to general values of α the sum in (6′) remains W-invariant. The proof of the lemma is finished.

We are now prepared to compute some coefficients explicitly.

THEOREM 5.3. *If*

$$\gamma=\pm1/2 \quad \textit{and} \quad \alpha,\beta>-1,$$

or

$$\gamma=0, \quad \beta=-1/2 \quad \textit{and} \quad \alpha>-1$$

or

$$\gamma=3/2, \quad \beta=1 \quad \textit{and} \quad \alpha>-1$$

the following explicit formulas hold.

$$(1') \qquad \varphi_N^{\alpha,\beta,\gamma}=\sum a\big(S^{-1}(N+\rho/2)-\rho/2\big)\varphi^{\alpha+1,\beta,\gamma}_{N-\mu_l/4+S\mu_l/4},$$

where

$$a(\lambda)=\frac{c(\lambda,l,\alpha,\beta,\gamma)}{c(\lambda,l,\alpha+1,\beta,\gamma)}$$

and

$$\rho=\rho(\alpha,\beta,\gamma);$$

$$(2') \qquad \prod_{1\le i\le l}(x_i-1)\varphi_N^{\alpha+1,\beta,\gamma}=\sum b\big(S^{-1}(N+\rho/2)-\rho/2\big)\varphi^{\alpha,\beta,\gamma}_{N+\mu_l/4+S\mu_l/4},$$

where

$$b(\lambda)=2^{-l}\frac{c(\lambda,l,\alpha+1,\beta,\gamma)}{c(\lambda+\mu_l/2,l,\alpha,\beta,\gamma)}$$

and

$$\rho=\rho(\alpha+1,\beta,\gamma).$$

In both formulas, the sum ranges over all different S_{μ_l}, $S\in W$.

Proof. Let α,β and γ be as in the assumption. If $\gamma=0$ it suffices to prove the formulas for $\alpha=(n-2l+1)/2$. Full generality is then obtained by analytic continuation with respect to α. Thus for these values of the parameters formula (1) of Theorem 5.1 extends to

$$\phi_\lambda^{\alpha,\beta,\gamma}=\sum_{0\preccurlyeq\varepsilon\preccurlyeq\mu_l/2}A(\lambda,\varepsilon)\phi^{\alpha+1,\beta,\gamma}_{\lambda+2i\varepsilon-i\mu_l/2},$$

where

$$A(\lambda,\varepsilon)=a_1\big((i\lambda-\rho(\alpha,\beta,\gamma))/2,\varepsilon\big)$$

(cf. Lemma 5.2).

The W-invariance of ϕ_λ implies that

$$A(\lambda,\varepsilon)=A(S\lambda,\varepsilon(S)), \qquad S\in W,$$

where we have put

$$\varepsilon(S)=\mu_l/4+S(\varepsilon-\mu_l/4).$$

It follows that $A(\lambda,\varepsilon)$ is nonzero only if

$$0\preccurlyeq\varepsilon(S)\preccurlyeq\mu_l/2 \quad \text{for all } S\in W$$

and

$$\varepsilon(S)=\frac{1}{2}\sum_{1\le i\le l}k_i\mu_i, \qquad k_i \text{ integers.}$$

Select S_0 corresponding to the highest $\varepsilon(S)$. Then

$$0 \leqslant S_0(\varepsilon - \mu_l/4) \leqslant \mu_l/4$$

or

$$\mu_l/2 \leqslant 2\varepsilon(S_0) \leqslant \mu_l.$$

Since $2\varepsilon(S_0)$ is an integral linear combination of $\mu_1, \cdots, \mu_l$, we conclude that $2\varepsilon(S_0) = \mu_l$. Thus $A(\lambda, \varepsilon)$ is nonzero only if there is an element S in the Weyl group such that

$$\varepsilon = (\mu_l - S\mu_l)/4;$$

moreover

$$A(\lambda, \varepsilon) = A(S^{-1}\lambda, \varepsilon(S^{-1})) = A(S^{-1}\lambda, 0).$$

Putting

$$\lambda = -i(2N + \rho(\alpha, \beta, \gamma))$$

in the formula again we get

$$\varphi_N^{\alpha,\beta,\gamma} = \sum a_1(S^{-1}(N+\rho/2) - \rho/2, 0)\varphi_{N-\mu_l/4+S\mu_l/4}^{\alpha+1,\beta,\gamma}.$$

Here we recognize formula (1′) if we write $a(\lambda)$ instead of $a_1(\lambda, 0)$. Finally $a(\lambda)$ is determined by comparison of the leading terms.

The second formula, (2′), is obtained in an analogous way.

Remark 5.4. Formulas changing β can be obtained by use of the relation

$$p_N^{\alpha,\beta,\gamma}(-x) = (-1)^{n_1 + \cdots + n_l} p_N^{\beta,\alpha,\gamma}(x).$$

If we apply the same idea to the formulas changing γ, (5) and (6), we obtain

$$\varphi_N^{\alpha,\beta,\gamma} = \sum_{\tau,S} a(S^{-1}(N+\rho/2) - \rho/2, \tau)\varphi_{N-\delta+S\tau}^{\alpha,\beta,\gamma+1}$$

and

$$\pi^2 \varphi_N^{\alpha,\beta,\gamma+1} = \sum_{\tau,S} b(S^{-1}(N+\rho^+/2) - \rho^+/2, \tau)\varphi_{N+\delta+S\tau}^{\alpha,\beta,\gamma}$$

where τ ranges over $0 \leqslant \tau \leqslant \delta$ and S over all different $S\tau$. In these formulas we have put

$$\rho = \rho(\alpha, \beta, \gamma)$$

and

$$\rho^+ = \rho(\alpha, \beta, \gamma + 1).$$

Of course the proofs are valid only if both sides of the formulas have a W-invariant analytic continuation. The coefficients $a(\lambda, \tau)$ and $b(\lambda, \tau)$ are unknown except for $\tau = \delta$.

6. The case $l = 2$. The two-variable polynomials $p_{n,k}^{\alpha,\beta,\gamma}(x,y)$ were introduced by Koornwinder in [7]. The further analysis developed in [10] and [9] depends on the existence of certain raising and lowering differential operators.

$$D_-^\gamma p_{n,k}^{\alpha,\beta,\gamma} = k(n+\gamma+1/2)p_{n-1,k-1}^{\alpha+1,\beta+1,\gamma},$$
$$D_+^{\alpha,\beta,\gamma} p_{n-1,k-1}^{\alpha+1,\beta+1,\gamma} = (k+\alpha+\beta+1)(n+\alpha+\beta+\gamma+3/2)p_{n,k}^{\alpha,\beta,\gamma},$$

$$E_{-}^{\alpha,\beta} p_{n,k}^{\alpha,\beta,\gamma} = (n-k)(n+k+\alpha+\beta+1) p_{n-1,k}^{\alpha,\beta,\gamma+1},$$
$$E_{+}^{\alpha,\beta,\gamma} p_{n-1,k}^{\alpha,\beta,\gamma+1} = (n-k+2\gamma+1)(n+k+\alpha+\beta+2\gamma+2) p_{n,k}^{\alpha,\beta,\gamma}.$$

These operators were used to calculate the quadratic norm and the value of $p_{n,k}^{\alpha,\beta,\gamma}(1,1)$ and also to prove that $p_{n,k}^{\alpha,\beta,\gamma}(x,y)$ contains no other monomials $x^i y^j$ than those for which (cf. Corollary 4.8)

$$(i,j) \preccurlyeq (n,k),$$

i.e.,

$$i \leq n \quad \text{and} \quad i+j \leq n+k.$$

It turns out that the polynomials can be normalized according to (1) and (2) of Theorem 4.10 for all values of α, β, and γ. Moreover

$$\varphi_{n,k}^{\alpha,\beta,\gamma}(1,1) = 1.$$

THEOREM 6.1. *Formulas* (1′)–(2′) *in Theorem* 5.3 *are valid for* $\alpha > -1$, $\beta > -1$ *and* $\gamma \geq -1/2$. *For example, formula* (2′) *reads*

$$(x-1)(y-1)\varphi_{n,k}^{\alpha+1,\beta,\gamma} = b_{11}\varphi_{n+1,k+1}^{\alpha,\beta,\gamma} + b_{01}\varphi_{n,k+1}^{\alpha,\beta,\gamma} + b_{10}\varphi_{n+1,k}^{\alpha,\beta,\gamma} + b_{00}\varphi_{n,k}^{\alpha,\beta,\gamma},$$

where

$$b_{11} = \frac{4(\alpha+1)(\alpha+\gamma+3/2)(n+k+\alpha+\beta+2\gamma+3)}{(2k+\alpha+\beta+2)(n+k+\alpha+\beta+\gamma+5/2)(2n+\alpha+\beta+2\gamma+3)},$$

$$b_{01} = -\frac{4(\alpha+1)(\alpha+\gamma+3/2)(n-k)}{(2k+\alpha+\beta+2)(n-k+\gamma+1/2)(2n+\alpha+\beta+2\gamma+3)},$$

$$b_{10} = -\frac{4(\alpha+1)(\alpha+\gamma+3/2)(n-k+2\gamma+1)}{(2k+\alpha+\beta+2)(n-k+\gamma+1/2)(2n+\alpha+\beta+2\gamma+3)},$$

$$b_{00} = \frac{4(\alpha+1)(\alpha+\gamma+3/2)(n+k+\alpha+\beta+2)}{(2k+\alpha+\beta+2)(n+k+\alpha+\beta+\gamma+5/2)(2n+\alpha+\beta+2\gamma+3)}.$$

Proof. The coefficients in formula (1) can be obtained by repeated application of the operator D_{-}^{γ}. From these the remaining coefficients can be determined.

THEOREM 6.2. *In the formula changing* γ,

$$(x-y)^2\varphi_{n,k}^{\alpha,\beta,\gamma+1} = b_{20}\varphi_{n+2,k}^{\alpha,\beta,\gamma} + b_{10}\varphi_{n+1,k}^{\alpha,\beta,\gamma} + b_{00}\varphi_{n,k}^{\alpha,\beta,\gamma}$$
$$+ b_{11}\varphi_{n+1,k+1}^{\alpha,\beta,\gamma} + b_{1-1}\varphi_{n+1,k-1}^{\alpha,\beta,\gamma},$$

the coefficients are

$$b_{20} = \frac{16(\gamma+1)(\alpha+\gamma+3/2)(n+\alpha+\beta+\gamma+5/2)(n+\alpha+\gamma+5/2)}{(n+k+\alpha+\beta+\gamma+5/2)(n-k+\gamma+3/2)(2n+\alpha+\beta+2\gamma+5)(2n+\alpha+\beta+2\gamma+4)},$$

$$b_{10} = \frac{32(\alpha-\beta)(\alpha+\beta)(\gamma+1)(\alpha+\gamma+3/2)}{(2n+\alpha+\beta+2\gamma+5)(2n+\alpha+\beta+2\gamma+3)(2k+\alpha+\beta+2)(2k+\alpha+\beta)},$$

$$b_{00} = \frac{16(\gamma+1)(\alpha+\gamma+3/2)(n+\gamma+3/2)(n+\beta+\gamma+3/2)}{(n-k+\gamma+3/2)(n+k+\alpha+\beta+\gamma+5/2)(2n+\alpha+\beta+2\gamma+3)(2n+\alpha+\beta+2\gamma+4)},$$

$$b_{11}=-\frac{16(\gamma+1)(\alpha+\gamma+3/2)(k+\alpha+1)(k+\alpha+\beta+1)}{(n+k+\alpha+\beta+\gamma+5/2)(n-k+\gamma+3/2)(2k+\alpha+\beta+2)(2k+\alpha+\beta+1)},$$

$$b_{1-1}=-\frac{16(\gamma+1)(\alpha+\gamma+3/2)k(k+\beta)}{(n+k+\alpha+\beta+\gamma+5/2)(n-k+\gamma+3/2)(2k+\alpha+\beta)(2k+\alpha+\beta+1)}.$$

Proof. Let $\varphi_n^{\alpha,\beta}(x)$ be the Jacobi polynomial of degree n normalized such that $\varphi_n^{\alpha,\beta}(1)=1$. In view of Theorems 4.5 and 4.6 we have

$$\varphi_{n,k}^{\alpha,\beta,-1/2}(x,y)=1/2\left(\varphi_n^{\alpha,\beta}(x)\varphi_k^{\alpha,\beta}(y)+\varphi_k^{\alpha,\beta}(x)\varphi_n^{\alpha,\beta}(y)\right)$$

and

$$\varphi_{n,k}^{\alpha,\beta,1/2}(xy)=\frac{2(\alpha+1)}{(n-k+1)(n+k+\alpha+\beta+2)}\,\frac{\varphi_{n+1}^{\alpha,\beta}(x)\varphi_k^{\alpha,\beta}(y)-\varphi_k^{\alpha,\beta}(x)\varphi_{n+1}^{\alpha,\beta}(y)}{x-y}.$$

Using a formula for Jacobi polynomials

$$(x-1)\varphi_n^{\alpha,\beta}=A_n\varphi_{n+1}^{\alpha,\beta}+B_n\varphi_n^{\alpha,\beta}+C_n\varphi_{n-1}^{\alpha,\beta}$$

we obtain

$$\begin{aligned}(x-y)^2\varphi_{n,k}^{\alpha,\beta,1/2}=&\frac{4(\alpha+1)}{(n-k+1)(n+k+\alpha+\beta+2)}\\ &\cdot\left[A_{n+1}\varphi_{n+2,k}^{\alpha,\beta,-1/2}+(B_{n+1}-B_k)\varphi_{n+1,k}^{\alpha,\beta,-1/2}\right.\\ &\left.+C_{n+1}\varphi_{n,k}^{\alpha,\beta,-1/2}-A_k\varphi_{n+1,k+1}^{\alpha,\beta,-1/2}-C_k\varphi_{n+1,k-1}^{\alpha,\beta,-1/2}\right]\end{aligned}$$

If we now insert the values for A_n, B_n and C_n

$$\begin{aligned}A_n&=\frac{2(n+\alpha+\beta+1)(n+\alpha+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)},\\ B_n&=-A_n-C_n,\\ C_n&=\frac{2n(n+\beta)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta)},\end{aligned}$$

a simple calculation shows that the theorem holds true for $\gamma=-1/2$. By successive application of the operator E_+ to this formula the case $\gamma=-1/2+j$, j integer, is proved. Analytic continuation with respect to γ then yields the general case. Note that

$$E_+^{\alpha,\beta,\gamma-1}(x-y)^2=(x-y)^2E_+^{\alpha,\beta,\gamma}$$

and

$$E_+^{\alpha,\beta,\gamma}\varphi_{n,k}^{\alpha,\beta,\gamma+1}=8(\gamma+1)(\alpha+\gamma+3/2)\varphi_{n+1,k}^{\alpha,\beta,\gamma}.$$

COROLLARY 6.3. *For* $\gamma=\frac{1}{2}$, $\alpha=2n-2m-3$ *and* $\beta=2m-3$, *Theorem* 6.2 *provides an explicit expression for the intertwining functions on*

$$Sp(n-m)\times Sp(m)\backslash Sp(n)/Sp(n-2)\times Sp(2)$$

in terms of Jacobi polynomials.

7. The case $l=3$. In three variables the computations concerning raising and lowering operators are much more complicated than in two variables. However we have the following result.

THEOREM 7.1. *Let the operator D^{γ}_{-} be defined by*

$$D^{\gamma}_{-}=D_1D_2D_3+(\gamma+1/2)\left(\frac{D_3\pi}{\pi}D_2D_2+\frac{D_2\pi}{\pi}D_1D_3+\frac{D_1\pi}{\pi}D_2D_3\right)$$

$$+(\gamma+1/2)^2\left(\frac{D_2D_3\pi}{\pi}D_1+\frac{D_1D_3\pi}{\pi}D_2+\frac{D_1D_2\pi}{\pi}D_3\right),$$

where $D_i=\partial/\partial x_i$ and $\pi=(x_1-x_2)(x_1-x_3)(x_2-x_3)$. Then for $\gamma=-1/2,\ 0, 1/2$ and $3/2$

$$D^{\gamma}_{-}p^{\alpha,\beta,\gamma}_{n_1,n_2,n_3}=(n_1+2\gamma+1)(n_2+\gamma+1/2)n_3p^{\alpha+1,\beta+1,\gamma}_{n_1-1,n_2-1,n_3-1}.$$

Proof. Since the calculations are too long to write down here we just outline the idea. First let us add a constant multiple of the identity to the Laplace operator and put

$$D_2^{\alpha,\beta,\gamma}=D_1^{\alpha,\beta,\gamma}-|\rho|^2\mathrm{Id}.$$

By Corollary 4.4 the eigenvalue of $p_N^{\alpha,\beta,\gamma}$ under $D_2^{\alpha,\beta,\gamma}$ do not change when we replace N by $N-(1,1,1)$ and α,β by $\alpha+1$, $\beta+1$ simultaneously.

Next we transform $D_2^{\alpha,\beta,\gamma}$ and D^{γ}_{-} to the new variables

$$u_1=x_1+x_2+x_3,$$
$$u_2=x_1x_2+x_1x_3+x_2x_3,$$
$$u_3=x_1x_2x_3.$$

And finally by direct calculation we find that

$$D_2^{\alpha+1,\beta+1,\gamma}D^{\gamma}_{-}=D^{\gamma}_{-}D_2^{\alpha,\beta,\gamma}.$$

From all this we conclude that

(i) $D^{\gamma}_{-}p_N^{\alpha,\beta,\gamma}=\sum_{M\preccurlyeq N-(1,1,1)}c_Mp_M^{\alpha+1,\beta+1,\gamma}$,

(ii) $c_{N-(1,1,1)}=(n_1+2\gamma+1)(n_2+\gamma+1/2)n_3$,

(iii) $D^{\gamma}_{-}p_N^{\alpha,\beta,\gamma}$ is an eigenfunction of $D_2^{\alpha+1,\beta+1,\gamma}$ with the correct eigenvalue.

Then $D^{\gamma}_{-}p_N^{\alpha,\beta,\gamma}$ is completely determined by Corollary 4.8 and the following lemma.

LEMMA 7.2. *If $M\prec N$ the eigenvalues of p_M and p_N under D_2 are different.*

Proof. The inequality $M\prec N$ implies that $\langle M,x\rangle\le\langle N,x\rangle$ if x is a linear combination of $\mu_1,\cdots,\mu_l$ with nonnegative coefficients. For the eigenvalues $-|2M+\rho|^2$ and $-|2N+\rho|^2$ we then have

$$|2M+\rho|^2=\langle 2M+\rho,2M+\rho\rangle\le\langle 2N+\rho,2M+\rho\rangle\le\langle 2N+\rho,2N+\rho\rangle=|2N+\rho|^2$$

with equality only if $M=N$.

We also need the normalized version of Theorem 7.1.

COROLLARY 7.3.

$$D^{\gamma}_{-}\varphi_N^{\alpha,\beta,\gamma}=\lambda(N,\alpha,\beta,\gamma)\varphi_{N-(1,1,1)}^{\alpha+1,\beta+1,\gamma}$$

where

$$\lambda(N)=\prod_{1\le i\le 3}\left[\left(n_i+\frac{\rho_i}{2}\right)^2-\left(\frac{\alpha+\beta+1}{2}\right)^2\right]$$

Moreover

$$\lambda(S(N+\rho/2)-\rho/2)=\lambda(N)\quad \textit{for all } S\in W.$$

By applying D^{γ}_{-} to formula (1′), Theorem 5.3, we find that the coefficients are obtained from the leading one by the substitution

$$N\to S(N+\rho/2)-\rho/2$$

also for all $\beta>-1$. This leads to

THEOREM 7.4. *If $l=3$ the explicit formulas in Theorem 5.3 are valid for all α and β.*

It is now natural to ask if there is an analogue of the operator $E^{\alpha,\beta}_{-}$ too. Such an operator could be used to derive explicit formulas changing γ. The change from $\gamma=1/2$ to $\gamma=3/2$ is of particular interest since it provides explicit expressions for the functions in Theorem 4.2, Table 1, c), e) and f) in terms of Jacobi polynomials. Unfortunately no such operator seems to exist. However the coefficients in the formula can be determined as follows. Consider

$$\pi^2\varphi_n^{\alpha,\beta,3/2}=\sum b_M\varphi_M^{\alpha,\beta,1/2}\quad \text{(cf. Remark 5.4)}$$

where $M=N+\delta+S\tau$, τ equals $(2,1,0)$, $(2,0,0)$, $(1,1,0)$, $(1,0,0)$ or $(0,0,0)$, $S\in W$ and $\delta=(2,1,0)$. The W-invariance of the formula is clear from Theorem 7.4 so it is sufficient to compute the coefficients corresponding to $S=\text{id}$. For simplicity we also assume that $\alpha=\beta$. Then it follows from the relation

$$p_N^{\alpha,\alpha,\gamma}(-x)=(-1)^{n_1+n_2+n_3}p_N^{\alpha,\alpha,\gamma}(x)$$

that the only possible values of τ are $(2,1,0)$ and $(1,0,0)$. The first value corresponds to the leading term which is expressible in terms of c-functions. Thus there is just one coefficient left to be determined, namely the one for $\tau=(1,0,0)$ and $S=\text{id}$. To do this we use expansion with respect to the first column of the determinant in Theorem 4.6,

$$\begin{aligned}p_N^{\alpha,\alpha,1/2}(x)=\frac{1}{(x_1-x_2)(x_1-x_3)}\Big(&p_{n_1+2}^{\alpha,\alpha}(x_1)p_{n_2,n_3}^{\alpha,\alpha,1/2}(x_2,x_3)\\ &-p_{n_2+1}^{\alpha,\alpha}(x_1)p_{n_1+1,n_3}^{\alpha,\alpha,1/2}(x_2,x_3)\\ &+p_{n_3}^{\alpha,\alpha}(x_1)p_{n_1+1,n_2+1}^{\alpha,\alpha,1/2}(x_2,x_3)\Big).\end{aligned}$$

By use of this expression we see that if we put $x_3=x_2=t$ in the formula and then multiply both sides by $(x_1-t)^2$, the left side will vanish while the right side can be considered as a linear combination of Jacobi polynomials, $p_k^{\alpha,\alpha}(x_1)$, with coefficients depending on t. Since the Jacobi polynomials are linearly independent the coefficients must vanish for all t. In particular, for $p_{n_1+5}^{\alpha,\alpha}(x_1)$ we get

$$\sum b_M c(M,3,\alpha,\alpha,1/2)p_{m_2,m_3}^{\alpha,\alpha,1/2}(t,t)=0,$$

where $M=N+\delta+\varepsilon$, and ε take five different values

$$\varepsilon=(1,\pm2,0),\ (1,0,\pm2),\ (1,0,0).$$

Putting $t=1$ and solving for $b_{N+\delta+(1,0,0)}$ we obtain

THEOREM 7.5. *In the case of three variables we have the following explicit formula* (*cf. Remark* 5.4):

$$\pi^2\varphi_N^{\alpha,\alpha,3/2}=\sum_{\tau,S} b\left(S^{-1}(N+\rho^+/2)-\rho^+/2,\tau\right)\varphi_{N+\delta+S\tau}^{\alpha,\alpha,1/2}$$

where $\tau=\delta$ *or* $\mu_1/2$ *and* $S\in W$ *runs over all different* $S\tau$. *Moreover*

$$b(\lambda,\delta)=2^{-3}\frac{c(\lambda,3,\alpha,\alpha,3/2)}{c(\lambda+2\delta,3,\alpha,\alpha,1/2)}$$

and

$$b(\lambda,\mu_1/2)=-\sum b(S^{-1}(\lambda+\rho^+/2)-\rho^+/2,\delta)\cdot\frac{c(\lambda+\delta+S\delta,3,\alpha,\alpha,1/2)}{c(\lambda+\delta+\mu_1/2,3,\alpha,\alpha,1/2)}$$
$$\cdot\frac{c((\lambda+\delta)^*,2,\alpha,\alpha,1/2)}{c((\lambda+\delta+S\delta)^*,2,\alpha,\alpha,1/2)}.$$

Here

$$S(\lambda_1,\lambda_2,\lambda_3)=(\lambda_2,\pm\lambda_1,\lambda_3)\quad\text{or}\quad(\lambda_2,\lambda_3,\pm\lambda_1)$$

and

$$(\lambda_1,\lambda_2,\lambda_3)^*=(\lambda_2,\lambda_3).$$

8. Another rank two case. Consider a compact symmetric space of rank two for which the restricted root system has Dynkin diagram $0-0$ and multiplicity $2\gamma+1$. Let W denote the Weyl group, α_1 and α_2 the simple roots and moreover let μ_1 and μ_2 be the fundamental weights defined by $\langle\mu_i,\mu_j\rangle=\delta_{ij}\langle\alpha_j,\alpha_j\rangle$ (see Fig. 1).

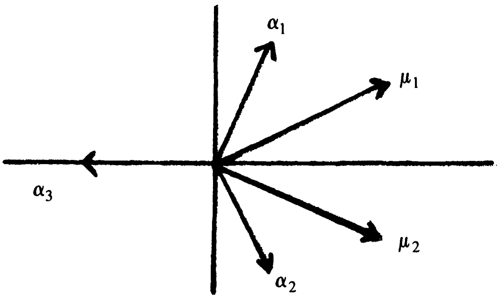


FIG. 1

Possible cases are $SU(3)/SO(3)$ ($\gamma=0$), $SU(3)$ ($\gamma=1/2$), $SU(6)/Sp(3)$ ($\gamma=3/2$) and E IV ($\gamma=7/2$). The elementary spherical functions $\varphi_{n,k}^\gamma$ corresponding to the highest weight $n\mu_1+k\mu_2$ is a trigonometric polynomial

$$\varphi_{n,k}^\gamma=\sum_{\substack{S\in W\\(i,j)\preccurlyeq(n,k)}} c_{i,j}^\gamma e^{S(i\mu_1+j\mu_2)}$$

where $(i,j)\preccurlyeq(n,k)$ means that

$$\langle i\mu_1+j\mu_2,\mu_m\rangle\le\langle n\mu_1+k\mu_2,\mu_m\rangle,\qquad m=1,2,$$

i.e.

$$2i+j\le 2n+k \quad\text{and}\quad i+2j\le n+2k.$$

$\varphi^\gamma_{n,k}$ can also be written as an algebraic polynomial

$$\varphi^\gamma_{n,k}=\sum_{(i,j)\preccurlyeq(n,k)} a_{i,j}z^i\bar z^j,$$

where

$$z=\frac{1}{3}\varphi^\gamma_{1,0}=e^{\mu_1}+e^{-\mu_2}+e^{-\mu_1+\mu_2}.$$

The orthogonality relations transform to

$$(\varphi^\gamma_{n,k},\varphi^\gamma_{i,j})=\frac{\int_\Omega\varphi^\gamma_{n,k}\overline{\varphi^\gamma_{i,j}}w^\gamma\,dz\,d\bar z}{\int_\Omega w^\gamma\,dz\,d\bar z}=\begin{cases}0 & \text{if } (n,k)=(i,j),\\ \dfrac{1}{d_{n,k}} & \text{if } (n,k)\neq(i,j).\end{cases}$$

Here $d_{n,k}$ denotes the dimension of the representation with highest weight $n\mu_1+k\mu_2$, w is defined by

$$w=-z^2\bar z^2+4z^3+4\bar z^3-18z\bar z+27,$$

and Ω is the region in the complex plane for which $w\ge 0$. Expressing $d_{n,k}$ in terms of Harish-Chandra's c-function as in §4 we get

$$\|\varphi^\gamma_{n,k}\|^2=\frac{1}{d_{n,k}}$$
$$=\frac{(\gamma+3/2)_{n+k}n!\,k!\,(\gamma+1/2)(\gamma+1/2)(2\gamma+1)}{(3\gamma+3/2)_{n+k}(2\gamma+1)_n(2\gamma+1)_k(n+\gamma+1/2)(k+\gamma+1/2)(n+k+2\gamma+1)}$$

The leading coefficient of $\varphi^\gamma_{n,k}$ is

$$a^\gamma_{n,k}=\frac{(\gamma+1/2)_n(\gamma+1/2)_k(2\gamma+1)_{n+k}}{(2\gamma+1)_n(2\gamma+1)_k(3\gamma+3/2)_{n+k}},$$

and the recurrence formula for elementary spherical functions derived in [11] takes the form

$$\bar z\varphi^\gamma_{n,k}=b_{0,1}\varphi^\gamma_{n,k+1}+b_{-1,0}\varphi^\gamma_{n-1,k}+b_{1,-1}\varphi^\gamma_{n+1,k-1},$$

where

$$b_{0,1}=\frac{(k+2\gamma+1)(n+k+3\gamma+3/2)}{(k+\gamma+1/2)(n+k+2\gamma+1)},$$
$$b_{-1,0}=\frac{n(n+k+\gamma+1/2)}{(n+\gamma+1/2)(n+k+2\gamma+1)},$$
$$b_{1,-1}=\frac{k(n+2\gamma+1)}{(k+\gamma+1/2)(n+\gamma+1/2)}.$$

In [7, parts III and IV] Koornwinder considered these polynomials for general $\gamma > -5/6$. His polynomials, denoted by $p^{\gamma}_{n,k}$, were normalized to have the leading coefficient equal to one. Thus

$$\varphi^{\gamma}_{n,k}(z,\bar{z}) = \frac{p^{\gamma}_{n,k}(z,\bar{z})}{p^{\gamma}_{n,k}(3,3)}.$$

To show that the formulas for the quadratic norm, the leading coefficient and the recurrence formula still remain valid for $\gamma > -5/6$, one can proceed as follows. First the value of $p^{\gamma}_{n,k}(3,3)$ is obtained with the aid of a generating function (see [8, p. 483]) and repeated application of a raising operator E^{γ}_{+} (see below). Next the recurrence formula is verified by explicit calculations of suitable monomials of $p^{\gamma}_{n,k}$. Finally this formula also yields a recurrence relation for the quadratic norm

$$\|\varphi^{\gamma}_{n,k+1}\|^2 = \|\varphi^{\gamma}_{n,k}\|^2 \frac{(n+k+\gamma+3/2)(n+k+2\gamma+1)(k+\gamma+1/2)(k+1)}{(n+k+3\gamma+3/2)(n+k+2\gamma+2)(k+\gamma+3/2)(k+2\gamma+1)}.$$

As indicated above there are lowering and raising operators E^{γ}_{-} and E^{γ}_{+}. They are defined by

$$E^{\gamma}_{-} = \frac{\partial^3}{\partial z^3} + \frac{\partial^3}{\partial \bar{z}^3} + z\frac{\partial^3}{\partial z^2 \partial \bar{z}} + \bar{z}\frac{\partial^3}{\partial z \partial \bar{z}^2} + \left(\gamma + \frac{5}{2}\right)\frac{\partial^2}{\partial z \partial \bar{z}}$$

and

$$E^{\gamma}_{+} = w^{-\gamma}(E^{\gamma}_{-})^* w^{\gamma+1}$$

Put also

$$L^{\gamma}_{+} = -w^{-\gamma}\frac{\partial}{\partial \bar{z}} w^{\gamma+1}.$$

Then

$$E^{\gamma}_{-}\varphi^{\gamma}_{n,k} = \frac{2nk(n+k+\gamma+1/2)(n+2\gamma+1)(k+2\gamma+1)(n+k+3\gamma+3/2)}{(6\gamma+5)(6\gamma+6)(6\gamma+7)}\varphi^{\gamma+1}_{n-1,k-1},$$

$$E^{\gamma}_{+}\varphi^{\gamma+1}_{n,k} = \frac{1}{2}(6\gamma+5)(6\gamma+6)(6\gamma+7)\varphi^{\gamma}_{n+1,k+1},$$

$$L^{\gamma}_{+}\varphi^{\gamma+1}_{n,k} = \frac{(6\gamma+5)(6\gamma+6)(6\gamma+7)}{2}$$

$$\cdot\left(\frac{1}{(n+\gamma+3/2)(k+\gamma+3/2)}\varphi^{\gamma}_{n,k+2} - \frac{1}{(k+\gamma+3/2)(n+k+2\gamma+3)}\varphi^{\gamma}_{n+1,k}\right.$$

$$\left. - \frac{1}{(n+\gamma+3/2)(n+k+2\gamma+3)}\varphi^{\gamma}_{n+2,k+1}\right).$$

We can now prove

THEOREM 8.1. *For $\gamma > -5/6$ there holds*

$$-w\varphi_{n,k}^{\gamma+1} = \sum a_{i,j}\left(\varphi_{n+i,k+j}^{\gamma} - \varphi_{n+1,k+1}^{\gamma}\right),$$

where the only nonzero coefficients are

$$a_{2,2} = \frac{(n+k+3\gamma+9/2)\cdot C}{(n+\gamma+3/2)(k+\gamma+3/2)(n+k+2\gamma+3)(n+k+2\gamma+4)},$$

$$a_{3,0} = -\frac{(n+2\gamma+3)\cdot C}{(k+\gamma+3/2)(n+\gamma+3/2)(n+\gamma+5/2)(n+k+2\gamma+3)},$$

$$a_{0,3} = \frac{(k+2\gamma+3)\cdot C}{(n+\gamma+3/2)(k+\gamma+3/2)(k+\gamma+5/2)(n+k+2\gamma+3)},$$

$$a_{-1,2} = -\frac{n\cdot C}{(n+\gamma+1/2)(n+\gamma+3/2)(k+\gamma+3/2)(n+k+2\gamma+3)},$$

$$a_{2,-1} = \frac{k\cdot C}{(k+\gamma+1/2)(k+\gamma+3/2)(n+\gamma+3/2)(n+k+2\gamma+3)},$$

$$a_{0,0} = \frac{(n+k+\gamma+3/2)\cdot C}{(k+\gamma+3/2)(n+\gamma+3/2)(n+k+2\gamma+2)(n+k+2\gamma+3)},$$

$$C = \frac{1}{2}(6\gamma+5)(6\gamma+6)(6\gamma+7).$$

Proof. Apply the operator L_+^γ to $\bar{z}\varphi_{n,k}^{\gamma+1}$.

$$-w\varphi_{n,k}^{\gamma+1} = L_+^\gamma \bar{z}\varphi_{n,k}^{\gamma+1} - \bar{z}L_+^\gamma\varphi_{n,k}^{\gamma+1}.$$

Then use the formulas for $\bar{z}\varphi_{n,k}^{\gamma+1}$ and for $L_+^\gamma\varphi_{n,k}^{\gamma+1}$. As a special case of this formula we obtain for $\gamma = 1/2$ an explicit expression for the elementary spherical functions on $SU(6)/Sp(3)$ in terms of the characters on $SU(3)$. For $\gamma = -1/2$ the formula expresses the characters in terms of the functions

$$\varphi_{n,k}^{-1/2} = \frac{1}{6}\sum_{S\in W} e^{S(n\mu_1 + k\mu_2)}$$

Finally, by repeated application of the formula the elementary spherical functions on EIV ($\gamma = 7/2$) are explicitly determined.

REFERENCES

[1] F. A. BEREZIN AND F. I. KARPELEVIČ, *Zonal spherical functions and Laplace operators on some symmetric spaces*, Dokl. Akad. Nauk SSSR, (N.S.), 118 (1958), pp. 9–12. (In Russian.)

[2] HARISH-CHANDRA, *Spherical functions on a semisimple Lie group* I, Amer. Math., 80 (1958), pp. 241–310.

[3] S. HELGASON, *Differential Geometry and Symmetric Spaces*, Academic Press, New York, 1962.

[4] A. T. JAMES, *Normal multivariate analysis and the orthogonal group*, Ann. Math. Statist., 25 (1954), pp. 40–75.

[5] ———, *Special functions of matrix and single argument in statistics*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 497–520.

[6] A. T. JAMES AND A. G. CONSTANTINE, *Generalized Jacobi polynomials as spherical functions of the Grassmann manifold*, Proc. London Mat. Soc., (3) 29 (1974), pp. 174–192.

[7] T. H. Koornwinder, *Orthogonal polynomials in two variables which are eigenfunctions of two algebraically independent partial differential operators*, I–IV, Indag. Math., 36 (1974), pp. 48–66, 357–381.

[8] ______, *Two-variable analogues of the classical orthogonal polynomials*, in Theory and Applications of special functions, R. Askey, ed., Academic Press, New York, 1975, pp. 435–495.

[9] T. H. Koornwinder and I. G. Sprinkhuizen-Kuyper, *Generalized power series expansions for a class of orthogonal polynomials in two variables*, this Journal, 9 (1978), pp. 457–483.

[10] I. G. Sprinkhuizen-Kuyper, *Orthogonal polynomials in two variables. A further analysis of the polynomials orthogonal on a region bounded by two lines and a parabola*, this Journal, 7 (1976), pp. 501–518.

[11] L. Vretare, *Elementary spherical functions on symmetric spaces*, Math. Scand., 39 (1976), pp. 343–358.

[12] G. Warner, *Harmonic Analysis on Semisimple Lie Groups* I *and* II, Springer-Verlag, Berlin, 1972.

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics
016

A SEQUENCE OF PIECEWISE ORTHOGONAL POLYNOMIALS*

Y. Y. FENG† AND D. X. QI‡

Abstract. In this paper we study the construction of an orthogonal sequence $(U_{k,n}^{(i)})$ of piecewise polynomials of degree k which is complete in L_2. Explicit expressions of $(U_{k,n}^{(i)})$ for $k=1,2,3$ are given.

We also study the sign-change properties of this sequence and consider the convergence of the corresponding Fourier series. The results generalize those obtained earlier for Walsh system.

AMS-MOS subject classification (1980). Primary 41A15

Key words. polynomial, piecewise polynomial, Legendre polynomial, series expansion, orthonormal function

1. Introduction. The study of orthogonal functions is very important. A number of function sets are known which are found to be orthogonal and hence can be used for series representation. The Fourier series, a system of sine and cosine functions, is the basis for development in many of these areas.

Additionally certain polynomials can be made orthogonal. These orthogonal polynomials form a series, $\varphi_n(x)$ $(n=0,1,2,\cdots)$, where n is the degree of the polynomial. This class contains many special functions commonly encountered in practical applications, e.g. Chebyshev, Hermite, Laguerre, Jacobi, Legendre polynomials.

None of these have the essential simplicity of the Walsh and Haar functions, the most important examples of nonsinusoidal functions, which form complete sets of orthogonal functions for the Hilbert space $L_2[0,1]$. Having this property, they provide an effective tool in Fourier analysis. With the application of digital techniques and semiconductor technology this kind of complete system of orthogonal functions has been considered and applied [1]. This system may also have other advantages rendering more directly it useful for some applications.

In 1910, Alfred Haar [6] proposed a set of orthogonal functions, taking essentially only two values, such that the formal expansion of a given continuous function in the new functions converges uniformly to the given function. The Walsh functions defined in 1923 by J. L. Walsh [7] form a complete orthogonal set taking only the values $+1$ and -1, and have been found to have many properties similar to the sinusoidal series.

From the point of view of approximation theory, it is important to construct a set such that functions in this set cannot be only piecewise constant. The Schauder basis was obtained by integration of the Haar function. Applying the Schmidt orthonormalization procedure to the Schauder basis, Ciesielski (1968) introduced an orthonormal uniformly bounded sequence of polygonals [4] which was a development of the Franklin orthonormal set discovered in 1928 [5]. The functions in this set are implicit.

In this paper we study the construction of an orthogonal sequence $(U_{k,n}^{(i)})$ of piecewise polynomials of degree k which is complete in L_2. Explicit expressions of $(U_{k,n}^{(i)})$ for $k=1,2,3$ are given.

*Received by the editors September 1, 1981, and in revised form November 19, 1982. This research was sponsored by the U.S. Army under contract DAAG29-80-C-0041.

†Department of Mathematics, China University of Science and Technology, Hofei, China, and Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706.

‡Department of Mathematics, Jilin University, Changchun, China, and Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706.

We also study the sign-change properties of this sequence and consider the convergence of the corresponding Fourier series. The results generalize those obtained earlier for Walsh systems.

2. An orthonormal sequence of piecewise linear functions. The sequence U, which we will study in this section, consists of the following functions:

(2.1)

$$U_0(x):=1, \qquad U_1(x):=\sqrt{3}\,(1-2x), \qquad 0\le x\le 1,$$

$$U_2^{(1)}(x):=\begin{cases}\sqrt{3}\,(1-4x),\\ \sqrt{3}\,(4x-3),\end{cases} \qquad U_2^{(2)}(x):=\begin{cases}1-6x, & 0\le x<\frac{1}{2},\\ 5-6x, & \frac{1}{2}<x\le 1,\end{cases}$$

$\cdots\ \cdots$

$$U_{n+1}^{(2k-1)}(x):=\begin{cases}U_n^{(k)}(2x), & 0\le x<\frac{1}{2},\\ U_n^{(k)}(2-2x), & \frac{1}{2}<x\le 1,\end{cases}$$

$$U_{n+1}^{(2k)}(x):=\begin{cases}U_n^{(k)}(2x), & 0\le x<\frac{1}{2},\\ -U_n^{(k)}(2-2x), & \frac{1}{2}<x\le 1,\end{cases} \qquad k=1,2,3,\cdots,2^{n-1}, \quad n=2,3,\cdots,\infty.$$

At a point of discontinuity, let these functions be the average of the two one-sided limits. The first eight of these functions are shown in Fig. 1.

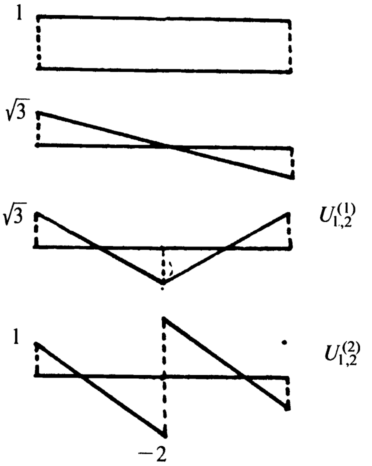


FIG. 1

Now we consider the orthogonality of the sequence U. We have the following theorem.

THEOREM 2.1. *The sequence of functions* $\{U_n^{(i)}\}$ *is normal and orthogonal; i.e.,*

$$\int_0^1 U_n^{(k)}(x)\,U_m^{(j)}(x)\,dx=\delta_{n,m}\delta_{k,j} \tag{2.2}$$

for $n,m=0,1,2,\cdots,\ k=1,2,3,\cdots,2^{n-1}, j=1,2,3,\cdots,2^{m-1}$, *with*

$$\delta_{ij}=\begin{cases}1, & i=j,\\ 0, & i\neq j.\end{cases}$$

Proof. It is easy to prove this theorem by mathematical induction, we leave it as an exercise for the reader.

We denote the collection of all piecewise polynomials of order $k+1$ with partition Δ_n by

$$\mathbb{P}_{k+1,\Delta_n},$$

where Δ_n is the uniform partition on 2^{n-1} intervals. It is obvious that

$$\dim \mathbb{P}_{k+1,\Delta_n} = (k+1)2^{n-1}. \tag{2.3}$$

Let

$$M_2 n := \operatorname{span}\left(U_0, U_1, \cdots, U_n^{(1)}, \cdots, U_n^{(2^{n-1})}\right).$$

When $k=1$, we get

$$M_2 n = \mathbb{P}_{2,\Delta_n} \tag{2.4}$$

since $\dim M_2 n = \dim \mathbb{P}_{2,\Delta_n} = 2^n$ and $M_2 n \subseteq \mathbb{P}_{2,4n}$. From (2.4) we obtain the following theorem.

THEOREM 2.2. *If f is a piecewise linear function whose breakpoints can only appear at q/p, where q is an integer and p is a power of two, then f can be exactly expressed by finitely terms of the series $\Sigma\alpha_i U_i$.*

Before studying convergence properties we consider the number of sign changes of the functions in the sequences U. First we define

$$S^-(f) := \sup\{n : \exists t_1 < t_2 < \cdots < t_{n+1}, f(t_i) f(t_{i+1}) < 0\}$$

to be the number of the sign changes of f on $[0, 1]$. It is easy to see that

$$S^-(U_0)=0, \quad S^-(U_1)=1, \quad S^-\left(U_2^{(1)}\right)=2, \quad S^-\left(U_2^{(2)}\right)=3.$$

By the method of construction of the sequence U,

$$S^-\left(U_{n+1}^{(2k-1)}\right) = 2S^-\left(U_n^{(k)}\right)$$

and

$$S^-\left(U_{n+1}^{(2k)}\right) = 2S^-\left(U_n^{(k)}\right) + 1;$$

thus

$$S^-\left(U_n^{(k)}(x)\right) = 2^{n-1} + k - 1,$$

since this formula holds for $n=2$ and follows for the general case by induction. Hence, each function $U_n^{(k)}$ has one more sign change than the preceding one. Therefore, it is convenient to use the notation $U_0, U_1, U_2, U_3, \cdots$ instead of $U_n^{(k)}$. When we study their sign changes from now on, we will use both $\{U_n^{(k)}\}$ and $\{U_N\}$ freely. Obviously

$$U_n^{(k)} = U_{2^{n-1}+k-1} \quad \text{for } n=2,3,\cdots, \quad k=1,2,3,\cdots,2^{n-1}.$$

Thus we get the following theorem.

THEOREM 2.3. $S^-(U_m) = m, m=0,1,2,3,\cdots$. *That is* $S^-(U_n^{(k)}) = 2^{n-1}+k-1$ *for* $n=1,2,3,\cdots,k=1,2,3,\cdots,2^{n-1}$.

Now we consider the convergence properties. The Fourier series of a given function F in terms of the functions U_i is

$$F \sim \sum_{i=0}^{\infty} \alpha_i U_i \tag{2.5}$$

with

$$\alpha_i := (F, U_i) = \int_0^1 F(x) U_i(x)\, dx. \tag{2.6}$$

Let

$$p_{n+1}F := \sum_{i=0}^{n} \alpha_i U_i \tag{2.7}$$

be the nth partial sum of the series (2.5). Then $p_{n+1}F$ is the best L_2-approximation to F from $M_{n+1} = \operatorname{span}(U_i)_0^n$. Hence it is convergent to F if F is in L_2, since the union of the M_n's is dense in L_2. Thus we get the following theorem.

THEOREM 2.4. *If $F \in L_2[0,1]$, then*

$$\lim_{n\to\infty} \|F - p_n F\|_2 = 0.$$

Next we will prove that $P_{2^n}F$ uniformly approximates $F \in C[0,1]$. It is well known [2] that

$$\|F - P_{2^n}F\|_\infty \le \left(1 + \|P_{2^n}\|\right) \operatorname{dist}_\infty(F, M_{2^n})$$

and we know

$$\|P_{2^n}\| = \|P_2\| < \infty$$

since the least-squares approximation for $M_{2^n} = \mathbb{P}_{2,\Delta_n}$ is local and $C[0,1]$ is in the closure of $\bigcup_{n=0}^{\infty} M_{2^n}$. Therefore we get the following theorem.

THEOREM 2.5. *Let $F \in C[0,1]$, P_{2^n} be an L_2-projector onto M_{2^n} on $C[0,1]$. Then*

$$\lim_{n\to\infty} \|F - P_{2^n}F\|_\infty = 0.$$

However, not every continuous function can be expanded in terms of a sequence U. We prove that there exists a continuous function whose expansion in terms of the U's does not converge at a point of the interval.

Suppose $\{\varphi_i\}_{i=1}^{\infty}$ is a complete orthonormal system on $[a,b]$. For a function f, the partial sum f_n of its formal Fourier series is defined by

$$f_n(s) := \int_a^b K_n(s,t) f(t)\, dt$$

with

$$K_n(s,t) := \sum_{i=1}^{n} \varphi_i(s)\varphi_i(t).$$

In our case the kernel is

$$K_n^{(j)}(x,y) := U_0(x)U_0(y) + U_1(x)U_1(y) + \cdots + U_n^{(j)}(x)U_n^{(j)}(y).$$

We have the following theorem.

THEOREM 2.6. *There exists a continuous function $f \in C[0,1]$ whose expansion*

$$\sum_{i=1}^{n} \int_0^1 f(x) U_i(x)\,dx U_i$$

in terms of $\{U_i\}$ does not converge to $f(x)$ uniformly.

Proof. According to the principle of uniform boundedness [9], from Theorem 2.5 for $j = 2^{n-1}$ and $x \in [0,1]$, $\int_0^1 |k_n^{(j)}(x,y)|\,dy$ is uniformly bounded for n. For general j let

$$K_n^{(j)}(x,y) = K_{n-1}^{(2n^{n-2})} + R_n^{(j)}(x,y),$$

i.e.,

$$R_n^{(j)}(x,y) := U_n^{(1)}(x) U_n^{(1)}(y) + \cdots + U_n^{(j)}(x) U_n^{(j)}(y).$$

Therefore it is sufficient to prove that the integral

$$C_n^{(j)}(\alpha) := \int_0^1 |R_n^{(j)}(\alpha,y)|\,dy$$

is not uniformly bounded for all n,j.

Table 1 shows the value of $C_n^{(k)}(0)$ for small value of n and for each value of $k \le 2^{n-1}$.

TABLE 1

$n=2$								$\frac{3}{2}$							
$n=3$				$\frac{3}{2}$								$\frac{3}{2}$			
$n=4$		$\frac{3}{2}$				$\frac{3}{2}$				$\frac{9}{4}$				$\frac{3}{2}$	
$n=5$	$\frac{3}{2}$		$\frac{3}{2}$		$\frac{9}{4}$		$\frac{3}{2}$		$\frac{21}{8}$		$\frac{9}{4}$		$\frac{21}{8}$		$\frac{3}{2}$

We have the general formulas

$$C_n^{(1)}(0) = C_n^{(2^{n-2})}(0) = \tfrac{3}{2},$$
$$C_n^{(2k)}(0) = C_{n-1}^{(k)}(0),$$
$$C_n^{(2k+1)}(0) = \tfrac{1}{2}\left(C_{n-1}^{(k)}(0) + C_{n-1}^{(k+1)}(0)\right) + \tfrac{3}{4}.$$

Let $k_3 := 1$, $k_n := 2k_{n-1} + (-1)^n$. Then $\lim_{n\to\infty} C_n^{(k_n)}(0) = \infty$. So $\int_0^1 |R_n^{(j)}(0,y)|\,dy$ is not uniformly bounded, and the proof is complete.

3. An orthonormal sequence of piecewise polynomials of degree k. In this section we study a general procedure for constructing a sequence of orthonormal polynomial functions. We use the following notation:

$$Z := \{0,1,2,\cdots\}, \qquad I_k := \{1,2,\cdots,k\},$$
$$O_n := \{1,3,5,\cdots,2n-1\}, \qquad E_n := \{0,2,4,\cdots,2n\}.$$
$$[x] := \max\{n : \text{integer}, n \le x\}, \qquad \langle f,g \rangle := \int_0^1 f(x) g(x)\,dx.$$

Suppose that $\{U_i\}_{i=0}^{k}$ is a sequence of orthonormal polynomials defined on $[0,1]$, even or odd with respect to the point $x = \frac{1}{2}$ and the degree of U_i is i. First we give the following theorem.

THEOREM 3.1. *There exist $k+1$ polynomials $Q_{k,i}(x)$ $(i\in I_{k+1})$, of exact degree k such that*

$$\frac{d^jQ_{k,i}(\frac{1}{2})}{dx^j}=0 \quad \textit{for } j=k-i-1, k-i-3\cdots, \tag{3.1}$$

and with the property that

$$U^{(i)}_{k,2}(x):=\begin{cases} Q_{k,i}(x), & 0\le x<\frac{1}{2}, \quad i\in I_{k+1}, \\ (-1)^{k+i}Q_{k,i}(1-x), & \frac{1}{2}<x\le 1, \end{cases} \tag{3.2}$$

satisfies

$$\left\langle U^{(i)}_{k,2}(x), x^j\right\rangle=0, \qquad j\in I_k\cup\{0\}, \quad i\in I_{k+1}, \tag{3.3}$$

$$\left\langle U^{(i)}_{k,2}(x), U^{(j)}_{k,2}(x)\right\rangle=\delta_{ij}, \qquad i,j\in I_{k+1} \tag{3.4}$$

with

$$\delta_{ij}=\begin{cases} 1, & i=j, \\ 0, & i\ne j. \end{cases}$$

Proof. Let $k=2m$ for $m\in Z$ and

$$Q_{k,i}(x):=\sum_{j=0}^{k} a^{(i)}_j x^j$$

on $[0,\frac{1}{2}]$. The coefficients $a^{(i)}_0$, $a^{(i)}_1,\cdots,a^{(i)}_{2m-1}$ $(i\in I_{2m+1})$ are defined by the following equations:

$$\begin{aligned} &\left\langle U^{(2i+1)}_{2m,2}, U_j\right\rangle=0, && j\in O_m, \\ &\left\langle U^{(2i+1)}_{2m,2}, U^{(j)}_{2m,2}\right\rangle=0, && j\in O_i, \quad i\in I_m\cup\{0\}, \\ &\left.\frac{d^jQ_{2m,2i+1}}{dx^j}\right|_{x=\frac{1}{2}}=0, && j\in E_{m-i-1}, \end{aligned} \tag{3.5}$$

with $O_0=\varnothing$, $E_{-1}=\varnothing$,

$$\begin{aligned} &\left\langle U^{(2i)}_{2m,2}, U_j\right\rangle=0, && j\in E_m, \\ &\left\langle U^{(2i)}_{2m,2}, U^{(j)}_{2m,2}\right\rangle=0, && j\in E_{i-1}\setminus\{0\}, \quad i\in I_m. \\ &\left.\frac{d^jQ_{2m,2i}}{dx^j}\right|_{x=\frac{1}{2}}=0, && j\in O_{m-i}. \end{aligned} \tag{3.6}$$

It is obvious that (3.5) and (3.6) have at least one solution $U^{(i)}_{k,2}$ for given $i\in I_{k+1}$. Indeed, the sequence U_1, $U_3,\cdots,U_{2m-1}$, $Q_{2m,1}$, $Q_{2m,3},\cdots,Q_{2m,2m+1}$ is obtained by orthogonalization of the sequence $(x-\frac{1}{2})$, $(x-\frac{1}{2})^3,\cdots,(x-\frac{1}{2})^{2m-1}$, $(x-\frac{1}{2})^{2m}$, $(x-\frac{1}{2})^{2m-2},\cdots,(x-\frac{1}{2})^2$, 1 on the interval $(0,\frac{1}{2})$ with respect to the constant weight function.

It is easy to see that the degree d of $Q_{k,i}$ satisfies $k\ge d\ge k-i+1$. From [2] (de Boor), we conclude

$$S^-\left(U^{(i+1)}_{k,2}\right)\ge k+i+1,$$

but from [2] we know

$$k \geq d \geq S^{-}\left(Q_{k,i}(0), \cdots, Q_{k,i}^{(d)}(0)\right) \geq Z\left(Q_{k,i}; (0, \tfrac{1}{2})\right) + S^{+}\left(Q_{k,i}(\tfrac{1}{2}), \cdots, Q_{k,i}^{(d)}(\tfrac{1}{2})\right)$$
$$\geq m + \left[\tfrac{1}{2}i\right] + m - \left[\tfrac{1}{2}i\right] = k.$$

Hence all inequalities are equalities. It follows that $Q_{2m,i}$ has exact degree $2m$. It is easy to check $U_{k,2}^{(i)}$ satisfies (3.1), (3.3) and (3.4). Let

$$M_{2(k+1)} := \operatorname{span}\{U_0, U_1, \cdots, U_k, U_{k,2}^{(1)}, \cdots, U_{k,2}^{(k+1)}\}. \tag{3.7}$$

From (2.3) and (3.7), we know

$$M_{2(k+1)} = \mathbb{P}_{k+1,\Delta_2}$$

since $\dim M_{2(k+1)} = \dim \mathbb{P}_{k+1,\Delta_2}$ and $M_{2(k+1)} \subseteq \mathbb{P}_{k,\Delta_2}$. Therefore the number of polynomials $Q_{k,i}$ is no more than $k+1$. We have proved the theorem for $k=2m$. When k is odd, the same kind of argument confirms the theorem.

After normalization, when $k=0, 1$, the functions $U_{k,2}^{(i)}$ $(i \in I_{k+1})$ defined by Theorem 3.1 are Walsh functions and piecewise linear functions ($U_2^{(1)}(x)$, $U_2^{(1)}(x)$ in Fig. 1 respectively); when $k=2, 3$, $U_{2,2}^{(i)}$ $(i \in I_3)$ and $U_{3,2}^{(i)}$ $(i \in I_4)$ are as follows:

$$\begin{aligned}
&U_{2,2}^{(1)} = \sqrt{5}\,(16x^2 - 10x + 1),\\
&U_{2,2}^{(2)} = \sqrt{3}\,(30x^2 - 14x + 1),\\
&U_{2,2}^{(3)} = 40x^2 - 16x + 1;\\
&U_{3,2}^{(1)} = \sqrt{7}\,(-64x^3 + 66x^2 - 18x + 1),\\
&U_{3,2}^{(2)} = \sqrt{5}\,(-140x^3 + 144x^2 - 24x + 1),\\
&U_{3,2}^{(3)} = \sqrt{3}\,(-224x^3 + 156x^2 - 28x + 1),\\
&U_{3,2}^{(4)} = -280x^3 + 180x^2 - 30x + 1.
\end{aligned}$$

The graphs of these functions are given in Fig. 2.

After getting $U_{k,2}^{(i)}$ $(i \in I_{k+1})$, we define in general

$$U_{k,n+1}^{(2l-1)}(x) := \begin{cases} U_{k,n}^{(l)}(2x), & 0 \leq x < \tfrac{1}{2},\\ U_{k,n}^{(l)}(2-2x), & \tfrac{1}{2} < x \leq 1, \end{cases} \tag{3.8}$$

$$U_{k,n+1}^{(2l)}(x) := \begin{cases} U_{k,n}^{(l)}(2x), & 0 \leq x < \tfrac{1}{2},\\ -U_{k,n}^{(l)}(2-2x), & \tfrac{1}{2} < x \leq 1, \end{cases} \tag{3.9}$$

$$l \in I_{2^{n-2}(k+1)}, \qquad n \in z \setminus \{0, 1\}.$$

We have the following theorem about the orthogonality of the sequence $\{U_{k,n}^{(i)}\}$.

THEOREM 3.2. *The sequence of functions $\{U_{k,n}^{(i)}\}$ is normal and orthogonal; i.e.*

$$\left\langle U_{k,n}^{(i)}, U_{k,m}^{(j)} \right\rangle = \delta_{n,m}\delta_{i,j}$$

with $U_{k,1}^{(l+1)} := U_l$, $l \in I_k U\{0\}$; $i \in I_\mu$, $j \in I_\nu$, *where* $\mu = (k+1)2^{\max(n-2,0)}$, $\nu = (k+1)2^{\max(m-2,0)}$.

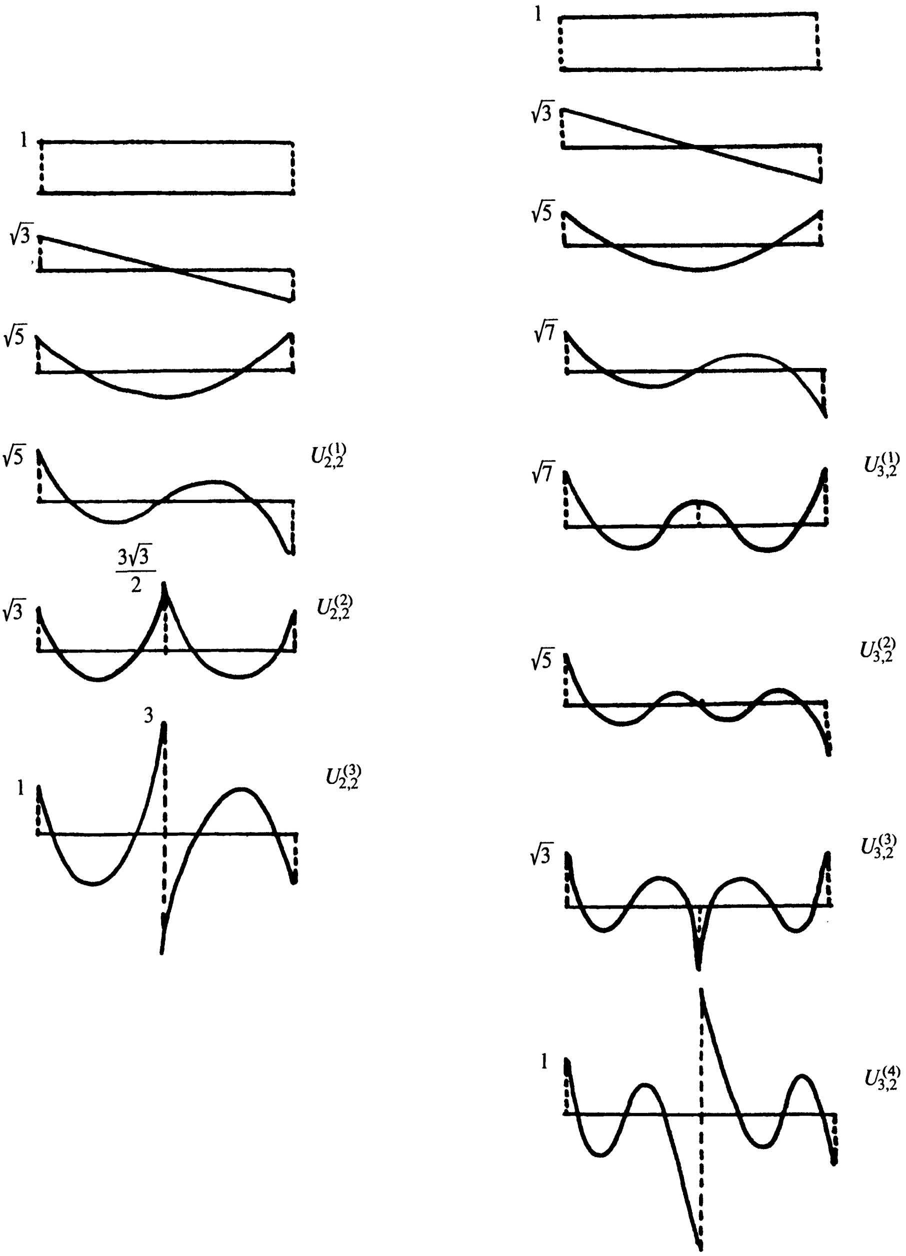


FIG. 2

Proof. The same kind of argument as in the proof of Theorem 2.1 confirms this theorem.

It is easy to see that

$$U_{k,m}^{(j)} \in \mathbb{P}_{k+1,\Delta_n}, \qquad m \in I_n, \quad j \in I_\nu.$$

Let

$$M_{(k+1)2^{n-1}} := \operatorname{span}\left(U_0, U_1, \cdots, U_{k,n}^{(1)}, \cdots, U_{k,n}^{((k+1)2^{n-2})}\right).$$

It is obvious that

$$M_{2^{n-1}(k+1)}=\mathbb{P}_{k+1,\Delta_n},$$

Therefore we have the following theorem.

THEOREM 3.3. *If f is a piecewise polynomial of degree k with breakpoints only at q/p, where q is integer and p is a power of two, then f can be exactly expressed by finite terms of the series $\Sigma_{i,j}\alpha_{i,j}U_{k,i}^{(j)}$.*

Let $S^+(a_1,\cdots,a_n)$ denote the maximum number of sign changes in the sequence $a_0, a_1,\cdots,a_n$ obtainable by giving any zero element the value $+1$ or -1, and define

$S^-(a_0,\cdots,a_n):=$ the number of sign changes in the sequence $a_0, a_1,\cdots,a_n$, where zeros are ignored.

Because $\{U_i\}$ $(i\in I_k\cup\{0\})$ is orthogonal on $[0,1]$, it is well known that

$$Z(U_i;[0,1])=i, \qquad i\in I_k\cup\{0\}$$

with $Z(f;[a,b])$ denoting the number of zeros of f on $[a,b]$.

In order to study the sign changes of $U_{k,2}^{(i)}$ $(i\in I_{k+1})$ on $[0,1]$ we need the following lemma.

LEMMA 1 (de Boor [1]). *If $\mathbf{t}=(t_i)_1^{n+k}$ is nondecreasing in $[a,b]$, with $t_i<t_{i+k}$ all i, and $f\in L_1[a,b]$ is orthogonal to $\mathbb{S}_{k,\mathbf{t}}$ on $[a,b]$, then there exists $\boldsymbol{\xi}=(\xi_i)_1^{n+1}$, strictly increasing in $[a,b]$ with $t_i\le\xi_i\le t_{i+k-1}$ (any equality holding iff $t_i=t_{i+k-1}$), $i\in I_{n+1}$, such that f is also orthogonal to $\mathbb{S}_{1,\boldsymbol{\xi}}$. Here $\mathbb{S}_{k,\mathbf{t}}$ denotes the collection of splines of order k with known sequence* $\mathbf{t}$.

In particular, if f is continuous, then it must vanish at the n points of some strictly increasing sequence $(\eta_i)_1^n$ with $t_i<\eta_i<t_{i+k}$ for all i.

It is easy to see that

$$\mathbb{S}_{k+1,\Delta_2^{(i)}}=M_{k+1+i}=\operatorname{span}\left(U_0,U_1,\cdots,U_{k,2}^{(i)}\right),$$

where $\Delta_2^{(i)}$ is the knot sequence $(t_j)_1^{2(k+1)+i}$,

$$t_j:=\begin{cases}0, & j\le k+1,\\ \frac{1}{2}, & k+1<j\le k+i+1,\\ 1, & j\ge k+2+i.\end{cases} \tag{3.10}$$

Using Lemma 1, we get

$$S^-\left(U_{k,2}^{(i+1)}\right)=k+1+i, \qquad i\in I_k\cup\{0\}, \tag{3.11}$$

since

$$\left\langle U_{k,2}^{(i+1)},S\right\rangle=0, \qquad S\in\mathbb{S}_{k+1,\Delta_2^{(i)}}$$

and $U_{k,2}^{(i+1)}\in\mathbb{S}_{k+1,\Delta_2^{(i+1)}}$.

We would like to study some further properties of the piecewise polynomials $\{U_{k,2}^{(i)}\}$. First, from the Budan–Fourier theorem [8], we know that if P is a polynomial of exact degree k, then

$$Z(P;(a,b))\le S^-\left(P(a),\cdots,P^{(k)}(a)\right)-S^+\left(P(b),\cdots,P^{(k)}(b)\right). \tag{3.12}$$

For convenience, suppose $k=2m$. From (3.2), (3.11) we know

$$Z\left(Q_{k,i};\left(0,\tfrac{1}{2}\right)\right)=m+\left\lceil\tfrac{i}{2}\right\rceil; \tag{3.13}$$

by (3.5), (3.6)

$$S^{+}\left(Q_{k,i}\left(\tfrac{1}{2}\right), Q'_{k,i}\left(\tfrac{1}{2}\right), \cdots, Q_{k,i}^{(k)}\left(\tfrac{1}{2}\right)\right) \geq m - \left\lceil \tfrac{i}{2} \right\rceil. \tag{3.14}$$

Because of (3.12), (3.13) and (3.14), we get

$$S^{-}\left(Q_{k,i}(0), \cdots, Q_{k,i}^{(k)}(0)\right) = k. \tag{3.15}$$

Therefore, from Descartes' rule, we know that the coefficients of the polynomial $Q_{k,i}$ strictly alternate in sign.

A similar discussion shows that (3.15) holds when k is odd. Thus we have the following lemma.

LEMMA 2. 1. $S^{-}(U_{k,2}^{(l)}) = k+l$, $l \in I_{k+1}$.

2. *The coefficients of the polynomial* Q_{k+1} *strictly alternate in sign.*

By the method of construction of the sequence $\{U_{k,n}^{(i)}\}$ (3.8), (3.9), we know

$$S^{-}\left(U_{k,n+1}^{(2l-1)}\right) = 2S^{-}\left(U_{k,n}^{(l)}\right),$$

$$S^{-}\left(U_{k,n+1}^{(2l)}\right) = 2S^{-}\left(U_{k,n}^{(l)}\right) + 1;$$

thus

$$S^{-}\left(U_{k,n}^{(l)}\right) = (k+1)2^{n-2} + l - 1,$$

since this formula holds for $n=2$, and follows for the general case by induction. Hence each function $U_{k,n}^{(l)}$ has one more sign change than the preceding one. It is convenient to use the notation $U_{k,0}$, $U_{k,1}, \cdots$ instead of $U_{k,n}^{(l)}$ when we study their sign changes. From now on, we will use both $\{U_{k,n}^{(l)}\}$ and $\{U_{k,n}\}$ freely with $U_{k,i} = U_i$ for $i \leq k$; obviously

$$U_{k,n}^{(l)} = U_{(k+1)2^{n-2}+l-1} \quad \text{for } n \in Z \setminus \{0,1\}, \quad l \in I_{(k+1)2^{n-2}}. \tag{3.16}$$

THEOREM 3.4. $S^{-}(U_{k,m}) = m$, $m \in Z$. *That is,*

$$S^{-}(U_i) = i, \qquad i \in I_k U \{0\},$$

$$S^{-}\left(U_{k,n}^{(l)}\right) = (k+1)2^{n-2} + l - 1, \qquad n \in Z \setminus \{0,1\}, \qquad l \in I_{(k+1)2^{n-2}}.$$

Now we begin to consider the convergence properties. The Fourier series of a given function F in terms of the functions $U_{k,i}$ is

$$F(x) \sim \sum_{i=0}^{\infty} \alpha_i U_{k,i}(x) \tag{3.17}$$

with

$$\alpha_i := \langle F(x), U_{k,i}(x) \rangle. \tag{3.18}$$

Let

$$\mathcal{P}_n F := \sum_{i=0}^{n-1} \alpha_i U_{k,i}(x)$$

be the nth partial sum of the series (3.17).

Then $\mathcal{P}_n F$ is the best L_2-approximation to F from $M_n := \operatorname{span}(U_{k,i})_0^{n-1}$. Hence it is convergent to F if F is in L_2, since the union of the M_n's is dense in L_2. We get the following theorem.

THEOREM 3.5. *If $f \in L_2[0,1]$, then* $\lim_{n\to\infty}\|F-\mathcal{P}_n F\|_2=0$.

Next we will prove that $\mathcal{P}_{(k+1)2^{n-1}}F$ uniformly approximates $F\in C[0,1]$.

It is well known [2] that

$$\left\|F-\mathcal{P}_{(k+1)2^{n-1}}F\right\|_\infty \le \left(1+\left\|\mathcal{P}_{(k+1)2^{n-1}}\right\|\right)\operatorname{dist}_\infty\left(F, M_{(k+1)2^{n-1}}\right),$$

and we know

$$\left\|\mathcal{P}_{(k+1)2^{n-1}}\right\| = \|\mathcal{P}_k\| < \infty,$$

since the least-square approximation for $M_{(k+1)2^{n-1}}=\mathbb{P}_{k+1,\Delta_n}$ is local and $C[0,1]$ is in the closure of $\bigcup_{n=1}^{\infty} M_{(k+1)2^{n-1}}$. Therefore we have

THEOREM 3.6. *Let* $F\in C[0,1]$. $\mathcal{P}_{(k+1)2^{n-1}}$ *be an* L_2*-projector onto* $M_{(k+1)2^{n-1}}$ *on* $C[0,1]$. *Then*

$$\lim_{n\to\infty}\left\|F-\mathcal{P}_{(k+1)2^{n-1}}F\right\|_\infty=0.$$

The same kind of argument as in the proof of Theorem 2.6 shows that the following theorem holds.

THEOREM 3.7. *There exists a continuous function* $f\in C[0,1]$ *whose expansion*

$$\sum_{i=0}^{n}\langle f(x), U_{k,i}(x)\rangle U_{k,i}$$

in terms of $\{U_{k,i}\}$ *does not converge to* f *uniformly when* $n\to\infty$.

Acknowledgment. We would like to thank Professor Carl de Boor for many helpful discussions and for his careful reading and correcting of the manuscript.

REFERENCES

[1] K. G. BEAUCHAMP, *Walsh Functions and Their Applications*, Academic Press, New York, 1975.

[2] C. DE BOOR, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.

[3] C. DE BOOR, MRC Tech. Summ. Rep. #1667, Mathematics Research Center, Univ. Wisconsin, Madison, 1976, p. 36.

[4] Z. CIESIELSKI, *Properties of the orthonormal Franklin system*, Studia Mathematica, 23 (1963), pp. 141–157, 27 (1966), pp. 289–323, *A bounded orthonormal system of polygonals*, 31 (1968), pp. 339–346.

[5] P. FRANKLIN, *A set of continuous orthogonal functions*, Math. Annal., 100 (1928), pp. 522–529.

[6] A. HAAR, *Zur Theorie der orthogonalen Funktionensysteme*, Math. Annal, 69 (1910), pp. 331–371.

[7] J. L. WALSH, *A closed set of normal orthogonal functions*, Amer. J. Math., 45 (1923), pp. 5–24.

[8] S. KARLIN, *Total Positivity*, Stanford Univ. Press, Stanford, CA, 1968.

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

SIAM J. MATH. ANAL.
Vol. 15, No. 4, July 1984

© 1984 Society for Industrial and Applied Mathematics

0017

TECHNIQUE FOR EVALUATING INDEFINITE INTEGRALS INVOLVING PRODUCTS OF CERTAIN SPECIAL FUNCTIONS*

JEAN C. PIQUETTE[†‡] AND A. L. VAN BUREN[†]

Abstract. A new technique is described for evaluating a general class of indefinite integrals involving products of many of the special functions of physics such as Bessel functions, Legendre functions, Hermite functions, etc. The technique is a generalization of the method used by Sonine to evaluate certain indefinite integrals of Bessel functions. It involves replacing the integral to be evaluated by a coupled set of linear, inhomogeneous differential equations. A particular solution of the set of differential equations is then sufficient to express the result of integration. Several examples are given to illustrate the technique.

1. Introduction. We present a technique of integration involving a special but very broad class of integrals. These are indefinite integrals of the general form

$$I=\int dx f(x)\prod_{i=1}^{m} R^{(i)}_{\mu_i}(x), \tag{1}$$

where $R^{(i)}_{\mu_i}(x)$ is the ith type of special function of order μ_i obeying the following set of recurrence relations:

$$R^{(i)}_{\mu+1}(x)=a_\mu(x)R^{(i)}_{\mu}(x)+b_\mu(x)R^{(i)}_{\mu-1}(x), \tag{2a}$$

$$DR^{(i)}_{\mu}(x)=c_\mu(x)R^{(i)}_{\mu}(x)+d_\mu(x)R^{(i)}_{\mu-1}(x). \tag{2b}$$

Here a_μ, b_μ, c_μ, and d_μ are known functions corresponding to $R^{(i)}_\mu$. The symbol D represents d/dx. The function $f(x)$ and the product $\prod R^{(i)}_\mu$ are both assumed bounded and continuous (or with at most a finite number of discontinuities) over an interval $[x_1, x_2]$, insuring that the integral I exists in the same interval.

Recurrence relations (2) may be combined to show that the functions $R^{(i)}_\mu$ satisfy the differential equation

$$\begin{aligned} D^2R^{(i)}_\mu &+\left[\frac{a_{\mu-1}d_{\mu-1}}{b_{\mu-1}}-c_\mu-c_{\mu-1}-\frac{1}{d_\mu}(Dd_\mu)\right]DR^{(i)}_\mu \\ &+\left[c_\mu c_{\mu-1}-Dc_\mu+\frac{c_\mu}{d_\mu}Dd_\mu-\frac{d_{\mu-1}}{b_{\mu-1}}(d_\mu+c_\mu a_{\mu-1})\right]R^{(i)}_\mu=0. \end{aligned} \tag{3}$$

Equation (3) is a special case of the Sturm-Liouville differential equation

$$D[\rho(x)D\psi(x)]+[S(x)+\gamma r(x)]\psi(x)=0, \tag{4}$$

where $r(x)=0$,

$$\rho(x)=\exp\left\{\int dx\left[\frac{a_{\mu-1}d_{\mu-1}}{b_{\mu-1}}-c_\mu-c_{\mu-1}-\frac{1}{d_\mu}Dd_\mu\right]\right\},$$

* Received by the editors May 26, 1981, and in revised form December 9, 1982.

[†] Naval Research Laboratory, Orlando, Florida 32856.

[‡] This paper is based on a part of this author's doctoral thesis for Stevens Institute of Technology, Hoboken, New Jersey.

$$S(x)=\rho(x)\left[c_{\mu-1}-Dc_\mu+\frac{c_\mu}{d_\mu}Dd_\mu-\frac{d_{\mu-1}}{b_{\mu-1}}(d_\mu+c_\mu a_{\mu-1})\right],$$

and $\psi(x)=R_\mu^{(i)}(x)$.

If either (2a) or (2b) is a two-term recurrence relation (i.e., if b_μ or d_μ is equal to zero for all μ), then the above expressions are undefined and $R_\mu^{(i)}$ does not satisfy the Sturm–Liouville differential equation. In this case $R_\mu^{(i)}$ satisfies instead a first-order differential equation and is in the form of an exponential. This may be readily seen by letting $d_\mu=0$ in (2b), in which case

$$R_\mu^{(i)}(x)=\exp\left[\int dx\, c_\mu\right]. \tag{5}$$

On the other hand, if $b_\mu=0$, we obtain from (2a):

$$R_\mu^{(i)}=a_{\mu-1}R_{\mu-1}^{(i)},$$

which when combined with (2b) yields

$$R_\mu^{(i)}(x)=\exp\left\{\int dx\left[c_\mu+\frac{d_\mu}{a_{\mu-1}}\right]\right\}. \tag{6}$$

An extensive search of the literature indicated that functions satisfying the recurrence relations (2) have not previously been named. For the purposes of this article we shall refer to them as birecurrent functions. Most of the special functions of physics fall into this category (including all Bessel functions, Legendre functions, Hermite functions, etc.). We exclude the special cases given in (5) and (6) from this category, preferring to call them exponential terms instead.

The integration technique presented in this article involves a generalization of the method (described by Watson [1]) used by Sonine [2] to evaluate certain indefinite integrals of Bessel functions. The integral to be evaluated in Sonine's method is replaced by a differential equation. A particular solution of the differential equation is then sufficient to express the result of integration. In the present work we generalize the method to include all functions obeying the relations of (2). In addition, we describe an approach for obtaining and solving the appropriate differential equations.

2. The technique. We assume the integral of (1) may be expressed in the form

$$I=\sum_{p_1=0}^{1}\sum_{p_2=0}^{1}\cdots\sum_{p_m=0}^{1}A_{p_1,p_2,\cdots,p_m}(x)\prod_{i=1}^{m}R_{\mu_i+p_i}^{(i)}, \tag{7}$$

where the 2^m coefficients $A_{p_1,p_2,\cdots,p_m}(x)$ are functions to be determined. For convenience, we will represent the multiple summation and the coefficients in (7) by the shorthand notations $\Sigma_{\{p\}}$ and A_p, respectively. In order to determine the functions A_p, we differentiate (7), substitute for DI the integrand from (1), and obtain

$$f(x)\prod_{i=1}^{m}R_{\mu_i}^{(i)}=\sum_{\{p\}}\left[A_pD\prod_{i=1}^{m}R_{\mu_i+p_i}^{(i)}+(DA_p)\prod_{i=1}^{m}R_{\mu_i+p_i}^{(i)}\right]. \tag{8}$$

Due to the recurrence relations (2), it is always possible to express the first sum on the right-hand side of (8) in the form

$$\sum_{\{p\}} A_p \sum_{q_1=0}^{1} \sum_{q_2=0}^{1} \cdots \sum_{q_m=0}^{1} B_{qp} \prod_{i=1}^{1} R^{(i)}_{\mu_i+q_i},$$

or

$$\sum_{\{p\}} \sum_{\{q\}} B_{pq} A_q \prod_{i=1}^{m} R^{(i)}_{\mu_i+p_i}, \tag{9}$$

where the 2^{2m} coefficients $B_{pq} \equiv B_{p_1,p_2,\cdots,p_m\, q_1,q_2,\cdots,q_m}(x)$ are known functions resulting from repeated applications of the relations of (2) and the regrouping of terms in the form $\prod_{i=1}^{m} R^{(i)}_{\mu_i+p_i}$.

Using (9) we can rewrite (8) to obtain

$$f(x) \prod_{i=1}^{m} R^{(i)}_{\mu_i} = \sum_{\{p\}} \left[DA_p + \sum_{\{q\}} B_{pq} A_q \right] \prod_{i=1}^{m} R^{(i)}_{\mu_i+p_i}. \tag{10}$$

We can now obtain a coupled set of differential equations for the functions A_p by imposing the sufficient condition that the coefficients of like special functions on each side of (10) be equal. Doing this, we obtain the following coupled set of linear inhomogeneous differential equations of first order

$$f(x)\delta_{0,p} = DA_p + \sum_{\{q\}} B_{pq} A_q, \tag{11}$$

where δ is a Kronecker delta defined equal to zero unless $p_1 = p_2 = \cdots = p_m = 0$.

In solving the set (11) of 2^m equations in the 2^m unknown functions A_p, one normally proceeds by differentiation and algebraic manipulation to uncouple a particular function from the remainder. This results in a differential equation of order 2^m. A particular solution of this uncoupled equation involves a particular choice of 2^m constants. Since this is exactly the number of arbitrary constants that the original set (11) involves, one must be careful not to introduce any further arbitrary constants. In this case we obtain the remaining functions by expressing them in terms of derivatives of the initial function that has been calculated, rather than in terms of integrals of it. Regardless of the method used in obtaining a particular solution of (11), one must avoid introducing more than 2^m arbitrary constants. Otherwise, the solution so obtained will neither satisfy (11) nor provide, via (7), a proper representation of the integral of (1).

When the integrand of (1) contains more than one birecurrent function, it may be desirable to move one (or possibly more) of the birecurrent functions out of the product term and treat it as part of $f(x)$. Each birecurrent function appearing in the product term doubles the number of unknown coefficient functions A_p and hence doubles the number of coupled differential equations to be solved. Thus, we halve the number of differential equations each time we move a birecurrent function out of the product term and group it with $f(x)$. However, each function so grouped will appear in the inhomogeneous term of the final set of differential equations. Conversely, none of the birecurrent functions grouped in the product term will appear explicitly in the final set of differential equations.

Any particular solution of (11) will give a set of functions A_p that can be used in (7) to express the result of the integration. We can see this if we differentiate the expression that results by substitution of this particular set into (7). The resulting (8) is obviously satisfied since the A_p are a *particular* solution. The fact that only a particular, rather than a general, solution is required is a powerful aspect of the technique.

The coupled set (11) is a standard form of linear inhomogeneous differential equations of first order that may be solved by well-known methods. A particular solution of (11) is easier to obtain than one may suspect since each equation contains exactly one term involving the derivative of a particular function A_p and the derivative of each of the functions A_p appears in only one equation.

The technique described above is equally applicable, with slight modifications, when one or more of the birecurrent functions in the product term of (1) is replaced by an exponential term of the form of (5) or (6). Because the exponential terms satisfy two-term recurrence relations, we do not have to include $p=1$ terms for them in (7). This reduces the number of unknown coefficients A_p and the resulting coupled differential equations by a factor of 2. Moving the exponential term from the product term into $f(x)$ does not change the number of differential equations to be solved.

3. An example. To illustrate the technique, we obtain the result to the following well-known integral: $I=\int dx\, x \sin\mu x$. In this case, $f(x)=x$ and $R_\mu(x)=\sin\mu x$.

The simplest way to apply the technique to this problem is to consider $\sin\mu x$ as the imaginary part of the exponential $R_\mu(x)=\exp(i\mu x)$ and assume that $I=A(x)\exp(i\mu x)$. Only one term, and hence only one unknown coefficient $A(x)$, is required in I in this case because differentiation of the exponential does not produce new functions (i.e., recurrence relation (2b) reduces to a two-term equation relating DR_μ to R_μ).

We shall use instead a somewhat more complicated approach that requires two unknown coefficients in order to illustrate several important aspects of the technique. This approach is based on the fact that the set of two functions $\sin\mu x$ and $\cos\mu x$ is closed under differentiation so that it is convenient to choose $R_\mu=\sin\mu x$ and $R_{\mu+1}=\cos\mu x$. We now proceed to obtain the integral following a step-by-step procedure:

a. We assume I may be expressed in the form

$$I=A_0(x)\sin\mu x+A_1(x)\cos\mu x. \tag{12}$$

b. Differentiation produces $DI=\mu A_0(x)\cos\mu x+[DA_0(x)]\sin\mu x-\mu A_1(x)\sin\mu x+[DA_1(x)]\cos\mu x$.

c. Equating DI to the integrand $x\sin\mu x$ and separately equating coefficients of $\sin\mu x$ and $\cos\mu x$, we obtain the following differential equations:

$$x=DA_0-\mu A_1, \tag{13a}$$

$$0=\mu A_0+DA_1. \tag{13b}$$

d. We now uncouple A_0 and A_1 by substituting into (13a) the expression for A_0 obtained from (13b). This gives

$$D^2A_1+\mu^2A_1=-\mu x. \tag{14}$$

e. A particular solution of (14) is

$$A_1(x)=-\frac{x}{\mu}. \tag{15}$$

Substitution of (15) into (13b) yields

$$A_0(x)=\frac{1}{\mu^2}. \tag{16}$$

f. Substitution of (15) and (16) into (12) gives the result

$$\int dx\, x\sin\mu x=\frac{\sin\mu x-\mu x\cos\mu x}{\mu^2}.$$

In order to investigate the consequences of using a different particular solution from the one chosen, we first obtain the general solution of (14). It is

$$A_1(x)=C_1\sin\mu x+C_2\cos\mu x-\frac{x}{\mu}, \tag{17}$$

where C_1 and C_2 are arbitary constants. Substitution of (17) into (13b) yields

$$A_0(x)=-C_1\cos\mu x+C_2\sin\mu x+\frac{1}{\mu^2}. \tag{18}$$

Substitution of these general solutions for A_0 and A_1 into (12) produces the following expression for the desired integral

$$\int dx\, x\sin\mu x=\frac{\sin\mu x-\mu x\cos\mu x}{\mu^2}+C_2. \tag{19}$$

Equation (19) involves only a single arbitrary constant to which the indefinite integral under consideration is entitled. Since the completely general solution to (14) was used in obtaining (19), it is clear that any particular solution to (14) would have sufficed, with only the constant C_2 in (19) being affected by a different choice.

4. A more complicated integrand. Although the first example provides a succinct illustration of the present integration technique, an integration-by-parts approach would certainly have been more straightforward. We present a second example that is less susceptible to standard techniques. This integral, which arose in a problem involving the scattering of sound by sound, is

$$I=\int dr\, r^{\mu}Z_{\nu}(r)\exp(ir), \tag{20}$$

where $Z_\nu(r)$ is an arbitrary Bessel function of one of the first three kinds of real argument r, and the range of integration is restricted to $r>0$. For generality, both the order ν and the exponent μ are chosen to be complex.

To begin the technique, we assume that I may be written as

$$I=A_0(r)Z_\nu(r)+A_1(r)Z_{\nu+1}(r). \tag{21}$$

Differentiating (21), expressing the resulting Bessel function derivatives in terms of Z_ν and $Z_{\nu+1}$ by use of the appropriate recurrence relations, equating the result to the integrand of (20), and imposing the sufficient condition that coefficients of like-order Bessel functions be equal, we obtain the following coupled set of differential equations:

$$\begin{aligned}0&=A_0-DA_1+\left[\frac{\nu+1}{r}\right]A_1,\\ r^{\mu}\exp(ir)&=DA_0+\frac{\nu}{r}A_0+A_1.\end{aligned} \tag{22}$$

These equations may be uncoupled to yield

$$D^2A_1-\frac{1}{r}DA_1+\left[1+\frac{1-\nu^2}{r^2}\right]A_1=r^\mu\exp(ir). \tag{23}$$

We now define a function $\rho(r)$ such that

$$A_1(r)=r\rho(r). \tag{24}$$

Substitution into (23) yields

$$r^2D^2\rho+rD\rho+[r^2-\nu^2]\rho=r^{\mu+1}\exp(ir). \tag{25}$$

One particular solution to (25) is an associated Bessel function as defined by Luke [3], namely, $\rho(r)=i^{\mu+1}H_{\mu,\nu}(-ir)$, so that

$$A_1(r)=i^{\mu+1}rH_{\mu,\nu}(-ir). \tag{26}$$

Use of the known properties of the associated Bessel functions results in

$$A_0(r)=i^{\mu+1}r\frac{[(\mu-\nu)(\mu-\nu-1)H_{\mu-1,\nu+1}(-ir)+(-ir)^\mu\exp(ir)]}{2\mu+1}, \tag{27}$$

when the expression for A_1 given by (26) is substituted into the first of (22). We thus have

$$\begin{aligned}\int dr\,r^\mu Z_\nu(r)\exp(ir) &= i^{\mu+1}r\left\{\frac{[(\mu-\nu)(\mu-\nu-1)H_{\mu-1,\nu+1}(-ir)+(-ir)^\mu\exp(ir)]}{2\mu+1}\right\}Z_\nu(r)\\ &\quad+i^{\mu+1}rH_{\mu,\nu}(-ir)Z_{\nu+1}(r).\end{aligned} \tag{28}$$

This result is identical to that obtained by Luke [4] using a specialized integration technique developed by McLachlan and Meyers [5] for certain integrals involving Bessel and Struve functions. Luke [3] provides formulas by which the associated Bessel functions appearing in (28) may be evaluated.

The above example resulted in a differential equation that was recognizable. However, the technique is still applicable even if no previously known solution to the differential equation exists. We first try to obtain a particular solution to our inhomogeneous differential equation by using standard methods (see, e.g. [6]) such as the method of Lagrange or the method by Cauchy. If none of these methods proves satisfactory, we can always obtain a solution in the form of an infinite series. As an example of this, we again return to (23) and assume that A_1 may be expressed as

$$A_1=\exp(ir)\sum_{m=-\infty}^{\infty}B_mr^{m+\mu}, \tag{29}$$

where the coefficients B_m are constants to be determined. To simplify subsequent calculations, we chose the form of the expansion to be compatible with the inhomogeneous term. If (29) is substituted into (23) and coefficients of like powers of r are equated, the following recursion relation is obtained for the B_m:

$$(m+\mu+\nu)(m+\mu-\nu)B_{m+1}+i[2(m+\mu)-1]B_m=\delta_{m,1}. \tag{30}$$

We can obtain a particular solution to (30) and hence to (23) by setting $B_m=0$ for $m>1$ and solving for the nonzero coefficients B_m, $m=1,0,-1,\cdots$, The resulting descending power series representation for A_1 can be expressed in terms of a hypergeometric function as follows:

$$A_1=i^{\mu+1}r\left[(-1)^{\mu+1}(ir)^{\mu}\exp(ir)\frac{{}_3F_1(1,-\mu+\nu,-\mu-\nu;\frac{1}{2}-\mu;\frac{1}{2}ir)}{2\mu+1}\right]. \tag{31}$$

In view of (26), it is not surprising that the quantity in brackets is identical to the series representation of $H_{\mu,\nu}(-ir)$ given by Luke [3]. The solution (31) is not defined if $\mu=-\frac{1}{2}$. It is also not defined if μ is an odd multiple of $\frac{1}{2}$ unless both $\mu\pm\nu$ are positive integers or zero. The solution is a terminating series if either $\mu+\nu$ or $\mu-\nu$ is a positive integer. The infinite series obtained otherwise is an asymptotic representation of A_1 that is valid for $r\to\infty$. Although this series is then divergent, it is useful nonetheless provided it is truncated properly.

We can obtain a second particular solution for A_1 by setting $B_m=0$ for $m\leq 1$ and solving (30) for B_m, $m=2,3,\cdots$. The hypergeometric representation of the resulting ascending power series in r is

$$A_1=i^{\mu+1}r\left\{(-ir)^{\mu+1}\frac{{}_2F_2(1,\mu+\frac{3}{2};\mu-\nu+2,\mu+\nu+2;-2ir)}{(\mu-\nu+1)(\mu+\nu+1)}\right\}. \tag{32}$$

The quantity appearing in braces in (32) is the series representation of the associated Bessel function $h_{\mu,\nu}(-ir)$ as defined by Luke [3]. This function is a second particular solution to the differential equation satisfied by $H_{\mu,\nu}$ and hence can be used in place of H in the solution (28) to the original integral.

The solution (32) is a terminating series if μ is a positive odd multiple of $-\frac{1}{2}$ (other than $-\frac{1}{2}$) and if both $\mu\pm\nu$ are not positive integers. It is not defined if either $\mu+\nu$ or $\mu-\nu$ is a negative integer and μ is not a positive odd multiple of $-\frac{1}{2}$ (other than $-\frac{1}{2}$).

5. Additional illustrative examples. The two previous examples illustrate the power and versatility of the current integration technique, but two objections may be raised: The first example can be handled trivially, and the second example is one involving products of Bessel functions and, hence, is amenable to the original approach proposed by Sonine. The examples that follow will serve to illustrate the applicability of the current technique to integrals that are not of the Bessel function type. Although some of these may be solved by standard techniques, they nonetheless illustrate the broad range of integrands that can successfully be handled via this technique (and to the authors' knowledge, several of these integrals have not been previously tabulated).

a. *Some integrals involving Legendre functions.* We now consider some examples of integrals of the general form

$$I=\int dx\,P_\nu(x)f(x), \tag{33}$$

where $P_\nu(x)$ is the Legendre function of order ν and $f(x)$ has the same meaning as in (1). We assume the integral I may be represented in the usual way as $A(x)P_\nu(x)+B(x)P_{\nu+1}(x)$. Following the procedure outlined in §2, we obtain two coupled equations for A and B which can be uncoupled to yield

$$A(x)=-xB(x)+\frac{(1-x^2)}{(\nu+1)}B'(x), \tag{34}$$

where

$$(1-x^2)B''(x)-2xB'(x)+\nu(\nu+1)B(x)=(\nu+1)f(x). \tag{35}$$

As a first example of an integral of the form (33), we consider the case where $f(x)=1$. A particular solution to (35) for this case is $B(x)=(1/\nu)$. Equation (34) now determines $A(x)$ to be $A(x)=-(x/\nu)$. Substitution of A and B into the representation for I now produces

$$\int dx\, P_\nu(x)=-\frac{x}{\nu}P_\nu+\frac{1}{\nu}P_{\nu+1}, \qquad \nu\neq 0. \tag{36}$$

We next consider a more challenging integrand (i.e., one which cannot be handled by direct manipulation of the recurrence relations for $P_\nu(x)$). Let $f(x)$ be $\ln(1\pm x)$. In obtaining a particular solution to (35) we use the inhomogeneous term as a guide and assume $B(x)=K_1\ln(1\pm x)+K_2$, where K_1 and K_2 are undetermined constants. Direct substitution into (35) gives $K_1=1/\nu$ and $K_2=1/[\nu^2(\nu+1)]$. Equation (34) may next be used to show that $A(x)=\{-(x/\nu)\ln(1\pm x)\pm 1/[\nu(\nu+1)]-x/\nu^2\}$. Substitution of A and B into the representation for I results in

$$\begin{aligned}\int dx \ln(1\pm x)P_\nu(x)=&\left[-\frac{x}{\nu}\ln(1\pm x)\pm\frac{1}{\nu(\nu+1)}-\frac{x}{\nu^2}\right]P_\nu(x)\\&+\left[\frac{1}{\nu}\ln(1\pm x)+\frac{1}{\nu^2(\nu+1)}\right]P_{\nu+1}(x), \qquad \nu\neq 0,-1.\end{aligned} \tag{37}$$

b. *Some integrals involving Hermite functions.* We next consider integrals of the general form

$$I=\int dx\, H_\nu(x)f(x), \tag{38}$$

where $H_\nu(x)$ is the Hermite function of order ν, and once again $f(x)$ has the same meaning as was used in connection with (1). We assume the integral of (38) may be represented as $A(x)H_\nu(x)+B(x)H_{\nu-1}(x)$. (Note that H_ν and $H_{\nu-1}$ are used to represent the integral as opposed to H_ν and $H_{\nu+1}$. This difference is inconsequential. Any two orders separated by one integral value will be adequate to implement the procedure.) Omitting the details, we uncouple the resulting coupled set to obtain

$$A(x)=-\frac{x}{\nu}B(x)-\frac{B'(x)}{2\nu}, \tag{39}$$

where

$$2\nu f(x)=-B''(x)-2xB'(x)-2(\nu+1)B(x). \tag{40}$$

As a first example of this general form, we let $f(x)=e^{-x^2}$. (This is the usual weighting function used in the orthogonality integral for $H_\nu(x)$.) If we assume a particular solution of (40) of the form $B(x)=Ke^{-x^2}$ (with K an undetermined constant), direct substitution gives $K=-1$. Equation (39) gives $A(x)=0$ so that

$$\int dx\, e^{-x^2}H_\nu(x)=-e^{-x^2}H_{\nu-1}(x). \tag{41}$$

As a second example we let $f(x)=x^{-\mu}$, where for the moment μ is an arbitrary exponent. If we assume a particular solution of the form $B(x)=K_1x^{-K_2}$, where K_1 and K_2 are constants, direct substitution into (42) produces

$$2\nu x^{-\mu}=-K_1K_2(K_2+1)x^{-(K_2+2)}+2K_1[K_2-(\nu+1)]x^{-K_2}. \tag{42}$$

There are two sets of values for the constants K_1, K_2, and μ for which (44) is satisfied: $K_1=-2\nu/[(\nu+1)(\nu+2)]$, $K_2=\nu+1$, $\mu=\nu+3$; and $K_1=-\nu/(\nu+2)$, $K_2=-1$, $\mu=-1$. Using the first set of values, we obtain for $\nu\neq-1,-2$

(43)

$$\int dx\,H_\nu(x)x^{-(\nu+3)}=\left[\frac{2x^{-\nu}}{(\nu+1)(\nu+2)}-\frac{x^{-(\nu+2)}}{(\nu+2)}\right]H_\nu(x)-\frac{2\nu}{(\nu+1)(\nu+2)}x^{-(\nu+1)}H_{\nu-1}(x).$$

We obtain from the second set of values

$$\int dx\,xH_\nu(x)=\left(\frac{1+2x^2}{2(\nu+2)}\right)H_\nu(x)-\frac{\nu x}{(\nu+2)}H_{\nu-1}(x), \qquad \nu\neq-2. \tag{44}$$

For a final example involving Hermite functions, we let $f(x)=xe^{i\gamma x}$, where γ is a constant initially assumed to be arbitrary.

A particular solution to (40) can be obtained with $B(x)=Ke^{i\gamma x}$, where K is an unknown constant. Direct substitution shows that a solution exists for $\gamma=\sqrt{2(\nu+1)}$ and $K=i\nu/\sqrt{2(\nu+1)}$. Using the resulting solution for B in (39) to obtain A, we then have

$$\text{(45)}\quad \int dx\,xe^{i\sqrt{2(\nu+1)}x}H_\nu(x)$$

$$=e^{i\sqrt{2(\nu+1)}x}\left[\left(\frac{-ix}{\sqrt{2(\nu+1)}}+\frac{1}{2}\right)H_\nu(x)+\frac{i\nu}{\sqrt{2(\nu+1)}}H_{\nu-1}(x)\right], \qquad \nu\neq-1.$$

c. *Some examples involving Laguerre functions.* We now consider integrals of the general form

$$I=\int dx f(x)L_\nu(x), \tag{46}$$

where $L_\nu(x)$ is the Laguerre function of order ν, and $f(x)$ has the same meaning as in (1). As usual, we represent I in the form $A(x)L_\nu(x)+B(x)L_{\nu-1}(x)$ and obtain the following uncoupled equations

$$A(x)=\left(\frac{x}{\nu}-1\right)B(x)+\frac{x}{\nu}B'(x), \tag{47}$$

where

$$\nu f(x)=xB''(x)+(x+1)B'(x)+(\nu+1)B(x). \tag{48}$$

As an example of this case we let $f(x)=xe^{-(\nu+1)x}$. To obtain a particular solution of (48) we assume $B(x)=Ke^{-(\nu+1)x}$. Direct substitution gives $K=1/(\nu+1)$. Using this value yields the integral

$$(49)\quad \int dx\, x e^{-(\nu+1)x} L_\nu(x) = \frac{e^{-(\nu+1)x}}{(\nu+1)}\left[-(1+x)L_\nu(x)+L_{\nu-1}(x)\right], \qquad \nu\neq -1.$$

As a final example, we let $f(x)=x(1+x)^{-(\nu+3)}$. A particular solution to (48) can then be obtained assuming $B(x)=K(1+x)^{-(\nu+1)}$, where again K is an unknown constant. Direct substitution into (48) yields $K=\nu/[(\nu+1)(\nu+2)]$ so that

$$(50)\quad \int dx\, x(1+x)^{-(\nu+3)}L_\nu(x) = \frac{(1+x)^{-(\nu+1)}}{(\nu+2)}\left[\left(\frac{x-\nu}{\nu+1}-\frac{x}{1+x}\right)L_\nu(x)+\left(\frac{\nu}{\nu+1}\right)L_{\nu-1}(x)\right], \qquad \nu\neq -1,-2.$$

d. *Some final results.* The following is a tabulation of some additional results obtained using the integration technique in this paper. For the sake of brevity the derivations have been omitted.

$$(51)\quad \int dx\, x e^{-x^2} H_\nu(x)H_\mu(x) = e^{-x^2}\left\{\frac{(\mu+\nu+1)}{2[(\mu-\nu)^2-1]}H_\mu(x)H_\nu(x)+\left(\frac{\mu}{\nu-\mu+1}\right)H_\nu(x)H_{\mu-1}(x) + \left(\frac{\nu}{\mu-\nu+1}\right)H_{\nu-1}(x)H_\mu(x)+\left[\frac{2\mu\nu}{(\mu-\nu)^2-1}\right]H_{\nu-1}(x)H_{\mu-1}(x)\right\}$$

where H is a Hermite function and $\mu-\nu\neq\pm 1$.

$$(52)\quad \int dx[P_{1/2}(x)]^2=\frac{x}{2}\left\{[P_{1/2}(x)]^2+[P_{-1/2}(x)]^2\right\}-P_{1/2}(x)P_{-1/2}(x),$$

$$(53)\quad \left[\frac{\mu(\mu+1)-\nu(\nu+1)}{(\nu+1)}\right]\int dx\, P_\nu(x)x^\mu = \frac{\mu(\mu-1)}{(\nu+1)}\int dx\, P_\nu(x)x^{\mu-2}-x^\mu P_{\nu+1}(x) + \left[x^{\mu+1}-\frac{\mu(1-x^2)x^{\mu-1}}{(\nu+1)}\right]P_\nu(x), \qquad \nu\neq -1.$$

Acknowledgments. The authors wish to express their appreciation to Dr. P. H. Rogers of the Office of Naval Research, Dr. W. J. Wolfe of Martin-Marietta, Denver, and Dr. P. Gardner for valuable discussions during the course of this investigation.

REFERENCES

[1] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Cambridge Univ. Press, London, 1966, pp. 132–134.

[2] N. J. SONINE, Math. Ann., XVI, (1880), pp. 1–80.

[3] Y. L. LUKE, *Integrals of Bessel Functions*, McGraw-Hill, New York, 1962.

[4] ______, *An associated Bessel function*, J. Math. Phys., 31 (1952), pp. 131–138.

[5] N. W. MCLACHLAN AND A. L. MEYERS, *Integrals involving Bessel and Stuve functions*, Philos. Mag., 21 (1936), pp. 437–448.

[6] See e.g., G. M. MURPHY, *Ordinary Differential Equations and Their Solutions*, D. Van Nostrand, Princeton, NJ, 1960.

[7] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1972, p. 557 Eq. 15.2.6 and p. 558 Eq. 15.2.27.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics
001

HETEROCLINIC PHENOMENA IN THE ISOSCELES THREE-BODY PROBLEM*

R. MOECKEL[†]

Abstract. When two of the three particles have equal masses, the three-dimensional three-body problem has a subsystem consisting of motions for which the configuration of the particles is always an isosceles triangle. This subsystem has only two degrees of freedom. Geometrical methods are used to construct an invariant set containing a variety of periodic orbits which exhibit close approaches to triple collision and wild changes of configuration. Furthermore, orbits heteroclinic between these periodic orbits as well as oscillation and capture orbits are found. The whole invariant set is described using symbolic dynamics.

Introduction. The three-body problem in $\mathbb{R}^3$ is a dynamical system with nine degrees of freedom. By making use of the ten well-known constants of motion it can be reduced to a problem with four degrees of freedom. In the special case where two of the three masses are equal there is an invariant subset of the phase space consisting of motions for which the positions and velocities of these two particles remain symmetric about an axis in $\mathbb{R}^3$. This subsystem, which is called the isosceles three-body problem, can be reduced to only two degrees of freedom.

Two special cases of this problem have received considerable attention. The limiting case obtained as the third mass tends to zero is known as Sitnikov's problem [10], [12]. In this case, orbits can be found which tend parabolically (limiting velocity zero) to infinity in both time directions. Near such a homoclinic orbit lies an invariant set, described by the methods of symbolic dynamics, containing among other things oscillation and capture orbits. The case of zero angular momentum has also been well studied recently in connection with the problem of triple collision [2], [5], [6], [9], [11]. In this case the motion takes place in a fixed plane. Orbits occur which both begin and end in triple collision and near these homoclinic orbits one finds an invariant set whose orbits pass repeatedly through a neighborhood of the singularity, approaching arbitrarily near without actually colliding. These orbits are also constructed by means of symbolic dynamics.

The goal of this paper is to study the case of small but nonzero angular momentum as a perturbation of the zero angular momentum case. It is an important fact that in the case of nonzero angular momentum, triple collision is impossible. However, there are interesting invariant sets whose orbits repeatedly pass as close to collision as their angular momenta allow.

The usual method of treating triple collision [7] is to first set the angular momentum to zero and then introduce rescalings whose effect is to extend the vectorfield to a limiting "collision manifold" which forms a boundary to the noncompact zero angular momentum manifold. The orbits on this collision manifold are limits of orbits with zero angular momentum which pass close to collision. A different limiting object is obtained if we take limits of orbits in the phase space whose angular momenta tend to zero and which are near triple collision. Of course the usual triple collision manifold is a subset of this new manifold, but we find much more. We get not just a boundary for the zero angular momentum space but rather a manifold of the same dimension as the zero

*Received by the editors October 26, 1982, and in revised form September 7, 1983.

[†]School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. This work was supported in part by the National Science Foundation under grant MCS 82-00765.

angular momentum space itself. In Fig. 4 the surface represents the triple collision manifold, the exterior represents the zero angular momentum manifold, and the interior represents the new part of the limiting set. Intuitively, the orbits inside describe the behavior exhibited by small angular momentum orbits close to collision, but not exhibited by zero angular momentum orbits close to collision.

The formulation of our main result (Proposition 4.4) involves the notion of a connection graph (see for example (5.1)). We view the behavior of an orbit which repeatedly approaches triple collision as being made up of a sequence of behaviors close to collision, represented in the graph by upward pointing arrows, and behaviors between successive close approaches, represented by downward pointing arrows. The main result states that any possible sequence of behaviors that can be imagined actually occurs for some orbit. To describe some of the implications of this we need to understand what the possible behaviors close to collision and between close approaches might be.

In the zero angular momentum case there are three special orbits with the property that the triangle formed by the three bodies never changes shape. The three possible shapes are collinear (along an axis perpendicular to the axis of symmetry) and equilateral in either orientation. The orbits begin with a triple collision, expand to some maximum size and then contract to another triple collision. If we add a small amount of angular momentum the behavior between close approaches will be virtually the same. When the particles are very close, however, their behavior is entirely different than that of the zero angular momentum orbits; they will spin around the axis of symmetry and avoid collision. The exact behavior during this close approach is not well understood.

The special "homothetic" behaviors described above are represented by downward arrows in the connection graph while each method of spinning around to avoid collision is represented by an upward arrow. Other downward arrows represent the following behavior between close approaches: the size of the triangle becomes extremely large with the middle particle travelling one direction along the axis of symmetry and the outer particles travelling the other way while spinning, close together, around the axis. The limit of this type of behavior is that the particles go away and never return for another close approach.

With these possibilities in mind we can use Proposition 4.4 to construct orbits which exhibit striking changes of shape. For example, if any sequence of the three homothetic behaviors is specified, an orbit can be found which has infinitely many close approaches to collision exhibiting the required sequence of shapes between approaches. The behavior is indistinguishable from the homothetic behavior except while the particles are extremely close together. Furthermore we can find orbits which exhibit such remarkable changes of configuration and then abruptly escape to infinity, or else very nearly escape and then return to resume the sequence of close approaches. In summary, the orbits near collision, those near infinity and the orbits heteroclinic between them are combined in one symbolic dynamical description.

1. The isosceles three-body problem. The three-body problem in $\mathbb{R}^3$ concerns the motion of three point particles with masses m_k, positions q_k and momenta p_k; $k=1,2,3$. Suppose $m_1=m_2=m$. Then there is an invariant subset of the phase space consisting of motions for which m_3 remains on the z-axis in $\mathbb{R}^3$, while m_1 and m_2 remain symmetric with respect to this axis. If we require that the center of mass remains at the origin then we obtain a dynamical system with three degrees of freedom; the positions and velocities of all three particles can be found once those of m_1 are known.

Let $q_1=(x,y,z)\in\mathbb{R}^3$ and $\dot q_1=(\dot x,\dot y,\dot z)$. Then by symmetry $q_2=(-x,-y,z)$ and $\dot q_2=(-\dot x,-\dot y,\dot z)$. Since the center of mass remains at the origin we find $2mz+m_3z_3=0$ and $2m\dot z+m_3\dot z_3=0$, where $q_3=(0,0,z_3)$ and $\dot q_3=(0,0,\dot z_3)$. The kinetic energy T and potential energy U are:

$$T=m(\dot x^2+\dot y^2)+m(1+2\alpha)\dot z^2,$$

$$U=\frac{1}{2}mm_3\Big[\alpha(x^2+y^2)^{-1/2}+4\big(x^2+y^2+(1+2\alpha)^2z^2\big)^{-1/2}\Big],$$

where $\alpha=m/m_3$ is the mass ratio.

Let M be the 3×3 matrix $\operatorname{diag}(m,m,m(1+2\alpha))$ and define

$$\xi=M^{1/2}\begin{pmatrix}x\\y\\z\end{pmatrix}\quad\text{and}\quad \eta=M^{1/2}\begin{pmatrix}\dot x\\\dot y\\\dot z\end{pmatrix}.$$

These new variables satisfy Hamilton's equations with Hamiltonian functions $H(\xi,\eta)=|\eta|^2-U(\xi)$ where

$$U(\xi)=\frac{1}{2}m^{3/2}m_3\Big[\alpha(\xi_1^2+\xi_3^2)^{-1/2}+4\big(\xi_1^2+\xi_2^2+(1+2\alpha)\xi_3^2\big)^{-1/2}\Big].$$

To study orbits near triple collision it is convenient to introduce new variables $r=|\xi|$, $s=r^{-1}\xi$, $z=r^{1/2}\eta$ and to multiply the resulting equations by $r^{3/2}$. The result is:

$$\begin{aligned}
r'&=(s\cdot z)r,\\
s'&=z-(s\cdot z)s,\\
z'&=\nabla U(s)+\frac{1}{2}(s\cdot z)z.
\end{aligned}\tag{1.1}$$

By definition, s is the "angular part" of the position coordinates and it satisfies $|s|=1$. Consequently we can introduce spherical coordinates using the formula $s=(s_1,s_2,s_3)=(\cos\theta\cos\varphi,\ \sin\theta\cos\varphi,\ \sin\varphi)$. Now the vectors

$$\begin{aligned}
u_1&=s,\\
u_2&=\frac{\partial s}{\partial\theta}=(-\sin\theta\cos\varphi,\ \cos\theta\cos\varphi,0),\\
u_3&=\frac{\partial s}{\partial\varphi}=(-\cos\theta\sin\varphi,\ -\sin\theta\sin\varphi,\ \cos\varphi)
\end{aligned}$$

form an orthogonal basis for $\mathbb{R}^3$. If we write $z=vu_1+w_2u_2+w_3u_3$ (where $v=z\cdot u_1=s\cdot z$, etc.) and use (1.1) to find equations for the variables $(r,\theta,\varphi,v,w_2,w_3)$ we get (eventually):

$$\begin{aligned}
r'&=vr,\\
\theta'&=w_2,\\
\varphi'&=w_3,\\
v'&=\frac{1}{2}v^2+w_2^2\cos\varphi+w_3^2U(\varphi),\\
w_2'&=-\frac{1}{2}vw_2+2\tan\varphi w_2w_3,\\
w_3'&=U'(\varphi)-\frac{1}{2}vw_3-w_2^2\cos^2\varphi\tan\varphi,
\end{aligned}$$

where $U(\varphi)=\frac{1}{2}m^{3/2}m_3[\alpha\sec\varphi+4(1+2\alpha\sin^2\varphi)^{-1/2}]$.

In deriving these equations one must express $\nabla U(s)$ in the new basis. We have $\nabla U(s)\cdot u_1=\nabla U(s)\cdot s=-U(s)$, by homogeneity of $U(s)$. Also $\nabla U(s)\cdot u_2=\nabla U(s)\cdot \partial s/\partial\theta=\partial U/\partial\theta=0$ and $\nabla U(s)\cdot u_3=\partial U/\partial\varphi=U'(\varphi)$. The fact that the potential energy is independent of θ is related to the conservation of angular momentum. The equation for w_2' does not involve U and one can easily show that $\omega=r^{1/2}w_2\cos^2\varphi$ is a constant of the motion. Using this fact, we can safely ignore the variables θ and w_2. We have the following system with only two degrees of freedom:

$$\begin{aligned}
&r'=vr,\\
&\varphi'=w_3,\\
&v'=\frac{1}{2}v^2+w_3^2+\omega^2r^{-1}\sec^2\varphi-U(\varphi),\\
&w_3'=U'(\varphi)-\frac{1}{2}vw_3-\omega^2r^{-1}\sec^2\varphi\tan\varphi.
\end{aligned}$$

If we restrict attention to orbits of fixed energy h we find

$$\frac{1}{2}\left(v^2+w_3^2+\omega^2r^{-1}\sec^2\varphi\right)-U(\varphi)=rh.$$

We consider only the case $h<0$.

We will now make some familiar regularizing transformations to eliminate the singularities at $\varphi=\pm\pi/2$. First replace the troublesome term $\omega^2r^{-1}\sec^2\varphi$ by $2rh+2U(\varphi)-v^2-w_3^2$. Then replace w_3 by $w=w_3\cos\varphi$ and multiply the resulting vectorfield by $\cos\varphi$. Using $'$ to denote differentiation with respect to this new parameter we find:

$$\begin{aligned}
&r'=vr\cos\varphi,\\
&\varphi'=w,\\
&v'=U(\varphi)\cos\varphi-\frac{1}{2}v^2\cos\varphi+2rh\cos\varphi,\\
&w'=U'(\varphi)\cos^2\varphi-\frac{1}{2}vw\cos\varphi\\
&\qquad-\left(2U(\varphi)+2rh-v^2\right)\sin\varphi\cos\varphi
\end{aligned}\tag{1.2}$$

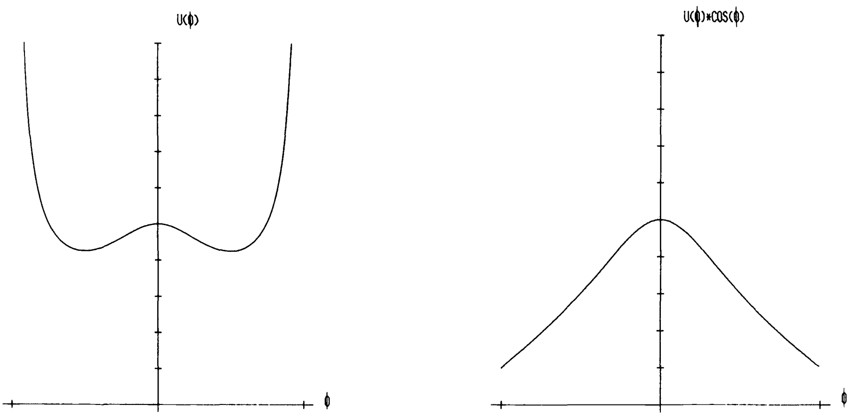


FIG. 1

with energy relation

$$\frac{1}{2}\left(v^2\cos^2\varphi+w^2+\omega^2r^{-1}\right)-U(\varphi)\cos^2\varphi=rh\cos^2\varphi. \tag{1.3}$$

The vectorfield (1.2) is analytic on $\mathbb{R}^4$ since the functions $U(\varphi)\cos\varphi$ and $U'(\varphi)\cos^2\varphi$ are analytic for all φ. Figure 1 shows the graphs of $U(\varphi)$ and $U(\varphi)\cos\varphi$.

2. The limiting variety. Equation (1.3) is a quadratic expression for $r(\varphi,v,w)$ on the manifold $\mathfrak{M}(h,\omega)$ consisting of orbits with energy h and angular momentum ω (we consider only $h<0$):

$$(2|h|\cos^2\varphi)r^2+\left(v^2\cos^2\varphi+w^2-2U(\varphi)\cos^2\varphi\right)r+\omega^2=0. \tag{2.1}$$

This equation has positive, real roots provided $\varphi\neq\pm\pi/2$ and

i) $2U(\varphi)\cos^2\varphi\geq v^2\cos^2\varphi+w^2$

and

ii) $(v^2\cos^2\varphi+w^2-2U(\varphi)\cos^2\varphi)^2\geq 8\omega^2|h|\cos^2\varphi$.

Taken together, these imply:

$$2U(\varphi)\cos^2\varphi-v^2\cos^2\varphi-w^2\geq\sqrt{8\omega^2|h|}\cos\varphi. \tag{2.2}$$

Inequality (2.2) should be viewed as defining the projection of $\mathfrak{M}(h,\omega)$ on (φ,v,w)-space. The manifold itself lies over its projection in two sheets provided strict inequality holds. These sheets join over the set where equality holds.

Note that (2.2) implies $U(\varphi)\cos\varphi\geq\sqrt{2\omega^2|h|}$. Now $U(\varphi)\cos\varphi=\frac{1}{2}m^{5/2}+O(\cos\varphi)$. Thus the projection of $\mathfrak{M}(h,\omega)$ extends from $-\pi/2$ to $\pi/2$ if and only if $m^5\geq 8\omega^2|h|$. For fixed m and h this will always hold for sufficiently small angular momenta. Figure 2 shows cross sections of the projection in this case. The projection is homeomorphic to a three-dimensional disc with two points removed and its boundary is a two-dimensional sphere with two points removed. In view of the way $\mathfrak{M}(h,\omega)$ lies over its projection we see that it is a union of two copies of the three-disc with deleted points joined together at the boundaries. A similar analysis for the case $m^5<8\omega^2|h|$ gives:

PROPOSITION 2.1. *For $m^5<8\omega^2|h|$, $\mathfrak{M}(h,\omega)$ is compact and is homeomorphic to S^3. For $m^5\geq 8\omega^2|h|$, $\mathfrak{M}(h,\omega)$ is noncompact and is homeomorphic to S^3 with two points deleted.*

In the limiting case $\omega=0$, (2.1) defines a reducible analytic variety instead of an analytic manifold. It is a union of the analytic manifolds $\{r=0\}$ and $\{2|h|r\cos^2\varphi+v^2\cos^2\varphi+w^2-2U(\varphi)\cos^2\varphi=0\}$. The only points with $r=0$ which are limits of points in $\mathfrak{M}(h,\omega)$ as $\omega\to 0$ are those satisfying (2.2) with $\omega=0$:

$$2U(\varphi)\cos^2\varphi\geq v^2\cos^2\varphi+w^2. \tag{2.3}$$

The only points in the second limiting manifold which are limits of the $\mathfrak{M}(h,\omega)$ are those with $r\geq 0$. Define

$$\begin{aligned}
&\mathfrak{M}_+=\{r\geq 0, v^2\cos^2\varphi+w^2-2U(\varphi)\cos^2\varphi=2rh\cos^2\varphi\},\\
&\mathfrak{M}_0=\{r=0, 2U(\varphi)\cos^2\varphi\geq v^2\cos^2\varphi+w^2\},\\
&\mathfrak{N}_0=\mathfrak{M}_+\cap\mathfrak{M}_0=\{r=0, 2U(\varphi)\cos^2\varphi=v^2\cos^2\varphi+w^2\},\\
&\mathfrak{M}=\mathfrak{M}_+\cup\mathfrak{M}_0.
\end{aligned}$$

We will refer to $\mathfrak{M}$ as the limiting variety of the $\mathfrak{M}(h,\omega)$ as $\omega\to 0$.

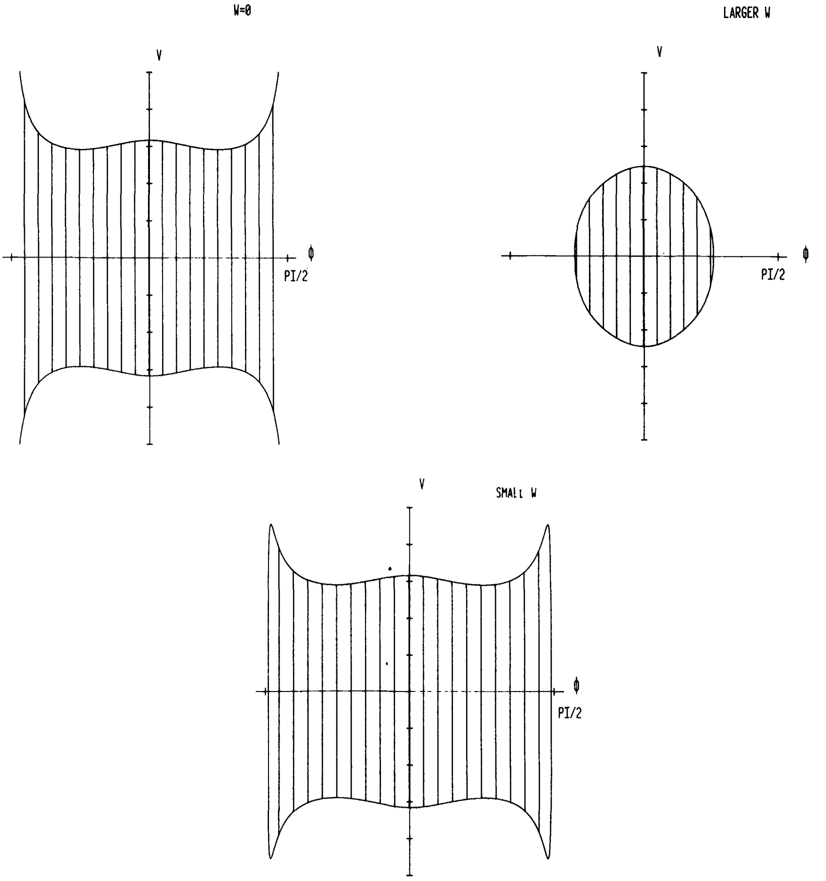


FIG. 2

$\mathcal{M}_0$ is the limit as $\omega \to 0$ of the projections of $\mathcal{M}(h,\omega)$ together with the lines $\{r=w=0,\ \varphi=\pm\pi/2\}$. Referring to Fig. 2 one finds that $\mathcal{M}_0$ is a three-disc with four deleted points. Its boundary, $\mathcal{N}_0$, is a two-sphere with four deleted points. Now $\mathcal{M}_+ \cap \{-\pi/2<\varphi<\pi/2\}$ projects homeomorphically to $\mathcal{M}_0 \cap \{-\pi/2<\varphi<\pi/2\}$ and so is a three-disc with two deleted points. But $\mathcal{M}_+$ also contains the half-planes $\{r\geq 0,\ w=0,\ \varphi=\pm\pi/2\}$. These fit together with the three-disc to produce a three-disc with two deleted arcs whose endpoints are the four deleted points in the boundary $\mathcal{N}_0$. Therefore $\mathcal{M}=\mathcal{M}_+ \cup \mathcal{M}_0$ is a three-sphere with two deleted points, just like the manifolds $\mathcal{M}(h,\omega)$. In both cases the deleted points are "at infinity", i.e., $r\to\infty$ as we approach them. Figure 3 shows in cross section the way $\mathcal{M}(h,\omega)$ converges to $\mathcal{M}$.

3. The flow on the limiting variety. The goal of this section is to locate certain invariant sets of the flow Φ defined by restricting (1.2) to $\mathcal{M}$ and to find "transverse" connections between these invariant sets. The question of how these features behave under perturbation to the nonzero angular momentum manifolds will be considered in §4.

We begin with the invariant surface $\mathcal{N}_0$. We have remarked that this is the so-called triple collision manifold of the $\omega=0$ isosceles problem. The flow on it has been well studied and most of the results of this section are direct consequences of previous work, mainly results in [2], [9], [11]. Many of the relevant features of this flow are summarized in Fig. 4. The flow is gradient-like with respect to the v coordinate, i.e.,

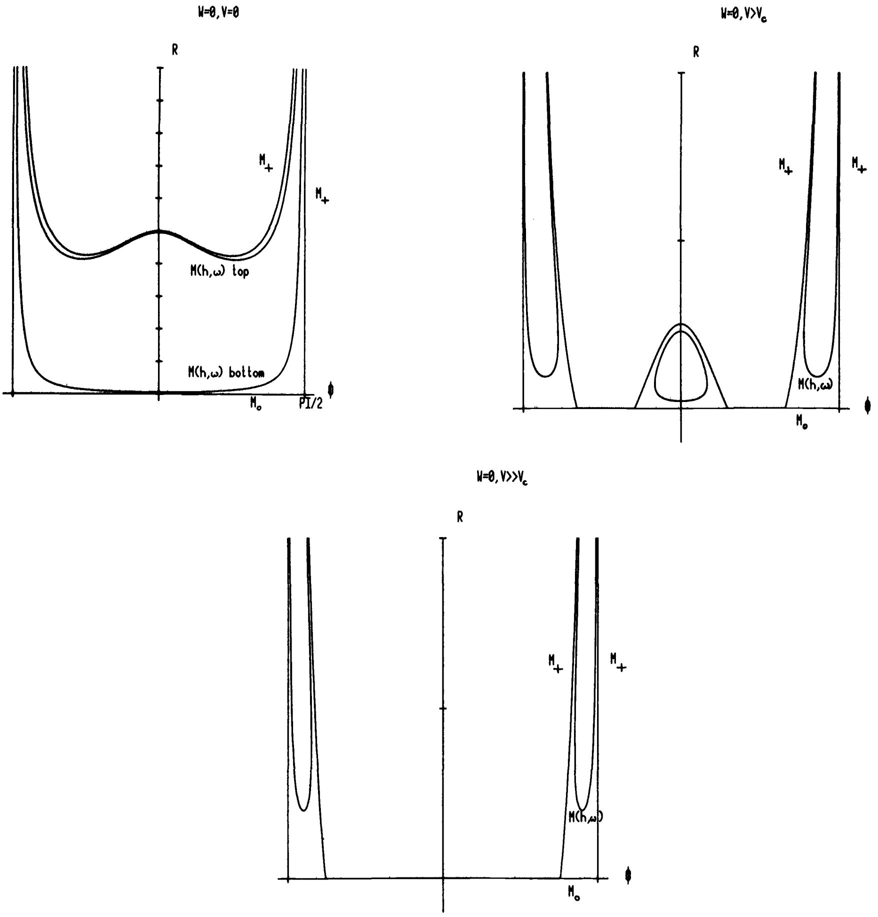

W=0,V=0
R
M+
M+
M(h,ω) top
M(h,ω) bottom
M₀
PI/2
φ
W=0,V>V_c
R
M+
M+
M(h,ω)
M₀
φ
W=0,V>>V_c
R
M+
M+
M(h,ω)
M₀
φ

FIG. 3

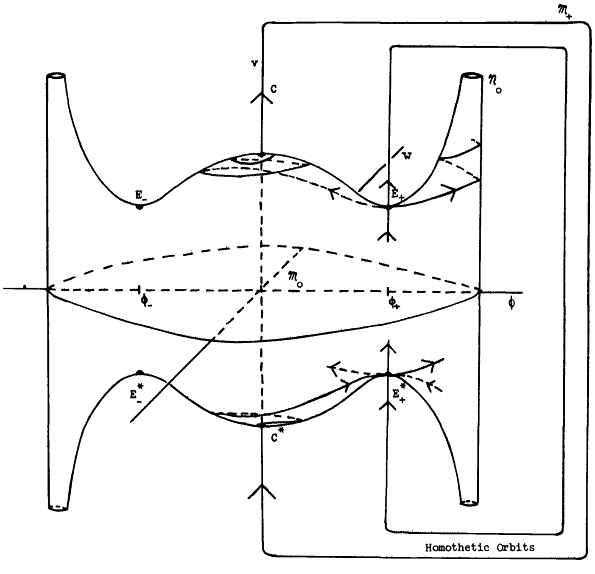

m+
v
C
η₀
w
E-
E+
m₀
φ-
φ+
φ
E*-
E*+
C*
Homothetic Orbits

FIG. 4′

v is strictly increasing on all solutions other than restpoints. There are six restpoints $(\varphi, v, w) = (\varphi_c, \pm v_c, 0)$, where $\varphi_c = 0$, φ_+ or φ_- is one of the three critical points of $U(\varphi)$ and $v_c = \sqrt{2U(\varphi_c)}$. The restpoints are denoted by C, C^*, $E_{+,-}$, $E^*_{+,-}$, the star indicating $v = -v_c$.

The configuration $\varphi = 0$ is the collinear one, while $\varphi = \varphi_{+,-}$ represents equilateral configurations with z_3 positive, negative. It is important to understand the local structure of these restpoints in the full four-dimensional domain of (1.2). To this end we examine the variational equations at $(r, v, \varphi, w) = (0, \pm v_c, \varphi_c, 0)$:

$$\begin{pmatrix} \delta r \\ \delta v \\ \delta\varphi \\ \delta w \end{pmatrix}' = \cos(\varphi_c) \begin{pmatrix} \pm v_c & 0 & 0 & 0 \\ 2h & \mp v_c & 0 & 0 \\ 0 & 0 & 0 & \sec(\varphi_c) \\ -2h\sin(\varphi_c) & 2v_c\sin(\varphi_c) & \cos(\varphi_c)U''(\varphi_c) & \mp\frac{1}{2}v_c \end{pmatrix} \begin{pmatrix} \delta r \\ \delta v \\ \delta\varphi \\ \delta w \end{pmatrix}.$$

The tangent space to $\mathfrak{N}_0$ at a restpoint is just the $(\delta\varphi, \delta w)$-plane and the eigenvalues of the restriction to this plane of the variational matrix are $\mp\frac{1}{4}v_c\cos\varphi_c \pm (\frac{1}{16}v_c^2\cos^2\varphi_c + U''(\varphi_c)\cos^2\varphi_c)^{1/2}$. Now $U''(0)$ is negative so the eigenvalues at C^* have positive real part while those at C have negative real part. A closer analysis of $U''(0)$ shows that these eigenvalues are real if and only if $\alpha \leq \frac{4}{55}$. At the equilateral configuration $U''(\varphi_{+,-})$ is positive so the restpoints $E_{+,-}$ and $E^*_{+,-}$ are all saddles when viewed in $\mathfrak{N}_0$. Evidently Fig. 4 depicts the flow for $\alpha > \frac{4}{55}$. Note also that the behavior of certain branches of the stable and unstable manifolds of the saddles is forced by the gradient-like structure.

The other two eigenvalues at the restpoints are $\pm v_c\cos\varphi_c$. It is easy to check that the $+v_c\cos\varphi_c$ eigenvector is tangent to $\mathfrak{N}_+$ while the $-v_c\cos\varphi_c$ eigenvector is tangent to $\mathfrak{N}_0$. Thus the restpoints $E^*_{+,-}$ viewed in $\mathfrak{N} = \mathfrak{N}_+ \cup \mathfrak{N}_0$ have two-dimensional stable manifolds lying in $\mathfrak{N}_+$ and two-dimensional unstable manifolds lying in $\mathfrak{N}_0$ (Fig. 5). The dimensions are the same for $E_{+,-}$ but this time the stable manifolds lie in $\mathfrak{N}_0$ and the unstable manifolds lie in $\mathfrak{N}_+$. The restpoint C^* has a one-dimensional stable manifold in $\mathfrak{N}_+$ and a three-dimensional unstable manifold $\mathfrak{N}_0$, while C has three-dimensional stable manifold in $\mathfrak{N}_+$ and one-dimensional unstable manifold in $\mathfrak{N}_0$.

We write $\mathrm{St}(p)$ and $\mathrm{Un}(p)$ to denote the stable and unstable manifolds of a restpoint p. The dimensional considerations above open the possibility of a rich network of heteroclinic connections among the equilateral restpoints. There may be transverse intersections $\mathrm{Un}(E^*_{+,-}) \cap \mathrm{St}(E_{+,-})$ in $\mathfrak{N}_0$ and transverse intersections $\mathrm{Un}(E_{+,-}) \cap \mathrm{St}(E^*_{+,-})$ in $\mathfrak{N}_+$. The collinear restpoints do not admit similar possibilities and we will concentrate on the equilateral restpoints from now on.

Besides the four equilateral restpoints there are two more landmarks in the limiting flow, namely the periodic orbits "at infinity". Equation (1.3) implies that as $r \to \infty$, $\varphi \to \pm\pi/2$. An orbit is said to approach infinity parabolically if $r \to \infty$ and $\dot{r} \to 0$ (recall that $\dot{}$ denotes differentiation in the original timescale). In [8], McGehee shows that the set of orbits tending parabolically to infinity as $t \to +\infty$ is a two-dimensional analytic submanifold of $\mathfrak{N}_+$ which can be viewed as the stable manifold of a (degenerately) hyperbolic periodic orbit at infinity. Similarly the set of orbits tending to infinity parabolically as $t \to -\infty$ form the unstable manifold of this orbit. We denote these

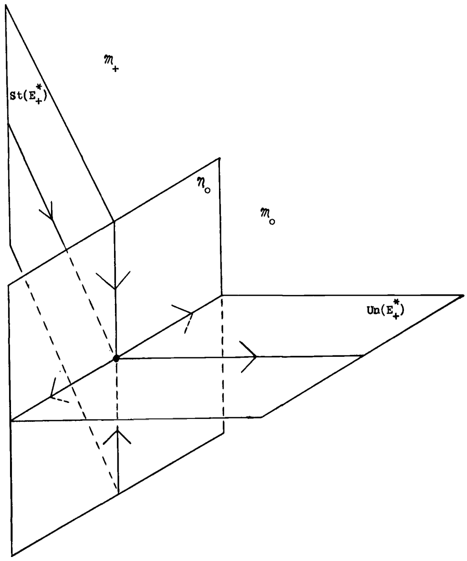


FIG. 5

limiting periodic orbits by ∞_+ or ∞_- accordingly as $\varphi = +\pi/2$ or $-\pi/2$. They should be viewed as invariant sets in another boundary manifold to $\mathcal{M}_+$. As the manifolds St($\infty_{+,-}$) and Un($\infty_{+,-}$) are all two-dimensional we can hope to include them in the network of transverse heteroclinic connections.

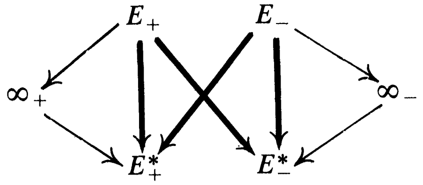
Before describing the connecting orbits any further we introduce a notion of transversality appropriate to analytic invariant manifolds of a flow. Let S_1 and S_2 be analytic two-dimensional invariant submanifolds of a flow on a three-dimensional manifold $\mathcal{M}$. We will say that S_1 and S_2 have an odd-order crossing along an orbit $\gamma \subset S_1 \cap S_2$ if in every local section Σ to the flow along γ, the analytic curves $S_1 \cap \Sigma$ and $S_2 \cap \Sigma$ have an odd-order crossing. Recall that two analytic curves in a plane $\sigma_1(t)$ and $\sigma_2(t)$ with $\sigma_1(0) = \sigma_2(0) = p$ are said to have an odd-order crossing at p if the series expansions agree up to but not including the terms of order $2n+1$ for some n. If S_1 and S_2 are stable and unstable manifolds of some invariant sets of the flow we call γ a transverse connecting orbit.

Whether or not connections exist between the invariant sets $E_{+,-}$, $E^*_{+,-}$, $\infty_{+,-}$ depends on the mass ratio α. We introduce the following device to keep track of these connections. By the connection graph for mass ratio α we will mean a directed graph with six vertices, labeled for the six distinguished invariant sets, and one directed edge for each transverse connecting orbit in the limiting flow Φ. Let $G(\alpha)$ denote this connection graph. The existence of connections can be inferred from what has been learned recently about the $\omega = 0$ case. The first result concerns connections in $\mathcal{M}_+$.

PROPOSITION 3.1. *For all* α, $G(\alpha)$ *contains the subgraph*

$$\begin{tikzcd}
 & E_+ \arrow[dl] \arrow[dd] & E_- \arrow[dr] \arrow[dd] & \\
\infty_+ \arrow[dr] & & & \infty_* \arrow[dl] \\
 & E^*_+ & E^*_- &
\end{tikzcd}$$

For all $\alpha > \frac{4}{55}$, $G(\alpha)$ contains the subgraph



(*Here bold arrows represent a countable infinity of distinct transverse connecting orbits.*)

Proof. All of the indicated connections occur in $\mathfrak{M}_+$. There is always a homothetic orbit which begins and ends in triple collision and this appears as the vertical arrow from E_+ to E_+^*. One checks easily that $(r(t), \varphi_{+,-}, v(t), 0)$ is a solution to (1.2) with configurations remaining equilateral, where

$$r(t) = \tfrac{1}{2}v_c^2|h|^{-1}\operatorname{sech}^2(-\tfrac{1}{2}v_c\cos\varphi_c t) \quad\text{and}\quad v(t) = -v_c\tanh(\tfrac{1}{2}v_c\cos\varphi_c t).$$

Clearly it connects $(0, \varphi_{+,-}, v_c, 0)$ to $(0, \varphi_{+,-}, -v_c, 0)$ as required. It is easily checked using variational equations that the intersection of $\mathrm{St}(E_{+,-}^*)$ and $\mathrm{Un}(E_{+,-})$ is transverse (even in the usual sense of the word).

We turn next to the connection from E_+ to ∞_+; the other connections in the top graph are handled similarly. Consider $\mathrm{Un}(E_+)$. We know that one orbit in $\mathrm{Un}(E_+)$ remains bounded for all time, namely the homothetic connecting orbit discussed above. It will be enough to show that there is another orbit in $\mathrm{Un}(E_+)$ which tends to infinity hyperbolically ($r \to \infty, \dot{r} > \varepsilon > 0$) with $\varphi \to +\pi/2$, for any connected manifold of orbits containing both bounded orbits and hyperbolic orbits must "cross" the manifold of parabolic orbits. As both manifolds are analytic we will get an odd-order crossing of $\mathrm{Un}(E_+)$ and $\mathrm{St}(\infty_+)$. Such a hyperbolic orbit in $\mathrm{Un}(E_+)$ is easily found near the branch of $\mathrm{Un}(E_+) \cap \mathfrak{M}_0$ which heads up the arm near $\varphi = \pi/2$ (Fig. 4). Following this branch we find orbits in $\mathrm{Un}(E_+)$ with v arbitrarily large and $r > 0$ but small. It is a result of McGehee that such orbits tend to infinity hyperbolically and we can even make the asymptotic value of $\dot{r}$ arbitrarily large [7].

The explanation of the lower graph which occurs for $\alpha > \frac{4}{55}$ lies in the eigenvalues at C and C^*. We have remarked that these are not real if $\alpha > \frac{4}{55}$. Furthermore there are connecting orbits in $\mathfrak{M}_0$ from $E_{+,-}$ to C and from C^* to $E_{+,-}^*$ (Fig. 4). Just as in the equilateral case there is a homothetic ($\varphi \equiv 0$) connection from C to C^* in $\mathfrak{M}_+$. Consider a neighborhood of this orbit as it crosses the section $v = 0$. Because of the connection from E_+ to C, $\mathrm{Un}(E_+)$ passes near C and follows along the homothetic connection. Because of the nonreal eigenvalues at C it meets the set $v = 0$ in a spiral about the point where the homothetic orbit hits the section (see Fig. 6; for more details consult [3], [9]). The symmetry of the isosceles problem can be used to show that the intersection of $\mathrm{Un}(E_-)$ with $\{v = 0\}$ is obtained by inversion of $\mathrm{Un}(E_+)$ through the origin and that then $\mathrm{St}(E_+^*)$ and $\mathrm{St}(E_-^*)$ are obtained from these by reflection through the φ axis. The result is a countable infinity of crossings of each of $\mathrm{St}(E_{+,-}^*)$ by each of $\mathrm{Un}(E_{+,-})$. Q.E.D.

We remark that among the infinitely many restpoint connections for $\alpha > \frac{4}{55}$ there are orbits which pass arbitrarily close to the collinear homothetic orbit.

The subgraphs of Proposition 3.1 contain no closed paths since connections from starred restpoints to unstarred ones cannot occur in $\mathfrak{M}_+$. The next result, however, concerns connections in $\mathfrak{M}_0$.

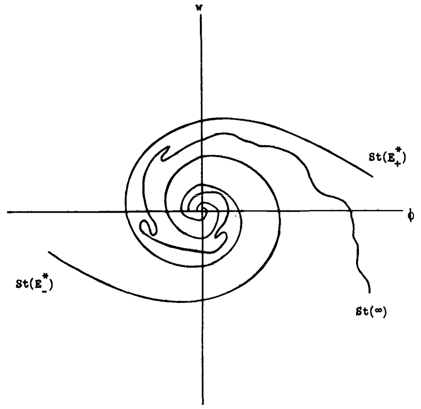


FIG. 6

PROPOSITION 3.2. *For all sufficiently small mass ratios* ($\alpha < \frac{4}{55} + \varepsilon$ *will do*), $G(\alpha)$ *contains the subgraph*

$$\begin{tikzcd} E_+ & E_- \\ E_+^* \arrow[u] & E_-^* \arrow[u] \end{tikzcd}.$$

For sufficiently large mass ratios, the following subgraph occurs:

$$\begin{tikzcd} E_+ & E_- \\ E_+^* \arrow[ur] & E_-^* \arrow[ul] \end{tikzcd}.$$

Proof. $\mathcal{M}_0$ is the region of (φ, v, w)-space inside $\mathcal{N}_0$. Viewed in $\mathcal{N}_0$, E_+^* is a saddle point, but in $\mathcal{M}_0$ there is an extra positive eigenvalue. We have remarked already that the flow on $\mathcal{N}_0$ is gradient-like with respect to v and in the interior we have $v' > 0$. So $\mathrm{Un}(E_+^*)$ can be followed up to the section $\{v=0\}$, where it forms an analytic curve with endpoints in $\mathcal{N}_0 \cap \{v=0\}$ (see Fig. 7). The location of these endpoints for various values of α is one of the key questions addressed in the studies of isosceles triple collision [2], [3], [9], [11]. Let p_+ denote the endpoint obtained by following the "front" branch of $\mathrm{Un}(E_+^*) \cap \mathcal{N}_0$ to the section and p_- denote the other one (Fig. 4). Figure 7 shows the case $\alpha \in (0, \frac{4}{55} + \varepsilon)$ for which it can be proved that p_+ lies in the second quadrant and p_- in the third [9]. Since $\mathrm{St}(E_+)$ is obtained from $\mathrm{Un}(E_+^*)$ by reflection through the φ-axis we get at least one odd-order crossing of $\mathrm{Un}(E_+^*) \cap \{v=0\}$ by $\mathrm{St}(E_+) \cap \{v=0\}$ (unless they are identical; this excludes at most a discrete set of mass ratios).

As $\alpha \to \infty$ one can prove that p_+ and p_- approach the positive and negative φ-axes respectively [11]. In this case the curve $\mathrm{Un}(E_+^*) \cap \{v=0\}$ necessarily crosses the w-axis and as $\mathrm{St}(E_-) \cap \{v=0\}$ is obtained by reflection in this axis we get at least one odd-order crossing of these curves. Q.E.D.

We remark that Simo's numerical work [11] indicates that the α-intervals implicit in Proposition 3.2 overlap, i.e., between the α values where only one or the other subgraph occurs, both subgraphs occur. In particular his work implies both subgraphs occur if $\alpha=1$ (three equal masses).

4. Perturbation of the limiting flow. Proposition 2.1 shows that the regularized manifolds $\mathfrak{M}(h,\omega)$ are homeomorphic to S^3 minus two points for all sufficiently small ω and the same is true for the limiting variety $\mathfrak{M}$. It is not difficult to realize these manifolds as the images of a family h_ω of embeddings of S^3 minus two points into $\mathbb{R}^4$ defined for $\omega\in(-\varepsilon,\varepsilon)$ and depending continuously on ω in the compact-open topology (here $\mathfrak{M}$ is the image of h_0). Using this family of embeddings we can pull-back the flow of vectorfield (1.2) to construct a family of flows $\Phi(\omega)$ on a single copy of S^3 minus two points depending continuously on ω in the compact-open topology on flows (of course, $\Phi(0)=\Phi$). All this just amounts to viewing the nonzero angular momentum case as a perturbed flow on the limiting variety rather than as a restriction of a big flow on $\mathbb{R}^4$ to nearby invariant manifolds. Both points of view are useful. Away from the "corners", i.e., away from $\mathfrak{N}_0$, we can even view the perturbed flow as depending analytically on ω.

Rather than attempting a variant of the usual smooth techniques for embedding symbolic dynamics we adopt the methods of window theory developed by Easton [4]. Easton explores the relation between windows in a flow and symbolic dynamics in great generality. In a three-dimensional space the following simple approach suffices.

Let $I=[-1,1]$ and define a triple $(B,b_+,b_-)=(I\times I,\ \{-1,1\}\times I, I\times\{-1,1\})$. We define a positive path in B to be a continuous curve $\sigma:(I,\{-1,1\})\rightarrow(B,b_+)$ with $\sigma(-1)$ and $\sigma(1)$ in different components of b_+. A negative path is defined in a similar way. By a window in $\mathfrak{M}$ we mean an embedding $w:B\rightarrow\mathfrak{M}$.

Consider two windows w_0 and w_1 and a flow Φ_t on $\mathfrak{M}$. It may happen that there is a flow-defined Poincaré map taking a subset of $w_0(B)$ along orbits of Φ to $w_1(B)$. Let $T:B\rightarrow\mathbb{R}^+$ be continuous and define Φ_{10} to be the composition $\Phi_{10}(\beta)=w_1^{-1}\circ\Phi_{T(\beta)}\circ w_0(\beta)$ for $\beta\in B$. This will be defined on some compact domain $D_{10}\subset B$ and map to some range $R_{10}\subset B$.

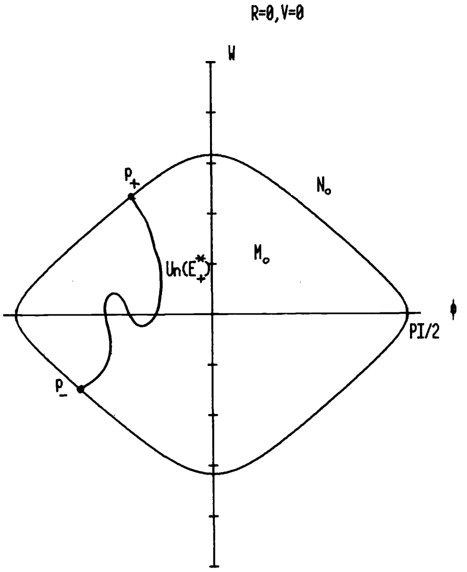


FIG. 7

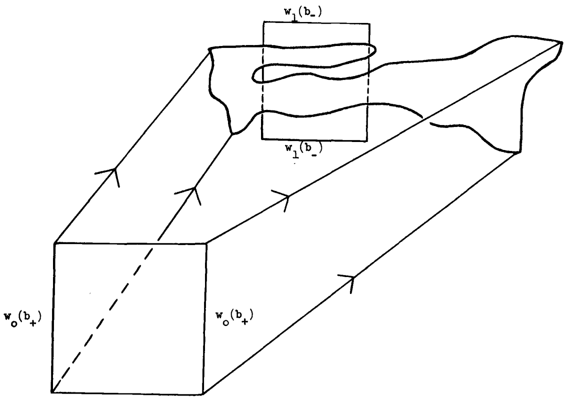


FIG. 8

We will say that Φ correctly aligns w_0 with w_1, if for some T as above, the following conditions hold:

i) Φ_{10}: $D_{10} \to R_{10}$ is a homeomorphism.

ii) $D_{10} \cap b_+ = R_{10} \cap b_- = \varnothing$.

iii) Every positive path in B contains a subpath in D_{10} mapping under Φ_{10} to a positive path while every negative path in B contains a subpath in R_{10} mapping under Φ_{10}^{-1} to a negative path.

These conditions are quite easy to verify in practice. Figure 8 shows two windows being correctly aligned by a flow and indicates the flexibility of the definition.

Now suppose we have a bi-infinite sequence of windows w_j, $j \in \mathbb{Z}$, such that Φ correctly aligns w_j with w_{j+1} for all j. Let D_{N0} be the domain of $\Phi_{N\,N-1} \circ \cdots \circ \Phi_{10}$ and let $R_{0\,-N}$ be the range of $\Phi_{0\,-1} \circ \cdots \circ \Phi_{-(N-1)-N}$. Using induction, we find that D_{N0} is compact, disjoint from b_+ and contains a negative path, while $R_{0\,-N}$ is compact, disjoint from b_- and contains a positive path. Consequently $D_{N0} \cap R_{0\,-N}$ is a nonempty compact subset of the interior of B. It follows that $\bigcap_N (D_{N0} \cap R_{0\,-N})$ is also a nonempty compact subset of the interior of B. Back in $\mathfrak{M}$ we find at least one orbit beginning in $w_0(B)$ which maps through all the windows in the appropriate order under the Poincaré maps.

One more definition will be needed. Let $S \subset \mathfrak{M}$ be a surface (usually a stable or unstable manifold). We will say that a window w is plus-transverse to S if $w(B)$ is contained in the domain of a submanifold chart for S which takes an open ball in $\mathfrak{M}$ to $\mathbb{R}^3$, S to $\mathbb{R}^2 \times 0$ $w(-1 \times I)$ to $\mathbb{R}^3_-$ and $w(1 \times I)$ to $\mathbb{R}^3_+$, where $\mathbb{R}^3_{+,-} = \{(x_1, x_2, x_3)$: $x_3 > 0, < 0\}$. In other words the components of $w(b_+)$ are on opposite sides of S. Define minus-transversality in a similar way. This is also easy to check in practice.

The following lemma allows us to construct suitable windows near each of the connecting orbits in the limiting flow.

LEMMA 4.1. *Suppose analytic invariant surfaces S_1 and S_2 have an odd-order crossing along an orbit γ. Let Σ be any local section to the flow along γ. Then there is a window w: $B \to \Sigma \subset \mathfrak{M}$ which is plus-transverse to S_1 and minus-transverse to S_2.*

Proof. $S_1 \cap \Sigma$ and $S_2 \cap \Sigma$ are analytic submanifolds of Σ, i.e., nonsingular analytic curves in Σ. We will find C^0 coordinates for Σ near $\gamma \cap \Sigma$ taking these curves to the coordinate axes. Then the square $I \times I$ in $\mathbb{R}^2$ is the appropriate window.

Choose analytic submanifold coordinates for $S_1 \cap \Sigma$ about $\gamma \cap \Sigma$. Calling these coordinates (x_1, x_2) we have $S_1 \cap \Sigma$, represented by the x_1-axis and $\gamma \cap \Sigma$ by $(0,0)$. Assume that $S_1 \cap \Sigma$ and $S_2 \cap \Sigma$ are not actually transverse since in this case the appropriate coordinate system can be found easily and it is even analytic. Thus $S_2 \cap \Sigma$

appears as the zero set of a function $f(x_1,x_2)=x_1^{2k+1}+x_2g(x_1,x_2)$, where $k\geq 1$ and $g(x_1,x_2)$ is an analytic function with $g(0,0)\neq 0$. Then the variables $(\xi_1,\xi_2)=(x_1,x_2g(x_1,x_2))$ are analytic coordinates near $(0,0)$. The further transformation $(\eta_1,\eta_2)=(\xi_1^{2k+1}+\xi_2,\xi_2)$ is a local homeomorphism with inverse $(\xi_1,\xi_2)=((\eta_1-\eta_2)^{1/2k+1},\eta_2)$. The composition $(\eta_1,\eta_2)=(f(x_1,x_2),\ x_2g(x_1,x_2))$ is our choice for the new coordinates. Clearly $S_2\cap\Sigma$ is represented locally by the η_2-axis and $S_2\cap\Sigma$ by the η_1-axis as desired. Q.E.D.

Consider a window w_0 near a connecting orbit which ends at one of the equilateral restpoints and another window w_1 near a connecting orbit which begins at the same restpoint. Unfortunately the limiting flow Φ does not determine a Poincaré map since $w_0(B)$ is held up in a neighborhood of the restpoint. However the nearby manifold $\mathcal{M}(h,\omega)$ does not contain the restpoint so the flows $\Phi(\omega)$ are not delayed in the corresponding neighborhood. We will show that these flows correctly align w_0 with w_1. This will require a somewhat closer analysis of the limiting process in a neighborhood of a restpoint.

Equation (2.1) which defines $\mathcal{M}(h,\omega)$ and $\mathcal{M}$ can be written as $rf(r,v,\varphi,w)=-\omega^2$, where $f(r,v,\varphi,w)=v^2\cos^2\varphi+w^2-2U(\varphi)\cos^2\varphi-2rh\cos^2\varphi$. Let $\lambda_1=rf$ and $\lambda_2=r^2-f^2$. The Jacobian of the coordinate change $(r,v,\varphi,w)\mapsto(\lambda_1,\lambda_2,\varphi,w)$ is $-4(r^2+f^2)v\cos^2\varphi$. We will use these coordinates only in a neighborhood of one of the restpoints in $\mathcal{M}$ and there $v\cos^2\varphi\neq 0$, but $r^2+f^2=0$ is exactly the equation for $\mathcal{N}_0$. Nevertheless, the new variables provide a C^0 coordinate system when restricted to $\{r\geq 0,f\leq 0\}$ containing the manifolds of interest. These are analytic coordinates on the complement of $\mathcal{N}_0$ and, trivially, on $\mathcal{N}_0$ itself. They are not globally analytic because they flatten the "corner" at $\mathcal{N}_0$. In fact, the new equation for $\mathcal{M}$ is $\lambda_1=0$ and $\mathcal{N}_0$ is given by $\lambda_1=\lambda_2=0$. $\mathcal{M}(h,\omega)$ appears as $\{\lambda_1=-\omega^2\}$.

We will need only a few facts about the vectorfield (1.2) in the new coordinates. First $\lambda_1'=0$ since λ_1 is $-\omega^2$. Since rf is constant we have $r'f+rf'=0$ from which we find $f'=-vf$. Then we find $\lambda_2'=2v\cos\varphi\,(r^2+f^2)$, so near a restpoint λ_2 is monotone on orbits in the complement of $\mathcal{N}_0$. Finally we note that the perturbed flows $\Phi(\omega)$ are the flows of vectorfields C^0 close to a vectorfield for Φ if ω is small. Before beginning the proof of the main perturbation lemma we make one more improvement in the coordinates. We can replace λ_2, φ, ω by λ_2, λ_3, λ_4 so that in the limiting flow near an equilateral restpoint (call it E) we have $\mathrm{St}(E)=\{\lambda_2\leq 0,\lambda_3=0\}$ and $\mathrm{Un}(E)\equiv\{\lambda_2\geq 0,\lambda_4=0\}$. We still have $\mathcal{N}_0=\{\lambda_2=0\}$. Figure 9 will be helpful throughout the proof. We choose a small ball D about E and suppose that the windows w_0 and w_1 have been followed along their associated connecting orbits to ∂D (this involves choosing he windows sufficiently small).

Lemma 4.2. *If w_0 and w_1 are sufficiently small windows in ∂D with w_0 plus-transverse to* St(E) *and w_1 minus-transverse to* Un(E), *then there is $\omega_{01}>0$ such that all the flows $\Phi(\omega)$ with $0<|\omega|<\omega_{01}$ correctly align w_0 with w_1 via the Poincaré map across D.*

Proof. There is a neighborhood of $\mathrm{St}(E)\cap\partial D$ in ∂D which is transverse to the vectorfield of Φ. We assume that $w_0(B)$ was chosen small enough that when it is followed forward to ∂D it lies in such a neighborhood. We make a similar assumption about $w_1(B)$ and a neighborhood of $\mathrm{Un}(E)$. If under $\Phi(\omega)$, $w_0(B)$ leaves D through a neighborhood of $\mathrm{Un}(D)$ which is transverse to the flow, then the time required to cross D will be a continuous function on $w_0(B)$ and the Poincaré map taking points of $w_0(B)$ to the point where their $\Phi(\omega)$ orbits leave D will be a homeomorphism. To verify correct alignment we must choose an especially nice neighborhood of $\mathrm{Un}(E)$.

Since $w_0(B)$ is plus-transverse to $\mathrm{St}(E)$ the two components of $w_0(b_+)$ lie on opposite sides of $\mathrm{St}(E)$. They leave D near the points of intersection of ∂D with the

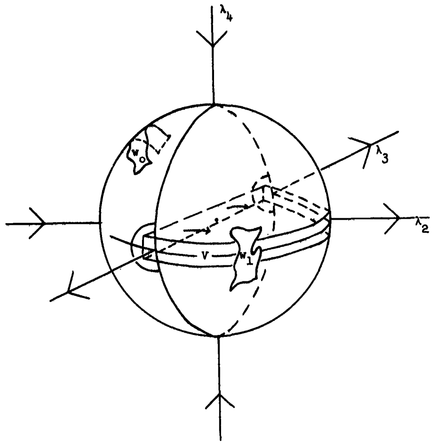


FIG. 9

λ_3-axis. Choose half-discs of the form $\partial D \cap \{\lambda_2 < 0,\ \lambda_3^2 + \lambda_4^2 < \delta\}$ in which the vectorfield for Φ points out of D. If $w_0(B)$ is a small enough window the orbits through $w_0(B)$ which leave D under Φ will do so through these half-discs. Furthermore the components of $w_0(b_+)$ will still leave D through the half-discs under the flows $\Phi(\omega)$ for $|\omega|$ sufficiently small.

Since $w_1(b)$ is minus-transverse to Un(E), there is an $\varepsilon > 0$ such that the two components of $w_1(b_-)$ lie in the sets $\{\lambda_4 > \varepsilon\}$ and $\{\lambda_4 < -\varepsilon\}$. The open band $\lambda_2 > 0$, $|\lambda_4| < \varepsilon$ in ∂D together with the half-discs constructed above form a neighborhood of Un$(E) \cap \partial D$, which we will call V. Choosing ε smaller, if necessary, we may assume that the vectorfield of φ points strictly out of D on V and that $\varepsilon < \delta/2$. We will show that for $|\omega|$ sufficiently small, orbits of $\Phi(\omega)$ which enter D through the window $w_0(B)$, leave D through V. To see this notice that $\{\lambda_2 > -\delta/2,\ |\lambda_4| < \varepsilon\}$ is positively invariant under Φ and also under $\Phi(\omega)$ for $|\omega|$ sufficiently small. Next notice that for the flow Φ there is a time T such that if $p \in w_0(B)$, the orbit segment $\Phi_t(p)$, $0 \le t \le T$, either leaves D through one of the half-discs or else enters the set $\lambda_2 > -\delta/2$, $|\lambda_4| < \varepsilon$. This condition persists (with the same T) for nearby all $\Phi(\omega)$ since it merely asserts that the image under Φ_T of the compact set $w_0(B)$ lies in the open set $\{\lambda_2 > -\delta/2,\ |\lambda_4| < \varepsilon\} \cup \Phi_{(0,T)}$ (half-discs). Now we have seen that $\lambda_2' > 0$ for the flows $\Phi(\omega)$ with $\omega \neq 0$ so every orbit entering D through $w_0(B)$ does leave D eventually and the above considerations show that such orbits must leave through V. Thus the Poincaré map across D stretches $w_0(B)$ through V taking the components of $w_0(b_+)$ to the half-discs at opposite ends of V. The way that V was chosen with respect to $w_1(B)$ makes it easy to verify correct alignment. First it is clear that the Poincaré map is a homeomorphism and so is the map Φ_{10}. Since $w_0(b_+)$ leaves D through the half-discs, $b_+ \cap \text{domain}(\Phi_{10}) = \varnothing$. Similarly $b_- \cap \text{range}(\Phi_{10}) = \varnothing$, since $w_1(b_-)$ lies in the complement of V. It is obvious that positive arcs in $w_0(B)$ are stretched through V from one half-disc to the other. But such a curve contains a subcurve connecting the components of $w_1(b_+)$ through $w_1(B)$. Similarly any negative path in $w_1(B)$ must connect the regions $\{\lambda_4 > \varepsilon\}$ and $\{\lambda_4 < -\varepsilon\}$ by crossing V and any such curve contains a subcurve connecting the images of the components of $w_0(b_-)$ through the image of $w_0(B)$. Q.E.D.

We will also need Poincaré maps between windows w_0 and w_1 where w_0 is plus-transverse to St$(\infty_{+,-})$ and w_1 is minus-transverse to Un$(\infty_{+,-})$. These are defined even for the limiting flow. Moreover we can get from w_0 to w_1 after going around the periodic orbit an arbitrary number of times. Perhaps the easiest way to keep track of the various possibilities is to introduce an extra window near the periodic orbit itself.

The next lemma is a window-theoretic version of results in [10], which the reader should consult for a more careful treatment of infinity. We will use only a result of McGehee [8] asserting the existence of a coordinate system with certain properties near the orbits at infinity. Specifically, there are coordinates (x_1, x_2, θ), where $x_1, x_2 \in \mathbb{R}$ and $\theta \in \mathbb{R}$ (mod 1), with the following properties:

i) An open subset of $\mathfrak{M}(h,\omega)$ in $\mathfrak{M}_+$ near infinity is mapped to the set of all (x_1, x_2, θ), where a certain function of the form $x_1 + x_2 + \cdots$ is positive (see Fig. 10; in particular, the first "quadrant" is in the set).

ii) The circle $x_1 = x_2 = 0$ represents a fictitious periodic orbit at infinity ($\theta' \approx 1$ near $x_1 = x_2 = 0$).

iii) The parabolic manifolds St(∞) and Un(∞) are represented by the cylinders $x_1 = 0$, $x_2 > 0$ and $x_2 = 0$, $x_1 > 0$ respectively.

iv) The Poincaré map of the first quadrant of the section $\theta = 0$ has the property that x_1 is strictly increased except on the x_2-axis and x_2 is strictly decreased except on the x_1-axis. These properties may be summarized by calling the periodic orbit degenerately hyperbolic.

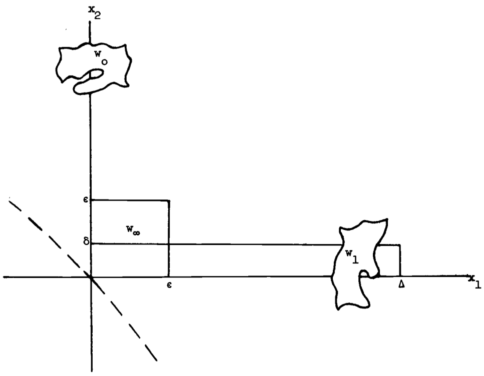


FIG. 10

Let Σ be a neighborhood of $(0,0)$ in the section $\theta = 0$. A window w_0 which is plus-transverse to St(∞) and is sufficiently small can be followed along orbits to Σ as can a window w_1 minus-transverse to Un(∞). Define a window w_∞ in Σ whose image is $\{0 \le x_1 \le \varepsilon,\ 0 \le x_2 \le \varepsilon\}$, with $w_\infty(b_+) \subset \{x_1 = 0 \text{ or } \varepsilon\}$ and $w_\infty(b_-) \subset \{x_2 = 0 \text{ or } \varepsilon\}$ (Fig. 10).

LEMMA 4.3. *Let w_0 and w_1 be windows in Σ with w_0 plus-transverse to* St(∞) *and w_1 minus-transverse to* Un(∞). *Let w_∞ be the "window at infinity". Then a sufficiently high iterate of the Poincaré map of Σ correctly aligns w_0 with each of w_∞ and w_1 and correctly aligns w_∞ with w_1.*

Proof. Choose $\delta < \varepsilon$ such that the components of $w_1(b_-)$ lie in $\{x_2 > \delta\}$ and $\{x_2 < 0\}$. Choose Δ such that $w_1(B)$ lies in $\{0 < x_1 < \Delta\}$. Condition iv) shows that a sufficiently high iterate of the Poincaré map takes $w_0(B)$ to $\{0 < x_2 < \delta\}$. Since w_0 is plus-transverse to St(∞), i.e., to the positive x_2-axis, a high enough iterate will take one component of $w_0(b_+)$ to $\{x_1 < 0\}$ and the other to $\{x_1 > \Delta\}$. As in the proof of Lemma 4.2, this type of geometric condition implies correct alignment of w_0 with each of w_∞

and w_1. Taking a higher iterate if necessary we have $w_\infty(B)$ mapping to $\{0<x_2<\delta\}$ with one component of $w_\infty(b_+)$ in $\{x_1=0\}$ and the other in $\{x_1>\Delta\}$ and so w_∞ is correctly aligned with w_1. Q.E.D.

Before stating the main result we augment the connection graph by adding a directed edge beginning and ending at vertex ∞_+ and one beginning and ending at vertex ∞_-. Denote the augmented graph by $\hat{G}(\alpha)$. The added vertices represent the windows near the periodic orbits at infinity constructed in Lemma 4.3. We can associate a window to each of the other edges in $\hat{G}(\alpha)$. Each edge represents an intersection of a two-dimensional stable manifold and a two-dimensional unstable manifold. Using Lemma 4.1 we can find arbitrarily small windows, transverse to the flow Φ, which are minus-transverse to the unstable manifold involved and plus-transverse to the stable manifold involved. We will choose them small enough that when they are followed along the connecting orbit to the neighborhood of the restpoint or periodic orbit, the appropriate lemma, 4.2 or 4.3, applies.

The limiting flow Φ with its six distinguished invariant sets and connecting orbits has, at this point, served its purpose, namely to indicate where to construct windows in $\mathfrak{M}$. As we have seen, there are no Poincaré maps defined between these windows for the limiting flow (the exception being windows along orbits to and from infinity). The homeomorphism of $\mathfrak{M}(h,\omega)$ with $\mathfrak{M}$ allows us to transfer the windows to $\mathfrak{M}(h,\omega)$. Four of the six invariant sets and the corresponding connecting orbits disappear when we consider nonzero angular momenta but the windows remain and are correctly aligned with one another by the flows $\Phi(\omega)$ thanks to Lemmas 4.2 and 4.3.

Edges in $\hat{G}(\alpha)$ represent windows in $\mathfrak{M}$. Two edges together with a vertex which is the terminal vertex of one and the initial vertex of the other represent the existence, for sufficiently small nonzero angular momenta, of a Poincaré map which correctly aligns the windows. This Poincaré map is obtained by following orbits of $\Phi(\omega)$ as they pass through one window, and move along near a connecting orbit of Φ through a neighborhood of one of the invariant sets of Φ and near another connecting orbit of Φ to the other window.

Recall that a path in a directed graph is a sequence of directed edges such that the terminal vertex of the nth edge is the initial vertex of the $(n+1)$st. We will say that a path P in $\hat{G}(\alpha)$ is realized by $\Phi(\omega)$ if the Poincaré maps represented by pairs of successive edges in P are defined for $\Phi(\omega)$, i.e., ω is sufficiently small that all of the necessary Poincaré maps are defined. As we have seen, the correct alignment of a sequence of windows guarantees the existence of at least one orbit which crosses the windows in the given order via the appropriate Poincaré maps.

Proposition 4.4. *Let Γ be a finite subgraph of $\hat{G}(\alpha)$. Then there is an $\omega(\Gamma)>0$ such that each of the flows $\Phi(\omega)$ with $0<|\omega|<\omega(\Gamma)$ realizes every path in Γ.*

Proof. It is only necessary to observe that since Γ is finite we can choose $\omega(\Gamma)>0$ so that all of the Poincaré maps represented by Γ are defined by $\Phi(\omega)$ if $0<|\omega|<\omega(\Gamma)$.

5. Heteroclinic phenomena. Proposition 4.4 asserts the existence of a complicated invariant set in the isosceles three-body problem for small nonzero angular momentum. There is a correspondence between orbits of this invariant set and paths in a subgraph Γ of the connection graph. The next proposition shows that cyclic paths in Γ are especially significant. A cycle in a graph is just a path which begins and ends at the same vertex. A cycle can be periodically extended to a bi-infinite path.

Proposition 5.1. *Let Γ be a finite subgraph of $\hat{G}(\alpha)$ and let $0<|\omega|<\omega(\Gamma)$. Let C be a cycle in Γ. The set $I(C)$ of orbits of $\Phi(\omega)$ which realize C is a nonempty compact isolated invariant set which carries a one-form* [1].

Proof. Let $w_0,\cdots,w_n$ be the windows represented by the edges of C. By composing the Poincaré maps between these windows we obtain a Poincaré map from a certain domain $D\subset w_0(B)$ to $w_0(B)$. We know that $I(C)\cap w_0(B)$ is a nonempty compact subset of the interior of D. Now the Poincaré map is obtained by following the orbits of $\Phi(\omega)$ through points of $w_0(B)$ for a finite amount of time described by a continuous function $T: w_0(B)\to\mathbb{R}^+$. So $I(C)=\{\Phi(\omega)_t p: t\in[0,T(p)], p\in I(C)\cap w_0(B)\}$ is also compact.

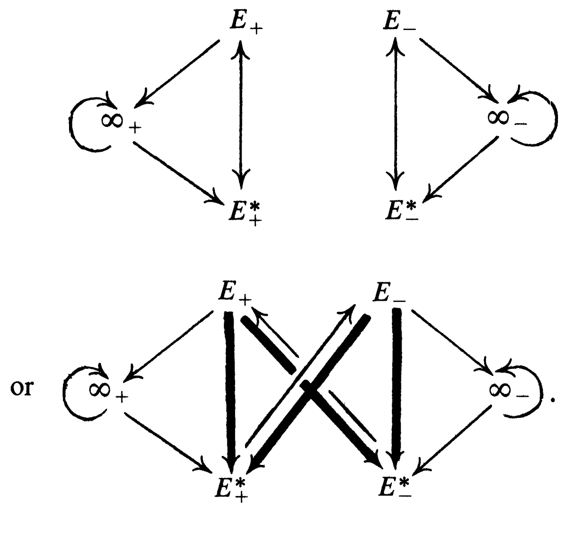
Recall that a compact invariant set is called isolated if it is the maximal invariant set in some neighborhood of itself. To construct a neighborhood for $I(C)$ consider $\{\Phi(\omega)_t p: p\in D, t\in[0,T(p)]\}$. Let V denote the interior of this set. Clearly $I(C)\subset V$. If $p\in D$ and if the orbit of p remains in V for all time then the Poincaré map of $w_0(B)$ repeatedly returns p to the interior of D and therefore the orbit of p realizes C. So $I(C)$ is the maximal invariant set in V.

Let I be an isolated invariant set and V an isolating neighborhood for I. Let $\alpha\in H'(V,\mathbb{R})$ be a real cohomology class on V. An orbit segment in V can be viewed as a one-dimensional simplex and so α can be evaluated on it. We say that the invariant set I carries α if there are constants $A>0$, $T>0$ such that any orbit segment σ which remains in V for $t>T$ units of time satisfies $\alpha(\sigma)>At$. Invariant sets which carry one-forms share certain properties with periodic orbits [1].

It is not difficult to find such a one-form in the neighborhood V of $I(C)$ constructed above. We can smoothly modify the time parametrization along orbits through D so that the Poincaré map of $w_0(B)$ becomes a time one map. Then to every point of V we can associate a time between zero and one. Define a one-form α as the integral of this time function. Choosing the constant A to be a Lipschitz constant for the time reparametrization, we see that $I(C)$ carries α. Q.E.D.

Propositions 3.1 and 3.2 show that $\hat{G}(\alpha)$ contains infinitely many distinct cycles for all sufficiently large or sufficiently small mass ratios α. In fact one of the following two subgraphs is present:

(5.1)



Each of these subgraphs contains a cycle involving only the restpoints which is linked to the cycle at infinity. Infinitely many other cycles can be constructed which shuttle between these two basic cycles.

It is interesting to consider the qualitative behavior of orbits in $I(C)$ when C is a simple cycle. As an example, consider the cycle C in the top subgraph of (5.1) specified by the sequence of vertices E_+, E_+^*, E_+, ∞_+, ∞_+, E_+^*, E_+. An orbit realizing this cycle behaves, initially, much like the homothetic connecting orbit; the configuration is approximately equilateral as the size first expands then contracts. As the particles approach they begin to behave more like the connecting orbit $E_+^* \to E_+$ known to exist in $\mathfrak{M}_0$. This behavior may be quite complicated but it all occurs while the particles are extremely close together. Following this relatively close encounter the third mass is propelled far up the z-axis while the other two masses circulate about one another a certain number of times. The particles again approach closely and emerge from the second encounter near the equilateral homothetic orbit again.

It is possible that the isolated invariant set $I(C)$ consists of a single hyperbolic periodic orbit. While our methods cannot detect this, it is possible to show that $I(C)$ contains at least one periodic orbit. Let w_0 be a window of C and consider the Poincaré map taking $w_0(B)$ around C and back to itself. Let Φ_{00} denote the induced map from a subset of B to B. If (β_1,β_2) are coordinates on B, $\beta_j \in [-1,1]$, we can write $\Phi_{00}(\beta_1,\beta_2) = (\hat{\beta}_1,\hat{\beta}_2)$ and define $A_1 = \{(\beta_1,\beta_2) : \hat{\beta}_1 = \beta_1\}$ and $A_2 = \{(\beta_1,\beta_2) : \hat{\beta}_2 = \beta_2\}$. Since the flow correctly aligns $w_0(B)$ with itself, every positive path in B contains a subpath mapping under Φ_{00} to a positive path. Hence every positive path in B meets A_1. Similarly, every negative path meets A_2. We will show that this forces $A_1 \cap A_2 \neq \varnothing$ and so Φ_{00} has a fixed point and $I(C)$ contains a periodic orbit.

The proof that $A_1 \cap A_2 \neq 0$ is an exercise in algebraic topology. Consider the exact cohomology sequence of the triple $(B, B \setminus A_1, b_+)$

$$\to H^1(B, B\setminus A_1) \xrightarrow{i^*} H^1(B, b_+) \xrightarrow{j^*} H^1(B\setminus A_1, b_+) \to .$$

Now $H^1(B,b_+)$ is generated by a single cohomology class, α_1, with the property that $\alpha_1(\sigma) = \pm 1$ if σ is a positive path in B and zero if σ is a one-chain with no positive subpath. This shows that $j^*(\alpha_1) = 0$ because there are no positive paths in $B \setminus A_1$. Hence i^* is surjective and there is a cohomology class $\hat{\alpha}_1 \in H^1(B, B\setminus A_1)$ with $i^*\hat{\alpha}_1 = \alpha_1$. Similarly we can find $\hat{\alpha}_2 \in H^1(B, B\setminus A_2)$ with $i^*\hat{\alpha}_2 = \alpha_2$ in $H^1(B, b_-)$. It follows that the cup product $\hat{\alpha}_1 \cup \hat{\alpha}_2$ is a nonzero cohomology class in $H^2(B, B\setminus(A_1 \cap A_2))$, which would be impossible if $A_1 \cap A_2 = \varnothing$.

Let C_1 and C_2 be cycles in a subgraph Γ of $\hat{G}(\alpha)$ and let P be a path in Γ which "begins" with an infinite string of C_1's and "ends" with an infinite string of C_2's. We will show that any orbit realizing P is heteroclinic between $I(C_1)$ and $I(C_2)$, i.e., the α limit set is contained in $I(C_1)$ and the ω limit set is contained in C_2.

Consider the usual Poincaré map from one of the windows $w_0(B)$ of C_2 to itself. $I(C_2) \cap w_0(B)$ is the intersection of the nested sequence of compact sets $D_N \cap R_N$, where D_N is the domain of the Nth iterate of the Poincaré map and R_N is its range. An orbit realizing P crosses $w_0(B)$ at a point $p \in \cap_N D_N$. The Nth iterate of p is therefore a point of $D_N \cap R_N$. It follows that $\omega(p) \subset \cap_N (R_N \cap D_N) = I(C_2)$. Similarly $\alpha(p) \subset I(C_1)$.

These heteroclinic orbits provide examples of solutions of the three-body problem with various types of asymptotic behavior as $t \to \pm\infty$. For example, the oscillation and capture phenomena found in Sitnikov's problem ($\alpha \approx \infty$, fairly large ω) occur for all mass ratios for which $\hat{G}(\alpha)$ contains one of the subgraphs (5.1), provided ω is sufficiently small. Indeed capture orbits realize paths heteroclinic between the cycle at infinity and any bounded cycle. Oscillation orbits realize paths which contain long repeating sequences of a cycle at infinity interspersed with bounded cycles.

The new feature present in our work is the control over the bounded part of such orbits. There are infinitely many bounded isolated invariant sets to which a capture orbit may tend or near which an oscillatory orbit may pass.

For large mass ratios the variety of phenomena is remarkable. The right-hand subgraph of (5.1) occurs and we can realize an arbitrarily large finite subgraph for $|\omega|$ sufficiently small. Since some of the connecting orbits from $E_{+,-}$ to $E^*_{+,-}$ are very near the collinear homothetic orbit and since the equilateral homothetic orbits are represented in the graph, we can construct bounded cycles which repeatedly behave like each of these three special orbits.

Acknowledgments. The author wishes to acknowledge helpful discussions with R. McGehee.

REFERENCES

[1] C. Conley, *Invariant sets which carry a one-form*, J. Differential Equations, 8 (1970), pp. 587–594.

[2] Robert L. Devaney, *Triple collision in the planar isosceles three-body problem*, Inv. Math., 60 (1980), pp. 249–267.

[3] ______, *Singularities in classical mechanical systems*, Ergodic Theory and Dynamical Systems, Birkhauser, Boston, 1981.

[4] R. W. Easton, *Isolating blocks and symbolic dynamics*, J. Differential Equations, 17 (1975), pp. 96–118.

[5] J. M. Irigoyen, *La variété de collision triple dans le cas isocèle du problème des trois corps*, C. R. Acad. Sci. Paris, 290 (1980), pp. B489–B492.

[6] Ernesto Lacomba and Lucette Losco, *Triple collision in the isosceles three-body problem*, Bull. Amer. Math. Soc., 3 (1980), pp. B489–B492.

[7] Richard McGehee, *Singularities in classical celestial mechanics*, Proc. of the International Congress of Mathematicians, Helsinki, 1978, pp. 827–834.

[8] ______, *A stable manifold theorem for degenerate fixed points with applications to celestial mechanics*, J. Differential Equations, 14 (1973), pp. 70–88.

[9] Richard Moeckel, *Orbits of the three-body problem which pass infinitely close to triple collision*, Amer. Jour. Math., 103 pp. 1323–1341.

[10] Jürgen Moser, *Stable and Random Motions in Dynamical Systems*, Ann. Math. Studies 77, Princeton Univ. Press, Princeton, NJ, 1973.

[11] C. Simo, *Analysis of triple collision in the isosceles problem*, in Classical Mechanics and Dynamical Systems, Marcel Dekker, New York, 1980.

[12] K. Sitnikov, *The existence of oscillatory motions in the three-body problem*, Soviet Physics Doklady, 5 (1961), pp. 647–650.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics
002

SCALING HAMILTONIAN SYSTEMS*

K. R. MEYER†

Abstract. This paper presents a detailed discussion of scaling techniques for Hamiltonian systems of equations. These scaling techniques are used to introduce small parameters into various systems of equations in order to simplify the proofs of the existence of periodic solutions. The discussion proceeds through a series of increasingly more complex examples taken from celestial mechanics. In particular, simple proofs are given for Lyapunov's center theorem, the continuation theorem of Hadjidemetriou, and several theorems on periodic solutions by the author.

1. Introduction. Perturbation analysts often argue over which general method is best—the methods of averaging, Lie transformations, two-timing, Lyapunov–Schmidt, etc., all have their strong advocates. However, no matter what perturbation technique is used, an important, fundamental and often overlooked question is the correct selection of the equations of the first approximation. In some cases it is so obvious what the first approximation is that there really is no choice, but in other cases the choice can drastically affect all subsequent analysis. In celestial mechanics the equations of the first approximation are called the main problem, and I shall use this term since it emphasizes the importance of these equations.

A historic example where the choice of the main problem had important consequences is found in lunar theory. Until the works of Hill were completely understood, researchers looking for a good approximate solution to the equations of celestial mechanics which described the motion of the moon used as their main problem two decoupled Kepler problems. The two Kepler problems were the equations of motion of the earth and moon about their combined center of mass, and the equations of motion describing the sun and the center of mass of the earth–moon system. Coupling terms were neglected in the main problem. Various perturbation techniques were used, but the approximate solutions failed to agree with the observational data over long periods of time. In a series of papers [5], Hill redefined the main problem of lunar theory by taking into account the fact that the motion of the moon is strongly affected by the sun. Hill's main problem took into account more terms, and as a result the perturbations were smaller and the series converged better numerically. In fact, for many years lunar ephemerides were computed from series developed by Brown, who used Hill's main problem. Even today searchers for more accurate lunar theories use Hill's main problem.

In this survey paper, I want to discuss a general procedure for deciding the correct definition of the main problem in various situations in celestial mechanics. The examples are taken from my own work and therefore consist mainly of problems of finding periodic solutions in Hamiltonian formalism.

The method present is certainly not new—in fact it is so old that I have no idea of when it originated. The method is also not obscure—in fact almost all perturbation techniques are based either explicitly or implicitly on this method. The method is simply that of scaling variables. Since the problems discussed here are written in Hamiltonian formalism, the scaling will be done so that the resulting equations are

*Received by the editors June 2, 1982, and in revised form June 23, 1983. This research was partially supported by the National Science Foundation under grant MCS-80-01851.

† Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio 45221.

again Hamiltonian and so the scaling is symplectic (canonical). Therefore I call the method symplectic scaling. There have been other discussions of scaling as a general procedure in applied mathematics; see for example [16].

Scaling is often presented as a triviality. Sometimes an author starts his discussion with a single statement like: "scale by $x \to \varepsilon x$ and $y \to \varepsilon y$" and then proceeds with page upon page of detailed calculation. Usually there is no discussion of why the equations were scaled nor whether this scaling is the best. In fact, I have seen many papers that could be greatly simplified if the author had used a different scaling (say $x \to \varepsilon x$ and $y \to \varepsilon^2 y$). The examples given below illustrate how to obtain the correct scaling for a particular problem in celestial mechanics.

2. Review of transformation theory. I shall deal exclusively with autonomous Hamiltonian systems. Even though this paper attempts to be reasonably self-contained, I assume that the reader has some background in differential equations and Hamiltonian mechanics. The excellent introductory book by Pollard [13] should be more than adequate. I shall not bog myself down with topological or smoothness questions, since the results given below are local in nature. All functions and vector fields will be assumed to be C^∞ on some open set in $\mathbb{R}^{2n}$ or even defined on all of $\mathbb{R}^{2n}$. Also vectors will be column vectors unless otherwise stated, but will be written as row vectors in the text for typographical reasons.

If $\phi: \mathbb{R}^l \to \mathbb{R}^k$ and $y = \phi(x)$ then $\partial\phi/\partial x$ or $\partial y/\partial x$ will denote the $k \times l$ Jacobian matrix. Thus if H: $\mathbb{R}^{2n} \to \mathbb{R}$, $x \in \mathbb{R}^{2n}$ then $\partial H/\partial x$ is a row vector. Define $\nabla_x H = \nabla H = (\partial H/\partial x)^T$ where the superscript T denotes the transpose.

An autonomous Hamiltonian system of n degrees of freedom in $\mathbb{R}^n$ is a system of ordinary differential equations of the form

$$\dot{x} = J \nabla_x H(x) \tag{2.1}$$

where $H: \mathbb{R}^{2n} \to \mathbb{R}^1$, $x \in \mathbb{R}^{2n}$, $\dot{} = d/dt$ and J is the $2n \times 2n$ constant matrix

$$J = \begin{pmatrix} 0 & I \\ -1 & 0 \end{pmatrix}$$

where 0 and I are the $n \times n$ zero and identity matrices. The independent variable t will be called time, the function H, the Hamiltonian and (2.1), the equations of motion. If $x = (q, p)$ where $q, p \in \mathbb{R}^n$ then the equations of motion take the classical form

$$\dot{q} = \frac{\partial H}{\partial p}, \qquad \dot{p} = -\frac{\partial H}{\partial q}. \tag{2.2}$$

Thus there is a well-defined prescription for obtaining the equations of motion from the Hamiltonian. This prescription is not invariant under all changes of variables. That is, if one changes variables in both the equations of motion and the Hamiltonian, then the new equations of motion may not be obtained from the new Hamiltonian by the prescription given in (2.1). Those changes of variables which preserve this prescription are known as symplectic or canonical. There is a vast literature on the subject of symplectic changes of variables, but fortunately only one basic fact will be needed for the subsequent discussion. The texts by Pollard [13], Wintner [17] and Abraham and Marsden [1] contain more details.

Consider the change of coordinates $x = \phi(y)$ for (2.1). These equations become

$$\dot{y} = P^{-1}(y) J \nabla_x H(\phi(y)) = P^{-1}(y) J \left\{ \frac{\partial H}{\partial x}(\phi(y)) \right\}^T \tag{2.3}$$

where $P(y)$ is the Jacobian $\partial\phi(y)/\partial x$. As noted above, these equations need not be in Hamiltonian form; i.e., the right-hand side is not of the form $J\nabla_y K(y)$ where $K:\mathbb{R}^{2n}\to\mathbb{R}$. However, if we assume that

$$J=\mu TJT^T \tag{2.4}$$

where μ is a nonzero constant, then

$$\dot{y}=T^{-1}J\left\{\frac{\partial H}{\partial x}\right\}^T=\mu JT^T\left\{\frac{\partial H}{\partial x}\right\}^T=\mu J\left\{\frac{\partial H}{\partial x}\frac{\partial \phi}{\partial y}\right\}^T=J\nabla_y\{\mu H(\phi(y))\},$$

or

$$\dot{y}=J\nabla_y K(y) \tag{2.5}$$

where

$$K(y)=\mu H(\phi(y)). \tag{2.6}$$

A change of variables $x=\phi(y)$ which satisfies (2.4) for all y and for some nonzero constant μ is called a symplectic transformation with multiplier μ. What was just shown is that these transformations preserve the Hamiltonian character of the equations and in particular transform (2.1) to (2.5).

If $\mu=1$ then the change of variables is simply called symplectic. Many elementary texts consider only this case, but the added generality of having a μ different from 1 is very important for scaling.

As an example, consider the problem of changing units in the N-body problem. Let $q_1,\cdots,q_N$ be the position vectors with respect to a Newtonian frame of N point masses moving in $\mathbb{R}^3$. Let $p_1,\cdots,p_N$ be the momentum vectors and $m_1,\cdots,m_N$ be masses of these point masses. Then the Hamiltonian for the N-body problem is

$$H=\sum_{i=1}^{N}\frac{\|p_i\|^2}{2m_i}-\sum_{1\le i<j\le N}\frac{km_im_j}{\|q_i-q_j\|} \tag{2.7}$$

where k is the universal gravitational constant. If $x=(q_1,\cdots,q_N,p_1,\cdots,p_N)$, then $x\in\mathbb{R}^{6N}$ and the equations of motion for the N-body problem are (2.1).

Scaling and changing units are essentially the same thing. Let's say for example that the quantities in this problem are all measured in the CGS system. Then $k=6.67\times 10^{-8}$. If we wish to change the unit of length, then we set $q_i=\alpha\bar{q}_i$, $p_i=\alpha\bar{p}_i$ where α is the conversion factor ($\alpha=100$ cm/m if the new unit of length is meters). This change of variables is symplectic with multiplier α^{-2}, and so the Hamiltonian becomes

$$H=\sum_{i=1}^{N}\frac{\|\bar{p}_i\|^2}{2m_i}-\sum_{1\le i<j\le N}\frac{k}{\alpha^3}\frac{m_im_j}{\|\bar{q}_i-\bar{q}_j\|}. \tag{2.8}$$

In this mixed system of units (MGS) the gravitational constant becomes $k/\alpha^3=6.67\times 10^{-14}$. In theoretical work it is convenient to use nonmetric units and to take $\alpha^3=k$ so that the gravitational constant is 1. We shall take $k=1$ in all subsequent discussions.

Since the bars over variables are not esthetic and are besides a lexicographical nuisance, it will be convenient to drop them in all subsequent discussions. The operation of first changing variables by $q_i=\alpha\bar{q}_i$, $p_i=\alpha\bar{p}_i$ and then dropping the bars is denoted by $q_i\to\alpha q_i$, $p_i\to\alpha p_i$. It should be carefully noted that this notation implies a change of variables and is only used to limit the proliferation of symbols.

Sometimes it is necessary to change the independent variable t also. If $t=\beta\tau$, where β is a constant, then (2.1) becomes

$$x'=J\nabla_x K(x) \tag{2.9}$$

where $'=d/d\tau$ and $K=\beta H$. Thus scaling time is equivalent to multiplying the Hamiltonian by a factor. In the scaling notation $t\to\beta t$ and $H\to\beta H$.

3. The noncritical case. Since the main application of scaling to be discussed in this paper is to establish the existence of periodic solutions, I shall summarize some of the known elementary results. Let $\phi(t,\xi,\lambda)$ be the solution of the Hamiltonian system

$$\dot{x}=J\nabla_x H(x,\lambda) \tag{3.1}$$

which satisfies $\phi(0,\xi,\lambda)=\xi$ where λ is a real parameter. Since (3.1) is autonomous, a necessary and sufficient condition for a particular solution $\phi(t,\xi_0,\lambda_0)$ to be T-periodic $(T>0)$ is

$$\phi(T,\xi_0,\lambda_0)=\xi_0. \tag{3.2}$$

This is easily proved by observing that both $\phi(t,\xi_0,\lambda_0)$ and $\phi(t+T,\xi_0,\lambda_0)$ are both solutions of (3.1) and that (3.2) implies that both these solutions satisfy the same initial condition. Thus the uniqueness theorem for ordinary differential equations assures that the two solutions are identical.

The necessary and sufficient condition (3.2) is interesting since it shows that the existence of periodic solutions of a differential equation is equivalent to solution of a system of (nondifferential) equations. In theory, at least, only finite dimensional methods could be used to establish the existence of periodic solutions. This is certainly not the case when we are trying to establish the existence of almost periodic solutions, invariant manifolds, etc. For these problems infinite dimensional methods are essential.

One approach to solving (3.2) is the use of the implicit function theorem. If (T,ξ_0,λ_0) satisfies (3.2) then the implicit function theorem would give nearby solutions, provided the Jacobian matrix

$$\frac{\partial\phi}{\partial\xi}(T,\xi_0,\lambda_0)-I \tag{3.3}$$

were nonsingular or equivalently that the Jacobian matrix

$$\frac{\partial\phi}{\partial\xi}(T,\xi_0,\lambda_0) \tag{3.4}$$

did not have the eigenvalue $+1$. The eigenvalues of (3.4) are so important in the study of periodic solutions that they are named the characteristic multipliers (or simply multipliers) of the periodic solution. Unfortunately, $+1$ is always a multiplier of a periodic solution of an autonomous system. Even worse, since (3.2) admits H as a first integral, the algebraic multiplicity of $+1$ as a multiplier is greater than or equal to 2. In the class of nonautonomous periodic equations, the usual case is that a periodic solution does not have the multiplier $+1$, and so this is usually called the noncritical case. In the class of autonomous equations, the usual case is that a periodic solution has the characteristic multiplier $+1$ with multiplicity precisely $=+1$ [7], and so for autonomous systems this is the noncritical case. In the case of autonomous Hamiltonian systems, the usual case is that a periodic solution has the characteristic multiplier $+1$ with algebraic multiplicity precisely equal to 2, and so for such systems this is the usual case [15].

In each of the cases listed above we have defined a noncritical case for each choice of our universe of discourse. The reason I call these the noncritical cases is that for each of these definitions of the noncritical case there is a theorem which states that in the noncritical case a small perturbation within the universe of discourse causes a slight perturbation in the periodic solution. Moreover, each of these theorems admits an elementary proof based on the implicit function theorem. The most satisfying discussion of these theorems is contained in Poincaré [14], but a clear, elementary discussion in modern notation can be found in Deprit and Henrard [2].

In the autonomous Hamiltonian case the precise statement of the theorem alluded to above is:

THEOREM 3.1. *Let $\lambda\in\mathbb{R}^k$, $k=0,1,2,\cdots,H:\mathbb{R}^{2n}\times\mathbb{R}^k\to\mathbb{R}^1$ be smooth and let $\phi(t,\xi,\lambda)$ be the solution of* (3.1) *so that $\phi(0,\xi,\lambda)=\xi$. Assume that (T,ξ_0,λ_0), $T>0$, satisfies*

i) $$\phi(T,\xi_0,\lambda_0)=\xi_0$$

and

ii) $$\operatorname{rank}\left\{\frac{\partial\phi}{\partial\xi}(T,\xi_0,\lambda_0)-I\right\}=2n-2.$$

Then the periodic solution $\phi(t,\xi_0,\lambda_0)$ is smoothly embedded in a $(k+2)$-parameter family of periodic solutions. That is, there are a neighborhood 0 *of λ_0 in $\mathbb{R}^k$, a neighborhood P of* $(0,0)$ *in $\mathbb{R}^2$, and smooth maps $\tau:P\times 0\to\mathbb{R}^n$ and $\xi:P\times 0\to\mathbb{R}^n$ such that $\tau(0,0,\lambda_0)=T$, $\xi(0,0,\lambda_0)=\xi_0$, and $\phi(t,\xi(\alpha,\beta,\lambda),\ \lambda)$ is a $\tau(\alpha,\beta,\lambda)$-periodic solution of* (3.1) *where $(\alpha,\beta)\in P$ and $\lambda\in 0$.*

Note that even when the equation does not depend on a parameter (i.e. $k=0$), the periodic solution is still embedded in a 2-parameter family of periodic solutions. These two additional parameters can be chosen as the value of the integral H on the periodic solution and the time from a well-chosen epoch along the periodic solution. In this case these periodic solutions locally fill a cylinder in $\mathbb{R}^{2n}$; see [1, Fig. 8.2-1]. Again we refer the reader to [2] for a simple clean proof of this theorem.

The remainder of this section is devoted to illustrating the method of symplectic scaling as a tool for reducing a given system to one to which the above theorem applies. Consider first the famous Lyapunov center theorem by glancing at the proofs given in [4], [6], [8]. In this theorem we assume that the equation (3.1) has an equilibrium point, say at $x=0$, and then expand the right-hand side in a Taylor series to get

$$\dot{x}=Ax+f(x) \tag{3.5}$$

where $f(0)=0$, $\partial f(0)/\partial x=0$ and A is a $2n\times 2n$ constant matrix. The Hamiltonian becomes

$$H(x)=\tfrac{1}{2}x^TSx+K(x) \tag{3.6}$$

where S is the Hessian of H at $x=0$, $A=JS$, K and the first and second partials of K vanish at $x=0$. By setting $f=0$ in (3.5) we obtain the linearization of the equations about the equilibrium point—an obvious candidate for the main problem. We would expect or at least hope that the solutions of the linear system are nearly the same as the solutions of the full equation when x is small, but how can we demonstrate this connection? To obtain a measure of x being small, scale by $x\to\varepsilon x$. This scaling is symplectic of order ε^2 and so (3.5) becomes

$$\dot{x}=Ax+O(\varepsilon) \tag{3.7}$$

and the Hamiltonian becomes

$$H(x)=\tfrac{1}{2}x^TSx+O(\varepsilon). \tag{3.8}$$

When $\varepsilon=0$, (3.7) becomes linear and the general solution is $\phi(t,\xi,0)=(\exp At)\xi$. In order to apply the theorem above, this system must have a periodic solution. Therefore let the eigenvalues of A be $\lambda_1,\lambda_2,\cdots,\lambda_{2n}$ and assume $\lambda_1=+i\omega$, $\lambda_2=-i\omega$, $\omega>0$. Let η and $\bar{\eta}$ be the corresponding eigenvectors so $A\eta=i\omega\eta$ and $(\exp At)\eta=(\exp i\omega t)\eta$. Thus $\operatorname{Re}(\exp At)\eta=(\exp At)\xi_0$ is a $T=2\pi/\omega$ periodic solution. The Jacobian matrix (3.4) becomes in this case

$$\frac{\partial\phi}{\partial\xi}(T,\xi_0,0)=\left(\exp\frac{2\pi}{\omega}A\right), \tag{3.9}$$

which has eigenvalues $\exp\pm2\pi i=1$, $\exp(\lambda_2 2\pi/\omega),\cdots,\exp(\lambda_{2n}2\pi/\omega)$. Thus for the second condition of the theorem to hold it must be assumed that

$$\frac{\lambda_k 2\pi}{\omega}\not\equiv 0 \bmod 2\pi i \quad \text{for } k=2,\cdots,2n, \tag{3.10}$$

or

$$\frac{\lambda_k}{\lambda_1} \text{ is not an integer} \quad \text{for } k=2,\cdots,2n. \tag{3.11}$$

If this condition applies, then (3.7) has a 3-parameter family of periodic solutions which are of the form $(\exp At)\xi_0+O(\varepsilon)$. The original equation (3.5) has a two-parameter family of periodic solutions of the form $\varepsilon(\exp At)\xi_0+O(\varepsilon^2)$. It may seem that we have proved that (3.5) has a 3-parameter family also, but this equation is independent of ε, and the theorem given above gives precisely a $k+2$ manifold of periodic solutions whose period is close to T. Thus one of the parameters is redundant. That proves Lyapunov's center theorem!

As a second example, consider the relationship between the full three-body problem and the restricted three-body problem. In the traditional derivation of the restricted three-body problem, one is asked to consider the motion of an infinitesimally small particle moving in the plane under the influence of the gravitational attraction of two finite particles which revolve around each other on a circular orbit of the Kepler problem. Although this description is picturesque, it hardly clarifies the relationship between the restricted three-body problem and the full problem. Consider the planar N-body problem where $N=2$ or 3 written in rotating coordinates [1]. The Hamiltonian is

$$H_N=\sum_{i=1}^{N}\frac{\|y_i\|^2}{2m_i}-x_i^TKy_i-\sum_{1\le i<j\le N}\frac{m_im_j}{\|x_i-x_j\|} \tag{3.12}$$

where m_i is the mass, x_i is the position and y_i is the momentum of the ith particle in a rotating coordinate system and $K=\left(\begin{smallmatrix}0&1\\-1&0\end{smallmatrix}\right)$. In order to consider the case when one particle is small, set $m_3=\varepsilon^\alpha$ where α is a positive number to be determined later and ε will be treated as a small parameter (we are not scaling at this point!). Making this substitution in (3.12) with $N=3$ and rewriting yields

$$H_3=\frac{\|y_3\|^2}{2\varepsilon^\alpha}-x_3^TKy_3-\sum_{i=1}^{2}\frac{\varepsilon^\alpha m_i}{\|x_i-x_3\|}+H_2. \tag{3.13}$$

Here the terms involving the third particle have been removed, leaving the Hamiltonian of the two-body problem as a remainder. Since ε is a small parameter which already measures the smallness of one mass, we should attempt to make ε also measure the deviation of the motion of the first two particles from a circular orbit. That is ε, or a power of ε, should measure not only the smallness of m_3, but also how close the first two particles come to a circular orbit. To accomplish this we must prepare the Hamiltonian H_2 so that one variable represents the deviation from a circular orbit. Actually part of this preparation has already been done, since in rotating coordinates a circular orbit appears as an equilibrium solutions. Let $Z=(x_1,x_2,y_1,y_2)$, so H_2 is a function of the 8-vector Z, and let $Z^*=(a_1,a_2,b_1,b_2)$ be a critical point of H_2, so $\nabla H_2(Z^*)=0$. (Later we shall give explicit values for the a's and b's, but for now it is enough to know that they exist.) By Taylor's theorem

$$H_2(Z)=H_2(Z^*)+\tfrac{1}{2}(Z-Z^*)S(Z-Z^*)+O\left(\|Z-Z^*\|^2\right) \tag{3.14}$$

where S is the Hessian of H_2 evaluated at Z^*. Since constant terms in the Hamiltonian drop out when the equations of motion are formed, we shall ignore $H_2(Z^*)$ by setting it to zero. If the motion of the first two particles is nearly circular, then $Z-Z^*$ should be small, so this suggests the scaling

$$Z-Z^*\to\varepsilon^{\beta}U \tag{3.15}$$

where U is a new variable and β is a positive number to be determined. So far we have implemented the assumptions that the third mass is small, that the deviation of the motion of the first two particles from a circular orbit is small, and that the smallness relationships is in the form of a power law. α and β have not been given yet, and so the precise relationship between the two small quantities is not yet established. This is the point at which symplectic scaling gives some guidance on how to proceed. Note first that (3.15) is a symplectic change of the U variables with multiplier $\varepsilon^{-2\beta}$; however, (3.15) is not a symplectic change of variables on the whole space since x_3 and y_3 have not been changed yet. The scaling (3.15) implies $x_1=a_1+O(\varepsilon^{\beta})$ and $x_2=a_2+O(\varepsilon^{\beta})$ where a_1 and a_2 are the constant vectors defined above, so x_1 and x_2 are order zero in ε. Since we are not interested in the case when x_3 is close to a_1 or a_2 (the collision problem) nor that when x_3 is large (the case of a comet), we shall take x_3 as order zero in ε also. Thus for a change of variables on the whole phase space to be symplectic, it is necessary that $y_3\to\varepsilon^{2\beta}\eta$. Thus we complete (3.15) with

$$x_3\to\xi,\qquad y_3\to\varepsilon^{2\beta}\eta. \tag{3.16}$$

Using (3.14), (3.15) and (3.16) in (3.13) yields

$$H_3=\varepsilon^{\alpha-2\beta}\frac{\|\eta\|^2}{2}-\xi^TK\eta-\varepsilon^{\alpha-2\beta}\sum_{i=1}^{2}\frac{m_i}{\|\xi-a_i\|}+\cdots+\frac{1}{2}U^TSU+\cdots. \tag{3.17}$$

In order to make the first and third terms in (3.17) just as important as the second and fourth, it is necessary to have $\alpha=2\beta$. Setting $\beta=1$, $\alpha=2$ gives a small integer solution of this relation. To summarize: if $m_3=\varepsilon^2$ then

$$x_3\to\xi,\quad y_3\to\varepsilon^2\eta,\quad Z-Z^*\to\varepsilon U \tag{3.18}$$

is a symplectic change of variables which reduces (3.13) to

$$H_3=\left\{\frac{\|\eta\|^2}{2}-\xi^TK\eta-\sum_{i=1}^{2}\frac{m_i}{\|\xi-a_i\|}\right\}+\frac{1}{2}U^TSU+O(\varepsilon). \tag{3.19}$$

The quantity in the braces above is the Hamiltonian of the restricted three-body problem if we take $m_1+m_2=1$, $m_1=\mu$, $m_2=1-\mu$, $a_1=(1-\mu,0)$ and $a_2=(-\mu,0)$. The quadratic term in U is simply the Hamiltonian of the linearization of the equations of motion of the two-body problem about the circular solution. For $\varepsilon=0$ the Hamiltonian H_3 is a sum of these two Hamiltonians, and so the equations of motion decouple. If $\xi=\phi(t)$, $\eta=\psi(t)$ is any solution of the restricted problem, then $\xi=\phi(t)$, $\eta=\psi(t)$, $U\equiv 0$ is a solution of the equations of motion defined by (3.18) with $\varepsilon=0$. Thus for bounded times, there are solutions of the full three-body problem of the form $\xi=\phi(t)+O(\varepsilon)$, $\eta=\psi(t)+O(\varepsilon)$ and $U=O(\varepsilon)$.

Looking at (3.18) we see that since y_3 is the momentum of the third particle and $m_3=\varepsilon^2$, the variable η is really the velocity of the third particle. Thus all the new quantities have been given physical meaning, and the relationship between the small quantities has been established.

The problem defined by the Hamiltonian (3.18) is still degenerate due to the fact that the original three-body problem admits symmetries and integrals. Specifically, the Hamiltonian H_3 is invariant under the full group of Euclidean motions of the plane and admits linear and angular momentum as integrals. Holding these integrals fixed and then identifying configurations which differ by a Euclidean motion only leads to a Hamiltonian on a reduced space. The details of this reduction are unimportant for the present discussion and are classical. It is enough to say that after this reduction is done the Hamiltonian (3.19) becomes

$$H_3=\left\{\frac{\|\eta\|^2}{2}-\xi^TK\eta-\sum_{i=1}^{2}\frac{m_i}{\|\xi-a_i\|}\right\}+\frac{1}{2}\{r^2+R^2\}+O(\varepsilon) \tag{3.20}$$

where r and R are scalar variables. See [9] for a complete discussion of this reduction. Thus if $\xi=\phi(t)$, $\eta=\psi(t)$ is a τ-periodic solution of the restricted problem with characteristic multipliers 1, 1, λ, λ^{-1}, then $\xi=\phi(t)$, $\eta=\psi(t)$, $r=R=0$ is a τ-periodic solution of the three-body problem defined by (3.20) with $\varepsilon=0$ with characteristic multipliers 1, 1, λ, λ^{-1}, $\exp\pm i\tau$. Thus if $\lambda\neq 1$ and $\tau\not\equiv 0 \bmod 2\pi$, this represents a nondegenerate τ-periodic solution of the three-body problem defined by (3.20) with $\varepsilon=0$.

Now the classical perturbation theorem applies to yield the theorem of Hadjidemetriou [3], namely, that any nondegenerate periodic solution of the restricted problem whose period is not a multiple of 2π can be continued into the full three-body problem.

There is another restricted three-body problem, known as Hill's lunar equations, which is derived under slightly different assumptions. The traditional description [5] of this equation is even more picturesque then the description of the restricted problem. One is asked to consider the motion of an infinitesimal body (the moon) which is attracted to a finite body (the earth) which is fixed at the origin of a rotating coordinate system. The coordinate system rotates so that the positive x-axis points to an infinite body (the sun) which is infinitely far away. The ratio of the two infinite quantities [sic] is taken so that the gravitational attraction of the sun on the moon is finite.

Briefly, I shall indicate how Hill's lunar equations can be derived from the three-body problem; for details see [12]. In this problem two masses m_1 and m_2 (the earth and moon) are small relative to the mass of the third (the sun). Also the distance between the earth and moon is small relative to the distance between their center of mass and the sun. The first assumption is easy to implement: simply set $m_1=\varepsilon^6\mu_1$, $m_2=\varepsilon^6\mu_2$ and $m_3=\mu_3$. (Here I have fixed the exponents since I already worked out what they should be.) In order to implement the second assumption, we must choose coordinates so that one variable represents the distance between the two bodies. A classical set of symplectic cordinates known as Jacobi coordinates has one coordinate which represents the distance between two of the bodies and so is the logical choice here. The Jacobi position vectors are u_0, the position of the center of mass of the triple; u_1, the position of particle 2 relative to particle 1; and u_2, the position of particle 3 relative to the center of mass of particles 1 and 2. The variables v_0, v_1 and v_2 are the corresponding momenta where v_0 is the total linear momentum of the system. Making the initial scaling

$$v_1 \to \varepsilon^6 v_1, \qquad v_2 \to \varepsilon^6 v_2 \tag{3.21}$$

as in the previous example and fixing the center of mass at the origin, $u_0=0$, and ignoring the total linear momentum v_0 leads to the following Hamiltonian for the full three-body problem:

$$\begin{aligned} &H_3=H'+H''+O(\varepsilon^6), \\ &H'=\frac{\|v_1\|^2}{2M_1}-u_1^TJv_1-\varepsilon^6\frac{\mu_0\mu_1}{\|u_1\|}, \\ &H''=\frac{\|v_2\|^2}{2M_2}-u_2^TJv_2-\frac{\mu_1\mu_2}{\|u_2-\nu_0u_1\|}-\frac{\mu_0\mu_2}{\|u_0+\nu_1u_1\|}. \end{aligned} \tag{3.22}$$

Here M_1, M_2, ν_0, ν_1 are all positive constants defined in terms of the original classes. The only property needed here is $\nu_0+\nu_1=1$.

The Hamiltonian H' contains only u_1 and v_1, the variables of the earth–moon pair, whereas the Hamiltonian H'' contains cross terms. Since u_1 is to be taken as a small quantity later, rewrite H'' as

$$\begin{aligned} &H''=H^*+H^{**}, \\ &H^*=\frac{\|v_2\|^2}{2M_2}-u_2^TJv_2-\frac{\mu_2(\mu_0+\mu_1)}{\|u_2\|}, \\ &H^{**}=\frac{\mu_2(\mu_0+\mu_1)}{\|u_2\|}-\frac{\mu_1\mu_2}{\|u_2-\nu_0u_1\|}-\frac{\mu_0\mu_1}{\|u_0+\nu_1u_1\|}. \end{aligned} \tag{3.23}$$

Now H^* contains only u_2 and v_2, the variables describing the motion of the earth–moon pair about the sun. Since this motion is assumed to be nearly circular, we set $Z=(u_2,v_2)$ and $Z^*=(a,b)$ as before so that

$$H^*(Z)=H^*(Z^*)+\tfrac{1}{2}(Z-Z^*)S(Z-Z^*)+\cdots. \tag{3.24}$$

Now the full set of physical assumptions can be affected by the following scaling:

$$u_1\to\varepsilon^2u_1, \qquad v_1\to\varepsilon^2v_1, \tag{3.25}$$

$$Z-Z_0\to\varepsilon^2U. \tag{3.26}$$

Scaling (3.25) says that the distance between the earth and the moon is small, and scaling (3.26) says that the earth–moon system moves about the sun in a nearly circular orbit. This scaling is symplectic with multiplier ε^{-4} and greatly simplifies the problem. The Hamiltonian H^{**} is not completely ignorable, though, so it must be expanded in a series of Legendre polynomials, the details of which are not appropriate here.

The end result is that the Hamiltonian of the full three-body problem becomes, under these assumptions,

$$H_3=\left\{\frac{\|\eta\|^2}{2}-\xi^TJ\eta-\frac{1}{\|\xi\|}+\left(3\xi_1^2-\|\xi\|^2\right)\right\}+\frac{1}{2}U^TSU+O(\varepsilon^2) \tag{3.27}$$

where ξ and η are essentially u_1 and v_1 (the variables describing the motion of the earth–moon system) and U measures the deviation from a circular orbit of the motion of the earth–moon system around the sun. The quantity in the braces in (3.27) is the Hamiltonian for Hill's lunar equations, and the last expression in parenthesis comes from H^{**}. As with the restricted problem, it is easy to prove that any nondegenerate periodic solution of Hill's lunar equation whose period is not a multiple of 2π can be continued into the full three-body problem. See [12] for a complete account of this derivation, including the details of the expansion of H^{**} in Legendre polynomials.

4. The critical cases. Scaling is particularly useful in the critical cases, since it is usually not obvious which terms in the Taylor expansion of the Hamiltonian are important for the perturbation analysis. The correct scaling not only defines the main problem, but also orders all the terms according to the strength of their influence on the problem at hand. Obviously, since the critical case is the complement of a nice case, further subdivision is necessary. Also, there will always be a system which is so degenerate that it does not fall into any of the previously defined subcases. This section defines what I consider to be the first critical subcase for Hamiltonian systems. This subcase is defined as all systems which can be analyzed by Lemma 4.1. The only new tool necessary to prove this lemma is the variation of constants formula, and so the proof is not much more difficult than the proof of Theorem 3.1.

The lemma deals with a Hamiltonian system of the form

$$\dot{z}=\nabla H(z,\varepsilon)=Az+\varepsilon f(z,\varepsilon) \tag{4.1}$$

where $z\in\mathbb{R}^{2n}$, $\varepsilon\in\mathbb{R}$, A is a $2n\times 2n$ nonsingular matrix such that $\exp AT=I$ for some $T>0$, and f is a smooth function. Since $\exp AT=I$, all solutions of (4.1) when $\varepsilon=0$ are T-periodic and all their characteristic multipliers are $+1$. Thus when $\varepsilon=0$ the system fails to satisfy the hypotheses of the perturbation theorem of the previous section. In order to restrict the level of degeneracy of this system, some condition must be placed on the higher order terms represented by f.

Let β be a real parameter and define

$$B(\beta,\zeta)=\beta A\zeta+\int_0^T e^{-As}f(e^{As}\zeta,0)\,ds. \tag{4.2}$$

The function B (sometimes called the describing function) is defined entirely in terms of the known quantities A and f and does not depend on the unknown solutions of (4.1). The first critical subcase is defined by:

LEMMA 4.1. *If there exist smooth functions $\zeta(\alpha)$, $\beta(\alpha)$, where α is real and $\zeta(\alpha)\in\mathbb{R}^n$, $\beta(\alpha)\in\mathbb{R}$, such that*

i) $$B(\beta(\alpha),\zeta(\alpha))=0,$$

$$\text{ii)} \qquad \operatorname{rank}\left(\frac{\partial B}{\partial \beta}, \frac{\partial B}{\partial \zeta}\right)(\beta(\alpha), \zeta(\alpha)) = 2n-1$$

for $|\alpha| \leq \alpha_0$, *then there exists a smooth 2-parameter family of periodic solutions of* (4.2), *denoted by* $\phi(t, \alpha, \varepsilon)$, *such that*

$$\text{iii)} \qquad \phi(t,\alpha,\varepsilon) \textit{ is } T(\alpha,\varepsilon) \textit{ periodic for } \varepsilon \textit{ small and } |\alpha| \leq \alpha_0,$$

$$\text{iv)} \qquad \phi(t,\alpha,0) = (\exp At)\zeta(\alpha),$$

$$\text{v)} \qquad T(\alpha,\varepsilon) = T + \varepsilon\beta(\alpha) + O(\varepsilon^2).$$

The details of the proof can be found in [11]. The essential step in the proof is a simple calculation of the general solution of (4.2). Let $\psi(t, \zeta_0, \varepsilon)$ be the solution of (4.1) which satisfies $\psi(0, \zeta_0, \varepsilon) = \zeta_0$ and seek the periodic solutions whose period is $T + \varepsilon\beta$. From the variation of constants formula

$$\psi(T+\varepsilon\beta, \zeta_0, \varepsilon) = \zeta_0 + \varepsilon B(\beta, \zeta_0) + O(\varepsilon^2),$$

so the problem of finding an initial condition leading to a periodic solution is just solving

$$B(\beta, \zeta_0) + O(\varepsilon) = 0.$$

One simply applies the implicit function theorem to this system of $2n$ equations to solve $2n-1$ of the equations and then uses the integral H to show that the last equation is also satisfied.

As the first example of how scaling can be used to reduce a problem to a system where this lemma applies, consider the restricted three-body problem where the small mass is far from the primaries. The Hamiltonian of the restricted three-body problem is

$$H = \frac{\|\eta\|^2}{2} - \xi^T K \eta - \sum_1^2 \frac{m_i}{\|\xi - a_i\|} \tag{4.3}$$

where the notation is the same as in (3.19). The equations of motion are

$$\dot{\xi} = K\xi + u, \qquad \dot{\eta} = K\eta + \sum_1^2 \frac{m_i(a_i - \xi)}{\|a_i - \xi\|^3}. \tag{4.4}$$

In order to study this problem for large ξ, scale by $\xi \to \varepsilon^{-2}\xi$ and $\eta \to \varepsilon\eta$. This is a symplectic scaling with multiplier ε, so the Hamiltonian becomes

$$H = -\xi^T K \eta + \varepsilon^3 \left\{ \frac{\|\eta\|^2}{2} - \frac{1}{\|\xi\|} \right\} + O(\varepsilon^5), \tag{4.5}$$

and the equations of motion become

$$\dot{\xi} = K\xi + \varepsilon^3 \eta + O(\varepsilon^5), \qquad \dot{\eta} = K\eta + \varepsilon^3 \|\xi\|^{-3} \xi + O(\varepsilon^5). \tag{4.6}$$

To lowest order in ε, these equations are linear and the general solution is $\xi = (\exp Kt)\xi_0$, $\eta = (\exp Kt)\eta_0$. So if $z = (\xi, \eta)$, $A = \operatorname{diag}(K, K)$, the system (4.6) is of the form (4.1) with ε^3 replacing ε. Since $\exp Kt$ is the rotation matrix by an angle t for small ε, the solutions are nearly circular with periods near 2π. Since rotating coordinates are being used, this means that near infinity the infinitesimal body mainly feels the effect of the Coriolis and centrifugal forces, and in a fixed frame it would be nearly at rest. The coefficient of the ε^3-term is the Hamiltonian of the Kepler problem where the central body has mass 1 (we have assumed that the sum of the masses of the primaries is 1). This can be interpreted as meaning that the next most important force felt by the infinitesimal body is the attraction of a fixed body at the center of mass of the two primaries whose mass is equal to the sum of the masses of the two primaries.

The function B in (4.2) is easy to compute. Setting $\zeta=(\xi,\eta)$, $B=0$ becomes

$$\beta K\xi+2\pi\eta=0, \qquad \beta K\eta-2\pi\frac{\xi}{\|\xi\|^3}=0. \tag{4.7}$$

It is not difficult to analyze these equations and show that Lemma 4.1 applies. (The details are found in [10].) The main conclusion of this analysis is that the restricted three-body problem has two families of nearly circular orbits of large radius.

The last example illustrates the proper method of scaling when one encounters nonelementary divisors in a matrix. The restricted three-body problem always has two equilibrium points which are at the vertices of an equilateral triangle, one of whose sides is the line segment joining the two primaries. The linearized equations about this equilibrium point consist of two harmonic oscillators when the mass ratio is small, and form a complex saddle when the mass ratio is near $\frac{1}{2}$. There is one specific value of the mass ratio where the linearized equation has two equal pairs of imaginary eigenvalues and the Jordan canonical form for the coefficient matrix has off-diagonal elements. When restricting to symplectic similarity transformations, the canonical form for the linearized system is

$$\begin{pmatrix} \omega i & 1 & 0 & 0 \\ 0 & \omega i & 0 & 0 \\ 0 & 0 & -\omega i & 0 \\ 0 & 0 & -1 & -\omega i \end{pmatrix} \tag{4.8}$$

with certain reality conditions.

After some preparation the Hamiltonian is of the form

$$H=i\omega(z_1z_3+z_2z_4)+z_2z_3+\left(a_1z_1^2z_3^2+a_2z_1^2z_3z_4+a_3z_1^2z_4^2\right)+\cdots \tag{4.9}$$

where the z's are complex coordinates satisfying the reality conditions $\bar{z}_1=-z_4$, $\bar{z}_2=z_3$. The best scaling for this problem will push the off-diagonal terms in the matrix into the higher order terms. Introducing a small parameter ε and scaling by

$$z_1\to\varepsilon z_1, \quad z_2\to\varepsilon^2z_2, \quad z_3\to\varepsilon^2z_3, \quad z_4\to\varepsilon z_4 \tag{4.10}$$

accomplishes this task. The scaling in (4.10) is symplectic with multiplier ε^{-3}, and so the Hamiltonian becomes

$$H=i\omega(z_1z_3+z_2z_4)+\varepsilon\left(z_2z_3+a_3z_1^2z_4^2\right)+\cdots. \tag{4.11}$$

The equations of motion implied by (4.11) are in the form (4.1), and the function B in (4.2) is easy to compute and analyze. The proper scaling in this problem has simplified not only the zeroth order terms but the first order terms as well, and this greatly simplifies the analysis. Applying the lemma to this problem establishes the existence of two families of periodic solution which emanate from this equilibrium point for the restricted three-body problem. The reader is referred to [11] where this problem and a more interesting one are discussed in detail.

There are many other interesting problems where scaling can greatly ease the analysis. The examples given here were chosen to illustrate a variety of different situations where scaling can help, without getting us too deeply involved in the technical aspects of the problem. The main point of this survey is to demonstrate how the correct scaling is obtained when suddenly the equations are greatly simplified, and suddenly (but after the fact) it is obvious why you should use that scaling.

REFERENCES

[1] R. ABRAHAM AND J. MARSDEN, *Foundations of Mechanics*, 2nd ed., Benjamin/Cummings, Reading, MA, 1978.

[2] A. DEPRIT AND J. HENRARD, *A manifold of periodic solutions*, Adv. Astron. Astrophys., 6 (1968), pp. 12 ff .

[3] J. D. HADJIDEMETRIOU, *The continuation of periodic orbits from the restricted to the general three-body problem*, Celestial Mech., 12 (1975), pp. 155–174.

[4] J. K. HALE, *Ordinary Differential Equations*, Wiley-Interscience, New York, 1969.

[5] G. W. HILL, *Researches in the lunar theory*, Amer. J. Math., 1 (1878), pp. 5–26, 129–147, 245–260.

[6] A. KELLY, *On the Liapunov sub-center manifold*, J. Math. Anal. Appl., 18 (1967), pp. 472–478.

[7] I. KUPKA, *Contribution à la théorie des champs génériques*, Contribution to Diff. Eqs., 2 (1963), pp. 457–484.

[8] V. V. NEMYTSKII AND V. V. STEPANOV, *Qualitative Theory of Differential Equations*, Princeton Univ. Press, Princeton, NJ, 1960.

[9] K. R. MEYER, *Periodic solutions of the N-body problem*, J. Differential Equations, 39 (1981), pp. 2–38.

[10] ———, *Periodic orbits near infinity in the restricted N-body problem*, Celestial Mech., 23 (1981), pp. 69–81.

[11] K. R. MEYER AND D. S. SCHMIDT, *Periodic orbits near L_4 for mass ratios near the critical mass ratio of Routh*, Celestial Mech., 4 (1971), pp. 99–109.

[12] ———, *Hill's lunar equations and the three-body problem*, J. Differential Equations, 44 (1982), pp. 1–10.

[13] H. POLLARD, *Mathematical Introduction to Celestial Mechanics*, Prentice-Hall, Englewood Cliffs, NJ, 1966.

[14] H. POINCARE, *Les méthodes nouvelles de la mécanique celeste*, Gauthier-Villar, Paris, 1892.

[15] C. ROBINSON, *Generic properties of conservative systems* I, II, Amer. J. Math., 92, pp. 562–603, 897–906.

[16] L. SEGEL, *Simplification and scaling*, SIAM Rev., 14 (1972), pp. 547–571.

[17] A. WINTNER, *The Analytic Foundations of Celestial Mechanics*, Princeton Univ. Press, Princeton NJ, 1941.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics
003

ASYMPTOTIC INTEGRABILITY AND PERIODIC SOLUTIONS OF A HAMILTONIAN SYSTEM IN 1:2:2-RESONANCE*

ELS VAN DER AA† AND FERDINAND VERHULST†

Abstract. A Hamiltonian system in 1 : 2 : 2-resonance, normalized to degree three admits three integrals, which represent three asymptotic integrals of the original system valid for all time. In a detuned system we find in this way only two integrals. On analyzing the periodic solutions we find an infinite set, a global bifurcation, which is expected to break up in higher order approximation. To demonstrate this phenomenon we study an example for which we calculate the normal form to degree four, i.e. the second order asymptotic approximation, which produces a break-up of the infinite set into four periodic solutions. In the last section we demonstrate that these results carry through for Hamiltonian systems with n degrees of freedom in $1:2:\cdots:2$-resonance.

1. Introduction. Some remarkable properties of three degrees of freedom potential problems have been discovered by Martinet, Magnenat and Verhulst (1981), results subsequently generalized by van der Aa (1983) for Hamiltonian systems. In these papers one studies a system in 1 : 2 : 2-resonance to find that the system is asymptotically integrable upon normalizing to degree three. Also one finds a global bifurcation i.e. an infinite set of periodic solutions for each (small) value of the energy. The term global bifurcation refers here to an iso-energetic family of periodic solutions as opposed to the local bifurcation of isolated periodic solutions which represents the generic case. In a different context global bifurcations are sometimes called vertical bifurcations. Both the asymptotic integrability and the global bifurcation are nongeneric phenomena which call for further investigation.

In §2 we present the general Hamiltonian in 1 : 2 : 2-resonance, normalized to degree three. We derive three independent integrals of the normalized system, which are asymptotic integrals of the original system valid for all time. The analysis is repeated in §3 for detuned systems (resonance ratios neighbouring 1 : 2 : 2). In this case we still have two asymptotic integrals. We find that for detuned systems in general no quadratic or cubic third integral can be found. In §4 we present the periodic orbits of the normalized system at exact resonance together with the asymptotic behaviour of the solutions with time. It is one of the advantages of combining normalization procedures with the theory of asymptotic approximations to have rigorous estimates of validity of the results. The various methods involve long but straightforward calculations which were carried out by hand.

The asymptotic integrals and the periodic solutions can be used to obtain a geometric picture of the phase flow on the energy manifold; this geometric analysis will not be carried out here. Such an analysis should contain all the possible bifurcations and its construction is by no means a simple exercise.

To understand the global bifurcation phenomena in these systems we introduce a particular potential problem, which is characteristic for what happens in general. We take

$$H=H_2+\varepsilon\left(\alpha_1x^2y+\alpha_2x^2z+\alpha_3xyz\right).$$

*Received by the editors December 7, 1982, and in revised form June 1, 1983.

† Mathematisch Instituut, Rijksuniversiteit Utrecht, 3508 TA Utrecht, the Netherlands.

On normalizing to degree four, in other words, on calculating the second order asymptotic approximation we find that the infinite set of periodic solutions breaks up into four isolated periodic solutions. This phenomenon is reminiscent of the critical inclination problem in celestial mechanics. In §6 we have a final surprise: the analysis of the preceding sections carries through for n degrees of freedom Hamiltonian systems in $1:2:\cdots:2$-resonance. We present the n asymptotic integrals in this case and the periodic solutions obtained from the system normalized to degree three. Note that a survey of the asymptotic analysis of Hamiltonian systems has been given by Verhulst (1983). In this survey paper one can also find a discussion of the asymptotic integrability of three degrees of freedom systems as it is known at present.

2. Asymptotic integrals. A remarkable property of a Hamiltonian system in $1:2:2$-resonance is that it is asymptotically integrable in the following sense. Suppose the Hamiltonian H is written in Euclidean coordinates $q(x,y,z)$ and corresponding momenta $p(p_x,p_y,p_z)$ while H can be expanded in homogeneous polynomials H_n, $n=2,3,\cdots$ as follows

$$H=H_2+\varepsilon H_3+\varepsilon^2H_4+\varepsilon^3H_5+\cdots$$

with nondegenerate

$$H_2=\frac{1}{2}(x^2+p_x^2)+\frac{1}{2}(4y^2+p_y^2)+\frac{1}{2}(4z^2+p_z^2). \tag{1}$$

In the sequel we shall use various coordinate systems all of which are defined by canonical transformations.

Normalization to order 3 leads to the Hamiltonian

$$\bar{H}=H_2+\varepsilon\bar{H}_3.$$

The system corresponding with $\bar{H}$ is integrable, i.e. three independent integrals (in involution) of the normalized system exist. For potential problems this result has been presented by Martinet, Magnenat and Verhulst (1981); for general Hamiltonian systems the proof has been given by van der Aa (1983). As we shall extend this result to systems with more than three degrees of freedom using the same method we summarize the proof.

We have a general homogeneous polynomial of the third degree which thus contains 56 terms. Almost each of these terms vanishes during the normalization process. We find that the flow of the dynamical system is, in first approximation, governed only by 12 different terms

$$\begin{aligned}H_3=&b_1x^2y+b_2x^2p_y+b_3xyp_x+b_4xp_xp_y+b_5yp_x^2+b_6p_x^2p_y\\&+b_7x^2z+b_8x^2p_z+b_9xzp_x+b_{10}xp_xp_z+b_{11}zp_x^2+b_{12}p_x^2p_z.\end{aligned} \tag{2}$$

On these terms we carry out the normalization process and $\bar{H}$ becomes

$$\begin{aligned}\bar{H}=H_2+\varepsilon\big[&2a_1\{y(x^2-p_x^2)+xp_xp_y\}+a_2\{p_y(p_x^2-x^2)+4xyp_x\}\\&+2a_3\{z(x^2-p_x^2)+xp_xp_z\}+a_4\{p_z(p_x^2-x^2)+4xzp_x\}\big],\end{aligned} \tag{3}$$

where

$$a_1=\frac{1}{8}(b_1+2b_4-b_5), \qquad a_3=\frac{1}{8}(b_7+2b_{10}-b_{11}),$$
$$a_2=-\frac{1}{8}(2b_2-b_3+2b_6), \qquad a_4=-\frac{1}{8}(2b_8-b_9+2b_{12}).$$

So, starting with a particular Hamiltonian system with an explicitly given cubic function H_3, one can immediately write down the corresponding normalized Hamiltonian by substitution of the parameter-values. We shall use expression (3) with parameters a_1, a_2, a_3, a_4 for further computations.

We have $\dot{x}=\partial H/\partial p_x$, $\dot{p}_x=-\partial H/\partial x$ etc. and the normalized equations of motion become

$$\begin{aligned}(4)\quad &\ddot{x}+x=-4\varepsilon[a_1(2xy+\dot{x}\dot{y})+a_2(2y\dot{x}-x\dot{y})+a_3(2xz+\dot{x}\dot{z})+a_4(2z\dot{x}-x\dot{z})],\\ &\ddot{y}+4y=-4\varepsilon[a_1(x^2-\dot{x}^2)+2a_2x\dot{x}],\\ &\ddot{z}+4z=-4\varepsilon[a_3(x^2-\dot{x}^2)+2a_4x\dot{x}].\end{aligned}$$

Two independent integrals of this system can be found immediately: $\overline{H}$ and H_2. $\overline{H}$ corresponds, by transformation, with the exact integral H of the original system and we have clearly

$$H-\overline{H}=O(\varepsilon^2) \quad \text{for } t\geq 0.$$

At the same time we note that

$$H-H_2=O(\varepsilon) \quad \text{for } t\geq 0.$$

Instead of the independent integrals $\overline{H}$ and H_2 of system (4) we may use $\overline{H}_3$ and H_2; these integrals are asymptotic integrals of the original system (before normalization), valid for all time. It is possible to find a third independent integral of system (4). Introduce the transformation

$$(5)\qquad u=a_1y-\frac{1}{2}a_2\dot{y}+a_3z-\frac{1}{2}a_4\dot{z}.$$

Differentiating twice and using equations (4) we find

$$(6)\qquad \begin{aligned}&\ddot{x}+x=-4\varepsilon(2xu+\dot{x}\dot{u}),\\ &\ddot{u}+4u=-4\varepsilon\alpha(x^2-\dot{x}^2),\end{aligned} \quad \text{with } \alpha=a_1^2+a_2^2+a_3^2+a_4^2.$$

Equations (6) constitute a two degrees of freedom system which can be made Hamiltonian by putting $\bar{x}=\alpha^{1/2}x$.

For such systems in normal form two independent integrals exist: the Hamiltonian $H^*(x,u,\dot{x},\dot{u})$ and a quadratic integral I, which plays the part of H_2 for system (6). Rewritten in the original coordinates it can be checked that $\overline{H}$, H^*, H_2 and I are not independent but $\overline{H}$, H_2 and I are. I reads

$$(7)\qquad I=\frac{1}{2}\alpha(x^2+\dot{x}^2)+\frac{1}{2}\left[4\left(a_1y-\frac{1}{2}a_2\dot{y}+a_3z-\frac{1}{2}a_4\dot{z}\right)^2 +(a_1\dot{y}+2a_2y+a_3\dot{z}+2a_4z)^2\right].$$

I is a third (asymptotic) integral of the original system which is conserved with error $O(\varepsilon)$ for all time. The validity for all time is connected with the existence of invariant manifolds, tori, in phase space around the periodic solutions. We shall discuss these periodic solutions in §4.

The existence of a third asymptotic integral valid for all time is even more surprising as we shall find that system (4) is structurally unstable in the sense that a certain family of periodic solutions can be perturbed away by introducing higher order terms, see §5.

3. Detuning (versal deformation) of the resonance. In general a system will never be at exact resonance and it is natural to allow for small detuning of the frequencies. For the left-hand sides of the equations of motion we put

$$\ddot{x}+x, \quad \ddot{y}+4(1+\delta_1(\varepsilon))y, \quad \ddot{z}+4(1+\delta_2(\varepsilon))z,$$

where $\delta_1(\varepsilon)$, $\delta_2(\varepsilon)=O(\varepsilon)$.

The normalization process leads to the Hamiltonian

$$\overline{H}_d=H_2+\varepsilon\overline{H}_{d3}$$

where H_2 is given again by (1), while

$$\overline{H}_{d3}=\overline{H}_3+\frac{\delta_1(\varepsilon)}{4\varepsilon}\left(4y^2+p_y^2\right)+\frac{\delta_2(\varepsilon)}{4\varepsilon}\left(4z^2+p_z^2\right). \tag{8}$$

$\overline{H}_3$ has been defined in §2.

As in §2, H_2 and $\overline{H}_{d3}$ are asymptotic integrals valid for all time. The third independent asymptotic integral I (found in §2) is not valid for this system. In fact no quadratic or cubic polynomial can be used as an additional independent integral, but of course it is still possible that another algebraic integral exists. Only in the special case that the linear vectorfield is detuned in the same way in the second and third degree of freedom (i.e. $\delta_1=\delta_2$), we find that the system is again asymptotically integrable. This case however may be rare in applications. The proof runs as follows.

Consider the system of second order differential equations

$$\begin{aligned} \ddot{x}+x&=-4\varepsilon\{a_1(2xy+\dot{x}\dot{y})+a_2(2y\dot{x}-x\dot{y}) \\ &\qquad +a_3(2xz+\dot{x}\dot{z})+a_4(2z\dot{x}-x\dot{z})\}, \\ \ddot{y}+4(1+\delta_1)y&=-4\varepsilon\{a_1(x^2-\dot{x}^2)+2a_2x\dot{x}\}, \\ \ddot{z}+4(1+\delta_2)z&=-4\varepsilon\{a_3(x^2-\dot{x}^2)+2a_4x\dot{x}\}, \end{aligned} \tag{9}$$

then we find, by differentiating I from (7)

$$\begin{aligned} \dot{I}=&-4\delta_1(a_1^2+a_2^2)y\dot{y}-4\delta_2(a_3^2+a_4^2)z\dot{z}+8(\delta_1-\delta_2)(a_2a_3-a_1a_4)yz \\ &-4\delta_1(a_1a_3+a_2a_4)y\dot{z}-4\delta_2(a_1a_3+a_2a_4)z\dot{y}. \end{aligned}$$

And so, as expected, for $\delta_1, \delta_2\neq 0$ I is not an integral anymore.

The next step is to check the existence of an integral which can be expanded in a finite series of polynomials. Such a calculation is easier to perform using complex coordinates.

We define

$$x=\frac{1}{2}(x_1+y_1),\qquad y=\frac{1}{4}\sqrt{2}\,(x_2+y_2),\qquad z=\frac{1}{4}\sqrt{2}\,(x_3+y_3),$$
$$p_x=-\frac{1}{2}i(x_1-y_1),\quad p_y=-\frac{1}{2}i\sqrt{2}\,(x_2-y_2),\quad p_z=-\frac{1}{2}i\sqrt{2}\,(x_3-y_3),$$

then by normalization

$$(10)\quad H_d=\frac{1}{2}x_1y_1+x_2y_2+x_3y_3 +\varepsilon\left[\frac{\delta_1}{2\varepsilon}x_2y_2+\frac{\delta_2}{2\varepsilon}x_3y_3+h_1x_1^2y_2+\bar{h}_1y_1^2x_2+h_2x_1^2y_3+\bar{h}_2y_1^2x_3\right]$$

where

$$h_1=\frac{1}{2}\sqrt{2}\,(a_1-ia_2)\quad\text{and}\quad h_2=\frac{1}{2}\sqrt{2}\,(a_3-ia_4).$$

Suppose there exists a *quadratic* asymptotic integral I_1; then I_1 must satisfy the two involution conditions

$$(11)\qquad \{I_1,H_2\}=0,\qquad \{I_1,\bar{H}_{d3}\}=0,$$

with H_2 and $\bar{H}_{d3}$ as defined above.

It is easy to verify that I_1 must be of the following form in order to satisfy the first condition

$$(12)\qquad I_1=\beta_1x_1y_1+\beta_2x_2y_2+\beta_3x_3y_3+\beta_4x_2y_3+\beta_5y_2x_3,$$

where $\beta_1,\cdots,\beta_5$ are arbitrary constants.

Checking out the second condition produces

$$\begin{aligned}\{I_1,\bar{H}_{d3}\}=&\,x_1^2y_2(-2\beta_1h_1+\beta_2h_1+\beta_5h_2)+y_1^2x_2(2\beta_1\bar{h}_1-\beta_2\bar{h}_1-\beta_4\bar{h}_2)\\&+x_1^2y_3(-2\beta_1h_2+\beta_4h_1+\beta_3h_2)+y_1^2x_3(2\beta_1\bar{h}_2-\beta_5\bar{h}_1-\beta_3\bar{h}_2)\\&+\frac{1}{2\varepsilon}x_2y_3\cdot\beta_4(\delta_1-\delta_2)-\frac{1}{2\varepsilon}y_2x_3\cdot\beta_5(\delta_1-\delta_2).\end{aligned}$$

It is clear that a necessary condition to find an independent integral is $\delta_1=\delta_2=\delta$. Putting the coefficients of the remaining four terms equal to zero produces equations for $\beta_1,\cdots,\beta_5$ which we can solve. The integral (12) becomes

$$(13)\qquad I_1=-h_2\bar{h}_2x_2y_2-h_1\bar{h}_1x_3y_3+\bar{h}_1h_2x_2y_3+h_1\bar{h}_2y_2x_3\quad\text{and}$$

$\bar{h}$ means here complex conjugate of h; in Euclidean coordinates

(13a)

$$I_1=-\frac{1}{4}(a_3^2+a_4^2)(4y^2+\dot{y}^2)-\frac{1}{4}(a_1^2+a_2^2)(4z^2+\dot{z}^2)+\frac{1}{2}(a_1a_3+a_2a_4)(4yz+\dot{y}\dot{z})+(a_1a_4-a_2a_3)(z\dot{y}-y\dot{z}).$$

It is also possible to derive in this case ($\delta_1=\delta_2$) a third integral $\tilde{I}_1$ directly from the equations of motion (9). Using the method described in §2, we put again $u=a_1y-1/2a_2\dot{y}+a_3z-1/2a_4\dot{z}$ and we find a detuned two degrees of freedom system. The energy integral corresponding with this system reads

$$\tilde{I}=\frac{1}{2}\left(a_1^2+a_2^2+a_3^2+a_4^2\right)\left(x^2+\dot{x}^2\right)+\frac{1}{2}\left\{4(1+\delta)u^2+\dot{u}^2\right\} \tag{14}$$

or

$$\begin{aligned}\tilde{I}=&\frac{1}{2}\left(a_1^2+a_2^2+a_3^2+a_4^2\right)\left(x^2+\dot{x}^2\right)\\&+\frac{1}{2}\left[4(1+\delta)\left(a_1y-\frac{1}{2}a_2\dot{y}+a_3z-\frac{1}{2}a_4\dot{z}\right)^2\right.\\&\left.+\left(a_1\dot{y}+2a_2(1+\delta)y+a_3\dot{z}+2a_4(1+\delta)z\right)^2\right].\end{aligned} \tag{14a}$$

Comparison of the integrals I_1 and $\tilde{I}$ shows their dependence

$$\tilde{I}=2\left(I_1+\frac{1}{4}\left(a_1^2+a_2^2+a_3^2+a_4^2\right)H_2\right)+O(\delta(\varepsilon)).$$

We have seen that if a third independent integral exists for $\delta_1\neq\delta_2$, it cannot start with independent quadratic terms. So we try a *cubic* polynomial I_2. From the first condition $\{I_2,H_2\}=0$ it follows that I_2 must have the same generators as $\overline{H}_3$

$$I_2=h_3x_1^2y_2+h_4y_1^2x_2+h_5x_1^2y_3+h_6y_1^2x_3, \tag{15}$$

where $h_3,\cdots,h_6$ must be chosen such that I_2 satisfies the second condition $\{I_2,\overline{H}_{d3}\}=0$. We have

$$\begin{aligned}\{I_2,\overline{H}_{d3}\}=&\frac{\delta_1}{2\varepsilon}h_4y_1^2x_2-\frac{\delta_1}{2\varepsilon}h_3x_1^2y_2+\frac{\delta_2}{2\varepsilon}h_6y_1^2x_3-\frac{\delta_2}{2\varepsilon}h_5x_1^2y_3\\&+x_1^2y_1^2\left(h_1h_4-\bar{h}_1h_3+h_2h_6-\bar{h}_2h_5\right)+x_1y_1x_2y_2\left(4\bar{h}_1h_3-4h_1h_4\right)\\&+x_1y_1x_3y_3\left(4\bar{h}_2h_5-4h_2h_6\right)+x_1y_1y_2x_3\left(4\bar{h}_2h_3-4h_1h_6\right)\\&+x_1y_1x_2y_3\left(4\bar{h}_1h_5-4h_2h_4\right).\end{aligned}$$

The right-hand side can only vanish if $\delta_1=\delta_2=0$.

An independent cubic third integral obviously does not exist either. Of course we could continue with a quartic polynomial and so on but it is quite unpredictable if a polynomial fits the conditions and if so, what degree will it have? If the question of a polynomial integral would be answered in a negative way, there would still remain the quest for an analytic integral and after that, maybe, nonanalytic integrals. These are important open problems, probably requiring a different approach.

4. Periodic orbits. Apart from invariant manifolds, periodic solutions play a key role in understanding the flow of a dynamical system. In discussing periodic orbits we should keep in mind that we have in fact one-parameter families of periodic solutions as periodic solutions are being discussed here for fixed values of the energy.

As remarked in §2, H_2, given by equation (1), represents an integral of system (4) and at the same time an $O(\varepsilon)$ approximation of the energy integral of the original problem, the flow induced by H.

We shall briefly discuss the periodic solutions found by van der Aa (1983), where we shall add the asymptotic approximations for the solutions as a function of time. After that we study the global bifurcation arising in this resonance problem.

In the approximations of the solutions with time given in the sequel we have the following estimate of asymptotic validity. Denoting the periodic solution associated with the normalized system (4) by $\bar{q}\in\mathbb{R}^3$ and the corresponding solution of the system induced by H before normalization by $q\in\mathbb{R}^3$, we have $q(t)-\bar{q}(t)=O(\varepsilon)$ on the time-scale $1/\varepsilon$.

a. Normal modes. There are two normal modes, one in the y-direction and another in the z-direction which are both unstable. As can be seen from system (4) these solutions are harmonic. We may represent them as

$$\text{(16a)}\qquad \bar{x}(t)=\bar{z}(t)=0,\qquad \bar{y}(t)=\frac{1}{2}\sqrt{2E}\sin(\varphi_2(0)+2t),$$

and

$$\text{(16b)}\qquad \bar{x}(t)=\bar{y}(t)=0,\qquad \bar{z}(t)=\frac{1}{2}\sqrt{2E}\sin(\varphi_3(0)+2t).$$

The similarity between these orbits clearly corresponds with the symmetry in the frequency-ratio.

b. Periodic orbits in general position. Two periodic solutions can be found with $x(t)$, $y(t)$, $z(t)$ not identically zero. We find the Euclidean coordinate $q(x,y,z)$ as a function of time

$$\text{(17)}\qquad \begin{aligned}\bar{x}(t)&=\frac{2}{3}\sqrt{3E}\sin\left(\varphi_1(0)+t+\frac{2}{3}\varepsilon t\sqrt{3(a_1^2+a_3^2)E}\cos k\pi\right),\\ \bar{y}(t)&=\frac{1}{6}a_1\sqrt{6(a_1^2+a_3^2)^{-1}E}\sin\left(\varphi_2(0)+2t+\frac{4}{3}\varepsilon t\sqrt{3(a_1^2+a_3^2)E}\cos k\pi\right),\\ \bar{z}(t)&=\frac{1}{6}a_3\sqrt{6(a_1^2+a_3^2)^{-1}E}\sin\left(\varphi_3(0)+2t+\frac{4}{3}\varepsilon t\sqrt{3(a_1^2+a_3^2)E}\cos k\pi\right),\end{aligned}$$

where

$$\text{(17a)}\qquad \begin{aligned}&\varphi_2(0)=-k\pi+2\varphi_1(0)-a_2,\\ &\varphi_3(0)=-k\pi+2\varphi_1(0)-a_4,\\ &k=0,1.\end{aligned}$$

At these orbits the integral $\bar{H}_3$ assumes its relative maximum and minimum value with respect to H_2. So $\bar{H}_3$ can be used as a Lyapunov function to prove the stability of these periodic solutions.

For more details on the normal modes and the orbits in general position the reader is referred to van der Aa (1983).

c. The global bifurcation. Inspection of system (4) reveals that we find a whole *family* of periodic solutions in the subspace $x=p_x=0$. This is remarkable as this happens in the general 1 : 2 : 2-Hamiltonian normalized till H_3 whereas up till now this phenomenon was only known in the case of more special Hamiltonians, where a small cubic Hamiltonian perturbation destroys the phenomenon. We mention two important examples. First the critical inclination problem; Cushman (1982) has shown that by using higher order normalization there is a break-up of the family into two stable and two unstable periodic solutions. The second example is also a two degrees of freedom problem, the Hénon–Heiles Hamiltonian. In this case the description of the global bifurcation has been given by Verhulst (1979); again this family breaks up into four periodic solutions as has been shown by Churchill, Kummer and Rod (1981).

In our case the family of periodic solutions is described by

$$
\begin{aligned}
&\bar{x}(t)=0,\\
&\bar{y}(t)=\frac{1}{2}\sqrt{2C_1}\sin(\varphi_2(0)+2t),\\
&\bar{z}(t)=\frac{1}{2}\sqrt{2C_2}\sin(\varphi_3(0)+2t),
\end{aligned}
\tag{18}
$$

with $C_1+C_2=E$.

The stability analysis of these solutions produces two eigenvalues zero, and in general one eigenvalue positive, one negative. So if the solutions exist, they are unstable. There is one exception with four eigenvalues zero; the coordinates are in this case

$$
\begin{aligned}
&\bar{x}(t)=0,\\
&\bar{y}(t)=\frac{1}{2}a_3\left\{2\left(a_1^2+a_3^2\right)^{-1}E\right\}^{1/2}\sin(\varphi_2(0)+2t),\\
&\bar{z}(t)=\frac{1}{2}a_1\left\{2\left(a_1^2+a_3^2\right)^{-1}E\right\}^{1/2}\sin(\varphi_3(0)+2t).
\end{aligned}
\tag{19}
$$

Unlike the normal modes and the general position periodic solutions, we expect this global bifurcation set to break up if we introduce higher-order, i.e. H_4 terms. We shall not carry out the normalization till H_4 for the general Hamiltonian. Note that H_3 already contains 56 terms, H_4 126 terms and though the calculation is straightforward, the general result does not look very attractive. Instead we shall carry out the normalization process in the next section for a particular potential problem. The break-up of the global bifurcation will be demonstrated explicitly there and it is then easy to perceive by inspection of the general Hamiltonian problem that the phenomena in this particular potential problem are characteristic for what happens generally in higher order.

5. A particular potential problem. Consider the Hamiltonian

$$
H=\frac{1}{2}\left(p_x^2+p_y^2+p_z^2\right)+U(x,y,z)
\tag{20}
$$

where U is a potential given by

$$
U(x,y,z)=\frac{1}{2}\left(x^2+4y^2+4z^2\right)+\varepsilon\left\{\alpha_1x^2y+\alpha_2x^2z+\alpha_3xyz\right\}.
\tag{21}
$$

This problem with $\alpha_3=0$ was studied by Martinet, Magnenat and Verhulst (1981) with normalization up till degree three. The results are then similar to those obtained in §§2 and 4. The motivation to introduce α_3 in potential (21) is that terms which are not discrete symmetric in x are to be normalized or averaged away in first order. A term like the one introduced with coefficient α_3 is then expected to show up in a higher order calculation; as the existence of the global bifurcation is tied in with these symmetries in x we expect the global bifurcation to break up in higher order. This idea turns out to be correct. The equations of motion derived from Hamiltonian (20) are

$$\begin{aligned} \ddot{x}+x &= -\varepsilon\{2\alpha_1 xy+2\alpha_2 xz+\alpha_3 yz\}, \\ \ddot{y}+4y &= -\varepsilon\{\alpha_1 x^2+\alpha_3 xz\}, \\ \ddot{z}+4z &= -\varepsilon\{\alpha_2 x^2+\alpha_3 xy\}. \end{aligned} \tag{22}$$

The Hamiltonian (20) normalized to degree four reads

(23)

$$\begin{aligned} \overline{H} = H_2 &+ \frac{1}{4}\varepsilon\left[\alpha_1\{y(x^2-p_x^2)+xp_xp_y\}+\alpha_2\{z(x^2-p_x^2)+xp_xp_z\}\right] \\ &- \frac{1}{32}\varepsilon^2\left[\frac{9}{8}(\alpha_1^2+\alpha_2^2)(x^2+p_x^2)^2+\left(\frac{1}{4}\alpha_1^2+\frac{1}{15}\alpha_3^2\right)(x^2+p_x^2)(4y^2+p_y^2)\right. \\ &\quad +\left(\frac{1}{4}\alpha_2^2+\frac{1}{15}\alpha_3^2\right)(x^2+p_x^2)(4z^2+p_z^2)+\frac{7}{60}\alpha_3^2(4y^2+p_y^2)(4z^2+p_z^2) \\ &\quad \left.+\frac{1}{2}\alpha_1\alpha_2(x^2+p_x^2)(4yz+p_yp_z)+\frac{1}{8}\alpha_3^2\{(4yz+p_yp_z)^2-4(yp_z-zp_y)^2\}\right]. \end{aligned}$$

This result was checked by performing the calculation both by Birkhoff normalization and by higher order averaging; the result was checked again by numerical calculations needed to construct the figures to be discussed later in this section. Note that α_3 does not occur in $\overline{H}_3$. We discuss again the periodic orbits for which various coordinate systems are useful. Introduce action-angle coordinates by

$$\begin{aligned} x&=\sqrt{2r_1}\sin\varphi_1, & y&=\sqrt{r_2}\sin\varphi_2, & z&=\sqrt{r_3}\sin\varphi_3, \\ p_x&=\sqrt{2r_1}\cos\varphi_1, & p_y&=2\sqrt{r_2}\cos\varphi_2, & p_z&=2\sqrt{r_3}\cos\varphi_3, \end{aligned}$$

with $r_i>0$ and $\varphi_i\in S^1$ $(i=1,2,3)$. $\overline{H}$ becomes in these variables

$$\begin{aligned} \overline{H} = r_1+2r_2+2r_3 &+ \frac{1}{2}\varepsilon r_1\left[\alpha_1\sqrt{r_2}\sin(2\varphi_1-\varphi_2)+\alpha_2\sqrt{r_3}\sin(2\varphi_1-\varphi_3)\right] \\ &- \frac{1}{4}\varepsilon^2\left[\frac{9}{16}(\alpha_1^2+\alpha_2^2)r_1^2+\left(\frac{1}{4}\alpha_1^2+\frac{1}{15}\alpha_3^2\right)r_1r_2+\left(\frac{1}{4}\alpha_2^2+\frac{1}{15}\alpha_3^2\right)r_1r_3\right. \\ &\quad \left.+\frac{7}{30}\alpha_3^2r_2r_3+\frac{1}{2}\alpha_1\alpha_2r_1\sqrt{r_2r_3}\cos(\varphi_2-\varphi_3)+\frac{1}{4}\alpha_3^2r_2r_3\cos 2(\varphi_2-\varphi_3)\right] \end{aligned} \tag{24}$$

and the equations of motion read

$$\begin{aligned} \dot{r}_1 &= -\varepsilon r_1\{\alpha_1\sqrt{r_2}\cos\psi_1+\alpha_2\sqrt{r_3}\cos\psi_2\}, \\ \dot{r}_2 &= \frac{1}{2}\varepsilon\alpha_1r_1\sqrt{r_2}\cos\psi_1-\frac{1}{8}\varepsilon^2\{\alpha_1\alpha_2r_1\sqrt{r_2r_3}\sin(\psi_2-\psi_1)+\alpha_3^2r_2r_3\sin 2(\psi_2-\psi_1)\}, \end{aligned} \tag{25}$$

$$\dot{r}_3=\frac{1}{2}\varepsilon\alpha_2 r_1\sqrt{r_3}\cos\psi_2+\frac{1}{8}\varepsilon^2\left\{\alpha_1\alpha_2 r_1\sqrt{r_2 r_3}\sin(\psi_2-\psi_1)+\alpha_3^2 r_2 r_3\sin 2(\psi_2-\psi_1)\right\},$$

$$\dot{\psi}_1=\frac{\varepsilon}{4\sqrt{r_2}}\left\{\alpha_1(4r_2-r_1)\sin\psi_1+4\alpha_2\sqrt{r_2 r_3}\sin\psi_2\right\}$$

$$-\frac{1}{2}\varepsilon^2\left\{\left(\alpha_1^2+\frac{9}{8}\alpha_2^2-\frac{1}{30}\alpha_3^2\right)r_1+\left(\frac{1}{4}\alpha_1^2+\frac{1}{15}\alpha_3^2\right)r_2+\left(\frac{1}{4}\alpha_2^2-\frac{1}{20}\alpha_3^2\right)r_3\right.$$

$$\left.+\frac{1}{8}\alpha_1\alpha_2\sqrt{\frac{r_3}{r_2}}(4r_2-r_1)\cos(\psi_2-\psi_1)-\frac{1}{8}\alpha_3^2 r_3\cos 2(\psi_2-\psi_1)\right\},$$

$$\dot{\psi}_2=\frac{\varepsilon}{4\sqrt{r_3}}\left\{4\alpha_1\sqrt{r_2 r_3}\sin\psi_1+\alpha_2(4r_3-r_1)\sin\psi_2\right\}$$

$$-\frac{1}{2}\varepsilon^2\left\{\left(\frac{9}{8}\alpha_1^2+\alpha_2^2-\frac{1}{30}\alpha_3^2\right)r_1+\left(\frac{1}{4}\alpha_1^2-\frac{1}{20}\alpha_3^2\right)r_2+\left(\frac{1}{4}\alpha_2^2+\frac{1}{15}\alpha_3^2\right)r_3\right.$$

$$\left.+\frac{1}{8}\alpha_1\alpha_2\sqrt{\frac{r_2}{r_3}}(4r_3-r_1)\cos(\psi_2-\psi_1)-\frac{1}{8}\alpha_3^2 r_2\cos 2(\psi_2-\psi_1)\right\},$$

where $\psi_1=2\varphi_1-\varphi_2$ and $\psi_2=2\varphi_1-\varphi_3$.

a. Normal modes. The normal modes are exact solutions of the equations of motion (22); putting $x=\dot{x}=y=\dot{y}=0$ we find harmonic solutions for z from $\ddot{z}+4z=0$ and an analogous result in the y-direction. These solutions correspond with the results given in equations (16a), (16b). For the stability analysis we cannot use action-angle coordinates in all degrees of freedom as for the actions the assumption $r_j>0$ holds. We shall check only the normal mode given by (16a) because of the symmetry between the second and third degree of freedom. Define

$$q_1=x,\qquad q_3=z\sqrt{2},$$
$$p_1=p_x,\qquad p_3=p_z/\sqrt{2},$$

while using action-angle coordinates in the (r_2,φ_2)-plane. The Hamiltonian $\bar{H}$ becomes

$$\text{(26)}\quad \bar{H}=\frac{1}{2}(q_1^2+p_1^2)+2r_2+(q_3^2+p_3^2)+\frac{1}{8}\varepsilon\Big[2\alpha_1\sqrt{r_2}\left\{(q_1^2-p_1^2)\sin\varphi_2+2q_1p_1\cos\varphi_2\right\}$$

$$+\alpha_2\sqrt{r_2}\left\{q_3(q_1^2-p_1^2)+2q_1p_1p_3\right\}\Big]$$

$$-\frac{1}{8}\varepsilon^2\Big[\frac{9}{32}(\alpha_1^2+\alpha_2^2)(q_1^2+p_1^2)^2+\left(\frac{1}{4}\alpha_1^2+\frac{1}{15}\alpha_3^2\right)r_2(q_1^2+p_1^2)$$

$$+\left(\frac{1}{8}\alpha_2^2+\frac{1}{30}\alpha_3^2\right)(q_1^2+p_1^2)(q_3^2+p_3^2)$$

$$+\frac{7}{30}\alpha_3^2 r_2(q_3^2+p_3^2)+\frac{1}{4}\alpha_1\alpha_2\sqrt{2r_2}(q_1^2+p_1^2)\{q_3\sin\varphi_2+p_3\cos\varphi_2\}$$

$$+\frac{1}{4}\alpha_3^2 r_2\{(p_3^2-q_3^2)\cos 2\varphi_2+2q_3p_3\sin 2\varphi_2\}\Big].$$

Defining $\mathcal{H}=\overline{H}-H_2$ and using the Lagrange-multipliers method we find that the normal mode must satisfy

$$d\mathcal{H}+\mu dH_2=0, \qquad H_2=E, \tag{27}$$

for $q_1=p_1=q_3=p_3=0$. Calculating (27) using (26) shows that $r_2=1/2E$ satisfies both equations if we choose the multiplier μ equal to zero. The stability of this solution is determined by the eigenvalue equation of the matrix $d^2\mathcal{H}$ along the orbit; we find for this matrix

$$\begin{bmatrix} \begin{array}{l} \frac{1}{2}\varepsilon\alpha_1\sqrt{r_2}\sin\varphi_2 \\ \quad -\frac{1}{4}\varepsilon^2 r_2\left(\frac{1}{4}\alpha_1^2+\frac{1}{15}\alpha_3^2\right) \end{array} & \frac{1}{2}\varepsilon\alpha_1\sqrt{r_2}\cos\varphi_2 & 0 & 0 \\ \frac{1}{2}\varepsilon\alpha_1\sqrt{r_2}\cos\varphi_2 & \begin{array}{l} -\frac{1}{2}\varepsilon\alpha_1\sqrt{r_2}\sin\varphi_2 \\ \quad -\frac{1}{4}\varepsilon^2 r_2\left(\frac{1}{4}\alpha_1^2+\frac{1}{15}\alpha_3^2\right) \end{array} & 0 & 0 \\ 0 & 0 & \begin{array}{l} -\frac{1}{240}\varepsilon^2\alpha_3^2 r_2 \\ \quad \cdot(14-15\cos 2\varphi_2) \end{array} & \frac{1}{16}\varepsilon^2\alpha_3^2 r_2\sin 2\varphi_2 \\ 0 & 0 & -\frac{1}{16}\varepsilon^2\alpha_3^2 r_2\sin 2\varphi_2 & \begin{array}{l} \frac{1}{240}\varepsilon^2\alpha_3^2 r_1 \\ \quad \cdot(14+15\cos 2\varphi_2) \end{array} \end{bmatrix} \tag{28}$$

and the equation for the eigenvalues

$$\left[\left\{\lambda+\frac{1}{4}\varepsilon^2 r_2\left(\frac{1}{4}\alpha_1^2+\frac{1}{15}\alpha_3^2\right)\right\}^2-\frac{1}{2}\varepsilon^2\alpha_1^2 r_2\right]\left[\left\{\lambda+\frac{7}{120}\varepsilon^2\alpha_3^2 r_2\right\}^2-\frac{1}{256}\varepsilon^4\alpha_3^4 r_2^2\right]=0,$$

producing

$$\begin{aligned} \lambda_1&=\frac{1}{4}\varepsilon\alpha_1\sqrt{2E}-\frac{1}{8}\varepsilon^2 E\left(\frac{1}{4}\alpha_1^2+\frac{1}{15}\alpha_3^2\right), \\ \lambda_2&=-\frac{1}{4}\varepsilon\alpha_1\sqrt{2E}-\frac{1}{8}\varepsilon^2 E\left(\frac{1}{4}\alpha_1^2+\frac{1}{15}\alpha_3^2\right), \\ \lambda_3&=\frac{1}{480}\varepsilon^2\alpha_3^2 E>0, \\ \lambda_4&=-\frac{29}{480}\varepsilon^2\alpha_3^2 E. \end{aligned} \tag{29}$$

All eigenvalues are real, while $\lambda_3>0$ so the normal mode is unstable. An analogous result holds for the other normal mode $r_3=1/2E$, $q_1=p_1=q_2=p_2=0$.

To illustrate the instability of the flow near a normal mode we plot the actions as a function of time for various initial conditions in Fig. 1; the energy is in all cases the same.

b. Periodic orbits in general position. Omitting the terms of $O(\varepsilon^2)$, the periodic orbits obtained from the critical points of equations (25) are given by

$$
\begin{aligned}
(30)\qquad & \bar{r}_1=\frac{2}{3}E,\\
& \bar{r}_2=\frac{1}{6}\alpha_1^2\left(\alpha_1^2+\alpha_2^2\right)^{-1}E,\\
& \bar{r}_3=\frac{1}{6}\alpha_2^2\left(\alpha_1^2+\alpha_2^2\right)^{-1}E,\\
& \bar{\psi}_1=\frac{1}{2}\pi+k\pi,\ k=0,1,\\
& \bar{\psi}_2=\bar{\psi}_1.
\end{aligned}
$$

We assume that correction-terms of order $O(\varepsilon)$ applied to (30) suffice to define a stationary solution of (25). Suppose

$$
\begin{aligned}
(31)\qquad & \tilde{r}_j=\bar{r}_j+\varepsilon\bar{\bar{r}}_j, && j=1,2,3,\\
& \tilde{\psi}_j=\bar{\psi}_j+\varepsilon\bar{\bar{\psi}}_j, && j=1,2
\end{aligned}
$$

and $(\tilde{r}_j,\tilde{\psi}_j)$ define a periodic solution of (25). Then we have $\bar{\bar{r}}_1+2\bar{\bar{r}}_2+2\bar{\bar{r}}_3=0$ as $\tilde{r}_j$ $(j=1,2,3)$ must satisfy the energy equation

$$
(32)\qquad r_1+2r_2+2r_3=E.
$$

Substitute (31) in (25) where we have put the left-hand sides equal to zero. We obtain four independent equations in $\bar{\bar{r}}_2$, $\bar{\bar{r}}_3$, $\bar{\bar{\psi}}_1$ and $\bar{\bar{\psi}}_2$. After expanding these in a power series in ε, the equations become linear in these four variables and we can solve them

$$
\begin{aligned}
(33)\qquad & \bar{\bar{r}}_1=-2\bar{\bar{r}}_2-2\bar{\bar{r}}_3,\\
& \bar{\bar{r}}_2=(3\alpha_1)^{-1}\bar{r}_2^{3/2}\left(\alpha_1^2+\alpha_2^2\right)^{-2}\sin\psi_1\left\{\frac{17}{4}\alpha_1^6+\frac{51}{4}\alpha_1^4\alpha_2^2+\frac{51}{4}\alpha_1^2\alpha_2^4-\frac{1}{15}\alpha_1^4\alpha_3^2\right.\\
& \qquad\left.+\frac{13}{120}\alpha_1^2\alpha_2^2\alpha_3^2+\frac{17}{4}\alpha_2^6-\frac{19}{24}\alpha_2^4\alpha_3^2\right\},\\
& \bar{\bar{r}}_3=(3\alpha_2)^{-1}\bar{r}_3^{3/2}\left(\alpha_1^2+\alpha_2^2\right)^{-2}\sin\psi_1\left\{\frac{17}{4}\alpha_1^6+\frac{51}{4}\alpha_1^4\alpha_2^2-\frac{19}{24}\alpha_1^4\alpha_3^2+\frac{51}{4}\alpha_1^2\alpha_2^4\right.\\
& \qquad\left.+\frac{13}{120}\alpha_1^2\alpha_2^2\alpha_3^2+\frac{17}{4}\alpha_2^6-\frac{1}{15}\alpha_2^4\alpha_3^2\right\}.
\end{aligned}
$$

The extreme values of $\overline{H}_3+\varepsilon\overline{H}_4$ are again found for the orbits given by (30), (31) and (33). We expand $\varepsilon\overline{H}_3+\varepsilon^2\overline{H}_4$ around each periodic orbit while only keeping the quadratic terms. For ε small enough they form a definite function which thus acts like a Lyapunov function. As $\varepsilon\overline{H}_3+\varepsilon^2\overline{H}_4$ and $\overline{H}_3$ are asymptotic integrals of system (25), the derivative $(d/dt)(\varepsilon H_3+\varepsilon^2 H_4)$ is of order $O(\varepsilon^3)$. We conclude that both orbits are stable.

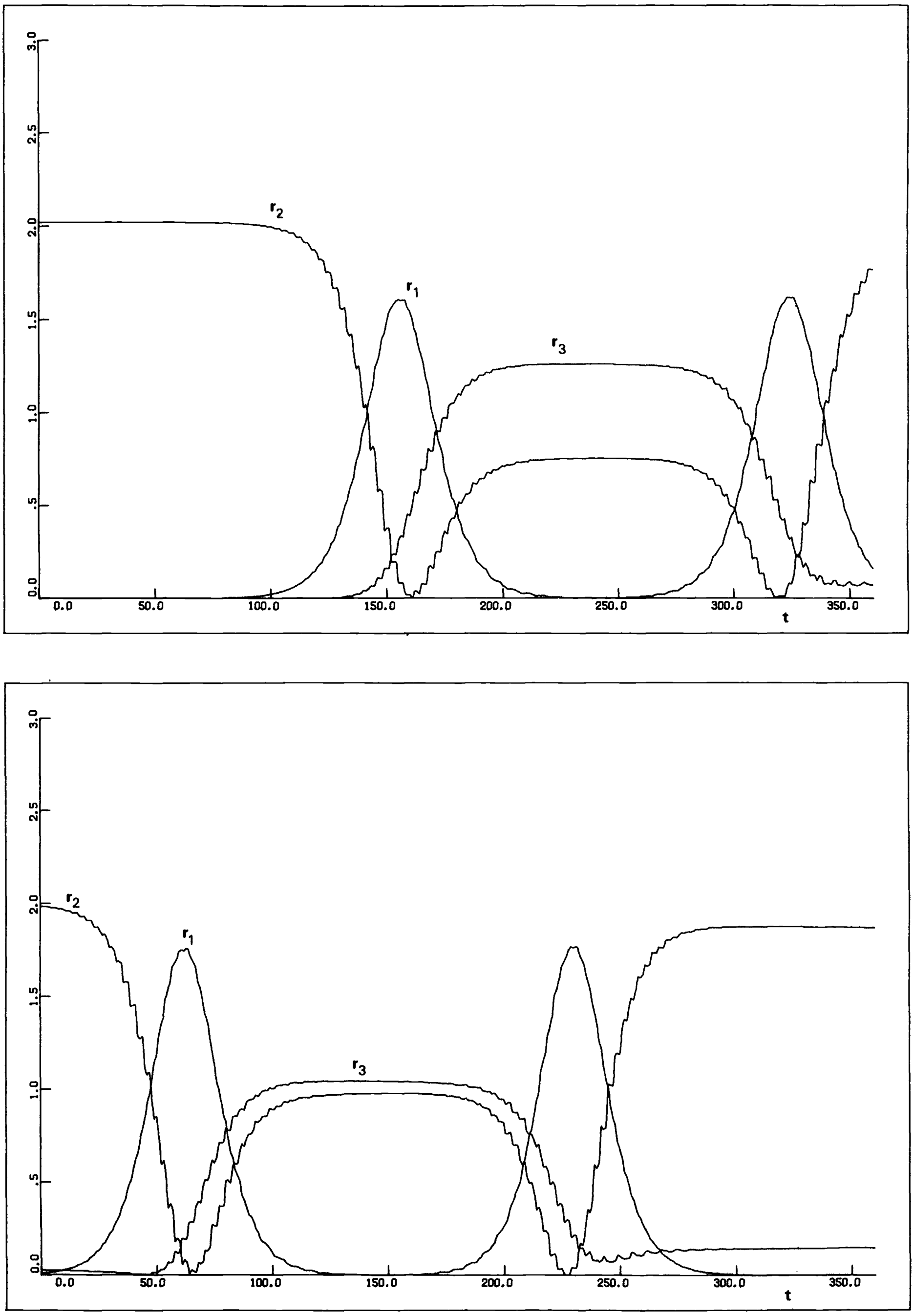
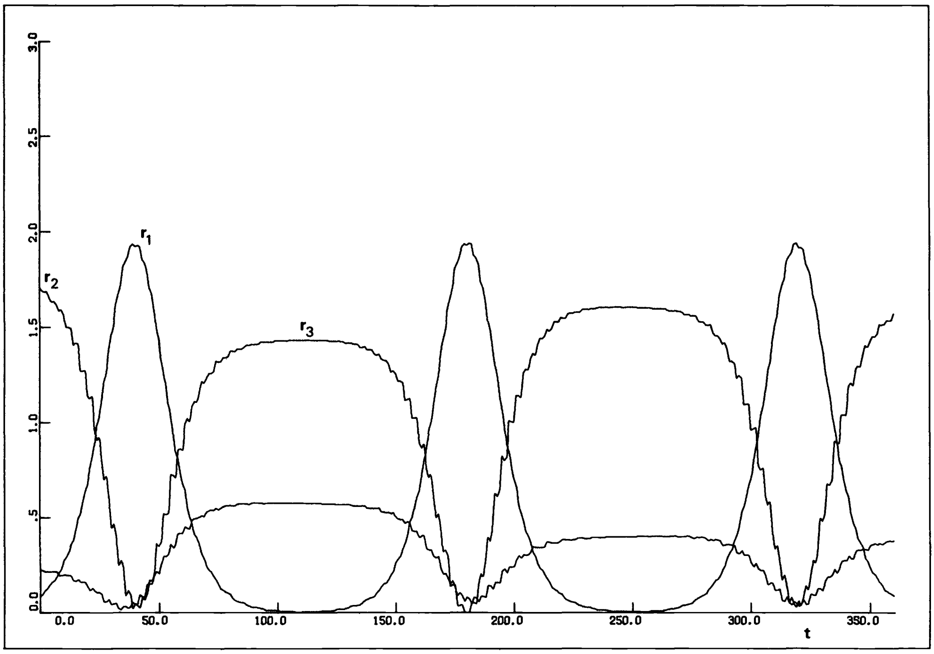
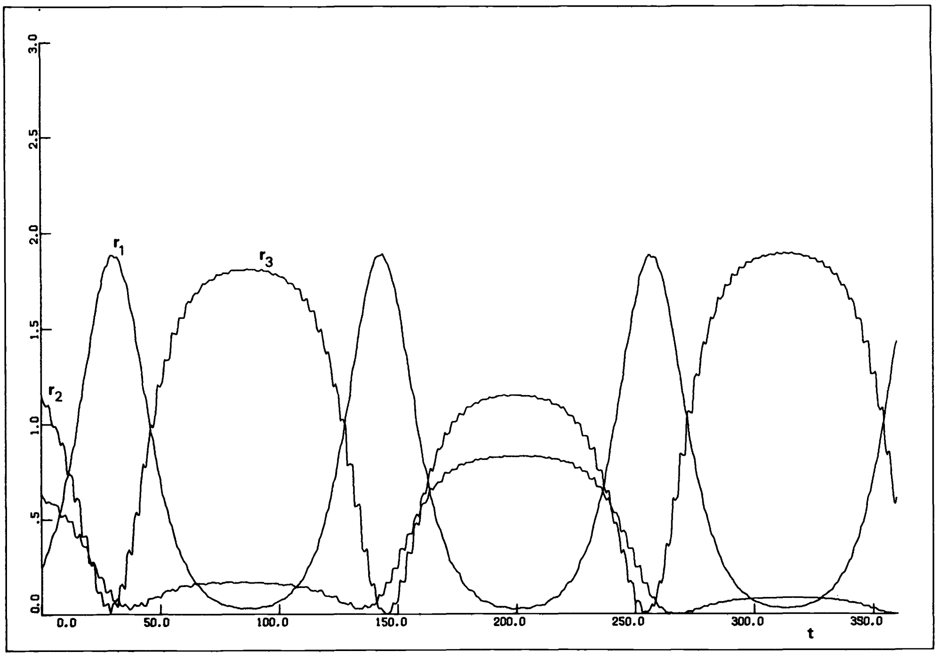


FIG. 1. *The actions r_1, $2r_2$ and $2r_3$ as functions of time near the normal mode in the y-direction. In the Hamiltonian* (20) *we put $\alpha_1 = -1$, $\alpha_2 = -.5$, $\alpha_3 = -.7$, $\varepsilon = .1$. The iso-energetic initial conditions are*

Fig. 1a $x(0) = p_x(0) = .001, p_y(0) = 0, z(0) = p_z(0) = .001$;
Fig. 1b .1 0 .1;
Fig. 1c .3 0 .3;
Fig. 1d .5 0 .5.





In all cases the initial value $y(0)$ is calculated from the energy-equation $1/2(x^2+p_x^2)+1/2(4y^2+p_y^2)+1/2(4z^2+p_z^2)=E$. The integration takes place from $t=0-360$ time units. Going from Figs. 1a–d *the distance from the normal mode increases which results in a shortening of the recurrence time of the exchange of energies between the modes. The recurrence time is characterized by a time-scale of order $1/\varepsilon$ periods, but this estimate holds at a $O(1)$ distance of the normal modes.*

c. The global bifurcation. We shall see that the set of periodic solutions given by (18) breaks up for $\alpha_3 \neq 0$. We use action-angle coordinates in the second and third degree of freedom. As r_1 equals zero we proceed to co-moving variables in the (q_1, p_1)-plane

$$(34)\qquad \begin{bmatrix} q_1 \\ p_1 \end{bmatrix} = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \begin{bmatrix} Q \\ P \end{bmatrix}.$$

The normalized Hamiltonian becomes

$$\begin{aligned}(35)\qquad \overline{H} = {} & \frac{1}{2}(Q^2+P^2)+2r_2+2r_3 \\ & + \frac{1}{4}\varepsilon\Big[\alpha_1\sqrt{r_2}\,\{(Q^2-P^2)\sin(\varphi_2-2t)+2QP\cos(\varphi_2-2t)\} \\ & \qquad + \alpha_2\sqrt{r_3}\,\{(Q^2-P^2)\sin(\varphi_3-2t)+2QP\cos(\varphi_3-2t)\}\Big] \\ & + \frac{1}{8}\varepsilon^2\Big[\frac{9}{32}(\alpha_1^2+\alpha_2^2)(Q^2+P^2)^2 + \left(\frac{1}{4}\alpha_1^2+\frac{1}{15}\alpha_3^2\right)r_2(Q^2+P^2) \\ & \qquad + \left(\frac{1}{4}\alpha_2^2+\frac{1}{15}\alpha_3^2\right)r_3(Q^2+P^2)+\frac{7}{15}\alpha_3^2 r_2 r_3 \\ & \qquad + \frac{1}{2}\alpha_1\alpha_2\sqrt{r_2 r_3}\,(Q^2+P^2)\cos(\varphi_2-\varphi_3)+\frac{1}{2}\alpha_3^2 r_2 r_3\cos 2(\varphi_2-\varphi_3)\Big].\end{aligned}$$

Define $\psi_1=\varphi_2-2t$ and $\psi_2=\varphi_3-2t$ then $\psi=\psi_1-\psi_2=\varphi_2-\varphi_3$ is independent of time. Consider the equations of motion

(36)

$$\begin{aligned}\dot{Q} = {} & \frac{1}{2}\varepsilon\Big[\alpha_1\sqrt{r_2}\,\{Q\cos\psi_1-P\sin\psi_1\}+\alpha_2\sqrt{r_3}\,\{Q\cos\psi_2-P\sin\psi_2\}\Big] \\ & - \frac{1}{4}\varepsilon^2 P\Big[\frac{9}{16}(\alpha_1^2+\alpha_2^2)(Q^2+P^2)+\left(\frac{1}{4}\alpha_1^2+\frac{1}{15}\alpha_3^2\right)r_2 \\ & \qquad + \left(\frac{1}{4}\alpha_2^2+\frac{1}{15}\alpha_3^2\right)r_3+\frac{1}{2}\alpha_1\alpha_2\sqrt{r_2 r_3}\cos\psi\Big], \\ \dot{P} = {} & -\frac{1}{2}\varepsilon\Big[\alpha_1\sqrt{r_2}\,\{Q\sin\psi_1+P\cos\psi_1\}+\alpha_2\sqrt{r_3}\,\{Q\sin\psi_2+P\cos\psi_2\}\Big] \\ & + \frac{1}{4}\varepsilon^2 Q\Big[\frac{9}{16}(\alpha_1^2+\alpha_2^2)(Q^2+P^2)+\left(\frac{1}{4}\alpha_1^2+\frac{1}{15}\alpha_3^2\right)r_2 \\ & \qquad + \left(\frac{1}{4}\alpha_2^2+\frac{1}{15}\alpha_3^2\right)r_3+\frac{1}{2}\alpha_1\alpha_2\sqrt{r_2 r_3}\cos\psi\Big], \\ \dot{r}_2 = {} & -\frac{1}{4}\varepsilon\alpha_1\sqrt{r_2}\,\{(Q^2-P^2)\cos\psi_1-2QP\sin\psi_1\} \\ & - \frac{1}{8}\varepsilon^2\Big[\frac{1}{2}\alpha_1\alpha_2\sqrt{r_2 r_3}\,(Q^2+P^2)\sin\psi+\alpha_3^2 r_2 r_3\sin 2\psi\Big],\end{aligned}$$

$$\dot{r}_3 = -\frac{1}{4}\varepsilon\alpha_2\sqrt{r_3}\,\{(Q^2-P^2)\cos\psi_2 - 2QP\sin\psi_2\}$$
$$+\frac{1}{8}\varepsilon^2\left[\frac{1}{2}\alpha_1\alpha_2\sqrt{r_2 r_3}\,(Q^2+P^2)\sin\psi + \alpha_3^2 r_2 r_3 \sin 2\psi\right],$$

$$\dot{\psi} = \frac{1}{8}\varepsilon\left[\frac{\alpha_1}{\sqrt{r_2}}\{(Q^2-P^2)\sin\psi_1 + 2QP\cos\psi_1\} - \frac{\alpha_3}{\sqrt{r_3}}\{(Q^2-P^2)\sin\psi_2 + 2QP\cos\psi_2\}\right]$$
$$-\frac{1}{8}\varepsilon^2\left[\frac{1}{4}(\alpha_1^2-\alpha_2^2)(Q^2+P^2) + \frac{7}{15}\alpha_3^2(r_3-r_2)\right.$$
$$\left.+\frac{1}{4}\alpha_1\alpha_2\frac{1}{\sqrt{r_2 r_3}}(r_3-r_2)(Q^2+P^2)\cos\psi + \frac{1}{2}\alpha_3^2(r_3-r_2)\cos 2\psi\right].$$

Putting the left-hand sides of (36) equal to zero while taking $Q=P=0$ produces the conditions on periodic solutions

$$(37) \qquad \bar{r}_2 = \bar{r}_3 = \frac{1}{4}E, \qquad \bar{\psi} = \frac{1}{2}k\pi, \qquad k = 0,1,2,3.$$

So we have four periodic orbits with the same action variables which differ $\pi/2$ in phase each. Note that if we put $\alpha_3 = 0$ in (36) the whole bifurcation-set is conserved. The set is perturbed away if we take $\alpha_3 \neq 0$ while four periodic orbits remain, given by (37). System (4) is thus structurally unstable as was already suggested in §2. We shall now examine the stability type of the four remaining orbits. The periodic solutions correspond with nondegenerate critical points of the vectorfield describing the flow in the (Q, P, r_2, ψ)-space. We perform linear stability analysis on these points: we translate the critical point to the origin by a linear transformation of the form $\tilde{u} = u - \bar{u}$ where $u = (Q, P, r_2, \psi)^T$. Linearization of the differential equations for $\tilde{u}$ produces

$$\frac{d}{dt}\tilde{u} = A\tilde{u} + F(\tilde{u}),$$

where $\lim_{\|\tilde{u}\|\to 0}\|F(\tilde{u})\|/\|\tilde{u}\| = 0$ and

$$A = \begin{bmatrix} \frac{1}{2}\varepsilon a & -\frac{1}{2}\varepsilon b - \varepsilon^2 c & 0 & 0 \\ -\frac{1}{2}\varepsilon b + \varepsilon^2 c & -\frac{1}{2}\varepsilon a & 0 & 0 \\ 0 & 0 & 0 & -\varepsilon^2\alpha_3^2\bar{r}_2^2\cos 2\bar{\psi} \\ 0 & 0 & \varepsilon^2\alpha_3^2\left(\frac{7}{15} + \frac{1}{2}\cos 2\bar{\psi}\right) & 0 \end{bmatrix}$$

with

$$a = \alpha_1\sqrt{\bar{r}_2}\cos\bar{\psi}_1 + \alpha_2\sqrt{\bar{r}_3}\cos\bar{\psi}_2,$$
$$b = \alpha_1\sqrt{\bar{r}_2}\sin\bar{\psi}_1 + \alpha_2\sqrt{\bar{r}_3}\sin\bar{\psi}_2,$$

$$c=\frac{1}{4}\left(\frac{1}{4}\alpha_1^2+\frac{1}{15}\alpha_3^2\right)\bar{r}_2+\frac{1}{4}\left(\frac{1}{4}\alpha_2^2+\frac{1}{15}\alpha_3^2\right)\bar{r}_3+\frac{1}{8}\alpha_1\alpha_2\sqrt{\bar{r}_2\bar{r}_3}\cos\bar{\psi}.$$

The eigenvalue equation becomes

$$\left\{\lambda^2-\frac{1}{4}\varepsilon^2(a^2+b^2)+\varepsilon^4c^2\right\}\left\{\lambda^2+\frac{1}{16}\varepsilon^4\alpha_3^4\bar{r}_2^2\left(\frac{7}{15}\cos 2\bar{\psi}+\frac{1}{2}\right)\right\}=0$$

and

$$\lambda_{1,2}^2=-\frac{1}{256}\varepsilon^4\alpha_3^4E^2\left(\frac{7}{15}\cos 2\bar{\psi}+\frac{1}{2}\right)<0,$$

$$\lambda_{3,4}^2=\frac{1}{16}\varepsilon^2E\left(\alpha_1^2+\alpha_2^2+2\alpha_1\alpha_2\cos\bar{\psi}\right)-\frac{1}{4}\varepsilon^4\left\{\frac{1}{32}E\left(\alpha_1^2+\alpha_2^2+2\alpha_1\alpha_2\cos\bar{\psi}\right)+\frac{1}{60}\alpha_3^2E\right\}^2.$$

We have $\bar{\psi}=k\pi/2$, $k=0,1,2,3$ and we shall treat the four orbits separately; note that we restrict ourselves to values of ε in a neighborhood of zero.

$k=1,3$

If $\alpha_1^2+\alpha_2^2>0$ then $\lambda_{3,4}^2>0$ so the orbits are unstable.

If $\alpha_1=\alpha_2=0$, $\alpha_3\neq 0$ all eigenvalues are purely imaginary and a linear stability analysis is not enough to determine stability.

$k=0$

If $\alpha_1+\alpha_2\neq 0$ then $\lambda_{3,4}^2>0$ so the orbits are unstable.

If $\alpha_1+\alpha_2=0$, $\alpha_3\neq 0$ all eigenvalues are purely imaginary and the question of stability is unresolved.

$k=2$

If $\alpha_1-\alpha_2\neq 0$ then $\lambda_{3,4}^2>0$ so the orbits are unstable.

If $\alpha_1-\alpha_2=0$, $\alpha_3\neq 0$ all eigenvalues are purely imaginary and the question of stability is unresolved.

In Fig. 2 we plot the actions as a function of time starting with initial conditions in the global bifurcation at $x=p_x=0$. We demonstrate the behaviour with time for solutions based on the normalization till degree three (§2) and on the complete Hamiltonian, which agrees with the solutions based on normalization till degree four.

Figure 3 summarizes the periodic solutions found for the Hamiltonian (20) normalized to degree three (Fig. 3a) and to degree four (Fig. 3b). We remark that Fig. 3a also represents the general Hamiltonian normalized to degree 3; cf. (3).

6. The 1:2:2: $\cdots$:2-resonance. Considering the fact that the 1:2:2-resonance system is asymptotically integrable (probably unique among genuine first order resonance systems) and remembering the method used to prove it, the step to a more than three degrees of freedom system is not a large one. Nevertheless it still is remarkable that a property like asymptotic integrability can be extended to systems with arbitrary many degrees of freedom. Let us look at the proof. We have a Hamiltonian H as function of the Euclidean coordinate-vector $\underline{x}\in\mathbb{R}^n$ and the corresponding momentum-vector $\underline{p}\in\mathbb{R}^n$. Suppose H can be expanded in homogeneous polynomials $H_n\,(n=2,3,\cdots)$ as follows (cf. §2)

$$\text{(38a)}\qquad H=H_2+\varepsilon H_3+\varepsilon^2H_4+\varepsilon^3H_5+\cdots$$

with nondegenerate

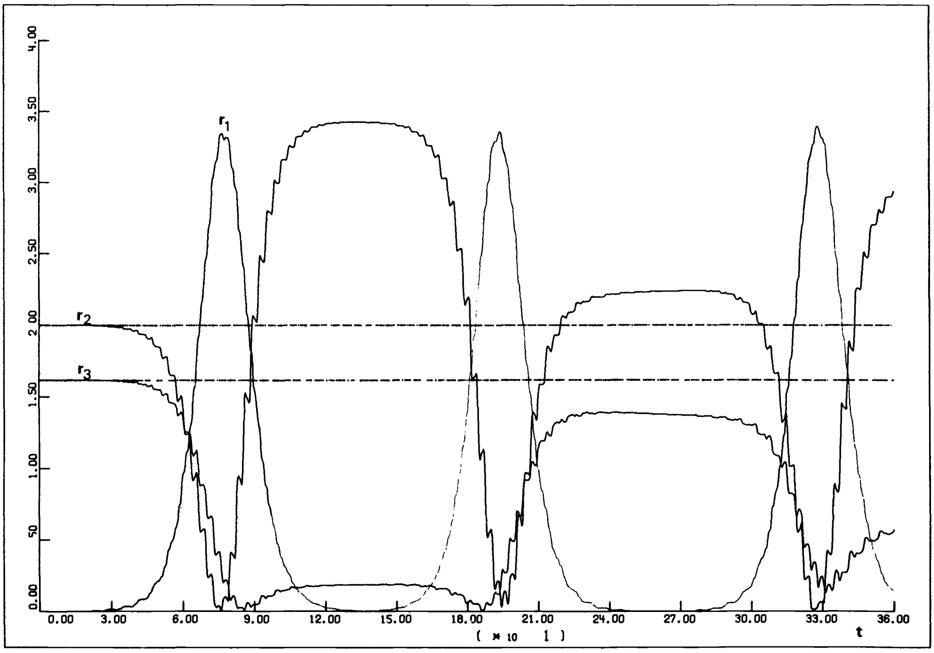


FIG. 2. *The actions r_1, $2r_2$ and $2r_3$ as functions of time $(0\leq t\leq 360)$ starting in the global bifurcation. In the Hamiltonian* (20) *we put $\alpha_1=-1$, $\alpha_2=-.5$, $\alpha_3=-.7$, $\varepsilon=.1$; the initial conditions are $x(0)=p_x(0)=0$, $y(0)=1$, $z(0)=.9$, $p_y(0)=p_z(0)=0$. The horizontal broken lines represent the periodic solution found for the system $H_2+\varepsilon\bar{H}_3$* (§4); *the other solutions are found by numerical integration of the equations corresponding with* (20) *which agrees with normalization to degree* 4 *as has been carried out in* §5.

(38b)
$$H_2=\sum_{j=1}^{n}\frac{1}{2}\left(\omega_j^2x_j^2+p_j^2\right).$$

Here $\underline{\omega}\in\mathbb{R}^n$ is the frequency-vector and the linear vectorfield indicates n harmonic oscillators with frequencies ω_j, $j=1,2,\cdots,n$ while $\omega_1=1$, $\omega_2=\omega_3=\cdots=\omega_n=2$. Normalization to order 3 leads to

(39a)
$$\bar{H}=H_2+\varepsilon\bar{H}_3$$

where

(39b)
$$\bar{H}_3=\sum_{j=2}^{n}\left[2a_{2j-3}\left\{x_j\left(x_1^2-p_1^2\right)+x_1p_1p_j\right\}+a_{2j-2}\left\{p_j\left(p_1^2-x_1^2\right)+4x_1x_jp_1\right\}\right].$$

The coefficients a_i are composed in the same way as the coefficients $a_1\cdots a_4$ in Hamiltonian (3).

We derive the equations of motion by the Hamilton equations

$$\dot{x}_j=\frac{\partial H}{\partial p_j},\qquad \dot{p}_j=-\frac{\partial H}{\partial x_j},\qquad j=1,\cdots,n.$$

The normalized system of second order differential equations becomes

(40)
$$\ddot{x}_1+x_1=-4\varepsilon\sum_{j=2}^{n}\left[a_{2j-3}\left\{2x_1x_j+\dot{x}_1\dot{x}_j\right\}+a_{2j-2}\left\{2\dot{x}_1x_j-x_1\dot{x}_j\right\}\right],$$

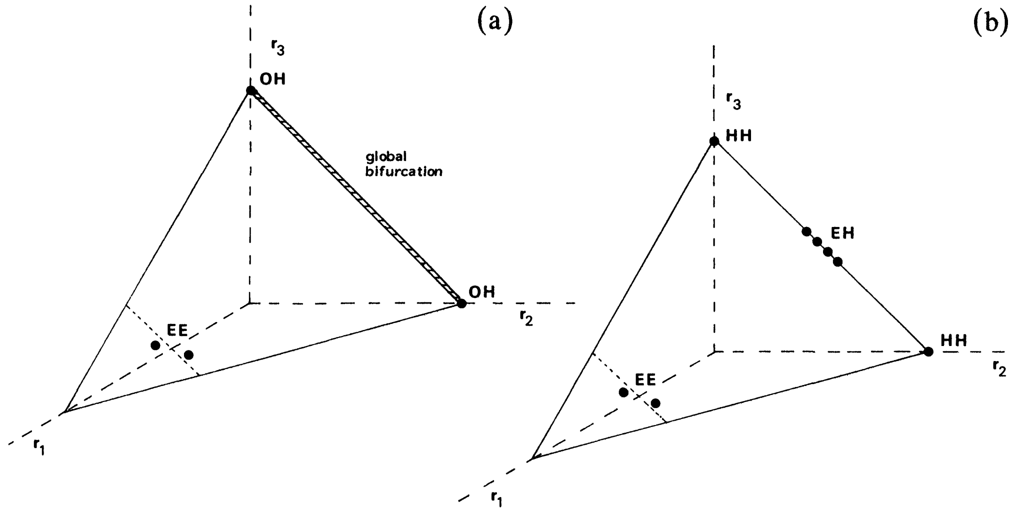


FIG. 3. *Energy simplices for the general Hamiltonian with three degrees of freedom near a nondegenerate equilibrium point in* 1 : 2 : 2-*resonance. Fig.* 3a *is based on the Hamiltonian normalized to degree* 3 *given by* (20). *Fig.* 3b *is based on the normalization to degree* 4. *The periodic solutions are indicated by a dot and are located in the plane* $r_1+2r_2+2r_3=$*constant. Stability has been indicated by characterizing the eigenvalues by letters*: *E* (*purely imaginary*), *H* (*one positive, one negative*), 0 (*zero eigenvalues*). *So the occurrence of at least one letter H implies instability*; *there are only* 4 *eigenvalues as the energy is fixed and one of the angles can be eliminated as explained in* §2.

$$\ddot{x}_j+4x_j=-4\varepsilon\{a_{2j-3}(x_1^2-\dot{x}_1^2)+2a_{2j-2}x_1\dot{x}_1\}, \qquad j=2,\cdots,n.$$

This system already has two independent integrals $\bar{H}$ and H_2. We shall now compute $n-2$ other independent integrals for system (40). Define for $k=3,4,\cdots,n$ the variable

$$u_k=-2a_{2k-2}x_2+a_{2k-3}\dot{x}_2+2a_2x_k-a_1\dot{x}_k;$$

then

$$\ddot{u}_k+4u_k=0.$$

So for each $k=3,4,\cdots,n$ we have a harmonic 2-oscillator in one degree of freedom. This gives us $n-2$ independent quadratic asymptotic integrals I_k, $k=3,4,\cdots,n$ given by

$$I_k=\frac{1}{2}(4u_k^2+\dot{u}_k^2). \tag{41}$$

These integrals are also independent of $\bar{H}$ and H_2, thus system (40) is integrable and the Hamiltonian (38a, b) is asymptotically integrable.

Note that all n integrals are valid for all time with an error of order ε (cf. §2).

We can also derive the periodic orbits as have been found in §4. We have

a. Normal modes. We can derive $n-1$ normal modes directly from (40). If we put $x_1=\dot{x}_1=0$ we find for each $j=2,\cdots,n$: $\ddot{x}_j+4x_j=0$ while we have $x_k=\dot{x}_k=0$ for $k=2,\cdots,n$ and $k\neq j$. The normal mode energy is given by

$$\frac{1}{2}(4x_j^2+\dot{x}_j^2)=E$$

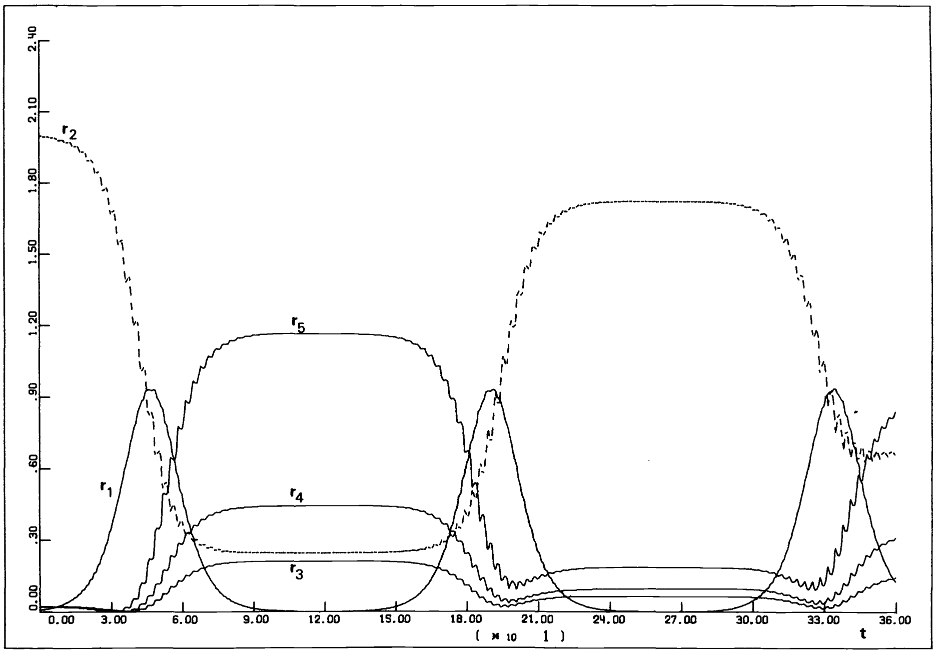


FIG. 4. *The actions* $r_1, \cdots, r_5$ *as functions of time* $(0 \leq t \leq 360)$ *for a system with five degrees of freedom in* 1 : 2 : 2 : 2 : 2*-resonance.*

$$H = H_2 - \varepsilon(x_2 + .6x_3 + .8x_4 + 1.2x_5)x_1^2.$$

The solutions were obtained by numerical integration with initial conditions in the neighbourhood of the unstable normal mode in the x_2*-direction*; *see* §6. *The picture represents the 5-dimensional analogue of Fig.* 1; *as initial conditions we have* .1 *except that* $x_2(0) = 1$, $\dot{x}_2(0) = 0$.

where E defines the total energy of the system. The stability analysis runs along the same lines as in §§2 and 4, cf. Fig. 4.

b. Periodic orbits in general position. As the calculations are easier to perform in action-angle coordinates we put

$$\begin{aligned} x_1 &= \sqrt{2r_1}\sin\varphi_1, & x_j &= \sqrt{r_j}\sin\varphi_j, \\ p_1 &= \sqrt{2r_1}\cos\varphi_1, & p_j &= 2\sqrt{r_j}\cos\varphi_j, \end{aligned} \qquad j = 2, \cdots, n,$$

and corresponding parameters

$$\begin{aligned} a_{2j-3} &= \alpha_{2j-3}\cos\alpha_{2j-2}, \\ a_{2j-2} &= \alpha_{2j-3}\sin\alpha_{2j-2}, \end{aligned} \qquad j = 2, \cdots, n.$$

Substitution in (39) produces for $\overline{H}$

$$\overline{H} = \sum_{j=1}^{n} \omega_j r_j + 4\varepsilon r_1 \sum_{j=2}^{n} \alpha_{2j-3}\sqrt{r_j}\cos(2\varphi_1 - \varphi_j - \alpha_{2j-2}). \tag{42}$$

Define

$$\psi_j = 2\varphi_1 - \varphi_j - \alpha_{2j-2}, \qquad j = 2, \cdots, n$$

then the Hamilton equations lead to a system of $2n-1$ differential equations

$$\dot{r}_1=8\varepsilon r_1\sum_{j=2}^{n}\left\{\alpha_{2j-3}\sqrt{r_j}\sin\psi_j\right\},\tag{43}$$

$$\dot{r}_j=-4\varepsilon\alpha_{2j-3}r_1\sqrt{r_j}\sin\psi_j,\qquad j=2,\cdots,n$$

$$\dot{\psi}_j=8\varepsilon\sum_{k=2}^{n}\left\{\alpha_{2k-3}\sqrt{r_k}\cos\psi_k\right\}-2\varepsilon\alpha_{2j-3}r_1r_j^{-1/2}\cos\psi_j,\qquad j=2,\cdots,n.$$

Periodic orbits are found by putting the left-hand sides of (43) equal to zero. So we have

$$\sin\psi_j=0,\qquad j=2,\cdots,n.\tag{44}$$

For the actions $r_j, j=1,\cdots,n$ we suppose an analogous relation as in three degrees of freedom systems

$$r_1=4\sum_{j=2}^{n}r_j.$$

The conditions $\dot{\psi}_j=0$ for $j=2,\cdots,n$ produce $n-1$ equations for the actions

$$\sum_{\substack{k=2\\k\neq j}}^{n}\left\{\alpha_{2k-3}r_j^{1/2}\cos\psi_k-\alpha_{2j-3}r_k^{1/2}\cos\psi_j\right\}r_k^{1/2}=0\quad\text{for } j=2,\cdots,n.\tag{45}$$

The solutions of (45) are given by

$$r_j=\frac{1}{6}E\left\{\sum_{k=2}^{n}\left(\frac{\alpha_{2k-3}}{\alpha_{2j-3}}\right)^2\right\}^{-1},\qquad j=2,\cdots,n\tag{46}$$

$$\cos\psi_j\cos\psi_k=1\qquad\qquad \forall j,k=2,\cdots,n$$

where the energy E is given by

$$\sum_{j=1}^{n}\omega_jr_j=E$$

and thus

$$r_1=\frac{2}{3}E.\tag{47}$$

So we have two periodic orbits defined by (44), (46) and (47); for one orbit $\cos\psi_j=1$ $(j=2,\cdots,n)$ while the second orbit is found for $\cos\psi_j=-1$ $(j=2,\cdots,n)$. The actions are the same for both orbits. Substitution of the solutions in (42) shows that on these orbits $\overline{H}_3$ achieves its relative extreme values with respect to H_2.

c. The global bifurcations. If we look again at system (40) we see that if we put $x_1=\dot{x}_1=0$ we have

$$\ddot{x}_j+4x_j=0\quad\text{for } j=2,\cdots,n,$$

and thus

$$\frac{1}{2}\left(4x_j^2+\dot{x}_j^2\right)=C_j=\text{constant for } j=2,\cdots,n.$$

Here the constants C_j are arbitrary, but they must satisfy

$$\sum_{j=2}^{n} C_j=E.$$

So we have again found an infinite set of periodic solutions for each value of the energy. In general these global bifurcations will be perturbed away by admitting higher order terms as in the case of the 1 : 2 : 2-resonance.

We finally remark that in the case of more than three degrees of freedom, this list of short-periodic orbits may be incomplete. For instance, normalizing up till H_4, one may find orbits of the $2:2:\cdots:2$ subsystem.

REFERENCES

R. Churchill, M. Kummer and D. Rod, (1981), *On averaging, reduction and symmetry in Hamiltonian systems*, preprint, Univ. Calgary, Calgary, Alberta, Canada.

R. Cushman, (1982), *Reduction, Brouwer's Hamiltonian, and the critical inclination*, preprint 243, Math. Inst., Rijksuniversiteit Utrecht, the Netherlands.

L. Martinet, P. Magnenat, and F. Verhulst, (1981), *On the number of isolating integrals in resonant systems with* 3 *degrees of freedom*, Celest. Mech., 29, pp. 93–99.

E. van der Aa, (1983), *First order resonances in three-degrees-of-freedom systems*, Celest. Mech., 31, pp. 163–191.

F. Verhulst, (1979), *Discrete-symmetric dynamical systems at the main resonances with applications to axi-symmetric galaxies*, Phil. Trans. Roy. Soc. London A, 290, pp. 435–465.

——— (1983), *Asymptotic Analysis of Hamiltonian Systems*, Lecture Notes in Mathematics 985, Springer-Verlag, Berlin.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics
004

SOME PROPERTIES OF SECOND ORDER LINEAR DIFFERENTIAL EQUATIONS AND PERTURBATIONS THAT PRESERVE THEM*

G. J. BUTLER† AND V. SREE HARI RAO‡

Abstract. We consider perturbations of linear second order ordinary differential equations that preserve certain properties of solutions. In particular, under the assumption that all solutions of the basic equation lie in a certain function space, we find conditions on forced perturbations that maintain the property that at most one solution is nonoscillatory. Under the assumption that all solutions of the basic equation are bounded and lie in a certain L^p space, we characterize those L^k linear perturbations that preserve this property. Results of Atkinson, Grimmer and Patula (Ann. Math. Purà Appl., 126 (1980), pp. 296–323), and Patula and Wong (Math. Ann., 197 (1972), pp. 9–28) are extended.

1. Introduction. The second order linear equation we shall consider is

$$(r(t)x'(t))'+q(t)x(t)=0, \tag{1.1}$$

and the types of perturbation we are interested are of the form

$$(r(t)y'(t))'+q(t)y(t)=f(t), \tag{1.2}$$

$$(r(t)z'(t))'+(q(t)+q_1(t))z(t)=0. \tag{1.3}$$

Here, r, q, q_1, f are real-valued, continuous functions on $[0,\infty)$ with $r(t)>0$; in fact all of our results could be obtained under local integrability conditions.

For both types of perturbed equation, our objective is to identify conditions on the perturbation that allow some property of the original equation (1.1) to be preserved. In [4], Grimmer and Patula obtained conditions on the forcing function f that guaranteed that (1.2) possesses at most one nonoscillatory solution. These results were extended in [1], where conditions were also given for (1.2) to possess a set of nonoscillatory solutions of dimension at most one. In these two papers, the standing hypotheses for the homogeneous equation (1.1) were that all solutions of (1.1) be bounded or (1.1) be in the limit-circle case (all solutions are in $L^2[0,\infty)$). Both these hypotheses ensure (under suitable hypotheses on r) that (1.1) is oscillatory, hence the results of [1], [4] may be interpreted as conditions on the perturbation that preserve the property of having at most one nonoscillatory solution. We shall henceforth refer to this property by saying that the corresponding equation is *essentially oscillatory*.

In §2 under the general hypothesis that all solutions of (1.1) belong to some prescribed function space, we shall find conditions for (1.2) to be essentially oscillatory. In this way, we obtain generalizations of [4, Thms. 1 and 2] and [1, Thm. 5]. The distances between successive zeros of solutions of (1.1) play an important role in our

*Received by the editors February 24, 1983, and in revised form June 1, 1983. This research was partially supported by the Natural Science and Engineering Research Council under grant A-8130.

† Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1.

‡ University of Alberta, Edmonton, Alberta, Canada T6G 2G1 and Iowa State University, Ames, Iowa 50010.

discussions in §2, and we pursue this from another point of view in §3, where we give a negative answer to a question raised by Patula in [8].

In §4, we consider linear perturbations of (1.1) of the form (1.3). Here we are motivated by [2] and [10] where the object was to extend Weyl's alternative theorem [11]. In [2] this approach was used to establish conditions on linear perturbations that preserved the limit-point, limit-circle classification of (1.1). We shall assume that solutions of (1.1) lie in some function space, and ask the question what linear perturbations preserve this property. Our results extend those of [10].

The special case of (1.1) for which $r(t)\equiv 1$ is of particular interest and we shall designate it by

$$x''(t)+q(t)x(t)=0. \tag{1.1$'$}$$

Similarly, we designate the equations

$$y''(t)+q(t)y(t)=f(t), \tag{1.2$'$}$$

$$z''(t)+(q(t)+q_1(t))z(t)=0. \tag{1.3$'$}$$

2. Essential oscillation. If (1.1) is oscillatory, the possibilities for (1.2) are:

(i) all solutions oscillate;

(ii) exactly one solution is nonoscillatory;

(iii) there is a one-dimensional set of nonoscillatory solutions;

(iv) there is a two-dimensional set of nonoscillatory solutions.

Essential oscillation of (1.2) means that (i) or (ii) occurs. When (ii) occurs, we might regard the unique nonoscillatory solution y_0 of (1.2) as a perturbation of the zero solution of (1.1). All other solutions of (1.2) "dominate" y_0 to the extent that they oscillate not only about y_0 but also about zero.

Even rapidly decaying forcing functions f may result in the existence of a nonoscillatory solution; for example $y''+y=e^{-t}$ has the unique nonoscillatory solution $y=\frac{1}{2}e^{-t}$.

The first thing we need to do is to establish a relation between the property that all solutions of (1.1) belong to some particular function space, and oscillation. The following lemma is well known (see [5] for the case $r(t)\equiv 1$).

LEMMA 2.1. *Equation* (1.1) *is oscillatory if and only if all nontrivial* (*i.e., not identically zero*) *solutions x satisfy*

$$1/rx^2\notin L^1[t_0,\infty) \quad \textit{for all } t_0\geq 0.$$

For a given function r, we shall denote the set of continuous functions x that satisfy $1/rx^2\notin L^1[t_0,\infty)$ for all $t_0\geq 0$ by $X[r]$, where any function that vanishes on some nondegenerate subinterval of $[t_0,\infty)$ for all $t_0\geq 0$ is automatically in $X[r]$. It is easily verified that $X[r]$ is a real vector space. Under appropriate conditions on r, the classical function spaces are vector subspaces of $X[r]$ (not topological subspaces; there is no suggestion of a topology for $X[r]$).

THEOREM 2.1. *Let* $0<p\leq\infty$ *and suppose that* $r^{-p/(p+2)}\in L^1[0,\infty)$ (*if* $p=\infty$, *suppose that* $r^{-1}\in L^1[0,\infty)$). *Assume that all solutions of* (1.1) *are in* $L^p[0,\infty)$. *Then* (1.1) *is oscillatory.*

Proof. We have to show that $L^p[0,\infty)\subset X[r]$. Let $x\in L^p[0,\infty)$. We have

$$(2.1)\qquad \int_{t_0}^{t}\frac{ds}{r^{p/(p+2)}(s)}=\int_{t_0}^{t}\left(\frac{1}{r^{p/(p+2)}(s)x^{2p/(p+2)}(s)}\right)x^{2p/(p+2)}(s)\,ds$$
$$\leq\int_{t_0}^{t}\left(\frac{ds}{(rx^2)(s)}\right)^{p/(p+2)}\left(\int_{t_0}^{t}|x(s)|^p ds\right)^{1/(p+2)},$$

by Hölder's inequality. By hypothesis, the left-hand side of (2.1) approaches $+\infty$ as $t\to\infty$, and $\int_{t_0}^{\infty}|x(s)|^p\,ds<\infty$. It follows that $1/(rx^2)\notin L^1[0,\infty)$.

Corollary 2.1. *If all solutions of* (1.1)′ *are in* $L^p[0,\infty)$, *where* $0<p\leq\infty$, *then* (1.1)′ *is oscillatory.*

Remarks. 2.1. We note that the values of p between 0 and 1 are included in Theorem 1 as well as the more usually considered values $1\leq p\leq\infty$. The case $p=1$ was given in [6] and the case $p=2$ was obtained in [9].

To obtain conditions for essential oscillation of (1.2) we shall exploit [1, Thm. 1], which we state here as:

Lemma 2.2. *Let* x_1, x_2 *be linearly independent solutions of* (1.1). *Let* (α_{nj},β_{nj}) *be a sequence of pairs of successive zeros of* x_j, *with* $\alpha_{nj}\to\infty$ *as* $n\to\infty$, *such that*

$$(2.2)\qquad \int_{\alpha_{nj}}^{\beta_{nj}}f(s)x_j(s)\,ds\to0\quad \text{as } n\to\infty,\quad j=1,2.$$

Then (1.2) *is essentially oscillatory.*

As a very simple consequence we have the following result:

Theorem 2.2. *Let* $B\subset X[r]$ *be any Banach function space consisting of locally Lebesgue integrable functions. Let* $\hat{B}$ *be the set of locally integrable functions* f *for which* $\int_0^\infty f(t)x(t)\,dt$ *exists* (*conditionally*) *as a finite-valued integral for all* $x\in B$. *Suppose that all solutions of* (1.1) *are in* B. *Then* (1.2) *is essentially oscillatory for all* $f\in\hat{B}$.

Proof. Immediate from the definition of $X[r]$ and Lemmas 2.1 and 2.2.

Corollary 2.2. *Let all solutions of* (1.1)′ *belong to* $L^p[0,\infty)$ $(1\leq p\leq\infty)$ *and suppose that* $f\in L^{p'}[0,\infty)$, *where* $1/p+1/p'=1$. *Then* (1.2)′ *is essentially oscillatory.*

The cases $p=2$, $p=\infty$ were obtained in [4]. To obtain more refined criteria for essential oscillation, we require information about the distances between successive zeros of solutions of (1.1). This is the content of our next result.

Theorem 2.3. *Let* $0<p\leq\infty$ *and suppose that* $r^{-p/p+2}\notin L^1[0,\infty)$. *In addition, assume that* $r\in L^\alpha[0,\infty)$ *for some* α *with* $0<\alpha\leq\infty$, *and that all solutions of* (1.1) *are in* $L^p[0,\infty)$. *Then if* $\{t_n\}$ *is the increasing sequence of zeros of any solution of* (1.1), *the sequence of successive distances between zeros* $\{t_{n+1}-t_n\}$ *is in the sequence space* l^k, *where* $k=(2\alpha+p+\alpha p)/\max(p,2\alpha)$. (*If* $p=\infty$, $\alpha<\infty$, *then* $k=1+\alpha$; *if* $p<\infty$, $\alpha=\infty$, *then* $k=1+p/2$; *if* $p=\alpha=\infty$, *then* $k=\infty$.)

Proof. Assume first that both p and α are finite. Let x_1 be a nontrivial solution of (1.1) whose zero sequence is $\{t_n\}$. If x_2 is a solution of (1.1) such that the Wronskian $W(x_1,x_2)=r(x_1x_2'-x_2x_1')\equiv1$, it is a well-known result (e.g. [6]) that

$$(2.3)\qquad \int_{t_n}^{t_{n+1}}\frac{dt}{r\left(x_1^2+x_2^2\right)}=\pi.$$

Let λ, μ, μ', ν, ν' satisfy $0<\lambda$, μ, $\mu'\leq\infty$, $\lambda\mu'=1$,

$$(2.4)\qquad \frac{1}{\mu}+\frac{1}{\mu'}=\frac{1}{\nu}+\frac{1}{\nu'}=1.$$

Using (2.3), two applications of Hölder's inequality give

$$\begin{aligned}(2.5)\qquad t_{n+1}-t_n&\le\left(\int_{t_n}^{t_{n+1}}\left(r\left(x_1^2+x_2^2\right)\right)^{-\mu'\lambda}dt\right)^{1/\mu'}\left(\int_{t_n}^{t_{n+1}}\left(r\left(x_1^2+x_2^2\right)\right)^{\mu\lambda}dt\right)^{1/\mu}\\ &=\pi^{1/\mu'}\left(\int_{t_n}^{t_{n+1}}r^{\mu\lambda}\left(x_1^2+x_2^2\right)^{\mu\lambda}dt\right)^{1/\mu}\\ &=\pi^{1/\mu'}\left(\int_{t_n}^{t_{n+1}}r^{\mu-1}\left(x_1^2+x_2^2\right)^{\mu-1}dt\right)^{1/\mu}\\ &\le\pi^{1/\mu'}\left(\int_{t_n}^{t_{n+1}}r^{\nu(\mu-1)}dt\right)^{1/\mu\nu}\left(\int_{t_n}^{t_{n+1}}\left(x_1^2+x_2^2\right)^{\nu'(\mu-1)}dt\right)^{1/\mu\nu'}.\end{aligned}$$

If we choose $\mu=1+\alpha p/(2\alpha+p)$, $\nu=1+2\alpha/p$, so that $\mu'=1+2/p+1/\alpha$, $\nu'=1+p/2\alpha$, raising both sides of (2.5) to the power $\mu\nu'=(2\alpha+p+\alpha p)/2\alpha$, we obtain

$$(2.6)\qquad (t_{n+1}-t_n)^{k_1}\le\pi^{p/2}\left(\int_{t_n}^{t_{n+1}}r^{\alpha}dt\right)^{p/2\alpha}\left(\int_{t_n}^{t_{n+1}}\left(x_1^2+x_2^2\right)^{p/2}dt\right),$$

where $k_1=(2\alpha+p+\alpha p)/2\alpha$. A well-known inequality gives

$$\left(x_1^2+x_2^2\right)^{p/2}\le c_p(|x_1|^p+|x_1|^p)$$

where

$$c_p=\begin{cases}2^{p-2}, & p\ge 2,\\ 1, & 0<p<2.\end{cases}$$

By hypothesis, $(\int_{t_n}^{t_{n+1}}r^{\alpha}dt)^{p/2\alpha}\to 0$ as $n\to\infty$ if $\alpha<\infty$, and is bounded as $n\to\infty$ if $\alpha=\infty$. It follows from (2.6) that $\{t_{n+1}-t_n\}\in l^{k_1}$. Raising (2.5) to the power $k_2=\mu\nu=(2\alpha+p+\alpha p)/p$ shows, in a similar fashion, that $\{t_{n+1}-t_n\}\in l^{k_2}$. Defining k to be $\min(k_1,k_2)$, now yields the result. For p or α or both $=\infty$, the argument is essentially the same with due regard being given to the interpretation of indeterminate expressions occurring in the preceding calculations and to the use of the appropriate forms of Hölder's inequality.

COROLLARY 2.3. *If all solutions of* (1.1)′ *are in* $L^p[0,\infty)$, *where* $0<p\le\infty$, *then the sequences of successive distances between zeros of solutions are in* $l^{1+p/2}$. *Thus for* $0<p<\infty$, *the distances between successive zeros tend to zero.*

Remark. For $p=2$, this result was obtained in [9]. For $p=\infty$, this result is not true, but it does hold under the stronger assumption that all solutions of (1.1) approach zero asymptotically.

Now we present the main result of this section, which is a generalization of [1, Thm. 5].

THEOREM 2.4. *Let* $1\le p<\infty$ *and* $1/p+1/p'=1$. *Suppose that* $r^{-p/p+2}\notin L^1[0,\infty)$ *and* $r\in L^{\alpha}[0,\infty)$ *for some* α *with* $0<\alpha\le\infty$. *Let* $k=(2\alpha+p+\alpha p)/\max(p,2\alpha)$, *and define* λ *to be* $pk/((p-1)(k-1))$. *Assume that all solutions of* (1.1) *are in* $L^p[0,\infty)$. *Then for any f satisfying*

$$(2.7)\qquad \int_0^t|f|^{p'}dt=O\left(t^{\lambda}(\log t)^{1/(p-1)}\right)\quad as\ t\to\infty,$$

(1.2) *is essentially oscillatory. If* $p=\infty$, *the above conclusion holds if* (2.7) *is replaced by*

$$\int_0^t |f|\,dt=o(t). \tag{2.8}$$

Proof. Following the proof [1, Thm. 5] it is enough to show that if x_1 is any solution of (1.1) with zeros $\{t_n\}$, then

$$\varliminf_{n\to\infty} n^{-1}\int_{t_n}^{t_{2n}} |fx_1|\,dt=0. \tag{2.9}$$

Now,

$$\int_{t_n}^{t_{2n}} |fx_1|\,dt\leq\left(\int_{t_n}^{t_{2n}} |f|^{p'}\,dt\right)^{1/p'}\left(\int_{t_n}^{t_{2n}} |x_1|^{p}\,dt\right)^{1/p}. \tag{2.10}$$

As shown in [1], we may choose an infinite sequence of values of n for which

$$\int_{t_n}^{t_{2n}} |x_1|^{p}\,dt<(\log n\log\log n)^{-1} \quad (p<\infty). \tag{2.11}$$

By Theorem 2.3, if $1/k+1/k'=1$, we have

$$\begin{aligned}|t_{2n}-t_0|\leq \sum_{m=0}^{2n-1} |t_{m+1}-t_m|&\leq\left(\sum_{m=0}^{2n-1} |t_{m+1}-t_m|^{k}\right)^{1/k}(2n)^{1/k'}\\ &\leq An^{1/k'} \quad \text{for some constant } A.\end{aligned} \tag{2.12}$$

From (2.7) and (2.10)–(2.12), we have an infinite sequence of values of n and a constant C for which

$$\begin{aligned}\int_{t_n}^{t_{2n}} |fx_1|\,dt&\leq Cn^{\lambda/k'p'}(\log n)^{1/(p-1)p'}\cdot(\log n)^{-1/p}(\log\log n)^{-1/p}\\ &=Cn(\log\log n)^{-1/p},\end{aligned}$$

since $\lambda=p'k'$. This yields (2.9) as desired. If $p=\infty$ and f satisfies (2.8), then $x_1(t)$ and $\{t_{n+1}-t_n\}$ are bounded, and so

$$\int_{t_n}^{t_{2n}} |fx_1|\,dt=O\left(\int_{t_n}^{t_{2n}} |f|\,dt\right)=o(t_{2n})=o(n),$$

and again we have (2.9). This completes the proof of the theorem.

Regardless of the value of α in the above theorem, λ will be at least one, so we have:

COROLLARY 2.4. *Let* $1\leq p<\infty$ *and* $r^{-p/(p+2)}\in L^1[0,\infty)$. *Suppose that* $r\in L^{\alpha}[0,\infty)$ *for some* α *with* $0<\alpha\leq\infty$, *and that all solutions of* (1.1) *are in* $L^p[0,\infty)$. *Then for any* f *in* $L^s[0,\infty)$, *where* $1/p+1/s\leq 1$ (*i.e.*, $s\geq p'$), (1.2) *will be essentially oscillatory. If* $p=\infty$, *the same conclusion holds for* f *in* $L^s[0,\infty)$, *where* $1\leq s<\infty$.

In particular, Corollary 2.4 applies to (1.1)$'$ and (1.2)$'$: if all solutions of (1.1)$'$ are in $L^p[0,\infty)$ $(1\leq p<\infty)$ and $f\in L^s[0,\infty)$ for some $s\geq p'$, then (1.2)$'$ is essentially oscillatory (with the appropriate modification when $p=\infty$).

We cannot, in general, allow $s=\infty$ when $p=\infty$; a simple example is the equation $y''+y=1$. With the stronger assumption that all solutions of (1.1) approach zero asymptotically, bounded forcing functions f will result in essential oscillation of (1.2). This is immediate from the remark following Corollary 2.3.

3. On a question concerning distances between zeros of solutions. We have seen that if all solutions of (1.1)$'$ are in $L^p[0,\infty)$, $0<p<\infty$, then the successive distances between zeros of solutions tend to zero. In the opposite direction, Patula [8] has shown that if the positive part q_+ of q is in $L^p[0,\infty)$, $1\le p<\infty$, the successive distances between zeros of solutions of (1.1)$'$ tend to infinity. He raised the question of whether this is true for $0<p<1$. The following example gives a negative answer.

Define $q(t)$ to be $4\pi^2n^{2\alpha}$ on $[n-n^{-\beta},n+n^{-\beta}]$, $n=1,2,\cdots$, and to be zero elsewhere. Here, α and β are positive integers to be chosen later.

Consider the solution x_n of (1.1)$'$ that satisfies $x_n(n)=1$, $x_n'(n)=0$.

On $[n-n^{-\beta},n+n^{-\beta}]$, we have

$$x_n''(t)+4\pi^2n^{2\alpha}x_n(t)=0,$$

and so $x_n(t)=\cos(2\pi^2n^{\alpha}t)$. Now $x_n'(n+n^{-\beta})=-2\pi n^{\alpha}\sin(2\pi n^{\alpha-\beta})$ and $x_n''(t)\equiv 0$ on $(n+n^{-\beta},n+1-(n+1)^{-\beta})$. Hence on this interval we have $x_n(t)=\cos(2\pi n^{\alpha-\beta})-2\pi n^{\alpha}(t+n^{-\beta}-n)\sin(2\pi n^{\alpha-\beta})$. So $x_n(n+\frac{1}{2}-n^{-\beta})=\cos 2\pi n^{\alpha-\beta}-\pi n^{\alpha}\sin(2\pi n^{\alpha-\beta})$. If $0<\alpha<\beta$, $\sin(2\pi n^{\alpha-\beta})\sim 2\pi n^{\alpha-\beta}$ for large n, and so $\sin(2\pi n^{\alpha-\beta})>\pi n^{\alpha-\beta}$ for $n\ge n_0$, say. For $n\ge n_0$, we have

$$x_n\left(n+\frac{1}{2}-n^{-\beta}\right)<1-\pi^2n^{2\alpha-\beta}<0$$

for n sufficiently large, provided that $2\alpha>\beta$. Similarly, for $n\ge n_1$ say, we have

$$x_n\left(n-\frac{1}{2}+n^{-\beta}\right)<0.$$

Thus x_n has two zeros in the interval $(n-\frac{1}{2},n+\frac{1}{2})$, if $n\ge n_1$. By the Sturm comparison theorem, every solution of (1.1)$'$ contains at least one zero in $(n-\frac{1}{2},n+\frac{1}{2})$, for $n\ge n_1$, and so (1.1)$'$ is oscillatory, and for any solution with zeros $\{t_k\}$, we have $t_{k+1}-t_k<2$ for k sufficiently large.

Now $\int_0^{\infty}(q_+)^p\,dt=\sum_{n=1}^{\infty}2\pi n^{2p\alpha-\beta}<\infty$ provided that $2p\alpha-\beta<-1$.

Thus we obtain a negative answer to Patula's question if we choose positive integers α, β such that $\alpha<\beta$, $2\alpha>\beta$, $2p\alpha<\beta-1$, i.e., with $\frac{\beta}{2}<\alpha<(\beta-1)/2p$, which we can always do since $0<p<1$.

4. Linear perturbations and extensions of Weyl's theorem. Weyl [11] showed that if $q_1\in L^{\infty}[0,\infty)$, then (1.3)$'$ has the same limit-point, limit-circle type as (1.1)$'$. In attempting to extend this result to L^k linear perturbations, Patula and Wong [10] showed that if all solutions of (1.1)$'$ are in $L^2[0,\infty)\cap L^{\infty}[0,\infty)$ and q_1 is in $L^k[0,\infty)$, $1\le k\le\infty$, then all solutions of (1.3)$'$ are in $L^2[0,\infty)\cap L^{\infty}[0,\infty)$.

Their conjecture that the boundedness assumption on solutions of (1.1)$'$ might be dropped was eventually disproved by Kwong [7].

In this section, we shall generalize the above result be assuming that all solutions of (1.1)$'$ lie in $L^p[0,\infty)\cap L^{\infty}[0,\infty)$ for some p, $1\le p\le\infty$. In fact the coefficient function $r(t)$ plays no role here; our results apply equally well to (1.1) and (1.3).

THEOREM 4.1. *Let $1 \leq p \leq \infty$, and suppose all solutions of* (1.1)′ *are in* $L^p[0,\infty) \cap L^\infty[0,\infty)$. *Let* $q_1 \in L^k[0,\infty)$, *where*

$$1 \leq k \leq \infty \quad \text{if } 1 \leq p \leq 2, \qquad 1 \leq k \leq \frac{p}{p-2} \quad \text{if } 2 < p \leq \infty.$$

Then all solutions of (1.3) *are in* $L^p[0,\infty) \cap L^\infty[0,\infty)$. *Moreover, for each value of p, the range of possible values of k is sharp.*

Proof. Our argument follows along the lines of [10, Thm. 5.1]. As with the proof of that theorem, our starting point is that solutions of (1.3)′ are given by

$$\begin{aligned} z(t) &= c_1 x_1(t) + c_2 x_2(t) \\ &\quad + x_2(t)\int_0^t x_1(s) q_1(s) z(s)\,ds + x_1(t)\int_0^t x_2(s) q_2(s) z(s)\,dx, \end{aligned} \tag{4.1}$$

where x_1, x_2 are solutions of (1.1)′ whose Wronskian is equal to one.

Let $1/p + 1/p' = 1/k + 1/k' = 1$. We have three cases to consider:

Case (a). $1 \leq k \leq p'$. Then $k' \geq p$. Since $x_i \in L^p[0,\infty) \cap L^\infty[0,\infty)$, we have $x_i \in L^{k'}[0,\infty)$ and so $x_i q_i \in L^1[0,\infty)$, $i = 1,2$. From (4.1), since $x_i \in L^\infty[0,\infty)$, there are constants M_1, M_2 such that

$$|z_1(t)| \leq M_1 + M_2 \int_0^t \{|x_1(s) q_1(s)| + |x_2(s) q_1(s)|\} |z(s)|\,ds,$$

i.e.,

$$|z(t)| \leq M_1 + \int_0^t |\phi(s)||z(s)|\,ds, \tag{4.2}$$

where $\phi \in L^1[0,\infty)$.

Applying Gronwall's inequality to (4.2), we obtain

$$|z(t)| \leq M_1 \exp\left(\int_0^t |\phi(s)|\,ds\right).$$

It follows that $z \in L^\infty[0,\infty)$. Since $x_i \in L^\infty[0,\infty)$, $x_i q_1 \in L^1[0,\infty)$, referring back to (4.1), we see that $z \in L^p[0,\infty)$.

Case (b). $1 \leq p \leq 2 \leq p' < k$. Let $A_t = \{s \in [0,t]: |q_1(s)| \leq 1\}$, $B_t = [0,t] - A_t$. Then

$$\begin{aligned} \int_{A_t} |x_i q_1 z|\,ds &\leq \left(\int_{A_t} |z|^p\,ds\right)^{1/p} \left(\int_{A_t} |x_i q_1|^{p'}\right)^{1/p'} \\ &\leq \left(\int_0^t |z|^p\,ds\right)^{1/p} \left(\int_0^t |x_i|^{p'}\right)^{1/p'} \\ &\leq M_3 \left(\int_0^t |z|^p\,ds\right)^{1/p} \left(\int_0^t |x_i|^{p}\right)^{1/p'} \end{aligned} \tag{4.3}$$

for some constant M_3, since $x_i \in L^\infty[0,\infty)$ and $p \leq p'$. Now

$$\int_{B_t} |x_i q_1 z|\,ds \leq M_4 \left(\int_{B_t} |z|^p\,ds\right)^{1/p} \left(\int_{B_t} |q_1|^{p'}\right)^{1/p'},$$

where M_4 is a bound for x_1 and x_2, and since $k>p'$, we have

$$\int_{B_t}|x_iq_1z|ds\le M_4\left(\int_0^t|z|^p ds\right)^{1/p}\left(\int_{B_t}|q_1|^k\right)^{1/p'}\le M_5\left(\int_0^t|z|^p ds\right)^{1/p} \tag{4.4}$$

for some M_5, since $q_1\in L^k[0,\infty)$. From (4.3) and (4.4), we have

$$\int_0^t|x_iq_1z|ds\le M_6\left(\int_0^t|z|^p ds\right)^{1/p},\qquad i=1,2 \tag{4.5}$$

for some constant M_6.

Raising (4.1) to the pth power, using (4.5) and the standard inequality $(a+b)^p\le 2^{p-1}(a^p+b^p)$ for $a,b\ge 0$, we have

$$\begin{aligned}|z(t)|^p&\le 2^{2p-2}\left\{|c_1x_1(t)|^p+|c_2x_2(t)|^p+|M_6x_2(t)|^p\int_0^t|z|^p ds+|M_6x_1(t)|^p\int_0^t|z|^p ds\right\}\\&=\phi(t)+w(t)\int_0^t|z|^p ds,\end{aligned}$$

where $\phi,w\in L^1[0,\infty)$. Again Gronwall's lemma yields $z\in L^p[0,\infty)\subset L^\infty[0,\infty)$.

Case (c). $1<p'<2<p,\ p'<k\le p/p-2$. For this case, we use Hölder's inequality to obtain

$$\begin{aligned}\int_{A_t}|x_iq_1|^{p'}ds&\le\left(\int_{A_t}|x_i|^p\right)^{1/p}\left(\int_{A_t}|q_i|^{pp'/(p-p')}\right)^{(p-p')/pp'}\\&\le\left(\int_{A_t}|x_i|^p\right)^{1/p}\left(\int_{A_t}|q_i|^k\right)^{(p-p')/pp'},\end{aligned}$$

since $k\le pp'/(p-p')=p/(p-2)$. Now we can proceed as in case (b). The resolution of cases (a), (b), (c) proves the first assertion of the theorem.

To show that the result is sharp, we show that for all pairs (p,k) not covered by the statement of the theorem, there exist an equation (1.1)′ all of whose solutions are in $L^p[0,\infty)$ and a function $q_1\in L^k[0,\infty)$ such that (1.3)′ has a solution which is not in $L^p[0,\infty)$. For convenience, we work on the interval $[1,\infty)$.

We require the following lemma which may be proved by straightforward computation:

LEMMA 4.1. *For nonnegative constants* α, β, γ, *define*

$$I(\alpha,\beta,\gamma)(t)=\int_1^t\tau^{-\alpha}(\sin\tau^\beta)^\gamma d\tau.$$

Then for any $M>0$ *and any* α, β, γ *with* $0\le\alpha$, β, $\gamma\le M$, *there exist constants* $c_1=c_1(M)$, $c_2=c_2(M)$, $c_0=c_0(\alpha,\beta,\gamma)$, *such that*

$$\begin{aligned}&\frac{c_1t^{1-\gamma}-c_0}{1-\gamma}\le I(\alpha,\beta,\gamma)(t)\le\frac{c_2t^{1-\gamma}-c_0}{1-\gamma}\qquad(\gamma\ne 1),\\&c_1\log t-c_0\le I(\alpha,\beta,1)(t)\le c_2\log t-c_0.\end{aligned} \tag{4.6}$$

As a consequence of the above lemma, we note that $t^{-\alpha}(\sin t^\beta)^\gamma\in L^m[1,\infty)$ if and only if $m>1/\alpha$.

Now let $p>2$, $k>p/(p-2)$, and let $c=c_1(3)$ as defined by Lemma 4.1.

We may choose $a>1/p$ but sufficiently close to it that

$$0<\frac{1}{\alpha}<p<\frac{1}{\alpha-c} \qquad k>\frac{1/\alpha}{1/\alpha-2}=\frac{1}{1-2\alpha}.$$

Now $x=t^{-\alpha}\sin(t^{1+2\alpha})$ and $x=t^{-\alpha}\cos(t^{1+2\alpha})$ are readily shown to be solutions of $x''+qx=0$, where $q=(1+2\alpha)^2t^{4\alpha}-\alpha(\alpha+1)t^{-2}$. By choice of α, all solutions of (1.1)′ with q as above are in $L^p[1,\infty)\cap L^\infty[1,\infty)$.

A simple calculation shows that

$$x_0(t)=\exp\left(\int_0^t\tau^{-1}(\sin\tau^{1+2\alpha})^2d\tau\right)\cdot t^{-\alpha}\sin t^{1+2\alpha}$$

solves the equation

$$x''+(q+q_1)x=0,$$

where q_1 is given by

$$q_1(t)=-4(1+\alpha)t^{-1+2\alpha}\sin(t^{1+2\alpha})\cos(t^{1+2\alpha})$$
$$+(1+2\alpha)t^{-2}(\sin t^{1+2\alpha})^2-t^{-2}(\sin t^{1+2\alpha})^4.$$

Since $k>1/(1-2\alpha)$, we have $q_1\in L^k[1,\infty)$. However, Lemma 4.1 shows that

$$|x_0(t)|\geq e^{-c_0}t^{c-\alpha}|\sin(t^{1+2\alpha})|$$

for some constant c_0.

If $c\geq\alpha$, x_0 is clearly not in $L^p[1,\infty)$; if $c<\alpha$, then x_0 is not in $L^p[1,\infty)$ because $p<1/(\alpha-c)$.

This completes the proof of the theorem.

Remark. Bellman [2] has also given an extension of Weyl's theorem. In the context of Theorem 4.1, he obtained the cases $1\leq p\leq 2$, $k=\infty$.

5. Concluding remarks. Conditions for all solutions to be in $L^p[0,\infty)$ are numerous for the case $p=2$. For general p, examples may easily be found, for instance, by employing [3, Thm. 13, p. 120]. We do not know whether Theorem 2.4 or Corollary 2.4 is sharp.

We have concentrated on the case where solutions of (1.1) are in one of the classical Banach spaces $L^p[0,\infty)$, but results may be obtained under the more general assumption that solutions are in some function space $B\subset X[r]$, provided that one can obtain information about the distances between zeros of solutions. Although we have not attempted to do so in this paper, it seems likely that some analogue of Theorem 2.3 can be given if B is a weighted L^p space.

REFERENCES

[1] F. V. Atkinson, R. C. Grimmer and W. T. Patula, *Nonoscillatory solutions of forced second order linear equations* II, Ann. Math. Pura Appl., 126 (1980), pp. 296–323.

[2] R. Bellman, *A stability property of solutions of linear differential equations*, Duke Math. J., 11 (1944), pp. 513–516.

[3] W. Coppel, *Stability and Asymptotic Behavior of Differential Equations*, Heath, Boston, 1965.

[4] R. C. Grimmer and W. T. Patula, *Nonoscillatory solutions of forced second-order linear equations*, J. Math. Anal. Appl., 56 (1976), pp. 452–459.

[5] P. Hartman, *Differential equations with non-oscillatory eigenfunctions*, Duke Math. J., 15 (1948), pp. 697–709.

[6] ______, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[7] M. K. Kwong, *L^p perturbations of second order linear differential equations*, Math. Ann., 215 (1975), pp. 23–34.

[8] W. T. Patula, *On the distance between zeroes*, Proc. Amer. Math. Soc., 52 (1975), pp. 247–251.

[9] W. T. Patula and P. Waltman, *Limit point classification of second order linear differential equations*, J. London Math. Soc.,8 (1974), pp. 209–216.

[10] W. T. Patula and J. S. W. Wong, *An L^p-analogue of the Weyl alternative*, Math. Ann., 197 (1972), pp. 9–28.

[11] H. Weyl, *Über gewohnliche Differentialgleichungen mit Singularitäten und die zugehörige Entwicklung Wilkurlicher Funktionen*, Math. Ann., 68 (1910), pp. 220–269.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics
005

COMPARISON THEOREMS FOR DISFOCALITY AND DISCONJUGACY OF DIFFERENTIAL EQUATIONS*

URI ELIAS†

Abstract. Pairs of ordinary differential equations are compared with respect to disfocality and disconjugacy.

1. Introduction. We consider two-term ordinary differential equations of the type

$$(1)\qquad y^{(n)}+p(x)y=0,$$

where $p(x)$ has a fixed sign. Much of the work about oscillation and disconjugacy of (1) is done by using the concept of $(k,n-k)$-disfocality: (1) is called $(k,n-k)$-disfocal on an interval I if for every $a,b\in I$, $a<b$, no solution of (1), except the trivial one, satisfies

$$(2)\qquad \begin{aligned} &y^{(i)}(a)=0, && i=0,\cdots,k-1,\\ &y^{(j)}(b)=0, && j=k,\cdots,n-1.\end{aligned}$$

Similarly, (1) is $(k,n-k)$-disconjugate if only the trivial solution satisfies

$$(3)\qquad \begin{aligned} &y^{(i)}(a)=0, && i=0,\cdots,k-1,\\ &y^{(j)}(b)=0, && j=0,\cdots,n-k-1.\end{aligned}$$

The above concepts are applicable to the study of (1) thanks to some of the following properties. First, if $p\geq 0$ (≤ 0) and $n-k$ is even (odd) then (1) is $(k,n-k)$-disfocal and $(k,n-k)$-disconjugate on every interval. Thus it is sufficient to consider the values of k such that

$$(4)\qquad (-1)^{n-k}p\leq 0.$$

Next, there are simple relations between disfocality and disconjugacy. $(k,n-k)$-disfocality implies $(k,n-k)$-disconjugacy on every interval and eventual $(k,n-k)$-disfocality (that is disfocality on some ray (c,∞)) is equivalent to eventual $(k,n-k)$-disconjugacy. Finally, $(k,n-k)$-disfocality is elegantly characterized: If $(-1)^{n-k}p\leq 0$, then (1) is $(k,n-k)$-disfocal on I if and only if there exists a solution y of (1) such that

$$(5)_k\qquad \begin{aligned} &y^{(i)}>0, && i=0,\cdots,k-1,\\ &(-1)^{j-k}y^{(j)}>0, && j=k,\cdots,n-1\end{aligned}$$

on I. References to these known facts and others may be found in [7], [2].

In [5], Jones proved that if (1) is eventually $(k,n-k)$-disfocal and $k\leq(n+1)/2$, then it is also eventually $(k-2,n-k+2)$-disfocal. In fact, Jones formulated his results in terms of eventual disconjugacy; however, his proof is more natural in the framework of disfocality. By this ordering theorem he reduced substantially the number of possible oscillation types of (1). In the course of his proof, Jones also compared disfocality types of equations of different orders. His proofs are based on the characterization of

*Received by the editors June 28, 1982, and in revised form May 31, 1983.

†Department of Mathematics, Technion, Haifa 32000 Israel.

disfocality by a solution which satisfies (5) and inequalities of Kiguradze which such solution satisfies.

This note, motivated by Jones' results, is aimed to compare pairs of differential equations with respect to various types of disfocality and disconjugacy. Our comparison theorems are based on Green's functions inequalities.

2. Disfocality. Green's function $g_{k,n-k}(x,t)$ of the operator d^n/dx^n and the boundary conditions (2) is explicitly known:

$$(6)\qquad g_{k,n-k}(x,t)=\begin{cases} -\sum\limits_{i=k}^{n-1}(x-a)^i(a-t)^{n-1-i}/i!(n-i-1)!, & a\le x<t\le b,\\ \sum\limits_{i=1}^{k-1}(x-a)^i(a-t)^{n-1-i}/i!(n-i-1)!, & a\le t\le x\le b.\end{cases}$$

Note that $g_{k,n-k}$ is independent of b and it is defined practically for $a\le x$, $t<\infty$. Also

$$(7)\qquad \begin{aligned} &(-1)^{n-k}g^{(i)}_{k,n-k}>0, && i=0,\cdots,k-1,\\ &(-1)^{n-i}g^{(i)}_{k,n-k}\ge 0, && i=k,\cdots,n-1 \end{aligned}$$

on (a,b) (with equality for $i\ge k$, $t\le x$). For $i\ge k$, (7) is immediate since $g^{(i)}_{k,n-k}(x,t)=-(x-t)^{n-i-1}/(n-i-1)!$ for $x<t$ and $g^{(k)}_{k,n-k}\equiv 0$ for $x>t$. Integration of $g^{(k)}_{k,n-k}$ from a to x yields (7) for $i\le k-1$.

THEOREM 1. *If $k>l$ then*

$$(8)\qquad \frac{l!(n-l-1)!}{k!(n-k-1)!}\le\frac{(-1)^{n-k}g_{k,n-k}(x,t)}{(x-a)^k(t-a)^{n-k-1}}\bigg/\frac{(-1)^{n-l}g_{l,n-l}(x,t)}{(x-a)^l(t-a)^{n-l-1}}\le\frac{(l-1)!(n-l)!}{(k-1)!(n-k)!},$$

the quotient bounded in (8) *increases with x and decreases with t and equalities are attained in* (8) *when $x=a$ and $x\to\infty$ respectively.*

Proof. We rewrite the quotient in (8) as a product

$$\left[-\frac{(t-a)g_{k,n-k}}{(x-a)g_{k-1,n-k+1}}\right]\left[-\frac{(t-a)g_{k-1,n-k+1}}{(x-a)g_{k-2,n-k+2}}\right]\cdots\left[-\frac{(t-a)g_{l+1,n-l-1}}{(x-a)g_{l,n-l}}\right]$$

and show that each factor increases. Note that if u/v is continuous and not monotone, then there exists a linear combination of u and v with two zeros. In our case, if $g_{k,n-k}/(x-a)g_{k-1,n-k+1}$ is not monotone, there exists a linear combination $h(x)=(x-a)g_{k-1,n-k+1}+cg_{k,n-k}$ with two zeros in $(0,\infty)$. Since h has a zero of multiplicity k at $x=a$, we obtain by Rolle's theorem that $h^{(k)}$ changes its sign twice in (a,∞). But according to (6), h is a polynomial of degree $k-1$ on (t,∞) and $h^{(k)}\equiv 0$ there. Thus, the two changes of sign of $h^{(k)}$ must be located in (a,t). But $h^{(k)}(x)=(x-a)g^{(k)}_{k-1,n-k+1}+kg^{(k-1)}_{k-1,n-k+1}+cg^{(k)}_{k,n-k}$ and as $g^{(i)}_{k,n-k}=-(x-t)^{n-i-1}/(n-i-1)!$ for $i\ge k$ and $x<t$, it is immediately seen that $h^{(k)}$ does not have two zeros in (a,t). Similarly, none of $h^{(i)}$ has two distinct zeros. This contradiction confirms the monotony of the first factor. By (6), $-(t-a)g_{k,n-k}/(x-a)g_{k-1,n-k+1}$ attains at $x=a$ the value

$(k-1)!(n-k)/k!(n-k-1)!=(n/k)-1$ and it tends to $(n/(k-1))-1$ as $x\to\infty$; hence it increases. This argument applied to each of the $k-l$ factors proves that the quotient in (8) increases on $[a,\infty)$. The bounds are obtained by taking $x=a$ and $x\to\infty$ respectively. The monotony with respect to t is proved similarly.

It is possible to prove the inequalities in (8) by replacing $g_{k,n-k}$ for $x<t$ and $x>t$ by the corresponding polynomials and direct manipulation.

THEOREM 2. *Let* $(-1)^{n-k}p\le 0$ *and suppose* (1) *is* $(k,n-k)$*-disfocal on* $[a,b]$. *If* $l\le k$ *then*

$$y^{(n)}+(-1)^{k-l}\left(\binom{n-1}{k}\Big/\binom{n-1}{l}\right)p(x)y=0 \tag{9}$$

is $(l,n-l)$*-disfocal on* $[a,b]$, *and if* $l\ge k$ *then*

$$y^{(n)}+(-1)^{k-l}\left(\binom{n-1}{k-1}\Big/\binom{n-1}{l-1}\right)p(x)y=0 \tag{10}$$

is $(l,n-l)$*-disfocal there.*

Proof. It is known that if (1) is $(k,n-k)$-disfocal on $[a,b]$ and $(-1)^{n-k}p\le 0$, then the unique solution of (1) which satisfies

$$(11)_{k-1}\qquad \begin{aligned} &y^{(i)}(a)=0, && i=0,\cdots,k-2,\\ &y^{(k-1)}(a)=1, &&\\ &y^{(j)}(b)=0, && j=k,\cdots,n-1,\end{aligned}$$

is positive and even satisfies $(5)_k$ on (a,b). Equations (1) and (11) are equivalent to the integral equation

$$y(x)=(x-a)^{k-1}/(k-1)!+\int_a^b g_{k,n-k}(x,t)[-p(t)]\,y(t)\,dt. \tag{12}$$

Put $u(x)=y(x)/[(x-a)^{k-1}/(k-1)!]$. Dividing (12) by $(x-a)^{k-1}/(k-1)!$ we get

(13)

$$u(x)=1+\int_a^b\left[(-1)^{n-k}((t-a)/(x-a))^{k-1}g_{k,n-k}(x,t)\right]\left[(-1)^{n-k-1}p(t)\right]u(t)\,dt$$

and the integrand is positive by (4) and (7). If $k\ge l$, then by (8) we have

(14)

$$u(x)\ge 1+\left(\binom{n-1}{k}\Big/\binom{n-1}{l}\right)$$

$$\cdot\int_a^b\left[(-1)^{n-l}((t-a)/(x-a))^{l-1}g_{l,n-l}(x,t)\right]\left[(-1)^{n-k-1}p(t)\right]u(t)\,dt.$$

Now, it is known that if inequality (14) has a positive solution u, then the corresponding integral equation

$$v=1+\mathcal{K}v, \tag{15}$$

where $\mathcal{K}$ denotes the integral operator on the right-hand side of (14), has a solution v such that $0 \leq v(x) \leq u(x)$. This may be verified by defining iterations $v_0 = u$, $v_i = 1 + \mathcal{K}v_{i-1}$. We multiply now (15) by $(x-a)^{l-1}/(l-1)!$ and put $\tilde{y}(x) = v(x)(x-a)^{l-1}/(l-1)!$ to obtain

$$\tilde{y}(x) = (x-a)^{l-1}/(l-1)! + (-1)^{k-l}\left(\binom{n-1}{k}\Big/\binom{n-1}{l}\right)$$
$$\cdot \int_a^b g_{l,n-l}(x,t)[-p(t)]\,\tilde{y}(t)\,dt,$$

which is equivalent to (9) and the boundary conditions $(11)_{l-1}$. By $(11)_{l-1}$ we see that the solution $\tilde{y}$ of (9) is not only positive but also satisfies $(5)_l$ on (a,b), hence (9) is $(l,n-l)$-disfocal on (a,b). It is disfocal on $[a,b]$ since $\tilde{y}$ satisfies $(11)_{l-1}$ and consequently no solution can satisfy the $(l,n-l)$-focal point boundary value conditions at a and b.

To treat the case $l \geq k$, we exchange the roles of l and k in (8) and use analogously the right-hand side of the inequality so obtained. Another approach is to note that $(k,n-k)$-disfocality of (1) is equivalent to $(n-k,k)$-disfocality of its adjoint and $\binom{n-1}{k-1} = \binom{n-1}{n-k}$.

The results of Jones [5] follow if $\binom{n-1}{k}/\binom{n-1}{l}$ (or $\binom{n-1}{k-1}/\binom{n-1}{l-1}$) is not smaller than 1 and we neglect it in the proof of Theorem 2 and in (9), (10). Thus, the $(k,n-k)$-disfocality of (1), where $(-1)^{n-k}p \leq 0$, implies also its $(l,n-l)$ disfocality when

$$l \leq k, \ \binom{n-1}{l} \leq \binom{n-1}{k} \quad \text{or} \quad l \geq k, \ \binom{n-1}{l-1} \leq \binom{n-1}{k-1},$$

that is for $l = 1,2,\cdots,k-1, n-k+1,\cdots,n-1$ or $l = 1,2,\cdots,n-k-1, k+1,\cdots n-1$. This can be written as $|l-n/2| > |k-n/2|$ when we do not exclude the values of l such that $l \not\equiv k \pmod 2$ (for which $(l,n-l)$-disfocality of (1) is trivial and is not a consequence of Theorem 2).

Following Jones we summarize:

THEOREM 3. *If* (1) *is* $(k,n-k)$*-disfocal on* $[a,b]$, $(-1)^{n-k}p \leq 0$, *then* (1) *is also* $(l,n-l)$*-disfocal on* $[a,b]$ *where* $|l-n/2| > |k-n/2|$ *and* $l \equiv k \pmod 2$. *The equation* $y^{(n)} - py = 0$ *is* $(l,n-l)$*-disfocal when* $|l-n/2| > |k-n/2|$ *and* $l \not\equiv k \pmod 2$.

The methods of Theorems 1 and 2 may be adopted to compare disfocality of equations of different order. Compare with [5, Thms. 1–4].

THEOREM 4. a) *If* $k < n < m$ *then*

$$(16) \quad (m-k-1)!/(n-k-1)! \leq (-1)^{m-n}(t-a)^{m-n} g_{k,n-k}(x,t)/g_{k,m-k}(x,t)$$
$$\leq (m-k)!/(n-k)!,$$

the quotient bounded in (16) *increases with* x *and equalities are obtained when* $x = a$ *and* $x \to \infty$, *respectively.*

b) *Let* $(-1)^{n-k}p \leq 0$ *and suppose* (1) *is* $(k,n-k)$*-disfocal on* $[a,b]$. *If* $m \geq n$ *then*

$$(17) \qquad y^{(m)} + (-1)^{m-n}((m-k-1)!/(n-k-1)!)(x-a)^{n-m}p(x)y = 0$$

is $(k,m-k)$*-disfocal on* $[a,b]$ *and if* $k < m < n$ *then*

$$(18) \qquad y^{(m)} + (-1)^{m-n}((m-k)!/(n-k)!)(x-a)^{n-m}p(x)y = 0$$

is $(k,m-k)$*-disfocal there.*

In order to prove (16) we show that $h(x)=g_{k,n-k}-cg_{k,n-k-1}$ has not two zeros in (a,∞). Next, by (16), the integral equation (12) implies

$$y(x)\geq(x-a)^{k-1}/(k-1)!+(-1)^{m-n}((m-k-1)!/(n-k-1)!)$$
$$\cdot\int_a^b g_{k,m-k}(x,t)\left[-(t-a)^{n-m}p(t)\right]u(t)\,dt$$

and (17) follows. Note that $g_{k,m-k}(x,t)(t-a)^{n-m}$ has no singularity at $t=a$ even if $n<m$.

By composing Theorems 2 and 4 the following results are obtained.

THEOREM 5. *Let* $(-1)^{n-k}p\leq 0$ *and suppose* $y^{(n)}+py=0$ *is* $(k,n-k)$*-disfocal on* $[a,b]$. *Then the equation*

$$y^{(m)}+(-1)^{(m-l)+(n-k)}A_{k,l}(x-a)^{n-m}p(x)y=0 \tag{19}$$

is $(l,m-l)$*-disfocal on* $[a,b]$ *where*

$$A_{k,l}=\begin{cases}\dfrac{l!(m-l-1)!}{k!(n-k-1)!} & \textit{if } m\geq n,\ k\geq l,\\[2mm] \dfrac{(n-k)(l-1)!(m-l)!}{(m-k)(k-1)!(n-k)!} & \textit{if } m\geq n,\ k\leq l,\\[2mm] \dfrac{(l-1)!(m-l)!}{(k-1)!(m-k)!} & \textit{if } m\leq n,\ k\leq l,\\[2mm] \dfrac{(m-l)l!(m-l-1)!}{(n-l)k!(n-k-1)!} & \textit{if } m\leq n,\ k\geq l.\end{cases} \tag{20}$$

To prove this for $m\geq n$ we pass from the given $(k,n-k)$- to a $(k,m-k)$- and finally to a $(l,m-l)$-disfocal equation. For the case $m\leq n$ we follow the scheme $(k,n-k)\rightarrow(l,n-l)\rightarrow(l,m-l)$.

3. Disconjugacy. Now we turn to disconjugacy and $(k,n-k)$-disconjugacy. Green's function $G_{k,n-k}(x,t)$ of the operator d^n/dx^n and the boundary conditions (3) is obtained from $g_{k,n-k}$ when we replace $(x-a)$ and $(t-a)$ by $(x-a)(b-t)/(b-a)$ and $(b-x)(t-a)/(b-a)$, respectively [7]. Consequently, we see from (8) that the quotient

$$H(x,t)=\frac{(-1)^{n-k}G_{k,n-k}(x,t)}{(x-a)^{k-1}(b-x)^{n-k}(t-a)^{k-1}(b-t)^{n-k}}\Big/$$
$$\frac{(-1)^{n-l}G_{l,n-l}(x,t)}{(x-a)^{l-1}(b-x)^{n-l}(t-a)^{l-1}(b-t)^{n-l}}$$

where $k>l$, increases with x and is bounded by $l!(n-l-1)!/k!(n-k-1)!$ and $(l-1)!(n-l)!/(k-1)!(n-k)!$. Indeed, if we denote the quotient in (8) by $h(x,t)$ then $H(x,t)=h(a+(x-a)(b-t)/(b-a),\ a+(b-x)(t-a)/(b-a))$ and $H_x=h_x\cdot(b-t)/(b-a)-h_t\cdot(t-a)/(b-a)>0$ since $h_x>0$ and $h_t<0$. Similarly, by (16),

$$(m-k-1)!/(n-k-1)!\leq(-1)^{m-n}((b-x)(a-t)/(b-a))^{m-n}$$
$$\cdot G_{k,n-k}(x,t)/G_{k,m-k}(x,t)$$
$$\leq(m-k)!/(n-k)!$$

when $m \geq n$. Also the integral equation

$$y(x)=(x-a)^{k-1}(b-x)^{n-k}/(b-a)^{n-k}(k-1)!+\int_a^b G_{k,n-k}(x,t)[-p(t)]\,y(t)\,dt$$

is equivalent to (1) and the boundary value conditions

$$\begin{aligned} &y^{(i)}(a)=0, && i=0,\cdots,k-2,\\ &y^{(k-1)}(a)=1, && \\ &y^{(j)}(b)=0, && j=0,\cdots,n-k-1 \end{aligned}$$

and (1) is $(k,n-k)$-disconjugate on $[a,b]$ iff this y is positive on (a,b). Repeating the proof of Theorem 2 with $u(x)=y(x)/[(x-a)^{k-1}(b-x)^{n-k}/(b-a)^{n-k}(k-1)!]$, we obtain

THEOREM 6. *Theorems* 2 *and* 3 *remain valid if the term disfocality is replaced everywhere by disconjugacy. Theorem* 5 *remains valid if the term disfocality is replaced by disconjugacy and the factor* $(x-a)^{n-m}$ *in* (19) *is replaced by* $((x-a)(b-x)/(b-a))^{n-m}$.

Recall that while eventual $(k,n-k)$-disfocality and eventual $(k,n-k)$-disconjugacy are equivalent, disfocality on $[a,b]$ implies disconjugacy there but is not implied by it. Therefore the last results are not direct consequences of Theorems 2 and 3.

Since $(k,n-k)$-disconjugacy (disfocality) on $[a,b]$ is equivalent to the absence of $(k,n-k)$-type conjugate (focal) point $\eta_{k,n-k}(a)(\zeta_{k,n-k}(a))$ on $[a,b]$, we can restate Theorems 3 and 6 as

THEOREM 6′. *If* $|l-n/2|>|k-n/2|$, $l\equiv k(\mathrm{mod}\,2)$ *then*

$$a<\eta_{k,n-k}(a)\leq\eta_{l,n-l}(a)\leq\infty,$$

$$a<\zeta_{k,n-k}(a)\leq\zeta_{l,n-l}(a)\leq\infty.$$

Various works deal with relations between disconjugacy (nonoscillation) of (1) on $[a,\infty)$ and that of various second order equations. For example, see [3], [6], [4]. In Theorem 6 of [2] we proved that if (1) is eventually disconjugate and $n>m$, then $y^{(n)}+((m-1)!/(n-1)!)p(x)(x-a)^{n-m}y=0$ is eventually disconjugate, too. (In fact $(m-1)!/(n-1)!$ is replaced by the smaller constant $m!/n!$ but the bigger constant is immediately available without any change in the proof). Now we improve this result and extend it.

THEOREM 7. *Let* (1) *be disconjugate on* $[a,b]$. *If* $m<n$, *then the two equations*

$$(21)\qquad y^{(m)}\pm\frac{[n/2][m/2]![(m+1)/2]!}{(n-[m/2])[n/2]![(n+1)/2]!}\left(\frac{(x-a)(b-x)}{b-a}\right)^{n-m}p(x)y=0$$

are disconjugate on $[a,b]$ *and if* $m>n$, *then the equations*

$$(22)\qquad y^{(m)}\pm\frac{[(n-1)/2][m/2]![(m-1)/2]!}{(m-[(n-1)/2])[n/2]![(n-1)/2]!}\left(\frac{(x-a)(b-x)}{b-a}\right)^{n-m}p(x)y=0$$

are disconjugate. (*Here* [ ] *denotes the integer part function*).

Analogous results hold when we replace $[a,b]$ *by* $[a,\infty)$ *and* $(x-a)(b-x)/(b-a)$ *by* $x-a$.

Proof. An equation of type (1) is disconjugate when it is $(i,n-i)$-disconjugate for every i, $1\leq i\leq n-1$, such that $(-1)^{n-i}p\leq 0$. In order to prove that this is the case, it is sufficient, by Theorems 3 and 6, to show that it is $(i, n-i)$-disconjugate either for $i=[n/2]$ or for $[n/2]+1$, according to the parity of n and the signature of p. Indeed, one of these values of i has the required parity and for it $|i-n/2|$ is minimal. In our case we have to show that (21) (or (22)) is $(l, m-l)$-disconjugate either for $l=[m/2]$ or for $l=[m/2]+1$. But there are many possibilities to deduce this from the disconjugacy of (1). According to Theorems 6 and 3, $(k, n-k)$-disconjugacy of (1), when $1 \leq k \leq n-1$, $(-1)^{n-k} p \leq 0$, implies $(l, m-l)$-disconjugacy of

$$y^{(m)}+(-1)^{m-l+n-k} A_{kl} p(x)((x-a)(b-x)/(b-a))^{n-m} y=0 \tag{23}$$

where A_{kl} is given in (20). Thus, we are in possession of several values of k by each of which we can deduce $(l, m-l)$-disconjugacy of an mth order equation. Since we did not assume what is the sign of $p(x)$ and the parity of n, we have to check two values of k of different parities and select from the two so-obtained mth order equations the one with the smaller coefficient A_{kl}. Since l will be either $[m/2]$ or $[m/2]+1$, it is easily seen that the best choice of a pair of values of k is $[(n-1)/2]$, $[(n+1)/2]$. First let $m<n$. If $m \leq n-2$, then $l \leq [m/2]+1 \leq [(n-1)/2] \leq k$ and by (20), $A_{kl}=(m-l)!l!/(n-l)k!(n-k-1)!$ is smaller for $k=[(n+1)/2]$. Among the two candidates for l, A_{kl} is smaller for $l=[m/2]$ and its value is

$$C_{mn}=\frac{[m/2]![(m+1)/2]!}{(n-[m/2])[(n+1)]![n/2-1]!}$$

(use $[n/2]+[(n+1)/2]=n$). This proves that both equations in (21) are disconjugate if $m \leq n-2$. If $m=n-1$, the result is still valid since by direct calculation we get an equation in which the numerical coefficient is even larger than that in (21).

When $m>n$, we select among $k=[(n-1)/2]$, $[(n+1)/2]$ the first one and from $l=[m/2]$, $[m/2]+1$ we choose the second to obtain (22). For $n=2$, only the + sign and $(1,1)$-disconjugacy are acceptable.

Let C_{mn} be defined by the numerical coefficient in (21) or (22), according as $m<n$ or $m>n$. Theorem 7 may be stated also as follows:

THEOREM 7′. *In order that $y^{(n)}+py=0$ be disconjugate on $[a, \infty)$, it is necessary that both $y^{(m)} \pm C_{mn}(x-a)^{n-m} p(x) y=0$ be disconjugate on $[a, \infty)$ and sufficient that one of $y^{(m)} \pm C_{nm}^{-1}(x-a)^{n-m} p(x) y=0$ be disconjugate on the interval (for $m=2$ take only + sign).*

Theorem 7 improves known results for $n \geq 4$ but is worse than those for $n=3$, $m=2$. Obviously the numerical constants in Theorem 7 are not the best possible. To estimate how good these constants are, we may compare the constants in the necessary and the sufficient conditions of Theorem 7′. For $n>m$, for example,

$$\frac{C_{mn}}{C_{nm}^{-1}}=\frac{[(m-1)/2](m-[m/2])(m-[(m-1)/2])}{[(n+1)/2](n-[m/2])(n-[(m-1)/2])}<1,$$

and the distance between this ratio and 1 give an idea how far are our results from the optimum.

When the sign of p and the parities of m and n are known, we can obviously get specific results which are better than the last theorem.

We can compare by similar method $y^{(n)}+py=0$ and $y^{(n)}-py=0$. If n is odd, these equations are adjoint and they are disconjugate together. However when n is even we obtain a nontrivial result. Following the proof of Theorem 7 we take in (23) $l=n/2$ and

$k=n/2+1$ if $(-1)^{n/2-1}p<0$ and $l=n/2+1$, $k=n/2$ if $(-1)^{n/2}p>0$. We obtain

THEOREM 8. *Let $y^{(2r)}+py=0$ be disconjugate. If $(-1)^r p\leq 0$, then $y^{(2r)}-py=0$ is disconjugate too. If $(-1)^r p>0$, then $y^{(2r)}-((r-1)/(r+1))py=0$ is disconjugate there.*

For $2r=4$ this result is far from being strict: We know that $y^{(4)}+Ax^{-4}y=0$ is disconjugate on $(0,\infty)$ for $0\leq A\leq 1$ while $y^{(4)}-Ax^{-4}y=0$ is disconjugate for $0\leq A\leq 9/16$.

4. Estimates of solutions.

THEOREM 9. *Suppose $(-1)^{n-k}p\leq 0$ and (1) is $(k,n-k)$-disfocal on $[a,b]$ and let $y_{k-1}(x,b)$ be the unique solution of (1) which satisfies the boundary conditions $(11)_{k-1}$. If $|l-n/2|>|k-n/2|$ and $(-1)^{n-l}p(x)\leq 0$, then the corresponding solution $y_{l-1}(x,b)$ of (1), $(11)_{l-1}$ satisfies $(5)_l$ and*

$$(24)\qquad 0<\frac{y_{l-1}(x,b)}{(x-a)^{l-1}/(l-1)!}\leq\frac{y_{k-1}(x,b)}{(x-a)^{k-1}/(k-1)!},$$

$$(25)\qquad 0\leq(-1)^{p-1}\frac{d^p}{dx^p}\left(\frac{y_{l-1}(x,b)}{(x-a)^{l-1}/(l-1)!}\right)\frac{\binom{n-1}{k+p-1}}{\binom{n-1}{l+p-1}}$$

$$\leq(-1)^{p-1}\frac{d^p}{dx^p}\left(\frac{y_{k-1}(x,b)}{(x-a)^{k-1}/(k-1)!}\right)\qquad p=1,\cdots,n-k-1,$$

on $[a,b]$.

In fact, we have already proved (24). In the proof of Theorem 2 it was seen that $u(x)=y(x)/[(x-a)^{k-1}/(k-1)!]$ and $v(x)=\tilde{y}(x)/[(x-a)^{l-1}/(l-1)!]$ satisfy $0\leq v(x)\leq u(x)$. If $|l-n/2|>|k-n/2|$ and we replace in the proof of Theorem 2 $\binom{n-1}{k}/\binom{n-1}{l}$ (or $\binom{n-1}{k-1}/\binom{n-1}{l-1}$) by the smaller number 1, we obtain (24) for the solutions y_{k-1}, y_{l-1} of (1). To prove (25), we need an extension of Theorem 1.

THEOREM 10. a) *For $q=0,\cdots,l$, $p=0,\cdots,n-l-1$ we have*

$$(26)\qquad (-1)^{(p-q)_+}\frac{\partial^p}{\partial x^p}\left(\frac{(-1)^{n-l}g_{l,n-l}(x,t)}{(x-a)^{l-q}}\right)\geq 0$$

where $i_+=\max\{i,0\}$.

b) *If $k>l$, then for $q=0,\cdots,l$, $p=0,\cdots,n-l-1$, the ratios*

$$(27)\qquad (-1)^{(p-q)_+}\frac{\partial^p}{\partial x^p}\left(\frac{(-1)^{n-k}g_{k,n-k}(x,t)}{(x-a)^{k-q}(t-a)^{n-k+q-1}}\right)\Big/$$

$$(-1)^{(p-q)_+}\frac{\partial^p}{\partial x^p}\left(\frac{(-1)^{n-l}g_{l,n-l}(x,t)}{(x-a)^{l-q}(t-a)^{n-l+q-1}}\right)$$

increase from $\binom{n-1}{k+(p-q)_+}/\binom{n-1}{l+(p-q)_+}$ to $\binom{n-1}{k+(p-q)_+-1}/\binom{n-1}{l+(p-q)_+-1}$ as x varies from a to ∞.

Once we know Theorem 10, it is easy to prove (25). Recall that $u=y_{k-1}(x,b)/[(x-a)^{k-1}/(k-1)!]$ satisfies (13) and $v=y_{l-1}/[(x-a)^{l-1}/(l-1)!]$ satisfies a similar integral equation. Differentiation of (13) yields

$$(28)\quad u^{(p)}(x)=\int_a^b \frac{\partial^p}{\partial x^p}\left\{(t-a)^{n-1}\frac{(-1)^{n-k}g_{k,n-k}}{(x-a)^{k-1}(t-a)^{n-k}}\right\}\{(-1)^{n-k-1}p(t)\}u(t)\,dt.$$

Equation (25) follows immediately if we apply Theorem 10(b) with $q=1$ and the inequality $u\geq v>0$ to (28).

Proof of Theorem 10. In [1, Thm. 1] it is shown that if for some j, k, $j\geq k\geq 0$, a function y fulfills

$$(29)\qquad y(a),\cdots,y^{(k-1)}(a)\geq 0,\qquad (-1)^{j-k}y^{(j)}(x)\geq 0\quad\text{on }[a,\infty),$$

then

$$(30)\qquad (-1)^{j-k}\left(y/(x-a)^k\right)^{(j-k)}\geq 0\quad\text{on }[a,\infty).$$

Equation (26) is a particular case of (30) with $k=l-q$, $j=l-q+p$ and $y(x)=(-1)^{n-l+p-(p-q)_+}g_{l,n-l}(x,t)$. Indeed, (29) holds since $y(a)=\cdots=y^{(l-q-1)}(a)=0$ and $(-1)^{j-k}y^{(j)}=(-1)^{n-l-(p-q)_+}g_{l,n-l}^{(l+p-q)}$ is positive by (7), either if $p-q\geq 0$ or $p-q<0$. Here we used $l\leq q$, $l+p-q\leq n-1$. Consequently, by (30),

$$0\leq(-1)^{j-k}\left(y/(x-a)^k\right)^{(j-k)}=(-1)^p\left((-1)^{n-l+p-(p-q)_+}g_{l,n-l}/(x-a)^{l-q}\right)^{(p)}$$
$$=(-1)^{(p-q)_+}\frac{\partial^p}{\partial x^p}\left((-1)^{n-l}g_{l,n-l}/(x-a)^{l-q}\right)$$

and (26) is proved.

The proof of part (b) makes use of the identity

$$(31)\qquad \left(\frac{d}{dx}\right)^i x^j\left(\frac{d}{dx}\right)^{j-i}x^{-i}y\equiv x^{j-i}\left(\frac{d}{dx}\right)^j y.$$

This equality between two differential operators of order j may be verified, for example, by applying both sides to the functions x^α, $-\infty<\alpha<\infty$.

We return now to the monotony of (27). If

$$(32)\qquad \frac{\partial^p}{\partial x^p}\left(g_{k,n-k}/(x-a)^{k-q}\right)\Big/\frac{\partial^p}{\partial x^p}\left(g_{k-1,n-k+1}/(x-a)^{k-q+1}\right)$$

is not monotone, then there is a linear combination

$$\frac{\partial^p}{\partial x^p}\left(g_{k-1,n-k+1}/(x-a)^{k-q+1}\right)+c\frac{\partial^p}{\partial x^p}\left(g_{k,n-k}/(x-a)^{k-q}\right)$$

with two zeros in (a,∞). When we multiply this combination by $(x-a)^{p-q+k}$, $(p-q+k\geq p\geq 0)$, we see that the function

$$(x-a)^{p-q+k}\left(\frac{\partial}{\partial x}\right)^p(x-a)^{-(k-q)}\left((x-a)g_{k-1,n-k+1}+cg_{k,n-k}\right)$$

has two zeros in (a,∞) and at least $(p-q+k)+(q-p)_+$ zeros, hence not less than k zeros, at $x=a$. Consequently, its $(k-q)$th derivative, $(q\leq l<k)$,

$$\left(\frac{d}{dx}\right)^{k-q}(x-a)^{p-q+k}\left(\frac{d}{dx}\right)^p(x-a)^{-(k-q)}h(x),$$

changes its sign at least twice in (a, ∞). But according to (31) this means that $(x-a)^p(d/dx)^{k-q+p}h(x)$ changes its sign twice in (a, ∞) while we have seen in the proof of Theorem 1 that $h(x)=(x-a)g_{k-1,n-k+1}+cg_{k,n-k}$ and its derivatives cannot have such zero distribution in (a, ∞). This proves the monotony of (32), and in turn establishes Theorem 11.

REFERENCES

[1] U. ELIAS, *Generalizations of an inequality of Kiguradze*, J. Math. Anal. Appl., to appear.

[2] ______, *Necessary conditions and sufficient conditions for disfocality and disconjugacy of differential equations*, Pacific J. Math., 81 (1979), pp. 379–397.

[3] G. J. ETGEN AND C. D. SHIH, *Disconjugacy and oscillation of third order differential equations with nonnegative coefficients*, Proc. Amer. Math. Soc., 38 (1973), pp. 577–582.

[4] R. GRIMMER, *Comparison theorems for third and fourth order linear operators*, J. Differential Equations, 25 (1977), pp. 1–10.

[5] G. D. JONES, *An ordering of oscillation types for* $y^{(n)}+py=0$, this Journal, 12 (1981), pp. 72–77.

[6] D. L. LOVELADY, *Oscillation and a class of odd order linear differential operators*, Hiroshima Math. J., 5 (1975), pp. 371–383.

[7] Z. NEHARI, *Green's functions and disconjugacy*, Arch. Rat. Mech. Anal., 62 (1976), pp. 53–76.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics
006

ASYMPTOTIC INTEGRATION OF LINEAR DIFFERENTIAL EQUATIONS SUBJECT TO MILD INTEGRAL CONDITIONS*

WILLIAM F. TRENCH[†]

Abstract. Sufficient conditions are given for a linear differential equation of order n to have a solution which behaves asymptotically like a given polynomial of degree $<n$. The integral smallness conditions on the coefficient and forcing functions are stated largely in terms of ordinary (rather than absolute) convergence, and the manner in which the solution behaves like the given polynomial is specified precisely.

1. Introduction and main theorem. We study the behavior as $t\to\infty$ of solutions of the scalar equation

$$x^{(n)}+P_1(t)x^{(n-1)}+\cdots+P_n(t)x=f(t),\qquad t>0(n\geq 2),\tag{1}$$

where $P_1,\cdots,P_n$, f, and x may be complex-valued. We regard (1) as a perturbation of the equation

$$y^{(n)}=0,\tag{2}$$

and give conditions which imply that (1) has a solution x_0 which behaves for large t like a given polynomial p of degree $<n$. Although this problem has already received much attention, we believe that our results are of interest because we specify bounds on the differences $x_0^{(r)}-p^{(r)}$ $(0\leq r\leq n-1)$ more precisely than is usually the case, and our integral smallness conditions on $P_1,\cdots,P_n$, and f are stated largely in terms of improper integrals which may converge conditionally rather than absolutely, as is usually required.

The main theorem is stated and proved in §2. Section 3 contains corollaries and examples. Section 4 is an appendix which contains the proof of a lemma used in §2.

2. The main theorem. Throughout this section, p is a given polynomial of degree $<n$. For convenience below, we rewrite (1) as

$$x^{(n)}+Mx=f,\tag{3}$$

where

$$Mx=\sum_{k=1}^{n}P_kx^{(n-k)},$$

and introduce the new unknown

$$h=x-p.\tag{4}$$

Since $p^{(n)}=0$, it is obvious that x is a solution of (3) (and therefore of (1)) if and only if h is a solution of

$$h^{(n)}=-Mh-g,\tag{5}$$

where

$$g=-f+Mp=-f+\sum_{k=1}^{n}P_kp^{(n-k)}.\tag{6}$$

*Received by the editors August 11, 1982, and in revised form February 10, 1983.

[†] Department of Mathematical Sciences, Drexel University, Philadelphia, Pennsylvania 19104.

Thus, g may be regarded as a measure of the extent to which p, a solution of the unperturbed equation (2), fails to be a solution of the perturbed equation (1).

The following is our main theorem.

THEOREM 1. *Let $P_1,\cdots,P_n$ and f be continuous on $(0,\infty)$, and let g be as defined in* (6), *where p is a given polynomial of degree $<n$. Suppose the integral $\int^\infty t^{n-m-1}g(t)\,dt$ converges, and*

$$(7)\qquad \int_t^\infty s^{n-m-1}g(s)\,ds=O(\phi(t)),$$

where m is an integer in $\{0,1,\cdots,n-1\}$ and ϕ is continuous, positive, and nonincreasing on $[\bar{T},\infty)$ for some $\bar{T}\geq 0$. Also, if $m\neq 0$, suppose $t^\gamma\phi(t)$ is nondecreasing on $[\bar{T},\infty)$ for some $\gamma<1$. Assume also that

$$(8)\qquad \int^\infty |P_1(t)|\,dt<\infty,$$

and that the integrals $\int^\infty P_k(t)\,dt$ $(2\leq k\leq n)$ converge and satisfy

$$(9)\qquad \int_t^\infty P_k(s)\,ds=o(t^{-k+1}),\qquad 2\leq k\leq n.$$

Finally, suppose also that

$$(10)\qquad \int_t^\infty s^{k-2}\phi(s)\left|\int_s^\infty P_k(\lambda)\,d\lambda\right|ds=o(\phi(t)),\qquad 2\leq k\leq n.$$

Then (1) *has a solution x_0 such that*

$$(11)\qquad x_0^{(r)}(t)=p^{(r)}(t)+O(\phi(t)t^{m-r}),\qquad 0\leq r\leq n-1.$$

Moreover, if (7) *holds with "O" replaced by "o", then so does* (11).

Remark. Under the stated assumptions on ϕ it is clear that if $\lim_{t\to\infty}\phi(t)>0$, then it may as well be assumed that $\phi=1$. In this case, of course, (7) holds with "O" replaced by "o," and therefore so does (11).

By way of motivation, we first outline the proof of Theorem 1.

From the remarks preceding the statement of Theorem 1, x_0 is a solution of (1) which satisfies (11) if and only if

$$x_0=p+h_0$$

(see (4)), where h_0 is a solution of (5) such that

$$(12)\qquad h_0^{(r)}(t)=O(\phi(t)t^{m-r}),\qquad 0\leq r\leq n-1.$$

We will show that (5) has a solution with these properties by exhibiting h_0 as the fixed point of a contraction mapping on the Banach space $H(t_0)$ of functions h in $C^{(n-1)}[t_0,\infty)$ such that

$$(13)\qquad h^{(r)}(t)=O(\phi(t)t^{m-r}),\qquad 0\leq r\leq n-1,$$

with norm

$$(14)\qquad \|h\|=\sup_{t\geq t_0}\left\{(\phi(t))^{-1}\sum_{r=0}^{n-1}t^{r-m}|h^{(r)}(t)|\right\}.$$

The contraction mapping will be obtained by converting (5) to an integral equation whose form is dictated by the integrability conditions that we have imposed. To guarantee that the mapping which we define in fact has the contraction property, we must assume that t_0 is sufficiently large, and the fixed point (function) h_0 is at first defined only on $[t_0, \infty)$. However, this presents no difficulty, since our assumptions clearly guarantee the continuability of any solution of (5) over $(0, \infty)$.

With g as in (6), let

$$G(t)=\int_t^\infty \frac{(t-s)^{n-1}}{(n-1)!}g(s)\,ds \quad \text{if } m=0, \tag{15}$$

or

$$G(t)=\int_{t_0}^t \frac{(t-\lambda)^{m-1}}{(m-1)!}d\lambda\int_\lambda^\infty \frac{(\lambda-s)^{n-m-1}}{(n-m-1)!}g(s)\,ds \quad \text{if } m=1,\cdots,n-1, \tag{16}$$

and notice that our integrability condition on g implies that the improper integral in (15) or (16) converges, by Dirichlet's theorem for improper integrals.

Now define the transformation L by

$$(Lh)(t)=\int_t^\infty \frac{(t-s)^{n-1}}{(n-1)!}(Mh)(s)\,ds \quad \text{if } m=0, \tag{17}$$

or by

$$(Lh)(t)=\int_{t_0}^t \frac{(t-\lambda)^{m-1}}{(m-1)!}d\lambda\int_\lambda^\infty \frac{(\lambda-s)^{n-m-1}}{(n-m-1)!}(Mh)(s)\,ds \quad \text{if } m=1,\cdots,n-1. \tag{18}$$

We will show that the mapping $\mathfrak{T}$ defined by

$$\mathfrak{T}h=G+Lh \tag{19}$$

maps $H(t_0)$ into itself, and is a contraction mapping if t_0 is sufficiently large. It will then follow that $\mathfrak{T}$ has a fixed point (function) h_0 in $H(t_0)$ such that

$$\mathfrak{T}h_0=h_0. \tag{20}$$

If $m=0$, then (15), (17), (19), and (20) imply that h_0 satisfies the integral equation

$$h_0(t)=\int_t^\infty \frac{(t-s)^{n-1}}{(n-1)!}[Mh_0(s)+g(s)]\,ds. \tag{21}$$

If $m=1,\cdots,n-1$, then (16), (18), (19), and (20) imply that h_0 satisfies the integral equation

$$h_0(t)=\int_{t_0}^t \frac{(t-\lambda)^{m-1}}{(m-1)!}d\lambda\int_\lambda^\infty \frac{(\lambda-s)^{n-1}}{(n-m-1)!}[Mh_0(s)+g(s)]\,ds. \tag{22}$$

In either case, routine differentiation shows that h_0 satisfies (5). Since $h_0 \in H(t_0)$, it automatically satisfies (12).

From these observations it should be clear that the proof reduces to showing that the mapping $\mathfrak{T}$ is a contraction mapping of $H(t_0)$ into itself if t_0 is sufficiently large, and that (12) can be replaced by

$$h_0^{(r)}(t)=o(\phi(t)t^{m-r}),\qquad 0\leq r\leq n-1, \tag{23}$$

if (7) holds with "O" replaced by "o." The following lemma is crucial for this proof.

LEMMA 1. *Let ϕ, m, and γ be as in Theorem 1, and suppose $w\in C[t_0,\infty)$ for some $t_0\geq \overline{T}$. Suppose also that $\int^\infty t^{n-m-1}w(t)\,dt$ converges, and*

$$\int_t^\infty s^{n-m-1}w(s)\,ds=O(\phi(t)), \tag{24}$$

and define

$$\rho(t)=\sup_{\tau\geq t}\left|(\phi(\tau))^{-1}\int_\tau^\infty s^{n-m-1}w(s)\,ds\right|. \tag{25}$$

Then the function v defined by

$$v(t)=\int_t^\infty \frac{(t-s)^{n-1}}{(n-1)!}w(s)\,ds\quad \text{if } m=0 \tag{26}$$

or by

$$v(t)=\int_{t_0}^t \frac{(t-\lambda)^{m-1}}{(m-1)!}d\lambda\int_\lambda^\infty \frac{(\lambda-s)^{n-m-1}}{(n-m-1)!}w(s)\,ds\quad \text{if } m=1,2,\cdots,n-1, \tag{27}$$

is in $C^{(n)}[t_0,\infty)$, and it satisfies the inequalities

$$|v^{(r)}(t)|\leq\frac{\rho(t_0)\phi(t)t^{m-r}}{(n-m-1)!\prod_{j=1}^{m-r}(j-\gamma)},\qquad 0\leq r\leq m-1, \tag{28}$$

$$|v^{(m)}(t)|\leq\frac{\rho(t)\phi(t)}{(n-m-1)!}, \tag{29}$$

and

$$|v^{(r)}(t)|\leq\frac{2\rho(t)\phi(t)t^{m-r}}{(n-r-1)!},\qquad m+1\leq r\leq n-1. \tag{30}$$

Moreover, if

$$\lim_{t\to\infty}\rho(t)=0, \tag{31}$$

then

$$v^{(r)}(t)=o(\phi(t)t^{m-r}),\qquad 0\leq r\leq n-1. \tag{32}$$

We leave the proof of this lemma for the appendix (§4). Since the lemma would be essentially trivial under the stronger assumption that $\int^\infty t^{n-m-1}|w(t)|\,dt<\infty$, it is important to observe that we are not assuming this. Notice that the lemma implies that the function v defined by (26) or (27) is in $H(t_0)$.

Proof of Theorem 1. First notice that, because of (7), Lemma 1 with $w=g$ implies that G, as defined by (15) or (16), is in $H(t_0)$ for any $t_0>0$. The next step, then, is to show that Lh (see (17) or (18)) is defined and in $H(t_0)$ whenever $h\in H(t_0)$. We start by showing that the improper integral in (17) or (18) converges if $h\in H(t_0)$. To this end, we first consider the integral

$$(33)\qquad J(t;h)=\int_t^{\infty} s^{n-m-1}(Mh)(s)\,ds=\sum_{k=1}^{n}\int_t^{\infty} s^{n-m-1}P_k(s)h^{(n-k)}(s)\,ds.$$

We will show that the integrals in this sum converge, and estimate them. In the following, let $s\geq t\geq t_0$.

From (14),

$$|s^{n-m-1}P_1(s)h^{(n-1)}(s)|\leq\|h\|\,|P_1(s)|\phi(s).$$

Therefore, (8) and the monotonicity of ϕ imply that the first integral on the right of (33) converges, and that

$$(34)\qquad \left|\int_t^{\infty} s^{n-m-1}P_1(s)h^{(n-1)}(s)\,ds\right|\leq\|h\|\phi(t)\int_t^{\infty}|P_1(s)|ds.$$

If $2\leq k\leq n$, then integration by parts yields

$$(35)\qquad \int_t^{\infty} s^{n-m-1}P_k(s)h^{(n-k)}(s)\,ds=t^{n-m-1}h^{(n-k)}(t)\int_t^{\infty}P_k(\lambda)\,d\lambda$$
$$+\int_t^{\infty}\left[s^{n-m-1}h^{(n-k)}(s)\right]'\left(\int_s^{\infty}P_k(\lambda)\,d\lambda\right)ds.$$

To justify this, observe that

$$\lim_{t\to\infty} t^{n-m-1}h^{(n-k)}(t)\int_t^{\infty}P_k(\lambda)\,d\lambda=0,$$

because of (9) and (13), and the integral on the right of (35) converges absolutely because of the convergence of the integral in (10) and the inequality

$$(36)\qquad |[s^{n-m-1}h^{(n-j)}(s)]'|\leq(n-m)\|h\|s^{k-2}\phi(s),$$

which follows from (14) and straightforward manipulation.

This proves that $J(t;h)$ converges. Moreover, from (10), (14), (34), (35), and (36),

$$(37)\qquad |J(t;h)|\leq\|h\|\phi(t)\sigma(t),$$

where

$$\sigma(t)=\int_t^{\infty}|P_1(\lambda)|d\lambda+\sum_{j=2}^{n}t^{k-1}\left|\int_t^{\infty}P_j(\lambda)\,d\lambda\right|$$
$$+(n-m)(\phi(t))^{-1}\sum_{j=2}^{n}\int_t^{\infty}s^{j-2}\phi(s)\left|\int_s^{\infty}P_j(\lambda)\,d\lambda\right|ds.$$

Now we can apply Lemma 1 with $w=Mh$ and $v=Lh$. (Compare (26) and (27) with (17) and (18).) Then (25) becomes

$$\rho(t)=\sup_{\tau\geq t}(\phi(\tau))^{-1}|J(\tau;h)|,$$

which, with (37), implies that

$$(38)\qquad \rho(t)\leq\|h\|\sup_{\tau\geq t}\sigma(\tau)=o(1).$$

Now (28), (29), and (30) with $w=Mh$ and $v=Lh$ imply that $Lh\in H(t_0)$ and

$$\|Lh\|\leq K\|h\| \sup_{\tau\geq t_0} \sigma(\tau),$$

where K is a universal constant.

Since G is also in $H(t_0)$, the transformation $\mathfrak{T}$ defined in (19) also maps $H(t_0)$ into itself. Moreover, if $h_1,h_2\in H(t_0)$, then

$$\|\mathfrak{T}h_1-\mathfrak{T}h_2\|=\|L(h_1-h_2)\|\leq K\|h_1-h_2\| \sup_{\tau\geq t_0} \sigma(\tau).$$

Therefore, $\mathfrak{T}$ is a contraction mapping if t_0 is so large that

$$\sup_{\tau\geq t_0} \sigma(\tau)<1/K,$$

which we now assume. (Recall that $\sigma(t)=o(1)$.) Consequently, $\mathfrak{T}$ has a fixed point (function) h_0 which satisfies

$$h_0=G+Lh_0, \tag{39}$$

which can also be written out as (21) if $m=0$, or as (22) if $m=1,\cdots,n-1$. Since (38) implies (31), Lemma 1 with $w=Mh_0$ and $v=Lh_0$ implies that

$$(Lh_0)^{(r)}(t)=o(\phi(t)t^{m-r}), \qquad 0\leq r\leq n-1. \tag{40}$$

Moreover, if we can replace "O" by "o" in (7), then Lemma 1 implies that

$$G^{(r)}(t)=o(\phi(t)t^{m-r}), \qquad 0\leq r\leq n-1. \tag{41}$$

But (39), (40), and (41) imply (23); that is, in this case we can replace "O" by "o" in (11). This completes the proof of Theorem 1.

3. Corollaries and examples. There are applications of Theorem 1 in which (8) is the only integral smallness condition on functions appearing in (1) which requires absolute convergence. The following corollary illustrates this.

COROLLARY 1. *Theorem* 1 *remains valid if* (10) *is replaced by*

$$\int_t^\infty \frac{\phi(s)}{s}\,ds=O(\phi(t)). \tag{42}$$

Proof. If (42) holds, then (9) implies (10).

The following corollary is of interest if (42) does not hold.

COROLLARY 2. *Theorem* 1 *remains valid if* (10) *is replaced by*

$$\int^\infty s^{k-2}\left|\int_s^\infty P_k(\lambda)\,d\lambda\right|ds<\infty, \qquad 2\leq k\leq n. \tag{43}$$

Proof. Since ϕ is nonincreasing, (43) implies (10).

COROLLARY 3. *Theorem* 1 *remains valid if* (10) *is replaced by*

$$\int^\infty t^{k-1}|P_k(t)|\,dt<\infty, \qquad 2\leq k\leq n. \tag{44}$$

Proof. We will show that (44) implies (43). If (44) holds, then the function

$$Q_k(t)=\int_t^\infty |P_k(s)|\,ds$$

is defined on $(0, \infty)$, and

$$Q_k(t) = o(t^{-k+1}). \tag{45}$$

Integration by parts yields

$$\int_{t_1}^{t_2} s^{k-2} Q_k(s)\, ds = \frac{1}{k-1} s^{k-1} Q_k(s)\Big|_{t_1}^{t_2} - \frac{1}{k-1}\int_{t_1}^{t_2} s^{k-1}|P_k(s)|\, ds.$$

From (44) and (45), we can let $t_2 \to \infty$ here and conclude that

$$\int^{\infty} s^{k-2} Q_k(s)\, ds < \infty.$$

Therefore, (43) holds, since

$$\left|\int_t^{\infty} P_k(\lambda)\, d\lambda\right| \leq Q_k(t).$$

To see that (43) is weaker than (44), notice that the function

$$P_k(t) = t^{-k+1/2} \sin t$$

satisfies (43), but not (44).

Example 1. Hartman [1, p. 315] has shown that if $P_1, \cdots, P_n \in C(0, \infty)$, and

$$\int^{\infty} t^{k-1+\alpha} |P_k(t)|\, dt < \infty, \qquad 1 \leq k \leq n, \tag{46}$$

for some $\alpha > 0$, then the homogeneous equation

$$x^{(n)} + P_1(t) x^{(n-1)} + \cdots + P_n(t) x = 0 \tag{47}$$

has a fundamental system $x_0, x_1, \cdots, x_{n-1}$ such that

$$x_\nu^{(r)}(t) = \begin{cases} t^{\nu-r}[1 + o(t^{-\alpha})]/(\nu - r)!, & 0 \leq r \leq \nu, \\ o(t^{\nu-r-\alpha}), & \nu + 1 \leq r \leq n-1. \end{cases} \tag{48}$$

The author [2] showed that this conclusion remains valid with (46) replaced by the assumption that

$$\int^{\infty} t^{\alpha} |P_1(t)|\, dt < \infty$$

and the integrals

$$\int^{\infty} t^{k-1+\alpha} P_k(t)\, dt, \qquad 2 \leq k \leq n,$$

converge, perhaps conditionally. The same conclusion can be obtained under the still weaker assumptions that

$$\int^{\infty} |P_1(t)|\, dt < \infty$$

and

$$\int_t^{\infty} P_k(s)\, ds = o(t^{-k+1-\alpha}), \qquad 1 \leq k \leq n. \tag{49}$$

To see this, let ν be any integer in $\{0,1,\cdots,n-1\}$ and let $p(t)=t^{\nu}/\nu!$. Then the function g in (6) becomes

$$g(t)=\sum_{k=n-\nu}^{n} P_k(t)\frac{t^{\nu-n+k}}{(\nu-n+k)!},$$

and (49) implies that

$$\int_t^{\infty} s^{n-m-1}g(s)\,ds=o(\phi(t)),$$

with $m=\max\{0,\nu-[\alpha]\}$ and $\phi(t)=t^{\nu-m-\alpha}$. Since (49) implies (43), Corollary 2 implies that (47) has a solution x_{ν} which satisfies (48).

Corollary 2 also implies that if (49) holds only with "O" (rather than "o") on the right, then the stated conclusion also holds with "O" rather than "o" on the right of (48).

Example 2. Consider the equation

$$y^{(n)}+\left[t^{-n-\nu+1}\phi(t)\sin t\right]y=t^{-n+1}\phi(t)\cos t, \tag{50}$$

where ν is an integer in $\{0,1,\cdots,n-1\}$ and ϕ is positive and continuously differentiable on $(0,\infty)$, $\phi'\le 0$, and $\lim_{t\to\infty}\phi(t)=0$. Here $P_1=\cdots=P_{n-1}=0$ and

$$\int_t^{\infty} P_n(s)\,ds=O(t^{-n-\nu+1}\phi(t))=o(t^{-n+1}), \tag{51}$$

which implies (8) and (9), and the function g defined by (6) is

$$g(t)=t^{-n+1}\phi(t)[t^{-\nu}p(t)\sin t-\cos t],$$

so

$$\int_t^{\infty} s^{n-1}g(s)\,ds=O(\phi(t)) \tag{52}$$

(and the convergence is conditional if $\int^{\infty}\phi(t)\,dt=\infty$), provided p is a polynomial of degree $\le\nu$. This implies (7) with $m=0$. Therefore, Theorem 1 implies that if p is any polynomial of degree $\le\nu$, then (50) has a solution x_0 such that

$$x_0^{(r)}(t)=p^{(r)}(t)+O(\phi(t)t^{-r}),\qquad 0\le r\le n-1,$$

provided

$$\int_t^{\infty} s^{-\nu-1}\phi^2(s)\,ds=o(\phi(t)), \tag{53}$$

since this implies (10), because of (51) and (52). However, (53) obviously holds for any nonincreasing function ϕ if $\nu>0$. If $\nu=0$ it holds, for example, if

$$\phi(t)=(1+\log t)^{-\alpha},$$

with $\alpha>1$.

Example 3. Consider the equation

$$y^{(n)}+\left[t^{\alpha}\sin(e^{t})\right]y=0, \tag{54}$$

where α is an arbitrary real number. By substituting $s=\log u$ it is easy to verify that

$$\int_t^\infty s^\alpha \sin(e^s)\,ds=O(t^\alpha e^{-t}),$$

where the convergence is conditional if $\alpha \geq -1$. Therefore, (54) satisfies (8), (9), and (43). For this equation the function g in (6) is

$$g(t)=t^\alpha p(t)\sin(e^t),$$

so if p is a polynomial of degree $\nu \leq n-1$, then

$$\int_t^\infty s^{n-1}g(s)\,ds=O(t^{n+\alpha+\nu-1}e^{-t}),$$

which implies (7) with $m=0$ and

$$\phi(t)=t^{n+\alpha+\nu-1}e^{-t}.$$

Therefore, Corollary 2 implies that (54) has a solution x_0 such that

$$x_0^{(r)}(t)=p^{(r)}(t)+O(t^{n+\alpha+\nu-r-1}e^{-t}), \qquad 0\leq r\leq n-1.$$

Example 4. Corollary 1 implies that the equation

$$y^{(n)}+[t^{-n+1/2}\sin t]y=\frac{t^{-1/2}\sin t}{(n-1)!}+t^{-1/2}\log t\cos t$$

has a solution x_0 such that

$$x_0^{(r)}(t)=[1+O(t^{-1/2}\log t)]t^{n-r-1}/(n-r-1)!, \qquad 0\leq r\leq n-1.$$

To see this, observe that here $P_1=\cdots=P_{n-1}=0$ and

$$\int_t^\infty P_n(s)\,ds=O(t^{-n+1/2}),$$

so (8) and (9) hold. With $p(t)=t^{n-1}/(n-1)!$, the function g in (6) is

$$g(t)=-t^{-1/2}\log t\cos t,$$

so

$$\int_t^\infty g(s)\,ds=O(t^{-1/2}\log t),$$

(with conditional convergence), which verifies (7) with $m=n-1$ and

$$\phi(t)=t^{-1/2}\log t.$$

Since this ϕ satisfies (42), Corollary 1 implies the conclusion.

4. Appendix. Proof of Lemma 1.

Proof. From Dirichlet's theorem, the convergence of the integral in (24) implies that the improper integral in (26) or (27) converges. Therefore, v is well-defined on $[t_0,\infty)$ by (26) or (27), and

$$(55) \qquad v^{(r)}(t)=\int_t^\infty \frac{(t-s)^{n-r-1}}{(n-r-1)!}w(s)\,ds, \qquad m\leq r\leq n-1.$$

With

$$Q(t)=\int_t^\infty s^{n-m-1}w(s)\,ds, \tag{56}$$

(55) can be rewritten as

$$v^{(r)}(t)=-\frac{1}{(n-r-1)!}\int_t^\infty\left(\frac{t}{s}-1\right)^{n-r-1}s^{m-r}Q'(s)\,ds,\qquad m\le r\le n-1. \tag{57}$$

If $m\le r\le n-2$, integrating (57) by parts yields

$$v^{(r)}(t)=\frac{1}{(n-r-1)!}\int_t^\infty Q(s)\frac{d}{ds}\left[\left(\frac{t}{s}-1\right)^{n-r-1}s^{m-r}\right]ds. \tag{58}$$

But

$$\left|\frac{d}{ds}\left[\left(\frac{t}{s}-1\right)^{n-r-1}s^{m-r}\right]\right|\le t^{m-r}\frac{d}{ds}\left(1-\frac{t}{s}\right)^{n-r-1}+(r-m)s^{m-r-1} \tag{59}$$

if $s\ge t$ and $r\ge m$. Since

$$|Q(s)|\le\rho(t)\phi(t)\quad \text{if}\, s\ge t \tag{60}$$

(see (25) and (56)), (58) implies (30) for $m+1\le r\le n-2$. If $r=n-1$, integrating (57) by parts yields

$$v^{(n-1)}(t)=t^{m-n+1}Q(t)+(m-n+1)\int_t^\infty s^{m-n}Q(s)\,ds. \tag{61}$$

If $m<n-1$, this and (60) imply (30) with $r=n-1$. Setting $r=m$ in (58) and (59) and invoking (60) yields (29) if $m<n-1$. If $m=n-1$, then (60) and (61) imply (29). If $0\le r\le m-1$, then we can differentiate (27) and substitute (55) with $r=m$ into the result to obtain

$$v^{(r)}(t)=\int_{t_0}^t\frac{(t-\lambda)^{m-r-1}}{(m-r-1)!}v^{(m)}(\lambda)\,d\lambda,\qquad 0\le r\le m-1.$$

Therefore, from (29),

$$|v^{(r)}(t)|\le\frac{1}{(n-m-1)!(m-r-1)!}\int_{t_0}^t(t-\lambda)^{m-r-1}\rho(\lambda)\phi(\lambda)\,d\lambda,\qquad 0\le r\le m-1. \tag{62}$$

Since ρ is nonincreasing and $t^\gamma\phi(t)$ is nondecreasing, this implies that

$$|v^{(r)}(t)|\le\frac{\rho(t_0)\phi(t)t^\gamma}{(n-m-1)!(m-r-1)!}\int_{t_0}^t(t-\lambda)^{m-r-1}\lambda^{-\gamma}\,d\lambda,\qquad 0\le r\le m-1.$$

Replacing t_0 by zero and integrating repeatedly by parts now yields (28). (Here we need the assumption that $\gamma<1$.)

From (29) and (30), (31) implies (32) for $m\le r\le n-1$. If $0\le r\le m-1$, then (62) and the monotonicity properties of ρ and ϕ imply that

$$|v^{(r)}(t)|\le\frac{t^{m-r-1+\gamma}\phi(t)}{(n-m-1)!(m-r-1)!}\int_{t_0}^t\rho(\lambda)\lambda^{-\gamma}\,d\lambda,\qquad 0\le r\le m-1. \tag{63}$$

But

$$\int_{t_0}^{t}\rho(\lambda)\lambda^{-\gamma}d\lambda=\int_{t_0}^{t_1}\rho(\lambda)\lambda^{-\gamma}d\lambda+\int_{t_1}^{t}\rho(\lambda)\lambda^{-\gamma}d\lambda$$
$$\leq\int_{t_0}^{t_1}\rho(\lambda)\lambda^{-\gamma}d\lambda+\rho(t_1)\frac{t^{1-\gamma}-t_1^{1-\gamma}}{1-\gamma}$$

if $t_1>t_0$. This and (63) imply that

$$\overline{\lim_{t\to\infty}}\, t^{-m+r}(\phi(t))^{-1}|v^{(r)}(t)|\leq\frac{\rho(t_1)}{(m-r-1)!(n-m-1)!(1-\gamma)}.$$

Since this holds for all $t_1\geq t_0$, (31) implies (32) for $0\leq r\leq m-1$. This completes the proof of Lemma 1.

REFERENCES

[1] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[2] W. F. TRENCH, *Asymptotic integration of linear differential equations subject to integral smallness conditions involving ordinary convergence*, this Journal, 7 (1976), pp. 213–221.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics
007

AN EXISTENCE THEOREM FOR A BOUNDARY VALUE PROBLEM RELATED TO THAT OF FALKNER AND SKAN*

BRUNO GABUTTI†

Abstract. We consider $z'''+zz''+z'^2-1=0$ together with the initial conditions $z(0)=z'(0)=0$, $z''(0)=k$ for any given $k\in\mathbb{R}$. We establish the existence of a value $\hat{k}<0$ such that $z'(\infty)=-1$. It is shown that if $k<\hat{k}$ then z' becomes unbounded, while if $k>\hat{k}$ then $\lim_{t\to\infty}z'(t)=1$ and $z(t)=t+k+O(t^{-1})$ as t tends to infinity.

1. Introduction and main results. In boundary layer theory the equation

$$f'''+\alpha ff''+\beta(1-f'^2)+\gamma f''=0 \tag{1}$$

is called the "equation of similar profiles" [11, p. 245]. The appropriate boundary conditions for this equation are [11, p. 246]

$$f(0)=f'(0)=0, \tag{2}$$

$$f'(\infty)=1. \tag{3}$$

This boundary value problem, which arises in Falkner–Skan approximate treatment of the laminar boundary layer in fluid dynamics, has been studied by several authors in the last forty years. Theorems of existence and uniqueness [3], [4], [5], [6], [13], [15] and asymptotic behavior of the solutions [8] were established twenty years ago. A resumè of the main results is given in [7, Chapt. 14, Part III]. For the physical significance of (1)–(3) we refer to [12].

The above-mentioned results consider solutions of the boundary value problem (1)–(3) together with the additional restriction

$$0<f'<1. \tag{4}$$

This assumption, which was originally suggested by physical considerations, has been recently discussed.

Rigorous results about the solutions of (1)–(3), which do not take into account (4) have been established in recent years by Hastings [10] and Troy [14].

From a purely mathematical point of view the results that we are now going to prove can be connected to the similar ones of Hastings (see [10, Thm. 2]).

THEOREM A. *Suppose f is a solution of* (1) *with* $\alpha=1$, $\gamma=0$ *such that* $f(0)=f'(0)=0$. *Then f satisfies* (3) *if either of the following conditions is satisfied*:

(i) $\beta<0, f''(0)=0$;

(ii) $-1\le\beta<0, f''(0)>k_*(\beta)$, *where*

$$k_*(\beta)=\inf\{k|\textit{if } f''(0)=k, \textit{ then for some } T_k>0, f'(T_k)=0\}.$$

Hastings also gives the conditions for which such a solution satisfies $f'(t)>1$ for some value of t.

Actually we can show that, at least in the case $\alpha=1$, $\beta=-1$, $\gamma=0$, the value $k_*(-1)$ can be characterized as the solution of another boundary value problem related

*Received by the editors November 30, 1982, and in revised from April 1, 1983.

†Instituto di Calcoli Numerici, Universita' degli Studi di Torino, Italy.

to (1) and stated by (5)–(7) of Theorem 1. Indeed, from Theorem A and from Theorem 2 below it follows immediately that $k_*(-1)=y''(0)$, where $y(t)$ denotes the unique solution of (5)–(7).

THEOREM 1. *The problem*

$$y'''+yy''+y'^2-1=0, \tag{5}$$

$$y(0)=y'(0)=0, \tag{6}$$

$$y'(\infty)=-1 \tag{7}$$

has a unique solution; that is, there exists one and only one function $y(t)$ satisfying the differential equation (5) *on* $(0,\infty)$ *and the boundary data* (6), (7). *This solution also satisfies*

$$-\infty<y(t)<0, \qquad t\in(0,\infty), \tag{8}$$

$$-1<y'(t)<0, \qquad t\in(0,\infty), \tag{9}$$

$$y''(t)<0, \qquad t\in[0,\infty). \tag{10}$$

Furthermore

$$-\frac{2}{\sqrt{3}}<y''(0)<-1. \tag{11}$$

The boundary value problem (5)–(7) seems to be of some interest in boundary layer theory. In [11, p. 251] it is stated that methods similar to those of Iglisch [4], [5] show the existence and uniqueness of solutions of the equation of similar profiles, with $\alpha=-1$, $\beta>0$, γ arbitrary, subject to the boundary conditions (2), (3) and the additional condition (4).

Now, by setting $y(t)=-f(t)$ the boundary value problem (5)–(7) reduces to that associated with the equation of similar profiles (1) with $\alpha=-1$, $\beta=1$, $\gamma=0$ and boundary conditions (2), (3). As a matter of fact the proof of uniqueness does not require the hypothesis $y'(t)>-1$ (see Lemma 3 below) and arguments similar to those used in the proof [10, Thm. 2], show that the assumption $y'(t)<0$ can also be dispensed with.

In the present paper the investigation of the problem (5)–(7) is mainly pursued because of its mathematical interest in view of the study of (12) below. This is also an example of an equation of similar profiles which arises by specifying $\alpha=1$, $\beta=-1$, $\gamma=0$ in (1). Indeed we consider the following initial value problem:

$$z'''+zz''+z'^2-1=0, \tag{12}$$

$$z(0)=z'(0)=0, \qquad z''(0)=k, \tag{13}$$

where k is an arbitrary real number.

The unique solution $y(t)$ of (5)–(7) plays an important role in the study of the behavior of the solutions of (12)–(13).

In fact we shall see that this solution separates the unbounded derivative solutions of (12)–(13) from solutions which satisfy

$$\lim_{t\to\infty} z'(t)=1. \tag{14}$$

More precisely we can prove:

THEOREM 2. *Let $\hat{k}=y''(0)<0$ where $y(t)$ is the unique solution of* (5)–(7). *Then*:

(i) *When $k<\hat{k}$, the unique solution of* (12)–(13) *is defined only in a finite interval* $[0,t_E)$, $t_E>0$; *moreover z, z', z'' are unbounded for $t\to t_E$.*

(ii) *When $k\in(\hat{k},0)$, the unique solution of* (12)–(13) *has bounded derivatives $z'(t)$ on* $[0,\infty)$. *The slope z' first decreases and has a relative minimum, after which it increases and has a relative maximum for some value of t where $z'(t)>1$.* (*Hence z' has one zero in* $(0,t)$.) *After this z' decreases monotonically and satisfies* (14). *See Fig.* 1 *where the solution z_1' corresponds to $k=-0.5$.*

(iii) *When $k\geq 0$, the unique solution of* (12)–(13) *has bounded derivative on* $[0,\infty)$. *The slope z' is increasing up to a maximum point $z'(t^*)>1$; then it decreases monotonically and satisfies* (14). *In Fig.* 1 *we report z_2' for $k=0.5$.*

Furthermore, in cases (ii) *and* (iii) *we have*

$$z(t)=t+k-\left(1+\frac{k^2}{2}\right)t^{-1}+o(t^{-1}). \tag{15}$$

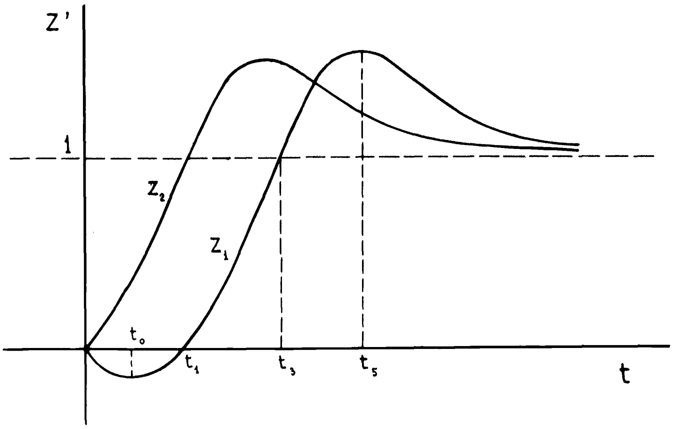


FIG. 1. *The points t_0, t_1, t_3, t_5 refer to the proof of Theorem* 2.

Remark 1. The boundary value problem (12)–(14) is a special case of (1)–(3).

This suggests the existence of possible results, analogous to those of Theorems 1 and 2, for the problem (1)–(3) with $\alpha=1$ and some $\beta<0$; however we do not consider this more general situation. Notice also that in the case $\alpha=1$, $\beta=-1$, $\gamma=0$, (1) can be integrated twice to yield a first order Riccati type equation (see (21) below). Some of the arguments of this paper make use of the integrated form, so that it is not clear that the detailed behavior in case (ii) above can be generalized to other values of β. In the affirmative case it should be interesting to know if the results established by Iglisch and Kemnitz in [6] can be obtained without the additional condition (4). In fact (12) is one of the class of equations considered in [6], where the existence and uniqueness of solutions of (1) with $\alpha=1$, $\beta<0$ subject to (2)–(4) is rigorously treated. Note also that the proof of Theorem 2 does not depend on condition (4).

From the physical point of view the solutions which appear most interesting seem to be those established by part (ii) of Theorem 2. In this case we have $z''(0)<0$, $z'(\infty)=1$ and the solutions of (12)–(13) are called solutions with "reversed flow" (see [10]).

All solutions of (12), (13) shown in parts (ii) and (iii) of Theorem 2 satisfy (14) and exhibit so-called "overshoot"; i.e., the derivative of the solution is greater than 1 for some $t>0$. The physical significance of these solutions is controversial. For a discussion about this see [14]. Here we remark that from a mathematical point of view our result is in agreement with that established by Stewartson in [13]. He proves that if f is any solution of (1) with $\alpha=1$, $\gamma=0$, satisfying (2), (3), and if $\beta<\beta_0=-0.1988$, then there exist values of $t>0$ for which f' shows "overshoot".

Remark 2. Parts (ii) and (iii) of Theorem 1 show the existence of infinite solutions of (1)–(3), with $\alpha=1$, $\beta=-1$, $\gamma=0$; each solution being characterized by $f''(0)>\hat{k}$, $\hat{k}<0$. Therefore, at least for $\beta=-1$, the boundary conditions (2)–(3) are insufficient to specify an unique solution of (1) with $\alpha=1$ and $\gamma=0$. This was evidenced by numerical experiments discussed in [9].

It is also interesting to compare the result stated in Theorem 2 with that of Troy [14]. By using previous results of Hastings [10], Troy first claims that there are values of $\tilde{k}<0$, $\tilde{\beta}<0$, such that if $f''(0)=\tilde{k}$ and $\beta=\tilde{\beta}$ then there are solutions of (1)–(3) with $\alpha=1$, $\gamma=0$, such that

$$(16)\qquad \begin{aligned} &-1<f'<1 \quad \text{on } (0,\infty),\\ &1-f'\sim d_0 t^{-1-2\beta}\exp\left(-\frac{t^2}{2}-d_1 t\right), \qquad t\to\infty \end{aligned}$$

for some $d_1\in\mathbb{R}$ and $d_0>0$.

Troy then states the following theorem.

THEOREM B. *There is a decreasing sequence $\{\beta_j\}, j\in\mathbb{N}$ of negative numbers such that for each $j\in\mathbb{N}$, if $\beta=\beta_j$ then the solution of* (1)–(2) *with $\alpha=1$, $\gamma=0$ and $f'(0)=\tilde{k}$ satisfies the boundary condition* (16) *as $t\to\infty$, where d_0, d_1 are replaced by some δ_j, ρ_j, $j\in\mathbb{N}$. Furthermore for each $j\in\mathbb{N}$ there are exactly j distinct positive values of t for which $f'-1=0$.*

We recall also (see [7, p. 534]) that if $\beta<0$ and if some solutions of (1) with $\alpha=1$, $\gamma=0$ exist, then these solutions satisfy either (16) or

$$(17)\qquad 1-f'\sim d_2 t^{2\beta}$$

as $t\to\infty$, where $d_2>0$ is a constant.

Therefore, if we denote by f_1 the solution of (1)–(3) with $\alpha=1$, $\beta=\beta_1$, $\gamma=0$, from Theorem B and Theorem 2 it follows that f_1' has the same qualitative behavior as all solutions z' of (12)–(13), with $k\in(\hat{k},0)$, and that f_1 and z have different behavior as $t\to\infty$. As a consequence of these observations we conjecture the existence of solutions of (1)–(3), with $\alpha=1$, $\beta<-1$, $\gamma=0$, such that the curve $g=f'(t)$ intersects the line $g=1$ more than once and such that $f'(t)$ tends to 1 algebraically.

From the point of view of applications, the solutions found by Troy are more important since the approach, as $t\to\infty$, of their derivatives to 1 is exponential. However our results can be used to complete the mathematical description of the model of Falkner and Skan.

2. Proofs. Before proving any theorems we consider a brief remark about the regularity of solutions of (1), (2). We observe that the usual class of classical solutions of (1), (2) is $C^3(0,t_*)\cap C^2[0,t_*)$, $t_*>0$, but in some proofs we explicitly use solutions with higher order derivatives. Without essential restrictions we can assume that any solution of (1), (2) can be extended to the class $C^\infty[0,t_*)$ for some $t_*>0$. This can be readily seen by rewriting (1) in system form (as an example see (23) for the case $\beta=-1$) and by using standard results of the theory of ordinary differential equations. See, for instance, [2, Thm. 8.1].

We prove Theorems 1 and 2 by using several lemmas.

LEMMA 1. *The boundary value problem* (5)–(7) *has at least one solution which satisfies* (8)–(10).

Proof. The proof is similar to that of Hartman [7, Thm. 6.1, p. 521]. As a preliminary to proving existence, let us observe that if a solution $y(t)$ of (4), (6) exists then it must satisfy the inequalities

$$-1\le y'(t)<0, \qquad t\in[0,\infty), \tag{18}$$

$$y''(t)\le 0, \qquad t\in[0,\infty). \tag{19}$$

These are weaker forms of (9), (10), which will be proved later.

It is straightforward to verify that (18) holds. Indeed, any relative extremum of y' on $(0,\infty)$, say at t_0, is a relative maximum if $[y'(t_0)]^2>1$ and a relative minimum if $[y'(t_0)]^2<1$. This follows directly from (5). Hence if $y'(t)<-1$ for some t then using (7) we find that the absolute minimum of y' on $[0,\infty]$ is simultaneously a relative maximum of y', a contradiction. This establishes the left-hand side of (18).

Similarly if $y'(t)$ is positive at some point but $y'(t)>1$ for all t then we obtain the contradiction that the (positive) absolute maximum of y' on $[0,\infty]$ is, by (5), simultaneously a relative minimum. But if $y'(t_0)=0$ for some $t_0>0$ then $y'(t)$ is strictly positive for some t. If $y''(t_0)\neq 0$ this follows from the mean value theorem, while if $y''(t_0)=0$ it follows from a two term Taylor expansion of y about t_0 and the fact that in this case $y'''(t_0)=1$ by (5). We may assume, without loss of generality, that $y''(t_0)>0$. Hence (in order to avoid the above contradiction stemming from $y'(t)>1$ for all t) the curve $Y=y'(t)$ must intersect the line $Y=1$. We now show that the intersection cannot occur at two different points. In order to verify this assertion we integrate (5) twice taking into account the initial condition (6)

$$y''=t-yy'+k, \tag{20}$$

$$y'=\frac{t^2}{2}-\frac{y^2}{2}+kt, \tag{21}$$

where $k=y''(0)$. Suppose, for the purpose of obtaining a contradiction, that there exists a $t_1>t_0$ such that $y'(t_1)=1$. Without loss of generality, $y''(t_1)\le 0$. From (20) it follows that $k+t_0\ge 0$ and $k-y(t_1)+t_1\le 0$; since $t_1>t_0>0$, we have $y(t_1)\ge k+t_1\ge 0$. By evaluating (21) at $t=t_1$ and using the last inequality, we get $1\le -k^2/2$, which is a contradiction. This completes the proof of (18). Inequality (19) follows immediately. Equation (5) and initial conditions (6) imply $y'''(0)=1$; by assuming $k\ge 0$ we have $y'(t)\ge 0$ on an interval $0<t<t_1$ for some $t_1>0$, but this violates (18).

A basic lemma which is required for our existence proof is the following (see [7, p. 520]; for the definition of egress, ingress and strict egress point see [7, p. 37] or [1]).

LEMMA 2. *Let u,f be d-dimensional vectors and $f(t,u)$ be continuous on an open (t,u) set Γ such that solutions of initial value problems associated with*

$$u'=f(t,u) \tag{22}$$

are unique. Let Γ_0 be an open subset of Γ with the properties that all egress points from Γ_0 are strict egress points and that the set Γ_e of egress points is not connected. Let Γ_i denote the set of ingress points of Γ_0 and let S be a connected subset of $\Gamma_0\cup\Gamma_e\cup\Gamma_i$ such that

$S\cap(\Gamma_0\cup\Gamma_i)$ contains two points (t_1,u_1) and (t_2,u_2), for which solutions $u_j(t)$ passing through (t_j,u_j) for $j=1,2$ leave Γ_0 with increasing t at points of different (connected) components of Γ_e. Then there exists at least one point (t_0,u_0) in $S\cap(\Gamma_0\cup\Gamma_i)$ such that the solution $u_0(t)$ of (22) *determined by $u_0(t_0)=u_0$ remains in Γ_0 on its (open) right maximal interval of existence.*

In order to apply Lemma 2 we rewrite (5) in system form. Let $u_1=y$, $u_2=y'$, $u_3=y''$; then $u_3'=y'''=1-u_2^2-u_1u_3$. The system equivalent to (5) is

$$u_1'=u_2,\quad u_2'=u_3,\quad u_3'=1-u_1u_3-u_2^2. \tag{23}$$

We choose $\Gamma=\{(t,u_1,u_2,u_3)|t,u_1,u_2,u_3\in\mathbb{R}\}$. Inequalities (18), (19) suggest that

$$\Gamma_0=\{(t,u_1,u_2,u_3)|t,u_1\in\mathbb{R},-1<u_2<0,u_3<0\}.$$

To determine the ingress and egress points, it is convenient to define the following boundary sets associated with Γ_0:

$$\begin{aligned}\Gamma_1&=\{(t,u_1,u_2,u_3)|t,u_1\in\mathbb{R},u_2=0,u_3<0\},\\ \Gamma_2&=\{(t,u_1,u_2,u_3)|t,u_1\in\mathbb{R},-1<u_2<0,u_3=0\},\\ \Gamma_3&=\{(t,u_1,u_2,u_3)|t,u_1\in\mathbb{R},u_2=-1,u_3<0\},\\ \Gamma_4&=\{(t,u_1,u_2,u_3)|t,u_1\in\mathbb{R},u_2=-1,u_3=0\},\\ \Gamma_5&=\{(t,u_1,u_2,u_3)|t,u_1\in\mathbb{R},u_2=0,u_3=0\}.\end{aligned}$$

The points in $\Gamma_i=\Gamma_1$ are ingress points, since in Γ_1 $u_2'=u_3<0$.

The set of egress points is $\Gamma_e=\Gamma_2\cup\Gamma_3$. If (t,u_1,u_2,u_3) is in Γ_3, then $u_2'=u_3<0$ and, hence, is a strict egress point. For (t,u_1,u_2,u_3) in Γ_2 we have $-1<u_2<0$ and $u_3=0$; from (23) it follows that $u_3'>0$, so Γ_2 consists of strict egress points.

The set Γ_4 is composed of solutions of $u_1=x+c$ ($c=$constant); therefore, points in Γ_4 are neither egress nor ingress points.

For points (t,u_1,u_2,u_3) in Γ_5, $u_3'=1$, hence the solution (u_1,u_2,u_3) through $(t_0,u_0,0,0)$, t_0, $u_0\in\mathbb{R}$ is not in Γ_0 since $u_2'=u_3=0$ implies that $u_2>0$ for small $|t-t_0|$.

The character of Γ_4 shows that Γ_e is not connected.

Let k be a fixed number satisfying $-\infty<k<0$; define the set $S=\{(t,u_1,u_2,u_3)|t=0,u_1=0,u_2=0,u_3=k\}$; S is a connected subset of $\Gamma_0\cup\Gamma_e\cup\Gamma_i$.

The point $(0,0,0,k_1)$, where k_1 is negative and small, is a strict ingress point of Γ_0 so, by continuity of the initial data, the solution of (23) with $u_1(0)=u_2(0)=0$, $u_3(0)=k$ leaves Γ_0 through the component Γ_2.

On the other hand, it will be shown that if $k_2<0$ is large enough, the solution of (13) with initial data $u_1(0)=0$, $u_2(0)=0$, $u_3(0)=k_2$ leaves Γ_0 through Γ_3. To verify this, we integrate the third equation of (23)

$$u_3=t-u_1u_2+k_2.$$

Since $-1\le u_2\le 0$, so that $-t\le u_1\le 0$ and u_1, $u_2\ge 0$, we have

$$u_3\le t+k_2.$$

Hence, if k_2 is sufficiently large and the solution of (23) through $(0,0,0,k_2)$ lies in Γ_0 on an interval $[0,t^*)$ for some $t^*>0$, then $u_3(t)$ is less than a given negative constant on $[0,t^*]$ and such a solution leaves Γ_0 through a point in Γ_3.

By applying Lemma 2, we obtain the existence of a point $(0,0,0,\hat{k})$ in $S\cap(\Gamma_0\cup\Gamma_1)$ such that the solution $(\hat{u}_1,\hat{u}_2,\hat{u}_3)$ of (23) with $\hat{u}_1(0)=\hat{u}_2(0)=0$, $\hat{u}_3=\hat{k}$ remains in Γ_0 on its right maximal interval of existence. By the structure of Γ_0, this right maximal interval of existence is necessarily $[0,\infty)$.

The solution $(\hat{u}_1,\hat{u}_2,\hat{u}_3)$ is in Γ_0 for $t\in(0,\infty)$. This implies that the limit $\lim_{t\to\infty}\hat{u}_2(t)$ exists and equals -1.

Suppose, for the purpose of obtaining a contradiction, that $\lim_{t\to\infty}\hat{u}_2(t)=\hat{u}_2(\infty)>-1$. Because of the structure of Γ_2 and the third equation of (23), we have

$$\lim_{t\to\infty}\hat{u}_3'=-\lim_{t\to\infty}\hat{u}_1\hat{u}_3-\lim_{t\to\infty}\left(\hat{u}_2^2-1\right)=1-\hat{u}_2^2(\infty)>0.$$

Let T be sufficiently large so that $-\hat{u}_1(t)\hat{u}_3(t)-\hat{u}_2(t)+1\geq(1-\hat{u}_2^2(\infty))/2$ for $t>T$; thus $\hat{u}_3'>(1-\hat{u}_2)/2>0$, $t>T$. A quadrature gives

$$\hat{u}_2'(t)-\hat{u}_2'(T)>\frac{1}{2}\left(1-\hat{u}_2^2(\infty)\right)(t-T).$$

When t tends to infinity in the above inequality we obtain the contradiction $\lim_{t\to\infty}\hat{u}_2(t)=\infty$ which proves that $\hat{u}_2(\infty)=-1$.

This completes the proof of the existence of a solution of the boundary value problem (5)–(7) and proves (9)–(10). After integrating (9) inequality (7) follows immediately, completing the proof of Lemma 2.

LEMMA 3. *The boundary value problem* (5)–(7) *has an unique solution which satisfies* (8)–(10).

Proof. First we observe that any solution of (5)–(7) satisfies (8)–(10). Suppose now that there are two solutions $y_1(t)$, $y_2(t)$. If $y_1\neq y_2$ then we may assume, without loss of generality, that $y_1'(t)>y_2'(t)$ on $(0,t_0)$ (for, if $y_1'(t)\equiv y_2'(t)$ on $[0,t_*)$ for some $t_*>0$, we merely introduce a new independent variable $\eta=t-t_*$) and $y_1'(t_0)=y_2'(t_0)$ where $0<t_0\leq\infty$. Then (6) implies that $y_1(t)>y_2(t)$.

The function $r(t)$, defined by $r(t)=y_1(t)-y_2(t)$, has the property that r' is positive on $(0,t_0)$ vanishes at 0 and t_0 and therefore has a relative maximum at some point t_1 in $(0,t_0)$. At $t=t_1$ we have $r'(t_1)>0$, $r''(t_1)=0$ and $r'''(t_1)\leq0$. However (5) leads to

$$r'''(t_1)=-r(t_1)y_2''(t_1)-r'(t_1)\left[y_1'(t_1)+y_2'(t_1)\right].$$

By (9)–(10) and the fact that $r(t)>0$ on $(0,t_0)$, the right-hand side is positive whereas the left-hand side is nonpositive. This contradiction implies that $y_1(t)\equiv y_2(t)$. This completes the proof of Lemma 3.

LEMMA 4. *The value $\hat{k}=y''(0)$, which characterizes the unique solution $y(t)$ of* (5)–(7), *satisfies*

$$-\frac{2}{\sqrt{3}}\leq\hat{k}\leq-1. \tag{24}$$

Proof. The proof of the lower bound given by (24) is based on the following property of the solution $y(t)$

$$\lim_{t\to\infty}y''(t)=0. \tag{25}$$

In the proof of Lemma 2 we showed that the solution $y(t)$, $t\in[0,\infty)$, cannot get out of Γ_0 through the boundary Γ_2. We now prove that the solution $y(t)$ leaves Γ_0 through the component Γ_4. In order to see this we assert that $y(t)$ cannot leave Γ_0 through Γ_3. Indeed, if $\lim_{t\to\infty}y''(t)>0$ then y' would be unbounded and thus would violate (9). This proves (25).

We now return to the proof of the lower bound of $\hat{k}$ by observing that inequalities (8), (10) imply $yy''\geq 0$ for $t\in(0,\infty)$; hence

$$y'''\leq 1-(y')^2. \tag{26}$$

Using (10) we can introduce $v=y'\leq 0$ as a new independent variable and then $w=v'$ as a new dependent variable. Then (26) becomes

$$w\frac{dw}{dv}\leq 1-v^2.$$

By integrating this we find that

$$w^2(0)-w^2(v)\leq\frac{2}{3}v^3-2v. \tag{27}$$

Next we observe that $w(v(t))=y''(t)$; hence from (25) we have $\lim_{t\to\infty}w(v(t))=0$. By using this and $\lim_{t\to\infty}v(t)=\lim_{t\to\infty}y'(t)=-1$ in (27) the left-hand side of (24) follows.

The upper bound can be obtained by using the further property of the solution $y(t)$ that

$$y'''(t)>0, \qquad t\in[0,\infty). \tag{28}$$

To prove (28) suppose, for the purpose of obtaining a contradiction, that $y'''(t_0)=0$ for some $t_0>0$. Differentiating (5) twice we obtain

$$y^{(\mathrm{IV})}+yy'''+3y'y''=0,$$
$$y^{(\mathrm{V})}+4y'y'''+3(y'')^2+yy^{(\mathrm{IV})}=0.$$

From (9), (10) it follows that $y^{(\mathrm{IV})}(t_0)=-3y'(t_0)y''(t_0)<0$; therefore $y'''(t_0)$ remains negative in a right neighborhood of t_0. Also $y'''(t)$ is a nonincreasing function of t for $t\geq 0$. Indeed, a possible $t_1>t_0$, where $y^{(\mathrm{IV})}(t_1)=0$ because $y^{(\mathrm{V})}(t_1)=-4y'(t_1)y'''(t_1)-3[y''(t_1)]^2<0$, would be a relative maximum for y'''. Thus $\lim_{t\to\infty}y'''(t)=y'''(\infty)<0$ which implies $y''(t)$ unbounded; this contradicts (25) and proves (28).

Now, the function $\frac{1}{2}y'^2-yy''$ has as its derivative $-yy'''$ and consequently is an increasing function of t. Hence $y'^2\geq 2yy''$ for $t\geq 0$. It follows that $w(v)=v'$ satisfies the differential inequality

$$w\frac{dw}{dv}\geq 1-\frac{3}{2}v^2.$$

An integration gives

$$w^2(0)-w^2(v)\geq v^3-2v,$$

which for $t\to\infty$ yields an upper bound for $\hat{k}$ and finishes the lemma.

Lemma 4 completes the proof of Theorem 1.

Remark 3. A numerical approach to the computation of $\hat{k}$ shows that $\hat{k}=-1.08637\ldots$.

Proof of Theorem 2 (i). We first show that

$$z''(t)<y''(t), \quad z'(t)<y'(t), \quad z(t)<y(t) \tag{29}$$

for t in the maximal interval of existence of z. From the assumptions of Theorem 2 we know that the function z satisfies inequality (29) on an interval $[0,t_0)$ for some $t_0>0$.

By subtracting (5) from (12), we get

$$z'''-y'''+zz''-yy''+z'^2-y'^2=0.$$

An integration leads to

$$z''-y''=k-\hat{k}+z(y'-z')+y'(y-z).$$

Suppose, $z''(t_0)=y''(t_0)$. From (8), (9) it follows that at $t=t_0$ the right-hand side is negative while the left-hand side is zero. This contradiction shows that (29) holds on the maximal interval of existence of z.

It is now asserted that every solution z of (12)–(13) exists only in a finite interval of t for any given $k<\hat{k}$. We suppose that the maximal interval of existence of z is $[0,\infty)$. If $z'(\infty)=-1$ then we have violated the uniqueness of the solution $y(t)$ (Theorem 1). Therefore from (29) it follows that $z'(\infty)<-1$.

Inequalities (8)–(10), (29) and equation (12) imply that $\lim_{t\to\infty}z'''(t)<0$ and so it follows that z', z'' are unbounded for $t\to\infty$. Also because of (12), (29) z''' tends to infinity for $t\to\infty$. The unboundness of z''' implies that there exists a point t_1 such that $z'''(t)<0$ for $t>t_1$. Thus

$$z(t_1)<0,\quad z'(t_1)<0,\quad z''(t_1)<0$$

and

$$z'''(t)<0 \quad \text{for } t\geq t_1. \tag{30}$$

A differentiation of (12) gives

$$z^{(\mathrm{IV})}=-zz'''-3z'z''. \tag{31}$$

From (29)–(30) it follows that, for $t\geq t_1$, the right-hand side of (31) is a nonincreasing function of t. In fact, we have, for instance, that the derivative with respect to z of the right-hand side of (31) is $-z'''>0$, etc.

Moreover, (31) has the elementary solution

$$z=\frac{2}{t-t_E},$$

where t_E is an arbitrary constant. We can choose $t_E>t_1$ so large that

$$\frac{2}{t_1-t_E}\geq z(t_1),\quad \frac{2}{t_1-t_E}\geq z'(t_1),\quad \frac{2}{t_1-t_E}\geq z''(t_1).$$

Then, from known results on differential inequalities (see for instance [7, Chapt. III, Exercise 4.1]), follows

$$\frac{2}{t-t_E}\geq z(t) \quad \text{for all } t\geq t_1. \tag{32}$$

Consequently the function $z(t)$ cannot be defined for all positive values of t, since the left-hand side of (32) tends to $-\infty$ when $t\to t_E-0$.

Finally if z' and z'' were bounded for $t\to t_E$ then z would also be bounded, whereas we have seen that this is impossible. This completes the proof of Theorem 2(i).

Proof of Theorem 2(ii). We start by showing that, under the present hypotheses, the maximal interval of existence of $z(t)$ is $[0,\infty)$. This follows from the inequality

$$y(t)<z(t)\leq\frac{t^3}{6},\qquad t\in(0,\infty),\tag{33}$$

where $y(t)$ is the unique solution of (5)–(7).

The right-hand inequality of (33) is easily proved by integrating (12) twice and using (13); we have

$$z''=k-zz'+t,\tag{34}$$

$$z'=kt+\frac{1}{2}t^2-\frac{1}{2}z^2,\tag{35}$$

whence $z'\leq t^2/2$.

By a further integration we get the upper bound of (33). To prove the lower bound we put $r(t)=z(t)-y(t)$.

Suppose, for the purpose of obtaining a contradiction, that there exists a point $t^*>0$ such that $y(t^*)=z(t^*)$, where t^* is the smallest value of t which enjoys this property. Then $k>\hat{k}$ and (13) implies that $r(0)=r(t^*)=0$ and $r(t)>0$ for $t\in(0,t^*)$. By Rolle's mean value theorem, there is a point t_M, $0<t_M<t^*$, where $r'(t_M)=0$ and, necessarily, $r''(t_M)\leq 0$. From the differential equations we obtain

$$r''=k-\hat{k}-ry'-zr'.\tag{36}$$

Thus

$$r''(t_M)=k-\hat{k}-y'(t_M)r(t_M).\tag{37}$$

Since $k>\hat{k}$ and by (9) it follows that the right-hand side of (37) is positive while the left-hand side is nonpositive. This contradiction shows that t^* does not exist and completes the proof of (33).

We now assert that z'' necessarily has a zero. The condition $z''(t)<0$ for all $t>0$ is incompatible with the uniqueness requirement of y. From (13), if $z''(t)<0$ for all $t\in(0,\infty)$ we would have

$$z'(t)<0,\quad z(t)<0\quad\text{for } t\in(0,\infty).\tag{38}$$

By subtracting (21) with $k=\hat{k}$ from (35), we get $z'-y'=(k-\hat{k})t+(y^2-z^2)/2$. By (33), (38), it follows that $z'(t)>y'(t)$ for $t\in(0,\infty)$. By using this and (38), (33), (8), (9) in (36), we obtain $y''(t)<z''(t)<0$ for $t\in[0,\infty)$. The last inequalities and (25) yield $\lim_{t\to\infty}z''(t)=0$; thus $\lim_{t\to\infty}z'''(t)=0$. Therefore (12) is satisfied only if $\lim_{t\to\infty}z'(t)=-1$, but this violates the uniqueness of y. This contradiction shows that z'' must have a zero; suppose that $z''(t_0)=0$. Without loss of generality, we may assume that t_0 is the first zero.

The function z'' satisfies the conditions

$$\begin{aligned}&z''(t)<0,\qquad -1<y'(t)<z'(t)<0,\\&y(t)<z(t)<0\quad\text{for } t\in(0,t_0).\end{aligned}\tag{39}$$

Then (12) shows that $z'''(t_0)>0$, and z' has a relative minimum in t_0; its derivative has the property that $z''(t)>0$ for t in some right neighborhood of t_0.

It will be shown, now, that z'' is positive for all $t>t_0$ such that z' remains less than 1. Suppose that $z''(t_*)=0$ for some $t_*>t_0$, then it follows that $z'''(t_*)\leq 0$. This inequality contradicts the inequality $z'''(t_*)=1-z'^2(t^*)>0$ and shows that z'' is positive.

We now assert that z' has a zero. Suppose that $z'(t)<0$ for all $t>t_0$. By integrating (12) from t_0, the zero of z'', to $t>t_0$, we obtain

$$z''(t)=-z(t)z'(t)+z(t_0)z'(t_0)+t-t_0 \tag{40}$$

from which, by using (39) and (8) we get

$$z''(t)\geq[z'(t_0)+1]t-t_0,$$

where $0<z'(t_0)+1<1$. By letting t approach infinity, we get $\lim_{t\to\infty}z''(t)=\infty$ and consequently $z'>0$ for large t. This contradiction shows that there is a $t>t_0$, call it t_1, such that $z'(t_1)=0$.

By putting $t=t_1$ in (40) it follows that

$$z''(t_1)>0. \tag{41}$$

This implies $z'>0$ in a right neighborhood of t_1.

It is now asserted that $z(t)$ has a zero. Integrating (12) from t_1 to t, $t_1<t$, leads to the equation

$$z''(t)=z''(t_1)-z(t)z'(t)+t-t_1. \tag{42}$$

By using (41) it follows that $z''\geq c_1$ for $t>t_1$ and some constant $c_1>0$, as long as $z(t)$ is negative and $z'(t)$ is positive. The inequality $z''\geq c_1$, $t\geq t_1$, implies that $z'(t)$ remains positive and that $z(t)$ is eventually positive. This contradiction shows that $z(t)$ has a zero, say $t=t_2$. At $t=t_2$ we have $z(t_2)=0$, $z''(t_2)>0$ and $z'(t_2)>0$. The last inequality is derived from integrating (42) from t_1 to t_2.

We now show that there is a point, say $t_3>t_2$, such that $z'(t_3)=1$. In order to see this suppose, for the purpose of obtaining a contradiction, that $z'(t)<1$ for all $t\geq t_2$. By the argument which we used in the proof of Lemma 1 it follows that $0<z'(t)\leq 1$ for $t\in[t_2,\infty)$. An integration gives $0<z(t)\leq(t-t_2)$; thus $z^2(t)\leq(t-t_2)^2$. On the other hand if we integrate (12) twice in $[t_2,t]$, we get

$$z'(t)=z'(t_2)+z''(t_2)(t-t_2)-\frac{1}{2}z^2(t)+\frac{1}{2}(t-t_2)^2.$$

Therefore, we infer that

$$z'(t)\geq z'(t_2)+z''(t_2)(t-t_2).$$

Recalling that $z'(t_2)$, $z''(t_2)>0$, this implies $\lim_{t\to\infty}z'(t)=\infty$ which contradicts the assumption that $z'(t)<1$ for $t\in[t_2,\infty)$ and proves the existence of t_3.

Clearly $z''(t_3)\geq 0$. We can show that $z''(t_3)>0$. If this is not so, it follows that the solution $z(t)$ satisfies: $z''(t_3)=0$, $z'(t_3)=1$, $z(t_3)=z_3$, where z_3 is a constant. However, by uniqueness of the initial value problem, the function $q(t)=t-t_3+z_3$ is the only solution of (12) that passes through (t_3,z_3) with slope 1 and zero second derivative. This contradiction shows that $z''(t_3)>0$ and that $z'(t)>1$ in some right neighborhood of t_3.

An argument similar to that used in the proof of Lemma 1 shows that the solution $z'(t)$ definitely remains over the line $z'=1$ for $t>t_3$. Setting $t_4=\max(t_2,t_3)$ yields $z(t)>t-t_4$ for $t>t_4$. Thus

$$z(t)>t-t_4,\quad z'(t)>1,\quad z''(t)>0,\quad t>t_4. \tag{43}$$

Depending on whether $t_3>t_2$ or $t_3<t_2$, the last inequality follows from $z''(t_3)>0$ or (42), respectively.

It is now asserted that $z''(t)$ has another zero. It is convenient to consider the cases where z' is bounded and z' is unbounded separately.

Suppose first that $z''(t)>0$ for $t\geq t_4$ and that z' is bounded. By considering (12) and (43) we see that there exists a constant $c_2>0$ such that

$$z'''\leq -c_2, \qquad t>t_4. \tag{44}$$

The assumption that z' is bounded implies that $\lim_{t\to\infty} z'''(t)=0$. The use of this in (44) leads to a contradiction. In this instance we see that z'' has a zero.

Next we show that $z''(t)>0$ for $t>t_4$, and z' unbounded are contradictory statements. By using these and the first two inequalities of (43) in (12) it follows that

$$\lim_{t\to\infty} z'''(t)=\lim_{t\to\infty}(-zz''+1-z'^2)\leq \lim_{t\to\infty}(1-z'^2)=-\infty. \tag{45}$$

This contradicts $z''(t)>0$, for $t>t_4$ and shows that $z''(t)$ must have another zero, say t_5, such that $t_5>t_4$.

At $t=t_5$, $z'''(t_5)<0$; thus t_5 is a relative maximum for z'.

After t_5 the function z' cannot have a relative minimum. This is a consequence of the property that z' cannot have relative minimum over the line $z'=1$ (see proof of Lemma 1). Thus, for $t>t_5$, z' is monotonically decreasing and satisfies

$$1<z'(t)<z'(t_5) \quad \text{if } t\in(t_5,\infty). \tag{46}$$

We finally show that of necessity

$$\lim_{t\to\infty} z'(t)=1. \tag{47}$$

Indeed, from the fact that

$$\lim_{t\to\infty} z'(t)=z'(\infty)=a>1,$$

it follows that

$$z(t)=at+o(t) \tag{48}$$

so that, using (35)

$$z'(t)=kt+\frac{1}{2}t^2(1-a^2)+o(t^2).$$

Hence

$$\frac{z'(t)}{t^2}-\frac{k}{t}+\frac{o(t^2)}{t^2}=\frac{1}{2}(1-a^2)$$

and since the left-hand side tends to zero, we have $a=\pm 1$. This and (46) establishes (47) and completes the proof of the lemma.

Proof of Theorem 2(iii). If $k=0$, then the initial value problem (12)–(13) is a particular case of that considered by Hastings; see Theorem A(i).

If $k>0$ then we have $z(t)>0$, $0<z'(t)<1$, $z''(t)>0$ in some right neighborhood of $t=0$. Therefore, apart from an irrelevant change of independent variables, we recover the same conditions of the solutions of (12)–(13) with $k\in(\hat{k},0)$. There we found in a right neighborhood of the point t_2 the first zero of these solutions; see proof of Theorem 2(ii). Thus the same procedure considered above, for $t>t_2$, can be used. We omit the details.

Proof of (15). From (47)–(48) we have

$$z(t)=t+o(t),\qquad z'(t)=1+o(1) \tag{49}$$

as t tends to infinity. Therefore from (35) it follows that

$$z=\left[t^2+2kt-2+o(1)\right]^{1/2}\quad \text{as } t\to\infty.$$

Consequently by well-known rules we obtain

$$\lim_{t\to\infty}\left[z(t)-t\right]=\lim_{t\to\infty}\frac{2kt-2+o(1)}{t+\left[t^2+2kt-2+o(1)\right]^{1/2}}=k.$$

Thus

$$z(t)=t+k+o(1).$$

Hence, setting

$$z(t)=t+k+s(t),$$

it follows (using (49)) that

$$s(t)=o(1),\qquad s'(t)=o(1) \tag{50}$$

and (35) becomes

$$s^2+2s(t+k)+2\left(1+\frac{k^2}{2}\right)+s'=0.$$

By solving this equation for s and by using (50), we get

$$s=-\frac{k^2+2+o(1)}{t+k+\left[t^2+2kt-2+o(1)\right]^{1/2}}$$

and by familiar rules we find

$$\lim_{t\to\infty} ts=-\left(1+\frac{k^2}{2}\right),$$

whence

$$s=-\left(1+\frac{k^2}{2}\right)t^{-1}+o(t^{-1}).$$

Returning to the variable z we obtain (15). This completes the proof of Theorem 2.

Remark 4. By iterating the procedure used in the proof (15), the expansion (15) can be improved. For instance, it is true that

$$z(t)=t+k-\left(1+\frac{k^2}{2}\right)t^{-1}+k\left(1+\frac{k^2}{2}\right)t^{-2}+o(t^{-2})\quad \text{as } t\to\infty.$$

Remark 5. An expansion analogous to (15) can be obtained for the solution $y(t)$ of the problem (5)–(7). We have

$$y(t)=-t-k-\left(1-\frac{k^2}{2}\right)t^{-1}+o(t^{-1}),$$

as $t\to\infty$. This can be also improved as (15) was in Remark 4.

Acknowledgments. I am grateful to the Professor David Kahaner who suggested this problem to me. I am also grateful to the referee for the illuminating and detailed report which allowed me to improve remarkably on the first versions of this paper.

REFERENCES

[1] J. BEBERNES AND V. LAKSMIKANTHAM, *An Introduction to Nonlinear Boundary Value Problems*, Academic Press, New York, 1973.

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

[3] W. A. COPPEL, *On a differential equation of boundary-layer theory*, Phil. Trans. Royal Soc., 253 (1960), pp. 101–136.

[4] R. IGLISCH, *Elementarer Existenzbeweis für die Strömung in der laminaren Grenzschicht zur Potential strömung $U=u_1x^m$ mit $m>0$ bei Absaugen und Ausblasen*, Z. Angew. Math. Mech., 33 (1953), pp. 143–147.

[5] ——, *Elementarer Beweis für dies Eindeutigkeit der Strömung in der laminaren Grenzschicht zur Potential strömung $U=u_1x^m$ mit $m\geq 0$ bei Absaugen und Ausblasen*, Z. Angew. Math. Mech., 34 (1954), pp. 441–443.

[6] R. IGLISCH AND F. KEMNITZ, *Über die in der Grenzschichttheorie auftretende Differentialglechung $f'''+ff''+\beta(1-f'^2)=0$ für $\beta<0$ bei gewissen Absauge und Ausblasegesetzen*, 50 Jahre Grenzschichtforschung, H. Görtler and W. Tollmien, eds., Vieweg, Braunschweig, 1955.

[7] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[8] ——, *On the asymptotic behavior of solutions of a differential equation in boundary layer theory*, Z. Angew. Math. Mech., 44 (1964), pp. 123–128.

[9] D. R. HARTREE, *On a equation occuring in Falkner and Skan's approximate treatment of the equations of the boundary layer*, Proc. Cambridge Philos. Soc., 33 (1937), pp. 223–239.

[10] S. P. HASTINGS, *Reversed flow solutions of the Falkner-Skan equation*, SIAM J. Appl. Math., 22 (1972), pp. 330-334.

[11] L. ROSENHEAD, ed., *Laminar Boundary Layers*, Clarendon Press, Oxford, 1963.

[12] H. SCHLICHTING, *Boundary Layer Theory*, Pergamon Press, London, 1955.

[13] K. STEWARTSON, *Further solutions of the Falkner-Skan equation*, Proc. Cambridge Philos. Soc., 50 (1954), pp. 454–465.

[14] W. C. TROY, *Non-Monotonic solutions of the Falkner-Skan boundary layer equation*, Quart. Appl. Math., 37 (1979), pp. 157–167.

[15] H. WEYL, *On the differential equations of the simplest boundary layer problem*, Ann. Math., 43 (1942), pp. 381–407.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics
008

REGULARIZING TRANSFORMATIONS FOR CERTAIN SINGULAR STURM-LIOUVILLE BOUNDARY VALUE PROBLEMS*

HANS G. KAPER[†], MAN KAM KWONG[†‡] AND ANTON ZETTL[†‡]

Abstract. It is shown that certain singular Sturm-Liouville boundary value problems can be transformed into regular problems by a simple transformation of the dependent variable.

1. Introduction. We consider the Sturm-Liouville differential expression τ,

$$\tau=-\frac{d}{dt}p(t)\frac{d}{dt}+q(t), \tag{1}$$

on (a,b), where $-\infty<a<b<\infty$. We assume throughout this article that the coefficients p and q are real-valued and measurable on (a,b), and that they satisfy the minimal conditions

$$p^{-1},q\in L^1_{\text{loc}}(a,b), \tag{2}$$

where $p^{-1}(t)=(p(t))^{-1}$ a.e. on (a,b). Furthermore, we assume that p is positive on (a,b),

$$p(t)>0 \quad \text{a.e. on } (a,b). \tag{3}$$

A function $\bar{y}$ is said to be a solution of the equation $\tau y=0$ if (i) $\bar{y}$ is absolutely continuous on (a,b), (ii) $p\bar{y}'$ is equal a.e. to an absolutely continuous function (which, with a slight abuse of notation, we denote by the symbol $p\bar{y}'$), and (iii) the identity $-(p\bar{y}')'(t)+q(t)\bar{y}(t)=0$ holds a.e. on (a,b).

From the theory of differential equations it is known that, given any point $t_0\in(a,b)$ and any pair of real numbers (c_0,c_1), there exists a unique solution $\bar{y}$ of $\tau y=0$, such that $\bar{y}(t_0)=c_0$ and $(p\bar{y}')(t_0)=c_1$. Hence, the set of solutions of the equation $\tau y=0$ forms a two-dimensional linear vector space. The Wronskian of two solutions $\bar{y}_1$ and $\bar{y}_2$ is defined by the expression $(p\bar{y}_1')(t)\bar{y}_2(t)-(p\bar{y}_2')(t)\bar{y}_1(t)$ and a variation-of-parameters formula holds. For details, see for example Naimark [1, §5.2].

The right endpoint b is said to be a regular endpoint for the differential expression τ if p^{-1} and q are integrable in a left neighborhood of b,

$$p^{-1},q\in L^1(c,b) \quad \text{for some } c\in(a,b). \tag{4}$$

Similarly, the left endpoint a is regular if

$$p^{-1},q\in L^1(a,c) \quad \text{for some } c\in(a,b). \tag{5}$$

If both a and b are regular endpoints, the differential expression τ is said to be regular; otherwise, it is called singular. Note that, for τ to be regular, p^{-1} and q need not be bounded on (a,b).

*Received by the editors October 19, 1982, and in revised form March 16, 1983. This work was supported by the Applied Mathematical Sciences Research Program (KC-04-02) of the Office of Energy Research of the U. S. Department of Energy under contract W-31-109-Eng-83.

[†] Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439.

[‡] Permanent address: Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115.

All solutions $\bar{y}$ of a regular Sturm–Liouville equation $\tau y=0$ are continuous on $[a,b]$, and the same property holds for the function $p\bar{y}'$. Hence, boundary value problems can be posed for such equations. More generally, the study of eigenvalue problems for operators associated with regular Sturm–Liouville differential expressions is meaningful. A central question in such a study, which is important in many applications, concerns the characterization of those boundary conditions that give rise to selfadjoint realizations of the differential expression. Other interesting and important questions concern the oscillatory properties of the eigenfunctions. For example, when $p(t)>0$ a.e. on (a,b), then the classical oscillation theory, including the Sturm comparison theorem, remains valid when p and q satisfy only the minimal conditions (2); see Everitt, Kwong and Zettl [2].

The study of equations involving singular Sturm–Liouville expressions is considerably more difficult, as their solutions exhibit singularities near the endpoints. Weyl [3] has developed a theory for the construction of selfadjoint realizations of singular differential expressions. The theory is based on a distinction between singularities of limit-circle type and those of limit-point type. The characterizations are, however, not concrete and therefore difficult to apply. Also, the oscillatory properties of the eigenfunctions are much harder to establish in the singular case.

In this article we show that an important class of singular problems involving singularities of both limit-circle and limit-point type can be reduced to regular problems by a transformation of the dependent variable. Thus, many of their properties can be deduced simply from the theory for regular problems. We give various illustrative examples.

2. The case of one singular endpoint. We first consider Sturm–Liouville differential expressions τ which are regular at one endpoint only. Suppose that the coefficients p and q satisfy the regularity condition (4), in addition to the minimal conditions (2), and the positivity condition (3). Suppose, furthermore, that p is such that

$$\int_a^b p^{-1}(t)\,dt=\infty. \tag{6}$$

Then b is a regular endpoint for τ, but a is not, so τ is singular.

Let the function ϕ be defined by the expression

$$\phi(t)=1+\int_t^b p^{-1}(s)\,ds, \qquad t\in(a,b), \tag{7}$$

and let the functions P and Q in turn be defined by the expressions

$$P(t)=p(t)\phi^2(t), \qquad Q(t)=q(t)\phi^2(t), \qquad t\in(a,b). \tag{8}$$

The functions P and Q define a Sturm–Liouville differential expression T on (a,b),

$$T=-\frac{d}{dt}P(t)\frac{d}{dt}+Q(t). \tag{9}$$

THEOREM 1. *Let τ be the differential expression* (1), *and let T be the expression derived from it according to the transformation* (7), (8). *Assume that the coefficients p and q of τ are such that* (4) *and* (6) *hold, but that the coefficients P and Q of T satisfy the conditions*

$$P^{-1},Q\in L^1(a,b). \tag{10}$$

Then τ is singular, but T is regular on (a,b). The function $\bar{y}$ is a solution of the singular equation $\tau y=0$ if and only if the function $\bar{Y}$ defined by

$$\bar{Y}(t)=\frac{\bar{y}(t)}{\phi(t)}, \qquad t\in(a,b), \tag{11}$$

is a solution of the regular equation $TY=0$. Moreover, if $\lim_{t\downarrow a}\bar{y}(t)$ exists and is finite, then $\lim_{t\downarrow a}\bar{Y}(t)=0$, and vice versa.

Proof. The regularity of T is clear from (10). The first part of the theorem is verified by a direct computation. As $\lim_{t\downarrow a}\phi(t)=\infty$, $\bar{Y}(t)$ must vanish as $t\downarrow a$ whenever $\bar{y}(t)$ tends to a finite limit. Conversely, if $\bar{Y}(t)$ tends to zero as $t\downarrow a$, then we have, for any $t\in(a,b)$,

$$\bar{y}(t)=\phi(t)\bar{Y}(t)=\phi(t)\int_a^t\bar{Y}'(s)\,ds=\phi(t)\int_a^tP\bar{Y}'(s)(\phi^{-1})'(s)\,ds.$$

The function $P\bar{Y}'$ is continuous on $[a,b]$. Hence,

$$\bar{y}(t)-P\bar{Y}'(a)=\phi(t)\int_a^t(P\bar{Y}'(s)-P\bar{Y}'(a))(\phi^{-1})'(s)\,ds.$$

Because both ϕ and $(\phi^{-1})'$ are positive on (a,b), the expression in the right member is estimated by

$$\phi(t)\int_a^t|P\bar{Y}'(s)-P\bar{Y}'(a)|(\phi^{-1})'(s)\,ds.$$

Given ε, we can choose t sufficiently small that $|P\bar{Y}'(s)-P\bar{Y}'(a)|<\varepsilon$ for all $s\in[a,t]$, so

$$|\bar{y}(t)-P\bar{Y}'(a)|<\varepsilon\phi(t)\int_a^t(\phi^{-1})'(s)\,ds=\varepsilon.$$

Thus, $\lim_{t\downarrow a}\bar{y}(t)$ exists and is finite. □

The condition (10) of the theorem is met, for example, if the coefficient q of τ is bounded near a, and if the coefficient p is such that

$$\int_a^b\left(\int_t^bp^{-1}(s)\,ds\right)^2dt<\infty, \tag{12}$$

i.e., if the left endpoint a is a limit-circle type singularity of τ.

A similar construction can be set up for the transformation of singular differential expressions τ which are regular at the left endpoint a, but singular at the right endpoint b. With the definitions

$$\phi(t)=1+\int_a^tp^{-1}(s)\,ds, \qquad t\in(a,b), \tag{7'}$$

$$P(t)=p(t)\phi^2(t), \qquad Q(t)=q(t)\phi^2(t), \qquad t\in(a,b), \tag{8'}$$

one obtains a differential expression T, as before. If the coefficients p and q of τ are such that (5) and (6) hold and, in addition, (10) is satisfied, then τ is singular, but T is regular on (a,b). This is the case, for example, if q is bounded near b and if p is such that

$$\int_a^b\left(\int_a^tp^{-1}(s)\,ds\right)^2dt<\infty, \tag{13}$$

i.e., if the right endpoint b is a limit-circle type singularity of τ. An analogue of Theorem 1 applies.

The existence of regularizing transformations can be exploited in the spectral analysis of singular differential operators and in the investigation of the oscillatory properties of eigenfunctions of such operators. (Note that the function ϕ is strictly positive on (a,b), so $\bar{y}$ and $\bar{Y}$ have the same oscillatory behavior.)

Example 1. The differential equation

$$-\frac{d}{dt}t^{\alpha}\frac{dy}{dt}+(\sin t)y(t)=0, \qquad t\in(0,1),$$

is regular when $\alpha<1$, but singular when $\alpha\geq 1$. The singularity at 0 is of limit-circle type if $1\leq\alpha<\frac{3}{2}$, and of limit-point type if $\alpha\geq\frac{3}{2}$. The conditions of Theorem 1 are satisfied for all $\alpha\in[1,2)$. If $\alpha=1$, the transformed equation is

$$-\frac{d}{dt}t(1-\ln t)^{2}\frac{dy}{dt}+(1-\ln t)^{2}(\sin t)y(t)=0.$$

If $\alpha\in(1,2)$ it is convenient to multiply the original equation by $(\alpha-1)^{-1}$ before applying the transformation. One thus obtains the transformed equation

$$-\frac{d}{dt}t^{2-\alpha}\frac{dy}{dt}+t^{-2(\alpha-1)}(\sin t)y(t)=0.$$

The transformed equations are indeed regular on (0, 1).

Example 2. The above procedure can be applied to the study of the eigenvalue problem for the Latzko equation:

$$-\frac{d}{dt}(1-t^{7})\frac{dy}{dt}=\lambda t^{7}y(t) \quad \text{on } (0,1),$$
$$y(0)=0, \qquad \lim_{t\uparrow 1}y(t) \text{ exists and is finite.}$$

Here $p(t)=1-t^{7}$, $q(t)=-\lambda t^{7}$. The differential expression is regular at 0, but not at 1. The singularity at 1 is of limit-circle type; the boundary condition at the right endpoint defines a selfadjoint realization of the differential expression. (The proof of the last statement is analogous to the proof of the corresponding result for the Legendre differential equation; see, for example, Akhiezer and Glazmann [4, App. 2].)

Defining ϕ in accordance with (7′),

$$\phi(t)=1+\int_{0}^{t}(1-s^{7})^{-1}ds, \qquad t\in(0,1),$$

and using ϕ to define a new dependent variable z,

$$z(t)=y(t)/\phi(t), \qquad t\in(a,b),$$

we find that z solves the eigenvalue problem

$$-\frac{d}{dt}(1-t^{7})\phi^{2}(t)\frac{dz}{dt}=\lambda t^{7}\phi^{2}(t)z(t) \quad \text{on } (0,1),$$
$$z(0)=0, \qquad z(1)=0.$$

The latter eigenvalue problem is regular on (0, 1).

3. The case of two singular endpoints. Transforming expressions with two singular endpoints into regular expressions is somewhat more involved. Our procedure is based on a reduction to the previous case.

Suppose that the coefficients p and q satisfy, in addition to the minimal conditions (2) and the positivity condition (3), the conditions

$$\int_a^c p^{-1}(t)\,dt=\infty,\qquad \int_c^b p^{-1}(t)\,dt=\infty,\quad \text{for some } c\in(a,b). \tag{14}$$

Then neither a nor b is a regular endpoint for the differential expression τ, so τ is singular.

Let $\varepsilon>0$ be chosen such that $c+\varepsilon<b$, and let the function ψ be defined by the expressions

$$\psi(t)=1+\int_t^c p^{-1}(s)\,ds,\qquad t\in(a,c], \tag{15-1}$$

$$\psi(t)=1+\int_c^t p^{-1}(s)\left(\frac{s-c}{\varepsilon}-1\right)ds,\qquad t\in(c,c+\varepsilon], \tag{15-2}$$

$$\psi(t)=\psi(c+\varepsilon),\qquad t\in(c+\varepsilon,b]. \tag{15-3}$$

(This definition is such that $p\psi'$ is a continuous piecewise linear function on (a,b).) Clearly, ψ and $p\psi'$ are absolutely continuous on compact subintervals of (a,b). Also, by decreasing ε if necessary, we can achieve the inequality $\psi(t)>0$ a.e. on (a,b). To see this, we observe that ψ satisfies the integral equation

$$\psi(t)=1+\int_c^t p^{-1}(s)(p\psi')(s)\,ds$$

on $(c,c+\varepsilon]$. Because $|p\psi'(t)|\leq 1$ for all t, we certainly have $|\int_c^t p^{-1}(s)(p\psi')(s)\,ds|\leq \int_c^t p^{-1}(s)\,ds$, so in particular for $t\in(c,c+\varepsilon]$,

$$\psi(t)\geq 1-\int_c^{c+\varepsilon} p^{-1}(s)\,ds.$$

By taking ε sufficiently small, we can certainly make the integral less than $\frac{1}{2}$, say.

Let the functions $\hat{p}$ and $\hat{q}$ be defined by the expressions

$$\hat{p}(t)=p(t)\psi^2(t),\qquad t\in(a,b), \tag{16-1}$$

$$\hat{q}(t)=q(t)\psi^2(t)-\psi(t)(p\psi')'(t),\qquad t\in(a,b). \tag{16-2}$$

We use these functions $\hat{p}$ and $\hat{q}$ to define a new differential expression $\hat{\tau}$ on (a,b),

$$\hat{\tau}=-\frac{d}{dt}\hat{p}(t)\frac{d}{dt}+\hat{q}(t). \tag{17}$$

The functions $\hat{p}$ and $\hat{q}$ satisfy the conditions (2) and (3). Assume that, in addition,

$$\hat{p}^{-1},\hat{q}\in L^1(a,c)\quad \text{for some } c\in(a,b). \tag{18}$$

Then a is a regular endpoint of $\hat{\tau}$; b is still a singular endpoint.

The differential expressions $\hat{\tau}$ fits into the framework of the previous section. We define a function ϕ,

$$\phi(t)=1+\int_a^t \hat{p}^{-1}(s)\,ds,\qquad t\in(a,b), \tag{19}$$

the functions P and Q,

$$P(t)=\hat{p}(t)\phi^2(t), \qquad Q(t)=\hat{q}(t)\phi^2(t), \qquad t\in(a,b), \tag{20}$$

and the differential expression T,

$$T=-\frac{d}{dt}P(t)\frac{d}{dt}+Q(t). \tag{21}$$

THEOREM 2. *Let τ be the differential expression* (1), *and let T be the differential expression derived from it according to the transformations* (15), (16) *and* (19), (20). *Assume that the coefficients p and q of τ are such that* (14) *holds, but that coefficients P and Q of T satisfy the conditions*

$$P^{-1},Q\in L^1(a,b). \tag{22}$$

Then τ is singular, but T is regular on (a,b). If $\bar{y}$ is a solution of the singular equation $\tau y=0$, which satisfies the boundary conditions

$$\lim_{t\downarrow a}y(t) \textit{ and } \lim_{t\uparrow b}y(t) \textit{ exist and are finite},$$

then $\bar{Y}$, defined by

$$\bar{Y}(t)=\bar{y}(t)/\phi(t)\psi(t), \qquad t\in(a,b), \tag{23}$$

is a solution of the regular equation $TY=0$, which satisfies the boundary conditions

$$\lim_{t\downarrow a}Y(t)=0, \qquad \lim_{t\uparrow b}Y(t)=0,$$

and vice versa.

Proof. The theorem is a direct consequence of the construction of the differential expression T, and Theorem 1 and its analogue. □

The condition (22) of the theorem is met, for example, if the coefficient q of τ is bounded on (a,b) and if the coefficient p is such that

$$\int_a^c\left(\int_t^c p^{-1}(s)\,ds\right)^2dt<\infty, \qquad \int_c^b\left(\int_c^t p^{-1}(s)\,ds\right)^2dt<\infty, \tag{24}$$

for some $c\in(a,b)$, i.e., if both endpoints are limit-circle type singularities of τ.

Theorem 2 is applicable, for example, to the Legendre differential expression on $(-1,1)$.

Example 3. The Legendre differential expression defines a singular eigenvalue problem on $(-1,1)$,

$$-\frac{d}{dt}(1-t^2)\frac{dy}{dt}=\lambda w(t)y(t) \quad \text{on } (-1,1), \tag{25-1}$$

where $w\in L^\infty(-1,1)$ and $w(t)>0$ a.e. on $(-1,1)$. Both endpoints are singular. The equation, supplemented by the boundary conditions

$$\lim_{t\downarrow -1}y(t) \text{ and } \lim_{t\uparrow 1}y(t) \text{ exist and are finite}, \tag{25-2}$$

is equivalent to a regular boundary value problem on $(-1,1)$ with Dirichlet boundary conditions. Hence, the boundary value problem (25) admits an infinite number of eigenvalues, which can be arranged in ascending order, $\lambda_0<\lambda_1<\lambda_2<\cdots$, with $\lambda_n\to\infty$, and to each eigenvalue λ_n corresponds a unique (up to a multiplicative constant) eigenfunction y_n which has exactly n zeros in $(-1,1)$. If $w(t)=1$ for all $t\in(-1,1)$, the eigenfunctions are multiples of the Legendre polynomials.

Acknowledgment. The authors wish to thank the referee, whose comments led to a generalization of the original version of Theorem 1.

REFERENCES

[1] M. A. NAIMARK, *Linear Differential Operators*, *Part* 2, Ungar, New York, 1968.

[2] W. N. EVERITT, MAN KAM KWONG AND A. ZETTL, *Oscillation of eigenfunctions of weighted regular Sturm–Liouville problems*, J. London Math. Soc., 27 (1983), pp. 106–120.

[3] H. WEYL, *Über gewöhnliche Differentialgleichungen mit Singularitäten und die zugehörigen Entwicklungen willkürlicher Funktionen*, Math. Ann., 68 (1910), pp. 220–269.

[4] N. I. AKHIEZER AND I. M. GLAZMANN, *Theory of Linear Operators in Hilbert Space*, Vol. 2, Pitman, London, 1981.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics
009

THE ANALYTIC CAUCHY PROBLEM FOR FOURTH ORDER ELLIPTIC EQUATIONS IN TWO INDEPENDENT VARIABLES*

ROBERT F. MILLAR†

Abstract. An explicit representation in terms of the Riemann function is derived for the solution to the analytic Cauchy problem for a class of fourth order elliptic equations in two independent variables. Two particular cases are considered. For the biharmonic equation, results of an elementary form are obtained and compatibility conditions on the Cauchy data are found that guarantee regularity of the solution throughout a given domain. Representations in terms of axial data are found for solutions to the generalized axially symmetric biharmonic equation and to the iterated generalized axially symmetric Helmholtz equation. In principle, the derivation can be extended to higher order elliptic equations in two independent variables.

1. Introduction. In a recent paper [1], a representation was obtained for the continuation of the solution to a boundary value problem for an analytic elliptic equation of second order in two independent variables. This result also provides the solution to the Cauchy problem on an analytic arc. The intent of the present paper is to extend these considerations to elliptic equations of higher order in the plane. In doing so, it turns out to be more natural to start with the Cauchy problem, rather than with a boundary value problem.

We shall consider an equation of the form

$$L^n[u]=0, \tag{1.1}$$

in which $L^n[u]:=L(L^{n-1}[u])$, $n=2,3,4,\cdots$, $L^1:=L$, and

$$L[u]:=\Delta u+au_x+bu_y+cu; \tag{1.2}$$

here Δ denotes the two-dimensional Laplacian operator and a, b, and c are analytic functions of the real variables x and y. Almost all of our attention will be directed to the case $n=2$, although in principle larger values of n and even more general equations of order $2n$ could be treated in the same manner.

For $n=2$, we give a representation, in terms of the Riemann function, for the solution to an analytic Cauchy problem for (1.1) on an analytic arc. In particular cases, explicit results may be found. We obtain a simple representation for the solution to the Cauchy problem for the biharmonic equation

$$\Delta^2 u=0. \tag{1.3}$$

This is used to discuss analytic properties of the solution in a domain intersected by an arc bearing the data. We also derive a Poisson-like representation, in terms of analytic axial values, for a solution to the generalized axially symmetric biharmonic equation in which $n=2$ and

$$L[u]=\Delta u+\left(\frac{2\alpha}{y}\right)u_y, \qquad \alpha>0. \tag{1.4}$$

From this we obtain the corresponding representation for solutions to the iterated generalized axially symmetric Helmholtz equation $(L+k^2)^2[u]=0$.

*Received by the editors December 7, 1982. This work was supported by the Natural Sciences and Engineering Research Council Canada under grant A 8808.

†Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1.

In previous work, Colton [2] has considered the analytic Cauchy problem for certain fourth order elliptic equations in the plane. In [2], the highest order terms also appear as $\Delta^2 u$, but lower order terms differ from those in $L^2[u]$. The complex Riemann function plays a major role in both Colton's and the present work. Here, however, the representation is explicit whereas in [2] the problem requires the solution of a system of Volterra integral equations. Work of a related nature, though more from the point of view of reflection across an analytic boundary, has been performed by Lewy [3] for second order equations and by Garabedian [4] for equations of second and fourth orders. For elliptic equations in which the highest-order terms appear as $\Delta^n u$, Yu [5], [6] has studied the reflection of solutions across a segment of the real axis. Hill [7] has considered the analytic Cauchy problem for systems of first order equations. For some additional references to earlier work, see [2], [6].

In the following section, from the Green's identity and a fundamental solution, we derive the representation (2.12) for the solution to the Cauchy problem and we observe how this also gives the continuation across an analytic boundary of the solution to a boundary value problem. The main result is summarized in the Representation Theorem. Explicit results for the biharmonic equation are obtained in §3, and for the generalized axially symmetric biharmonic equation in §4. Some concluding remarks are made in the final section.

2. Derivation of representation. Let C: $\xi=\xi(s)$, $\eta=\eta(s)$ (where s ranges over some real interval and $\xi'(s)^2+\eta'(s)^2\neq 0$) denote an analytic arc in the ξ, η-plane, and consider an analytic Cauchy problem on C for $L^n[u]=0$. Here $L[u]$ is given by (1.2) and the parameter s does not necessarily denote arc length. Suppose that N is some closed neighborhood of an interior point of C. According to the theorems of Cauchy-Kowalewski [8, pp. 39–56] and Holmgren [8, pp. 237–239], there exists a unique solution to $L^n[u]=0$ in N that assumes the given analytic Cauchy data on C, and is an analytic function of x and y for $(x,y)\in N$. We let $z:=x+iy$, $z^*:=x-iy$ be independent complex variables. An overbar will denote a conjugate domain, curve, or number; in particular $z^*\in\bar{N}$ if and only if $\bar{z}^*\in N$. Later, we shall require that the solution $u(\frac{1}{2}(z+z^*), -\frac{1}{2}i(z-z^*))$ be an analytic function of the two independent complex variables z, z^* for $(z,z^*)\dot{\in} N\times\bar{N}$. This will certainly be true if N is sufficiently small, provided that it is a fundamental domain of the differential equation in the sense of Vekua [9, p. 8]. By this is meant that $a(\frac{1}{2}(z+z^*), -\frac{1}{2}i(z-z^*))$ is analytic for $(z,z^*)\in N\times\bar{N}$; and similarly for b and c.

A normal vector to C is given by $\boldsymbol{\nu}:=(\eta'(s), -\xi'(s))$. Now the arc C separates N into two subsets D and D'. We shall denote by D that part of N for which $\boldsymbol{\nu}$ is an outward normal on C. The subarc of C that forms a portion of ∂D will be denoted by C', and C'' will refer to the remaining part of ∂D; thus $N=D\cup C'\cup D'$. It is assumed that ∂D is oriented in the counterclockwise sense and, without loss in generality, that C'' is smooth; then ∂D is smooth except, perhaps, at the intersections of C' and C''. We suppose that C' and C'' intersect at an angle in $(0,\pi]$, and we extend $\xi(s)$, $\eta(s)$ continuously so that ∂D is given by $\xi=\xi(s)$, $\eta=\eta(s)$ on some s-interval; $\xi(s)$ and $\eta(s)$ are piecewise smooth on ∂D, and are analytic on C'.

The function u is an analytic solution to (1.1) in $D\cup C'$. Thus, if v is any function that is C^{2n} in $D\cup C'$, we may apply Green's identity to $L^n[u]$ and v to give

$$\int_D \left(vL^n[u]-L^{n-1}[u]L^*[v]\right)d\xi\,d\eta=\int_{\partial D} M(v,L^{n-1}[u])\,ds.$$

Here L^* is the adjoint of L:

$$L^*[v]:=\Delta v-(av)_\xi-(bv)_\eta+cv.$$

For M we may take

$$M[v,w]:=v\partial w/\partial\nu-w\partial v/\partial\nu+vw(a\eta'(s)-b\xi'(s)),$$

in which $\partial w/\partial\nu:=\boldsymbol{\nu}\cdot\nabla w=\eta'(s)w_\xi-\xi'(s)w_\eta$. (It should be noted that this is the usual normal derivative only when $\boldsymbol{\nu}$ is a unit vector; that is, when s denotes arc length.)

Green's identities for $L^{n-m}[u]$ and $L^{*m}[v]$ $(m=1,2,\cdots,n)$ may be written down in turn. Addition of these results leads to

$$\int_D(vL^n[u]-uL^{*n}[v])\,d\xi\,d\eta=\sum_{m=1}^{n}\int_{\partial D}M(L^{*m-1}[v],L^{n-m}[u])\,ds.$$

(See Vekua [9, p. 182] for an analogous treatment for a different equation.)

For the case $n=2$, to which further attention is confined, we find that

$$(2.1)\quad \int_D(vL^2[u]-uL^{*2}[v])\,d\xi\,d\eta$$
$$=\int_{\partial D}\left\{v\frac{\partial}{\partial\nu}L[u]-L[u]\frac{\partial v}{\partial\nu}+vL[u](a\eta'(s)-b\xi'(s))\right.$$
$$\left.+L^*[v]\frac{\partial u}{\partial\nu}-u\frac{\partial}{\partial\nu}L^*[v]+uL^*[v](a\eta'(s)-b\xi'(s))\right\}ds.$$

It should be noted that the integral on the right in (2.1) is zero if $L^{*2}[v]=0$ throughout D.

We shall use (2.1) to obtain an expression for u at a point $(x,y)\in D$; from this a representation for the solution to the Cauchy problem on C' will be derived. In [1] we started with the solution to a boundary value problem in a domain D with an analytic boundary, and we found a representation for the continuation of the solution across the boundary. Here the same ideas are used, except that the solution in D arises from the Cauchy problem by virtue of the Cauchy–Kowalewski theorem.

It is known (see, for example, [10, Chapter III]) that L^{*2} has a fundamental solution with singularity at (x,y) that is of the form

$$(2.2)\qquad S(\xi,\eta;\,x,y)=-A(\xi,\eta;\,x,y)\log r+B(\xi,\eta;\,x,y).$$

Here $r:=[(\xi-x)^2+(\eta-y)^2]^{1/2}$ and A, the Riemann function, is normalized so that $L^{*2}[S]=\delta(r)$. This is the normalization used by John [10, p. 43], Vekua [9, p. 183], and by Weinacht [11]; it differs by a factor of 2π from the normalization adopted in [1]. The functions A and B are real analytic in their four arguments if r is sufficiently small.

On choosing $v=S$ in (2.1), we find that

(2.3)
$$u(x,y)=-\int\left\{S\frac{\partial}{\partial\nu}L[u]-L[u]\frac{\partial S}{\partial\nu}+L^*[S]\frac{\partial u}{\partial\nu}-u\frac{\partial}{\partial\nu}L^*[S]\right.$$
$$\left.+(SL[u]+uL^*[S])(a\eta'(s)-b\xi'(s))\right\}ds,\qquad (x,y)\in D.$$

Here, and subsequently unless the contrary is indicated, integration is over ∂D. The right-hand side of (2.3) is zero for $(x,y)\in D'$.

At this stage, it is useful to introduce the following notation:

$$z := x+iy, \qquad z^* := x-iy, \tag{2.4}$$

$$\zeta := \xi+i\eta, \qquad \zeta^* := \xi-i\eta, \tag{2.5}$$
$$Z(s) := \xi(s)+i\eta(s), \qquad \bar{Z}(s) := \xi(s)-i\eta(s).$$

Then, under the transformations (2.4) and (2.5),

$$u(x,y) \rightarrow U(z,z^*),$$

and

$$A(\xi,\eta;\, x,y) \rightarrow R(\zeta,\zeta^*;\, z,z^*).$$

For certain similar equations, Vekua [9, Chapter V] has shown that the solution and the Riemann function are analytic functions of their arguments, when ζ and z vary in a fundamental domain of the equation and ζ^*, z^* vary in the conjugate domain. Here we shall assume that the neighborhood N is sufficiently small that $U(z,z^*)$ is analytic for $(z,z^*)\in N\times\bar{N}$ and that $R(\zeta,\zeta^*;\, z,z^*)$ is analytic for $(\zeta,\zeta^*,\, z,z^*)\in N\times\bar{N}\times N\times\bar{N}$; this is certainly possible, since N is assumed to be a fundamental domain of the equation.

On employing (2.2), (2.3) becomes

$$\begin{aligned} U(z,z^*) &= \int F(s;\, z,z^*)\log r\, ds \\ &\quad + \int G(s;\, z,z^*)\frac{\partial}{\partial\nu}\log r\, ds + \int H(s;\, z,z^*)\, ds, \qquad (z,z^*)\in D\times\bar{D}, \end{aligned} \tag{2.6}$$

with

$$\begin{aligned} F(s;\, z,z^*) := {} & A\left\{\frac{\partial}{\partial\nu}L[u]+(a\eta'-b\xi')L[u]\right\}-L[u]\frac{\partial A}{\partial\nu} \\ & +L^*[A]\left\{\frac{\partial u}{\partial\nu}+(a\eta'-b\xi')u\right\}-u\frac{\partial}{\partial\nu}L^*[A], \end{aligned} \tag{2.7}$$

$$G(s;\, z,z^*) := -AL[u]-uL^*[A], \tag{2.8}$$

and

$$\begin{aligned} H(s;\, z,z^*) := {} & L[u]\frac{\partial B}{\partial\nu}-B\frac{\partial}{\partial\nu}L[u]-(a\eta'-b\xi')BL[u] \\ & +\left\{\frac{\partial u}{\partial\nu}+(a\eta'-b\xi')u\right\}\{(2A_\xi-aA)r_\xi/r+(2A_\eta-bA)r_\eta/r-L^*[B]\} \\ & -u\frac{\partial}{\partial\nu}\{(2A_\xi-aA)r_\xi/r+(2A_\eta-bA)r_\eta/r-L^*[B]\}. \end{aligned}$$

It is the intention to use (2.6) to continue U analytically across $C'\times\bar{C}'$. On doing so, we shall obtain a different representation for U which, moreover, is valid for $(z,z^*)\in N\times\bar{N}$.

After some algebraic manipulations, it is found that H may be written as

$$H = H_1+H_2+H_3,$$

in which

$$H_1(s;\,z,z^*):=\left\{u\frac{\partial K}{\partial\nu}-\left[\frac{\partial u}{\partial\nu}+(a\eta'-b\xi')u\right]K\right\}\frac{1}{Z(s)-z}+iuK\frac{Z'(s)}{[Z(s)-z]^2},$$

and

$$K:=\frac{1}{2}(a+ib)A-(A_\xi+iA_\eta).$$

The function H_2 is obtained by replacing K, $Z(s)$, $Z'(s)$ and z in H_1 by K^*, $\bar{Z}(s)$, $\bar{Z}'(s)$ and z^*, and changing the sign of the last term of H_1; here K^* is found from K by replacing i in the above expression by $-i$. The function H_3 comprises the terms in H that depend on B. Evidently H_1 may be singular for $Z(s)=z$, H_2 may be singular for $\bar{Z}(s)=z^*$, and H_3 is analytic.

Let us consider $\int H_1(s;\,z,z^*)\,ds$. The apparent pole of second order at $Z(s)=z$ may be reduced to one of first order by an integration by parts. The integrated terms vanish and we are led to a new integrand, H_1', say, with what seems to be a first order pole at $Z(s)=z$. A rather tedious examination shows that H_1' is not singular when $Z(s)=z$; this makes use of the fact that the Riemann function contains the factor r^2 (see, for example, [10, Chap. III]). Analogous conclusions may be drawn with respect to $\int H_2(s;\,z,z^*)\,ds$. Consequently, if z crosses C' and z^* crosses $\bar{C}'$, the continuation of the third integral in (2.6) is the integral itself.

On the other hand, the integrands of the first two integrals in (2.6) are singular for $z\in C'$ or $z^*\in\bar{C}'$. The continuation of these integrals may be effected in the manner described in [1]. We define

$$\Phi(t;\,z,z^*):=\int_0^t F(s;\,z,z^*)\,ds. \tag{2.9}$$

Then $\Phi(0;\,z,z^*)=0=\Phi(l;\,z,z^*)$, where ∂D is described as s runs from 0 to l; the second equality follows on setting $v=A$ in (2.1), since $L^{*2}[A]=0$ in D. After an integration by parts the first integral in (2.6) becomes

$$-\int_{\partial D}\Phi(s;\,z,z^*)\frac{\partial}{\partial s}\log r\,ds.$$

After simplification, it is found that

$$\begin{aligned} U(z,z^*)=&-\frac{1}{2}\int[\Phi(s;\,z,z^*)+iG(s;\,z,z^*)]Z'(s)/\Delta\,ds\\ &-\frac{1}{2}\int[\Phi(s;\,z,z^*)-iG(s;\,z,z^*)]\bar{Z}'(s)/\Delta^*\,ds\\ &+\int H(s;\,z,z^*)\,ds, \end{aligned} \tag{2.10}$$

wherein $\Delta:=Z(s)-z$, $\Delta^*:=\bar{Z}(s)-z^*$, and $\arg\Delta+\arg\Delta^*=0$ when $0\le s\le l$ and $z^*=\bar{z}$.

Denote the solutions to $Z(s)=\zeta$ and $\bar{Z}(s)=\zeta^*$ ($\zeta\in\partial D$, $\zeta^*\in\partial\bar{D}$, $0\le s\le l$), by $s=S(\zeta)$ and $s=\bar{S}(\zeta^*)$, respectively. Then $S(\zeta)$ is holomorphic and single-valued for ζ in a neighborhood of C', the analytic portion of ∂D; and similarly for $\bar{S}(\zeta^*)$ with respect to $\bar{C}'$. Consequently, (2.10) becomes

$$\begin{aligned}(2.11)\qquad U(z,z^*)=&-\frac{1}{2}\int_{\partial D}[\Phi(S(\zeta);z,z^*)+iG(S(\zeta);z,z^*)]\frac{d\zeta}{\zeta-z}\\&+\frac{1}{2}\int_{\partial\bar{D}}[\Phi(\bar{S}(\zeta^*);z,z^*)-iG(\bar{S}(\zeta^*);z,z^*)]\frac{d\zeta^*}{\zeta^*-z^*}\\&+\int H(s;z,z^*)\,ds,\qquad (z,z^*)\in D\times\bar{D},\end{aligned}$$

in which ∂D and $\partial\bar{D}$ are oriented in the counterclockwise sense.

If z crosses C' and z^* crosses $\bar{C}'$, then

$$\begin{aligned}U(z,z^*)=&\,\pi i[\Phi(\bar{S}(z^*);z,z^*)-\Phi(S(z);z,z^*)]\\&+\pi[G(\bar{S}(z^*);z,z^*)+G(S(z);z,z^*)]+I,\qquad (z,z^*)\in D'\times\bar{D}'.\end{aligned}$$

Here I denotes the right-hand side of (2.11). By reversing the transformations that led to (2.11), it is seen that I is equal to the right-hand side of (2.3) and is therefore zero.

On employing (2.7), (2.8), and (2.9), we find

$$\begin{aligned}(2.12)\qquad U(z,z^*)=&-\pi\{R(S(z);z,z^*)L[u]+u(S(z))L^*[R]\}\\&-\pi\{R(\bar{S}(z^*);z,z^*)L[u]+u(\bar{S}(z^*))L^*[R]\}\\&+\pi i\int_{S(z)}^{\bar{S}(z^*)}\Big(R(s;z,z^*)\Big\{\frac{\partial}{\partial\nu}L[u]+(a\eta'-b\xi')L[u]\Big\}\\&\qquad-L[u]\frac{\partial R}{\partial\nu}(s;z,z^*)\\&\qquad+L^*[R]\Big\{\frac{\partial u}{\partial\nu}+(a\eta'-b\xi')u\Big\}-u\frac{\partial}{\partial\nu}L^*[R]\Big)\,ds.\end{aligned}$$

In the integral, the arguments $(s;\ z,z^*)$ of R and its derivatives are an abbreviation for $(Z(s),\bar{Z}(s);\ z,z^*)$. The argument $(\xi(s),\eta(s))$ of u and its derivatives will be abbreviated to s. In the remaining terms of (2.12), the argument s of R and u is evaluated at $S(z)$ or at $\bar{S}(z^*)$, as the case may be. Thus (2.12) expresses $U(z,z^*)$ in terms of data on C'.

The right-hand side of (2.12) is analytic for (z,z^*) in a neighborhood of $C'\times\bar{C}'$. Thus, although derived for $(z,z^*)\in D'\times\bar{D}'$, (2.12) is valid throughout $N\times\bar{N}$ and provides the solution to the analytic Cauchy problem on C'. Moreover, since C' is any sufficiently small subarc of C, (2.12) gives the solution to the analytic Cauchy problem on C.

From the mode of derivation, based on the Cauchy–Kowalewski theorem, one may conclude that (2.12) automatically satisfies $L^2[u]=0$, and that it will reproduce the data when $z^*=\bar{z}$ and z approaches C; these points may also be verified directly from (2.12). Again, from (2.12) it is clear that appropriate Cauchy data are the values of u, $\partial u/\partial\nu$, $L[u]$ and $\partial L[u]/\partial\nu$, which may be prescribed as arbitrary analytic functions on C. In contrast to an earlier result [2], the representation (2.12) exhibits the solution explicitly in terms of the Riemann function and the Cauchy data.

We summarize these results in a theorem.

REPRESENTATION THEOREM. *Let C denote an analytic arc in the x, y-plane, and suppose that analytic data u, $\partial u/\partial \nu$, $L[u]$ and $(\partial/\partial\nu)L[u]$, are prescribed on C. Then the solution to the analytic Cauchy problem for $L^2[u]=0$ on C is given by* (2.12).

In the particular case for which C is an arc of the ξ-axis, we have $\xi(s)=s$, $\eta(s)=0$, $\partial/\partial\nu=-\partial/\partial\eta$, $Z(s)=s=\bar{Z}(s)$, $S(z)=z$, $\bar{S}(z^*)=z^*$, and (2.12) becomes

$$\begin{aligned}U(z,z^*)=-\pi\{&R(z,z;z,z^*)(L[u])(z)+R(z^*,z^*;z,z^*)(L[u])(z^*)\\&+u(z)(L^*[R])(z,z;z,z^*)+u(z^*)(L^*[R])(z^*,z^*;z,z^*)\}\\&-\pi i\int_z^{z^*}\left(R(s;z,z^*)\left\{\frac{\partial}{\partial\eta}L[u]+bL[u]\right\}+L[u]\frac{\partial R}{\partial\eta}(s;z,z^*)\right.\\&\left.+(L^*[R])(s;z,z^*)\left(\frac{\partial u}{\partial\eta}+bu\right)-u\left(\frac{\partial}{\partial\eta}L^*[R]\right)(s;z,z^*)\right)ds.\end{aligned}$$

This may be regarded as a natural generalization of Henrici's result [12, §5.3] for second order equations.

If u happens to be the solution to an analytic boundary value problem in a domain D, the boundary of which contains the analytic arc C, then (2.12) provides a representation for the continuation of u across C. In the earlier work [1], the boundary value problem was taken as the starting point; it is clear from the present work that the Cauchy–Kowalewski theorem provides a more natural point of departure.

The expression (2.12) is the analogue of one found in [1] for the equation $L[u]=0$. There a connection with results of Lewy [3] and Garabedian [4] was pointed out. It is apparent that (2.12) could be found by suitably extending their analyses. Moreover, since (2.12) in effect must be the result of summing the series arising in the Cauchy–Kowalewski theorem, a connection is likely to exist with the results of Hill [7].

Knowledge of a domain of analyticity of R and of the Cauchy data enables one to determine a corresponding domain of analyticity of U. Such results have been obtained by Henrici [12] for second order equations and by Colton [2] for a class of fourth order equations. Since the equations considered in the present work are of the type studied by Vekua [9, Chapter V], R will be analytic for $(\zeta,\zeta^*,z,z^*)\in D\times\bar{D}\times D\times\bar{D}$ if D is a fundamental domain. Thus, results similar to those in [2] could be easily obtained. We prefer, however, to describe some related results for the biharmonic equation. This is done in the following section.

3. The biharmonic equation. For the biharmonic equation (1.3), a normalized fundamental solution is given by John [10, p. 44], or by Vekua [9, p. 183, (36.6)] with a change of sign:

$$S(\xi,\eta;x,y)=(8\pi)^{-1}r^2\log r+B(\xi,\eta;x,y).$$

So, by (2.2),

$$A(\xi,\eta;x,y)=-(8\pi)^{-1}\left[(\xi-x)^2+(\eta-y)^2\right],$$

and

$$R(\zeta,\zeta^*;z,z^*)=-(8\pi)^{-1}(\zeta-z)(\zeta^*-z^*).$$

Since $L=\Delta$, a straight forward calculation together with (2.12) yields the following representation for the solution to the Cauchy problem:

$$\begin{aligned}(3.1)\quad U(z,z^*)=&\frac{1}{2}\left[u(S(z))+u(\bar{S}(z^*))\right]\\&-\frac{1}{8}\int_{S(z)}^{\bar{S}(z^*)}\Big(\Delta u\{(Z(s)-z)\bar{Z}'(s)-(\bar{Z}(s)-z^*)Z'(s)\}\\&\qquad+i(Z(s)-z)(\bar{Z}(s)-z^*)\frac{\partial\Delta u}{\partial\nu}+4i\frac{\partial u}{\partial\nu}\Big)\,ds.\end{aligned}$$

This result seems to be new.

In the particular case for which C is an arc of the ξ-axis, we find that

$$\begin{aligned}(3.2)\qquad U(z,z^*)=&\frac{1}{2}\left[u(z,0)+u(z^*,0)\right]\\&+\frac{1}{8}i\int_z^{z^*}\left\{2y\Delta u+\left[(s-x)^2+y^2\right]\frac{\partial\Delta u}{\partial\eta}+4\frac{\partial u}{\partial\eta}\right\}ds,\end{aligned}$$

where, in the integrand, u and its derivatives are evaluated at $(s,0)$. Other equivalent forms for U may be obtained in this case through integration by parts of terms that involve $u_{\xi\xi}$.

These results may be applicable to problems of two-dimensional elasticity theory. They could be used to generate complete sets of solutions to the biharmonic equation, analogous to those described by Vekua [9, Chapter II], or to obtain alternative representations for known solutions to the biharmonic equation. But, rather than proceed along these lines, we shall show how (3.1) may be used to obtain regularity results for u in a domain D intersected by an analytic arc C. (The function u is regular in D if $u\in C^4(D)$.)

It will be assumed that $u(x,y)$ is real, and that D is simply connected; this latter condition may be relaxed in some of the following. Then from (3.1) it is not difficult to show that

$$(3.3)\qquad u(x,y)=\operatorname{Re}[\bar{z}f(z)+g(z)];$$

here f and g are analytic in a neighborhood of C, and

$$(3.4)\quad f(z):=\frac{1}{4}\int_0^{S(z)}\left[Z'(s)\Delta u-i(Z(s)-z)\frac{\partial\Delta u}{\partial\nu}\right]ds,$$

$$\begin{aligned}(3.5)\quad g(z):=u(S(z))+&\frac{1}{4}\int_0^{S(z)}\Big[\{(Z(s)-z)\bar{Z}'(s)-\bar{Z}(s)Z'(s)\}\,\Delta u\\&\qquad+i(Z(s)-z)\bar{Z}(s)\frac{\partial\Delta u}{\partial\nu}+4i\frac{\partial u}{\partial\nu}\Big]\,ds.\end{aligned}$$

Without loss in generality, the lower limits of the integrals have been taken to be zero. The choice of another real value of s that corresponds to a point on C will change f and g by linear functions in z and leave $u(x,y)$ unaltered.

It is known that any solution to the biharmonic equation, regular in a simply-connected domain D, may be written in the form (3.3), with f and g analytic in D; see, for example [9, (35.22)]. Thus it is not surprising that (3.1) can be put in this form.

Let us now suppose that f and g can be continued analytically throughout D. For u to be regular in D, it is not necessary that each of the Cauchy data u, $\partial u/\partial \nu$, Δu, and $\partial\Delta u/\partial \nu$ (evaluated at $s=S(z)$) be analytic throughout D; but certain combinations of the data must be analytic. To obtain these conditions in their most simple form, we differentiate (3.4) and (3.5) twice with respect to z. It is found that

$$f''(z)=\frac{1}{4}\frac{d}{dz}\Delta u(S(z))+\frac{1}{4}iS'(z)\frac{\partial\Delta u}{\partial \nu}(S(z))=\frac{1}{4}\Delta(u_\xi-iu_\eta),$$

in which the latter expression is evaluated at $(\xi(S(z)),\eta(S(z)))$. In terms of U, this becomes

$$f''(z)=2U_{zzz^*}(z,T(z)), \tag{3.6}$$

where $T(z):=\bar{Z}(S(z))$, and the corresponding result for g'' is

$$g''(z)=2U_{zz}(z,T(z))-2T(z)U_{zzz^*}(z,T(z)). \tag{3.7}$$

All quantities in (3.6) and (3.7) can be found from the Cauchy data on C. It is evident that u is regular in D if the right-hand sides of (3.6) and (3.7) can be continued analytically throughout D.

The converse of this result is also true. If u is a regular biharmonic function in a simply-connected domain D then, in a neighborhood $N\times\bar{N}$ of $C\times\bar{C}$, the representation (3.1) is valid. From this follows (3.3), with f'' and g'' given by (3.6) and (3.7). Moreover, it is known that there exist functions F and G, analytic in D, such that $u(x,y)=\operatorname{Re}[\bar{z}F(z)+G(z)]$. Then, for $z\in N$, $F''(z)=f''(z)$ and $G''(z)=g''(z)$; thus F and G differ from f and g at most by linear functions of z. Hence f and g can be continued throughout D and the result follows.

The foregoing discussion is summarized in the following theorem.

THEOREM. *Let D be a simply-connected domain in the x, y-plane that is intersected by an analytic arc C. Consider a Cauchy problem for the biharmonic equation with real analytic data on C. Then the solution exists and is a regular biharmonic function in D if and only if the functions f'' and g'' of* (3.6) *and* (3.7), *obtainable from the Cauchy data, can be continued analytically throughout D.*

Expressions for f and g that are equivalent to (3.4) and (3.5), but which exhibit more clearly the dependence on the analytic combinations of Cauchy data can be obtained by integration of (3.6) and (3.7):

$$f(z)=2\int_{Z(0)}^{z}ds\int_{Z(0)}^{s}U_{zzz^*}(t,T(t))\,dt+\frac{1}{4}(z-Z(0))\Delta u(0),$$

$$g(z)=2\int_{Z(0)}^{z}ds\int_{Z(0)}^{s}[U_{zz}(t,T(t))-T(t)U_{zzz^*}(t,T(t))]\,dt$$

$$+\frac{z-Z(0)}{Z'(0)}\left[u'(0)+i\frac{\partial u}{\partial \nu}(0)-\frac{1}{4}\bar{Z}(0)Z'(0)\Delta u(0)\right]+u(0).$$

It is observed that the theorem does not require that the data individually be analytic in D, nor that D be conformally symmetric in the sense of Henrici [12, Def. 5.1]. But to prove that each of the data can be continued analytically into D, given that u is regular biharmonic in D, it is necessary to show that $\bar{Z}(S(z))\in\bar{D}$ when $z\in D$ and this requires that D be conformally symmetric with respect to C. Here $s=S(z)$ will play the role of the conformal transformation that maps C into an interval of the real axis and D into a domain that is symmetric with respect to the real axis.

If u is the solution to a boundary value problem in a domain D with boundary C on which analytic boundary conditions are prescribed, then the functions f and g must be analytic in D. This shows that in general it is not appropriate to prescribe Δu and $\partial\Delta u/\partial\nu$ on C, since the analyticity requirement on f may not be met. No inconsistency necessarily follows on this account if other pairs of data are specified.

Let us consider a simple example to illustrate the theorem. Suppose that u is a fundamental solution for the biharmonic equation, with singularity at $(0,-h)$, $h>0$:

$$u(x,y)=r^2\log r^2, \tag{3.8}$$

in which $r^2:=x^2+(y+h)^2$. Corresponding to (3.8), we have

$$U(z,z^*)=(z+ih)(z^*-ih)\log[(z+ih)(z^*-ih)].$$

For C we shall take $y=0$, so $z^*=z$ and $\bar{Z}(S(z))\equiv z$. It is evident that the Cauchy data are singular at $z=ih$ and $z=-ih$. But

$$U_{zzz^*}(z,z)=\frac{1}{z+ih},\qquad U_{zz}(z,z)-zU_{zzz^*}(z,z)=-\frac{ih}{z+ih},$$

and these are analytic except at $z=-ih$, that is, at $x=0$, $y=-h$. Consequently $u(x,y)$ is regular except at $(0,-h)$.

Remark. Analogous but more simple results obtain for solutions to the Laplace equation. From (7.2) of [1], we have

$$u(x,y)=\operatorname{Re}h(z), \tag{3.9}$$

where

$$h(z):=u(S(z))+i\int_0^{S(z)}\frac{\partial u}{\partial\nu}(s)\,ds.$$

It is apparent that a necessary and sufficient condition for u to be regular (that is, C^2) in a simply-connected domain D intersected by an analytic arc C is that $[u'(S(z))+i\partial u/\partial\nu(S(z))]S'(z)$ be analytic in D. This expression may be written more simply as $2U_z(z,T(z))$, and (3.9) as

$$u(x,y)=u(0)+2\operatorname{Re}\int_{Z(0)}^{z}U_z(t,T(t))\,dt.$$

If u happens to be the solution to a boundary value problem in D, then h is precisely the function obtained when the integral equation satisfied by u and $\partial u/\partial\nu$ on C is continued analytically into the complex domain; see [1, (6.2)]. It is likely that a similar correspondence exists in the biharmonic case.

4. The generalized axially symmetric biharmonic equation. As a second application, we shall consider the generalized axially symmetric biharmonic equation

$$L^2[u]=0, \tag{4.1}$$

in which

$$(4.2)\qquad L[u]:=\Delta u+\left(\frac{2\alpha}{y}\right)u_y,\qquad \alpha>0,$$

and we shall see how corresponding results for solutions to $(L+k^2)^2[u]=0$ are easily obtained.

Specific examples of (4.1) arise in three-dimensional elasticity problems in which case y denotes a radial variable. In particular, when a solution to such a problem is expressed in terms of Fourier components with respect to an angular variable θ $(0\leq\theta\leq 2\pi)$, the coefficients of $\sin n\theta$ and $\cos n\theta$ $(n=0,1,2,\cdots)$ are of the form $y^nU(x,y)$ where U is a solution to (4.1) for $\alpha=n+\frac{1}{2}$.

In this section, we shall specialize (2.12) in accordance with (4.2) and we shall obtain specific and rather simple results when C is an arc of $y=0$ on which the differential equation is singular but the solution is analytic.

To find the Riemann function, we use Weinacht's [11] integral representation for a fundamental solution. By taking $k=2\alpha$, $n=2$, $s=2$ in (3.1) and (3.5) of [11], we find that

$$(4.3)\qquad S(\xi,\eta;\,x,y)=\frac{\eta^{2\alpha}}{8\pi(\alpha-1)}\int_0^{\pi}\sigma^{2-2\alpha}\sin^{2\alpha-1}\theta\,d\theta,$$

with

$$\sigma^2:=(\xi-x)^2+\eta^2+y^2-2\eta y\cos\theta.$$

This representation is valid for $\alpha>0$, $\alpha\neq 1$. The excluded value $(\alpha=1)$ is one of the exceptional set $\{1,0,-1,-2,\cdots\}$ for which a solution to (4.1) that is analytic on $y=0$ is not necessarily an even function of y. We shall consider only even solutions and arguments of analytic continuation will be used later to extend our results to $\alpha=1$. Since S satisfies (4.1) as a function of x and y, it is easy to verify that it satisfies $L^{*2}[S]=0$ in ξ and η.

By elementary changes of variable in (4.3), S may be expressed in terms of the hypergeometric function F:

$$(4.4)\qquad S(\xi,\eta;\,x,y)=\frac{2^{2\alpha-4}\eta^{2\alpha}(\Gamma(\alpha))^2}{\pi(\alpha-1)q^{2\alpha-2}\Gamma(2\alpha)}F\left(\alpha-1,\alpha;\,2\alpha;\,\frac{4\eta y}{q^2}\right),$$

wherein

$$q^2:=(\xi-x)^2+(\eta+y)^2.$$

Then, on using an expansion [13, p. 559, 15.3.11] for F, the coefficient A of $\log(1/r)$ in (4.4) is found to be

$$A(\xi,\eta;\,x,y)=-\frac{2^{2\alpha-3}\eta^{2\alpha}r^2}{\pi q^{2\alpha}}F\left(\alpha,\alpha+1;\,2;\,\frac{r^2}{q^2}\right).$$

In what follows, we shall consider only the case in which C becomes, in the limit, an arc of the ξ-axis. Here $\xi(s)=s$, $\eta(s)=\eta(>0)$, $\partial/\partial\nu=-\partial/\partial\eta$, $Z(s)=s+i\eta$, $\bar{Z}(s)=s-i\eta$, $S(\zeta)=\zeta-i\eta$, $\bar{S}(\zeta^*)=\zeta^*+i\eta$, $a=0$, $b=2\alpha/\eta$. It is also convenient to employ the function A rather than R. Thus, (2.12) becomes

(4.5)

$$\begin{aligned}U(z,z^*)=-\pi\{&A(z-i\eta,\eta;\,x,y)(L[u])(z-i\eta,\eta)\\&+u(z-i\eta,\eta)(L^*[A])(z-i\eta,\eta;\,x,y)\\&+A(z^*+i\eta,\eta;\,x,y)(L[u])(z^*+i\eta,\eta)\\&\qquad+u(z^*+i\eta,\eta)(L^*[A])(z^*+i\eta,\eta;\,x,y)\}\\-\pi i\int_{z-i\eta}^{z^*+i\eta}\Bigg(&A\left\{\frac{\partial}{\partial\eta}+\frac{2\alpha}{\eta}\right\}L[u]-L[u]A\\&\qquad+L^*[A]\left\{u_\eta+\left(\frac{2\alpha}{\eta}\right)u\right\}-u\frac{\partial}{\partial\eta}L^*[A]\Bigg)\,ds.\end{aligned}$$

In the integrand, the arguments of A and its derivatives are $(s,\eta;\,x,y)$; the arguments of u and its derivatives are (s,η).

We require knowledge of the behavior of both u and A, and of certain derivatives, as $\eta\downarrow 0$. To this end, we note first that a solution $u(\xi,\eta)$ that is analytic in a neighborhood of $\eta=0$ is necessarily an even function of η if $\alpha\neq 1,0,-1,-2,\cdots$. One may see this by using power series arguments similar to those of Henrici [14] or Hyman [15]. We shall restrict our attention to analytic solutions that are even in η. Thus we expect that the representation for U will involve only the axial values $u(\xi,0)$ and $u_{\eta\eta}(\xi,0)$.

The axial behavior of the Riemann function is more complex. According to [13, p. 559, 15.3.6], one may write

$$\begin{aligned}(4.6)\quad A(\xi,\eta;\,x,y)=&-\frac{2^{2\alpha-3}\Gamma(1-2\alpha)}{\pi\Gamma(2-\alpha)\Gamma(1-\alpha)}r^2\left(\frac{\eta}{q}\right)^{2\alpha}F\left(\alpha,\alpha+1;\,2\alpha;\,\frac{4\eta y}{q^2}\right)\\&-\frac{2^{-2\alpha-1}\Gamma(2\alpha-1)y^{1-2\alpha}r^2\eta}{\pi\Gamma(\alpha)\Gamma(1+\alpha)q^{2-2\alpha}}F\left(2-\alpha,1-\alpha;\,2-2\alpha;\,\frac{4\eta y}{q^2}\right),\end{aligned}$$

for $\alpha>0$, $\alpha\neq p/2$ $(p=1,2,3,\cdots)$. Analytic continuation arguments will be used later to extend our results to these excluded values of α.

By employing (4.6) and series expansions of the hypergeometric functions for small values of $4\eta y/q^2$, we find, as $\eta\downarrow 0$, that

$$(4.7)\qquad A(s,\eta;\,x,y)\to 0,$$

$$(4.8)\qquad A_\eta-(2\alpha/\eta)A\to-\frac{2^{-2\alpha-1}\Gamma(2\alpha)}{\pi\Gamma(\alpha)\Gamma(\alpha+1)}y^{1-2\alpha}\rho^{2\alpha},$$

$$(4.9)\qquad (L^*[A])(z-i\eta,\eta;\,x,y)\to 0,$$

$$(4.10)\qquad (L^*[A])(z^*+i\eta,\eta;\,x,y)\to 0,$$

$$(4.11)\qquad \left(\frac{\partial}{\partial\eta}-\frac{2\alpha}{\eta}\right)(L^*[A])(s,\eta;\,x,y)\to\frac{\Gamma(\alpha+\frac{1}{2})}{\pi\Gamma(\frac{1}{2})\Gamma(\alpha)}y^{1-2\alpha}\rho^{2\alpha-2},$$

in which

$$\rho^2:=(x-s)^2+y^2.$$

Because of the factor r^2 in (4.6), the limits (4.7), (4.8), and (4.11) are uniform on the integration path. In evaluating some of these limits, it is necessary to assume that α is sufficiently large; but the final results will be true for all $\alpha>0$.

On letting $\eta\downarrow 0$ in (4.5), and inserting the results (4.7)–(4.11), we find that

$$(4.12)\quad U(z,z^*)=-i\frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(\alpha)}y^{1-2\alpha}\int_{x-iy}^{x+iy}\left\{L[u](s,0)\frac{\rho^{2\alpha}}{(4\alpha)}+u(s,0)\rho^{2\alpha-2}\right\}ds.$$

Here

$$L[u](s,0)=u_{\xi\xi}(s,0)+(1+2\alpha)u_{\eta\eta}(s,0),$$

since $u_\eta/\eta\to u_{\eta\eta}$ as $\eta\downarrow 0$. Thus U is determined by axial values of u and $u_{\eta\eta}$.

Although derived under the assumptions that α is sufficiently large and $\alpha\neq p/2$ $(p=1,2,3,\cdots)$, the validity of (4.12) can be extended to all $\alpha>0$ by analytic continuation arguments.

The substitution $s=x+iy\cos\theta$ leads to

$$(4.13)\quad U(z,z^*)=\frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(\alpha)}\left\{\int_0^\pi u(x+iy\cos\theta,0)\sin^{2\alpha-1}\theta\,d\theta\right.$$
$$\left.+\frac{y^2}{4\alpha}\int_0^\pi(L[u])(x+iy\cos\theta,0)\sin^{2\alpha+1}\theta\,d\theta\right\}.$$

These results are believed to be new. They are analogous to Weinstein's results [16] in generalized axially symmetric potential theory.

It is evident that $U(z,z^*)$ is even in y and is analytic on the axis $y=0$. Moreover, by direct calculation it may be verified that $L^2[U]=0$, and that U and $U_{\eta\eta}$ reduce correctly to u and $u_{\eta\eta}$ as $y\downarrow 0$. Thus, the previous somewhat formal analysis is justified.

Each of the integrals in (4.13) is a generalized axially symmetric potential function. More specifically, u is written as $u=v_1+y^2v_2$, v_1 and v_2 being solutions to the second order equation $L[v]=0$ for values α and $\alpha+1$, respectively, of the parameter. This form of the decomposition of a solution to $L^2[u]=0$ was given originally by Payne [17]. From (4.13), it is evident that $v_2\equiv 0$ if $L[u]=0$.

This decomposition suggests the form of the corresponding representation for solutions to the equation

$$(4.14)\qquad (L+k^2)^2[w]=0.$$

A comparison with Henrici's representation [14] for a solution to the generalized axially symmetric Helmholtz equation in terms of axial values indicates that $\rho^{2\alpha}$ in (4.12) should be replaced by $\rho^\alpha 2^\alpha k^{-\alpha}\Gamma(\alpha+1)J_\alpha(k\rho)$, and similarly for $\rho^{2\alpha-2}$. Then, on replacing u by w and $L[u]$ by $(L+k^2)[w]$, we obtain the following representation for $W(z,z^*)$:

(4.15)

$$W(z,z^*)=-ik^{-\alpha}2^{\alpha-2}\frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\frac{1}{2})}y^{1-2\alpha}$$
$$\cdot\int_{x-iy}^{x+iy}\{(L+k^2)[w](s,0)\rho^\alpha J_\alpha(k\rho)+2kw(s,0)\rho^{\alpha-1}J_{\alpha-1}(k\rho)\}\,ds.$$

Direct calculation shows that this is a solution to (4.14) that is analytic on $y=0$, even in y, and such that W and $W_{\eta\eta}$ reproduce the values $w(x,0)$ and $w_{\eta\eta}(x,0)$ as $y\downarrow 0$. The analogue of (4.13) is

(4.16)

$$W(z,z^*)=\frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\frac{1}{2})}\left\{\int_0^{\pi} w(x+iy\cos\theta,0)\sin^{2\alpha-1}\theta\left(\frac{ky\sin\theta}{2}\right)^{1-\alpha}J_{\alpha-1}(ky\sin\theta)\,d\theta\right.$$

$$\left.+y^2\int_0^{\pi}(L+k^2)[w](x+iy\cos\theta,0)\sin^{2\alpha+1}\theta\left(\frac{ky\sin\theta}{2}\right)^{-\alpha}J_{\alpha}(ky\sin\theta)\,d\theta\right\}.$$

Here the first integral is a solution to the second-order equation $(L+k^2)[w]=0$; the second integral is a solution to the same equation with α replaced therein by $\alpha+1$.

5. Concluding remarks. Although it has been applied only to an equation of the form $L^2[u]=0$, the method developed here is rather more general. It is only necessary that the equation possess a fundamental solution of the form (2.2) near the point in question. The existence of such a fundamental solution has been demonstrated; see, for example, [10, Chapter III]. Thus, at least in principle, a representation analogous to (2.12) for a general analytic elliptic equation of order $2n$ in two independent variables could be constructed.

Some explicit results concerning regularity of solutions to the biharmonic equation have been obtained. It is expected that similar results could be obtained in the general case. Each of the representations of biharmonic functions is of interest in itself, and may find application in elasticity theory. In particular, the representation (4.12) (or (4.13)) is analogous to the Poisson representation for solutions to the generalized axially symmetric potential equation; see, for example [16, (33)]. Therefore, it may be useful in solving axially symmetric boundary value problems for the biharmonic equation in the same way that the Poisson representation has been employed by Heins [18] to solve axially symmetric boundary value problems for the Laplace equation. For related work, see [19] and [20].

Finally, we observe that some results for a real, hyperbolic equation can be obtained from the present work. Under the transformation $y\to iy$, $\eta\to i\eta$, the Cauchy problem on $\eta=0$ for the elliptic equation goes into the Cauchy problem on $\eta=0$ for the hyperbolic equation. It is easily verified that this substitution in (3.2) yields the following representation for the solution to Cauchy's problem for $u_{xxxx}-2u_{xxyy}+u_{yyyy}=0$ on $y=0$:

$$u(x,y)=\frac{1}{2}[u(x+y,0)+u(x-y,0)]$$

$$-\frac{1}{8}\int_{x-y}^{x+y}\left\{2y(u_{\xi\xi}-u_{\eta\eta})-\left[(s-x)^2-y^2\right](u_{\xi\xi\eta}-u_{\eta\eta\eta})-4u_{\eta}\right\}ds.$$

Here the (real) arguments of u and its derivatives in the integrand are $(s,0)$. Similarly, the solution to $L^2[u]=0$, where $L[u]:=u_{xx}-u_{yy}-(2\alpha/y)u_y$, and for which u_y and u_{yyy} are zero on $y=0$, follows from (4.13):

$$u(x,y)=\frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(\alpha)}\left\{\int_0^{\pi}u(x-y\cos\theta,0)\sin^{2\alpha-1}\theta\,d\theta\right.$$

$$\left.-\frac{y^2}{4\alpha}\int_0^{\pi}(L[u])(x-y\cos\theta,0)\sin^{2\alpha+1}\theta\,d\theta\right\}.$$

REFERENCES

[1] R. F. MILLAR, *The analytic continuation of solutions to elliptic boundary value problems in two independent variables*, J. Math. Anal. Appl. 76 (1980), pp. 498–515.

[2] D. COLTON, *Cauchy's problem for a class of fourth order elliptic equations in two independent variables*, Applicable Anal., 1 (1971), pp. 13–22.

[3] H. LEWY, *On the reflection laws of second order differential equations in two independent variables*, Bull. Amer. Math. Soc., 65 (1959), pp. 37–58.

[4] P. R. GARABEDIAN, *Applications of analytic continuation to the solution of boundary value problems*, J. Rational Mech. Anal., 3 (1954), pp. 383–393.

[5] C. L. YU, *Reflection principle for solutions of higher order elliptic equations with analytic coefficients*, SIAM J. Appl. Math., 20 (1971), pp. 358–363.

[6] ———, *Integral representation, analytic continuation and the reflection principle under the complementing boundary condition for higher order elliptic equations in the plane*, this Journal, 5 (1974), pp. 209–223.

[7] C. D. HILL, *Linear functionals and the Cauchy-Kowalewski theorem*, J. Math. Mech., 19 (1969), pp. 271–277.

[8] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Interscience, New York, 1962.

[9] I. N. VEKUA, *New Methods for Solving Elliptic Equations*, North-Holland, Amsterdam, Wiley-Interscience, New York, 1967.

[10] F. JOHN, *Plane Waves and Spherical Means Applied to Partial Differential Equations*, Interscience, New York, 1955.

[11] R. J. WEINACHT, *Fundamental solutions for a class of singular equations*, Contributions to Differential Equations, 3 (1964), pp. 43–55.

[12] P. HENRICI, *A survey of I. N. Vekua's theory of elliptic partial differential equations with analytic coefficients*, Z. Angew. Math. Phys., 8 (1957), pp. 169–203.

[13] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, U.S. Department of Commerce, National Bureau of Standards, Applied Mathematics Series 55, Washington, 1966.

[14] P. HENRICI, *Zur Funktionentheorie der Wellengleichung*, Comment. Math. Helv., 27 (1953), pp. 235–293.

[15] M. A. HYMAN, *Concerning analytic solutions of the generalized potential equation*, Nederl. Akad. Wetensch. Indag. Math., 16 (1954), pp. 408–413.

[16] A. WEINSTEIN, *Generalized axially symmetric potential theory*, Bull. Amer. Math. Soc., 59 (1953), pp. 20–38.

[17] L. E. PAYNE, *Representation formulas for solutions of a class of partial differential equations*, J. Math. and Phys., 38 (1959), pp. 145–149.

[18] A. E. HEINS, *Axially-symmetric boundary-value problems*, Bull. Amer. Math. Soc., 71 (1965), pp. 787–808.

[19] A. IA. ALEKSANDROV, *One form of solution of three-dimensional axisymmetric problems of elasticity theory by means of functions of a complex variable and the solution of these problems for the sphere*, J. Appl. Math. Mech., 26 (1962), pp. 188–198.

[20] A. IA. ALEKSANDROV AND IU. I. SOLOV'EV, *The solution of the three-dimensional axisymmetric problem in the theory of elasticity by means of line integrals*, J. Appl. Math. Mech., 28 (1964), pp. 1106–1112.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics
010

NUMERICAL COMPUTATIONS OF THE SPECTRA OF THE LAPLACIAN ON 7-DIMENSIONAL HOMOGENEOUS MANIFOLDS $SU(3)/T(k,l)$*

HAJIME URAKAWA[†]

Abstract. The spectra of the Laplacian of a few compact Riemannian manifolds without boundary, e.g., flat tori, lens spaces and Riemannian symmetric spaces have been determined. In general, it is difficult to determine the spectra of the Laplacian of Riemannian manifolds. In this paper we compute numerically the spectra of 7-dimensional nonsymmetric Riemannian manifolds $SU(3)/T(k,l)$, using unitary representation theory.

1. Introduction. Let (M,g) be an n-dimensional compact Riemannian manifold without boundary. Let Δ be the Laplacian of (M,g) acting on the space $C^\infty(M)$ of complex valued C^∞ functions on M, that is,

$$\Delta=-\sum_{i,j=1}^{n} g^{ij}\left(\frac{\partial^2}{\partial x_i \partial x_j}-\sum_{k=1}^{n}\Gamma_{ij}^{k}\frac{\partial}{\partial x_k}\right),$$

where the g_{ij} are the components of g with respect to a local coordinate $(x_1,\cdots,x_n)$, (g^{ij}) is the inverse matrix of (g_{ij}) and Γ_{ij}^k is Christoffel's symbol. Then the spectrum $\mathrm{Spec}(M,g)$ of Δ consists of

$$0=\lambda_0<\lambda_1\leq\lambda_2\leq\lambda_3\leq\cdots\to+\infty.$$

Problem. Given a Riemannian manifold (M,g), calculate its spectrum $\mathrm{Spec}(M,g)$. This task seems to be impossible, in general, for nonhomogeneous Riemannian manifolds and difficult even for nonsymmetric homogeneous Riemannian manifolds. For a few Riemannian manifolds, e.g., flat tori, lens spaces and symmetric spaces, spectra have been calculated (cf. [2], [3], [5] and [7]).

In this paper, we state the results of an experiment in the computation of the spectra of some 7-dimensional Riemannian manifolds $SU(3)/T(k,l)$ which were treated by S. Aloff and N. R. Wallach [1].

2. Preliminaries. In this section, we present some results on the spectra for normal homogeneous Riemannian manifolds following [6], [7].

Let G be a compact connected Lie group and let K be a closed subgroup of G. A Riemannian manifold $(G/K,g)$ is called *normal homogeneous* if g is canonically induced from a bi-invariant metric on G. That is, let $(\cdot,\cdot)$ be an $\mathrm{Ad}(G)$-invariant inner product on the Lie algebra $\mathfrak{g}$ of G. Let $\mathfrak{m}$ be the orthogonal complement to the subalgebra $\mathfrak{k}$ of K in $\mathfrak{g}$ relative to $(\cdot,\cdot)$, so that $\mathfrak{g}=\mathfrak{k}+\mathfrak{m}$ and $\mathrm{Ad}(K)\mathfrak{m}=\mathfrak{m}$. The tangent space $T_o(G/K)$ of G/K at the origin $o=\{K\}$ can be identified with the subspace $\mathfrak{m}$ by $\mathfrak{m}\ni X\mapsto X_o\in T_o(G/K)$, where $X_of=d/dtf(\exp tX\cdot o)|_{t=0}$ for a C^∞ function f on G/K. An inner product g_o on $T_o(G/K)$ defined by $g_o(X_o,Y_o)=(X,Y)$, $X,Y\in\mathfrak{m}$, can be uniquely extended to a G-invariant Riemannian metric g on M.

The spectrum $\mathrm{Spec}(G/K,g)$ of the Laplacian for a normal homogeneous Riemannian manifold can be obtained as follows. Let $\mathfrak{t}$ be a maximal abelian subalgebra

* Received by the editors July 28, 1982.

[†] Department of Mathematics, College of General Education, Tohoku University, Sendai, 980, Japan.

of $\mathfrak{g}$. Since the weight of a finite unitary representation of G relative to $\mathfrak{t}$ has its value in purely imaginary numbers on $\mathfrak{g}$, we consider the weight as an element of $\sqrt{-1}\,\mathfrak{t}^*$, where $\mathfrak{t}^*$ denotes the real dual space of $\mathfrak{t}$. From the $\mathrm{Ad}(G)$-invariant inner product $(\cdot,\cdot)$ on $\mathfrak{t}$, a positive definite inner product on $\sqrt{-1}\,\mathfrak{t}^*$ is defined in the usual way and denoted by the same symbol $(\cdot,\cdot)$. Fixing a lexicographic order $>$ on $\sqrt{-1}\,\mathfrak{t}^*$, let P be the set of all positive roots of the complexification $\mathfrak{t}^{\mathbf{C}}$ of $\mathfrak{t}$ relative to $\mathfrak{t}$. We denote by δ half the sum of all elements in P: $\delta=\frac{1}{2}\Sigma_{\alpha\in P}\alpha$. Let $\Gamma(G)=\{H\in\mathfrak{t};\ \exp H=e\}$ and $I=\{\lambda\in\sqrt{-1}\,\mathfrak{t}^*;\ \lambda(H)\in\sqrt{-1}\,2\pi\mathbf{Z}$ for all $H\in\Gamma(G)\}$. An element in I is called a *G-integral form*. The elements of

$$D(G)=\{\lambda\in I;\ (\lambda,\alpha)>0 \text{ for all } \alpha\in P\}$$

are called *dominant* G-integral forms. Then there exists a natural bijection from $D(G)$ onto the set $\mathfrak{D}(G)$ of all nonequivalent finite dimensional irreducible unitary representation of G which map a dominant G-integral form $\lambda\in D(G)$ to an irreducible unitary representation (V_λ,π_λ) having highest weight λ. For $\lambda\in D(G)$, put $d(\lambda)$ the dimension of the representation V_λ. $d(\lambda)$ is given by

$$d(\lambda)=\prod_{\alpha\in P}\frac{(\lambda+\delta,\alpha)}{(\delta,\alpha)}.$$

A representation (V_λ,π_λ) in $\mathfrak{D}(G)$ is called *spherical* relative to K if there exists a nonzero vector $v\in V_\lambda$ such that $\pi_\lambda(k)v=v$ for all $k\in K$. Put

$$m(\lambda)=\dim\{v\in V_\lambda;\ \pi_\lambda(k)v=v \text{ for all } k\in K\}.$$

Let $\mathfrak{D}(G,K)$ be the set of all spherical representations in $\mathfrak{D}(G)$ relative to K and $D(G,K)=\{\lambda\in D(G);\ (V_\lambda,\pi_\lambda)\in\mathfrak{D}(G,K)\}$. Then we have the following:

PROPOSITION 1. *The spectrum* $\mathrm{Spec}(G/K,g)$ *of the Laplacian on* $(G/K,g)$ *is given as follows*:

eigenvalues: $4\pi^2(\lambda+2\delta,\lambda),\ \lambda\in D(G,K)$,

multiplicity: $m(\lambda)d(\lambda)$.

Proof. See [7], for example.

3. A computation of the spectrum of $SU(3)/T(k,l)$. We consider the following 7-dimensional homogeneous space $SU(3)/T(k,l)$ admitting positively curved Riemannian metrics, which was discovered by S. Aloff and N. R. Wallach [1].

We preserve the notation used in §2. Let $G=SU(3)$ and $\mathfrak{g}=\mathfrak{su}(3)$ the Lie algebra of $SU(3)$. Take as K,

$$T(k,l)=\{\mathrm{diag}[e^{2\pi ik\theta},e^{2\pi il\theta},e^{-2\pi i(k+l)\theta}];\ \theta\in\mathbf{R}\},$$

$|k|+|l|\neq 0$ $(k,l\in\mathbf{Z})$, $i=\sqrt{-1}$. Here $\mathrm{diag}[x,y,z]$ denotes a 3×3 diagonal matrix whose diagonal entries are x,y and z. Consider the coset manifold $M(k,l)=SU(3)/T(k,l)$, which is simply connected and $H^4(M(k,l),\mathbf{Z})\cong\mathbf{Z}/r\mathbf{Z}$ with $r=k^2+l^2+kl$, provided k,l are relatively prime. We assume that k,l are relatively prime in the following. The Lie algebra $\mathfrak{t}(k,l)$ of $T(k,l)$ is included in a maximal abelian subalgebra $\mathfrak{t}$ of $SU(3)$ given by

$$\mathfrak{t}=\{2\pi i\,\mathrm{diag}[x_1,x_2,x_3];\ x_j\in\mathbf{R}\,(j=1,2,3),\ x_1+x_2+x_3=0\}.$$

We give an $\mathrm{Ad}(G)$-invariant inner product $(\cdot,\cdot)$ on $\mathfrak{g}$ by

$$(X,Y)=-\mathrm{Trace}(XY),\ X,Y\in\mathfrak{g}.$$

Let g be the $SU(3)$-invariant Riemannian metric on $SU(3)/T(k,l)$ induced from this inner product $(\cdot,\cdot)$.

We will compute the spectrum of the Laplacian of $(SU(3)/T(k,l),g)$ making use of Proposition 1. For this, we denote by $x_j \in \mathfrak{t}^*$ $(j=1,2,3)$ the linear mapping

$$2\pi i\,\mathrm{diag}[x_1,x_2,x_3]\mapsto x_j.$$

Put $\lambda_j=\sqrt{-1}\,x_j\in\sqrt{-1}\,\mathfrak{t}^*$ $(j=1,2,3)$. We fix an order $>$ on $\sqrt{-1}\,\mathfrak{t}^*$ in such a way that $\lambda_1>\lambda_2>0>\lambda_3$. Then the set $D(SU(3))$ of all dominant integral forms on $SU(3)$ relative to $\mathfrak{t}$ is given by

$$D(SU(3))=\{\lambda=m_1\lambda_1+m_2\lambda_2;\ m_1\geq m_2\geq 0,\ m_j\in\mathbb{Z}(j=1,2)\}.$$

On the other hand, the elements $H_{x_j}\in\mathfrak{t}$ $(j=1,2,3)$ such that $x_j(H)=(H_{x_j},H)$ for all $H\in\mathfrak{t}$ are given as follows:

$$H_{x_1}=i(6\pi)^{-1}\,\mathrm{diag}[2,-1,-1],\quad H_{x_2}=i(6\pi)^{-1}\,\mathrm{diag}[-1,2,-1]\quad\text{and}$$
$$H_{x_3}=i(6\pi)^{-1}\,\mathrm{diag}[-1,-1,2].$$

Then the inner product $(\cdot,\cdot)$ on $\sqrt{-1}\,\mathfrak{t}^*$ is given by

$$(\lambda_i,\lambda_j)=(H_{x_i},H_{x_j})=\begin{cases}6^{-1}\pi^{-2} & (i=j),\\ -12^{-1}\pi^{-2} & (i\neq j).\end{cases}$$

The set P of all positive roots of $\mathfrak{g}^{\mathbb{C}}$ relative to $\mathfrak{t}$ is

$$P=\{\lambda_i-\lambda_j;\ 1\leq i<j\leq 3\},$$

so we have

$$\delta=\lambda_1-\lambda_3=2\lambda_1+\lambda_2.$$

Therefore we have

$$(1)\qquad 4\pi^2(\lambda+2\delta,\lambda)=4\pi^2((m_1+4)\lambda_1+(m_2+2)\lambda_2,\ m_1\lambda_1+m_2\lambda_2)$$
$$=\frac{2}{3}(m_1^2+m_2^2-m_1m_2+3m_1)$$

for $\lambda=m_1\lambda_1+m_2\lambda_2\in D(SU(3))$. Moreover we have

$$(2)\qquad d(\lambda)=\prod_{1\leq i<j\leq 3}\frac{(\lambda_i-\lambda_j,\lambda+\delta)}{(\lambda_i-\lambda_j,\delta)}=\frac{1}{2}(m_1-m_2+1)(m_1+2)(m_2+1)$$

for $\lambda=m_1\lambda_1+m_2\lambda_2\in D(SU(3))$.

Now we will compute

$$m(\lambda)=\begin{cases}m(\lambda), & \lambda\in D(SU(3),T(k,l)),\\ 0, & \lambda\notin D(SU(3),T(k,l)),\end{cases}$$

by the same method as in [4].

LEMMA 1. *For $\lambda=m_1\lambda_1+m_2\lambda_2\in D(SU(3))$, the character χ_λ of the representation (V_λ,π_λ) is decomposed into the following form on $T(k,l)$:*

$$\chi_\lambda(\mathrm{diag}[e^{2\pi ik\theta},e^{2\pi il\theta},e^{-2\pi i(k+l)\theta}])$$

$$=\sum_{p=m_2+1}^{m_1+1}\sum_{q=0}^{m_2}\sum_{r=0}^{p-q-1}e^{2\pi i(k(m_1+m_2+2-2p-q+r)+l(1-p+q+2r))\theta},$$

where $i=\sqrt{-1}$.

Proof. Due to Weyl's character formula, the character χ_λ of (V_λ,π_λ) is given by

$$\xi_\delta\chi_\lambda=\xi_{\lambda+\delta}$$

on exp(t). Here

$$\xi_\delta=\prod_{1\le i<j\le 3}\left(e\left(\frac{x_i-x_j}{2}\right)-e\left(-\frac{x_i-x_j}{2}\right)\right),\quad\text{and}$$

$$\xi_{\lambda+\delta}=\begin{vmatrix} e(p_1x_1) & e(p_1x_2) & e(p_1x_3)\\ e(p_2x_1) & e(p_2x_2) & e(p_2x_3)\\ 1 & 1 & 1\end{vmatrix},$$

where we denote $p_1=m_1+2$, $p_2=m_1+1$ and $e(x)=e^{2\pi ix}$. Moreover we have

$$\xi_{\lambda+\delta}=e((p_1+p_2)x_1)\begin{vmatrix} 1 & e(p_1(x_2-x_1)) & e(p_1(x_3-x_1))\\ 1 & e(p_2(x_2-x_1)) & e(p_2(x_3-x_1))\\ 1 & 1 & 1\end{vmatrix}$$

$$=e((p_1+p_2)x_1)\begin{vmatrix} 1 & 0 & 0\\ 1 & e(p_1(x_2-x_1))-1 & e(p_1(x_3-x_1))-1\\ 1 & e(p_2(x_2-x_1))-1 & e(p_2(x_3-x_1))-1\end{vmatrix}$$

$$=e((p_1+p_2)x_1)\begin{vmatrix} X^{p_1}-1 & Y^{p_1}-1\\ X^{p_2}-1 & Y^{p_2}-1\end{vmatrix}$$

$$=e((p_1+p_2)x_1)(X-1)(Y-1)\left(\sum_{p=0}^{m_1+1}\sum_{q=0}^{m_2}(X^pY^q-X^qY^p)\right),$$

where $X=e^{2\pi i(x_2-x_1)}$, $Y=e^{2\pi i(x_3-x_1)}$. Here we have

$$\sum_{p=0}^{m_1+1}\sum_{q=0}^{m_2}(X^pY^q-X^qY^p)=\sum_{p=m_2+1}^{m_1+1}\sum_{q=0}^{m_2}(X^pY^q-X^qY^p)$$

$$=(X-Y)\sum_{p=m_2+1}^{m_1+1}\sum_{q=0}^{m_2}\sum_{r=0}^{p-q-1}X^{q+r}Y^{p-1-r}.$$

Substituting $m_1 = m_2 = 0$, we have

$$\xi_\delta = e(3x_1)(X-1)(Y-1)(X-Y) = X^{-1}Y^{-1}(X-1)(Y-1)(X-Y).$$

Therefore we have

$$\chi_\lambda = e((m_1+m_2+3)x_1) \sum_{p=m_2+1}^{m_1+1} \sum_{q=0}^{m_2} \sum_{r=0}^{p-q-1} X^{q+r+1}Y^{p-r}$$

$$= \sum_{p=m_2+1}^{m_1+1} \sum_{q=0}^{m_2} \sum_{r=0}^{p-q-1} e((m_1+m_2+2-2p-q+r)x_1+(1-p+q+2r)x_2).$$

Substituting $x_1 = k\theta$, $x_2 = l\theta$, we have the desired formula. Q.E.D.

For every $m \in \mathbb{Z}$, the following homomorphism χ_m of the 1-dimensional group $T(k,l)$ into the multiplicative group $\{z \in \mathbb{C};\ |z| = 1\}$ is well-defined:

$$\chi_m\colon\ T(k,l) \ni \operatorname{diag}[e^{2\pi ik\theta}, e^{2\pi il\theta}, e^{-2\pi i(k+l)\theta}] \mapsto e^{2\pi im\theta}.$$

Hence $\chi_m (m \in \mathbb{Z})$ are characters of $T(k,l)$. In fact, we have

$$\operatorname{diag}[e^{2\pi ik\theta}, e^{2\pi il\theta}, e^{-2\pi i(k+l)\theta}] = \text{identity} \Leftrightarrow \theta \in \mathbb{Z},$$

since k, l are relatively prime.

Therefore, due to Lemma 1, we have

PROPOSITION 2. *Let (V_λ, π_λ) be an irreducible unitary representation of $SU(3)$ with the highest weight $\lambda = m_1\lambda_1 + m_2\lambda_2 \in D(SU(3))$. Then, as a representation of $T(k,l)$, V_λ is decomposed into $T(k,l)$-irreducible submodules as follows*:

$$V_\lambda = \sum_{p=m_2+1}^{m_1+1} \sum_{q=0}^{m_2} \sum_{r=0}^{p-q-1} V_{k(m_1+m_2+2-2p-2q+r)+l(1-p+q+2r)},$$

where V_m $(m \in \mathbb{Z})$ is the 1*-dimensional irreducible $T(k,l)$-submodule of V_λ with the character χ_m.*

Because of Proposition 2, the number $m(\lambda)$ is the one of the solutions (p,q,r) of the equation

(3) $$k(m_1+m_2+2-2p-q+r)+l(1-p+q+2r)=0,$$

satisfying the conditions

(4) $$m_2+1 \le p \le m_1+1, \quad 0 \le q \le m_2, \quad 0 \le r \le p-q-1,$$

for every k, l (relatively prime) and $m_1 \ge m_2 \ge 0$ $(m_1, m_2 \in \mathbb{Z})$. To compute $m(\lambda)$, we arrange (3) and (4).

Put

(5) $$n_1 = m_1 - m_2 \ge 0, \quad n_2 = m_2 \ge 0, \quad p' = p - n_2 - 1.$$

Then the ranges in which p', q and r vary are given by

(6) $$0 \le p' \le n_1, \quad 0 \le q \le n_2, \quad 0 \le r \le p' + (n_2 - q).$$

The equation which p', q and r satisfy is given by

$$kn_1 - ln_2 - (2k+l)p' + (-k+l)q + (k+2l)r = 0. \tag{7}$$

Put $\mathbb{Z}^+ = \{0, 1, 2, \cdots\}$. For every k, l and $(n_1, n_2) \in \mathbb{Z}^+ \times \mathbb{Z}^+$, we denote by $S^{k,l}_{n_1,n_2}$ the number of the solutions (p', q, r) of (7) satisfying condition (6). If there are no solutions of (6) and (7), put $S^{k,l}_{n_1,n_2} = 0$.Thus we have the following:

THEOREM 1. *Let us preserve the above conditions. Then the spectrum of the Laplacian of the Riemannian manifold* $(SU(3)/T(k,l), g)$ *is given as follows*:

eigenvalues: $\frac{2}{3}(m_1^2 + m_2^2 - m_1 m_2 + 3m_1)$,

multiplicity: $d(\lambda)m(\lambda)$,

where

$$d(\lambda) = \tfrac{1}{2}(m_1 - m_2 + 1)(m_1 + 2)(m_2 + 1), \qquad m(\lambda) = S^{k,l}_{n_1,n_2}$$

are the number of the solutions of (6) *and* (7), $n_1 = m_1 - m_2$ *and* $n_2 = m_2$. *Here* m_1 *and* m_2 *vary over all the integers subject to the condition* $m_1 \geq m_2 \geq 0$.

4. Numerical computations. M. Kasugawa wrote a program to compute the numbers $S^{k,l}_{n_1,n_2}$ making use of a Yokokawa-Hewlett Packard computer YHP 9825 A. In the tables below, we express $S^{k,l}_{n_1,n_2}$ by the number whose position is (n_1, n_2) (see Fig. 1).

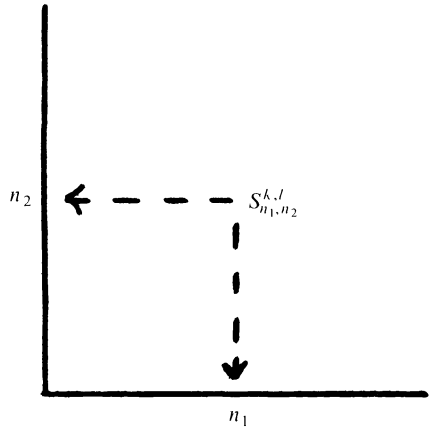


FIG. 1.

Remarks. Observing the tables of $S^{k,l}_{n_1,n_2}$, (Tables 1–7) it seems that: (I) Spec$(SU(3)/T(k,l))$=Spec$(SU(3)/T(k',l'))$ implies that $SU(3)/T(k,l)$ is isometric to $SU(3)/T(k',l')$, and (II) for every large number m, there exist (k,l) and (k',l') such that

(i) $SU(3)/T(k,l)$ is not homeomorphic to $SU(3)/T(k',l')$,

(ii) for $j \leq m$, the jth eigenvalues of $SU(3)/T(k,l)$ and $SU(3)/T(k',l')$ coincide with each other, but

(iii) Spec$(SU(3)/T(k,l), g) \neq$ Spec$(SU(3)/T(k',l'), g)$.

Acknowledgment. We express our hearty gratitude to Mr. Masa'atsu Kasugawa who wrote a program for the YHP 9825 A computer, which computes the number of solutions of certain equations.

TABLE 1

CASE $k=7$, $l=13$.

n_2 20	0	0	3	0	0	8	0	0	15	0	0	22	0	0	29	0	0	35	0	0	41	
19	0	2	0	0	6	0	0	12	0	0	19	0	0	26	0	0	32	0	0	38	0	
18	1	0	0	4	0	0	9	0	0	16	0	0	23	0	0	29	0	0	35	0	0	
17	0	0	3	0	0	$\underline{7}$	0	0	13	0	0	20	0	0	26	0	0	32	0	0	35	
16	0	2	0	0	5	0	0	10	0	0	17	0	0	23	0	0	29	0	0	32	0	
15	1	0	0	4	0	0	$\underline{8}$	0	0	14	0	0	20	0	0	26	0	0	29	0	0	
14	0	0	3	0	0	6	0	0	$\underline{11}$	0	0	17	0	0	23	0	0	26	0	0	29	
13	0	2	0	0	5	0	0	$\underline{9}$	0	0	$\underline{14}$	0	0	20	0	0	23	0	0	26	0	
12	1	0	0	4	0	0	7	0	0	$\underline{11}$	0	0	17	0	0	20	0	0	23	0	0	
11	0	0	3	0	0	6	0	0	9	0	0	$\underline{14}$	0	0	17	0	0	20	0	0	22	
10	0	2	0	0	5	0	0	8	0	0	11	0	0	14	0	0	17	0	0	19	0	
9	1	0	0	4	0	0	7	0	0	10	0	0	$\underline{11}$	0	0	14	0	0	16	0	0	
8	0	0	3	0	0	6	0	0	9	0	0	9	0	0	$\underline{11}$	0	0	13	0	0	$\underline{15}$	
7	0	2	0	0	5	0	0	8	0	0	8	0	0	$\underline{9}$	0	0	$\underline{10}$	0	0	$\underline{12}$	0	
6	1	0	0	4	0	0	7	0	0	7	0	0	7	0	0	$\underline{8}$	0	0	$\underline{9}$	0	0	
5	0	0	3	0	0	6	0	0	6	0	0	6	0	0	6	0	0	$\underline{7}$	0	0	8	
4	0	2	0	0	5	0	0	5	0	0	5	0	0	5	0	0	5	0	0	6	0	
3	1	0	0	4	0	0	4	0	0	4	0	0	4	0	0	4	0	0	4	0	0	
2	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	
1	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	
0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	n_1

TABLE 2

CASE $k=4$, $l=19$.

n_2 20	0	0	3	0	0	8	0	0	12	0	0	17	0	0	23	0	0	29	0	0	35	
19	0	2	0	0	6	0	0	10	0	0	14	0	0	20	0	0	26	0	0	32	0	
18	1	0	0	4	0	0	8	0	0	12	0	0	17	0	0	23	0	0	29	0	0	
17	0	0	3	0	0	$\underline{6}$	0	0	10	0	0	14	0	0	20	0	0	26	0	0	29	
16	0	2	0	0	5	0	0	8	0	0	12	0	0	17	0	0	23	0	0	26	0	
15	1	0	0	4	0	0	$\underline{7}$	0	0	10	0	0	14	0	0	20	0	0	23	0	0	
14	0	0	3	0	0	6	0	0	$\underline{9}$	0	0	12	0	0	17	0	0	20	0	0	23	
13	0	2	0	0	5	0	0	$\underline{8}$	0	0	$\underline{11}$	0	0	14	0	0	17	0	0	20	0	
12	1	0	0	4	0	0	7	0	0	$\underline{10}$	0	0	13	0	0	14	0	0	17	0	0	
11	0	0	3	0	0	6	0	0	9	0	0	$\underline{12}$	0	0	12	0	0	14	0	0	17	
10	0	2	0	0	5	0	0	8	0	0	11	0	0	11	0	0	12	0	0	14	0	
9	1	0	0	4	0	0	7	0	0	10	0	0	$\underline{10}$	0	0	10	0	0	12	0	0	
8	0	0	3	0	0	6	0	0	9	0	0	9	0	0	$\underline{9}$	0	0	10	0	0	$\underline{12}$	
7	0	2	0	0	5	0	0	8	0	0	8	0	0	$\underline{8}$	0	0	$\underline{8}$	0	0	$\underline{10}$	0	
6	1	0	0	4	0	0	7	0	0	7	0	0	7	0	0	$\underline{7}$	0	0	$\underline{8}$	0	0	
5	0	0	3	0	0	6	0	0	6	0	0	6	0	0	6	0	0	$\underline{6}$	0	0	8	
4	0	2	0	0	5	0	0	5	0	0	5	0	0	5	0	0	5	0	0	6	0	
3	1	0	0	4	0	0	4	0	0	4	0	0	4	0	0	4	0	0	4	0	0	
2	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	
1	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	
0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	n_1

TABLE 3

CASE $k=3$, $l=19$.

n_2	20	0	0	3	0	2	6	0	3	9	0	4	12	2	5	15	4	6	18	6	7	21	
	19	0	2	0	1	5	0	2	8	0	3	11	1	4	14	3	5	17	5	6	20	7	
	18	1	0	0	4	0	1	7	0	2	10	0	3	13	2	4	16	4	5	19	6	6	
	17	0	0	3	0	0	6	0	1	9	0	2	12	1	3	15	3	4	18	5	5	18	
	16	0	2	0	0	5	0	0	8	0	1	11	0	2	14	2	3	17	4	4	17	6	
	15	1	0	0	4	0	0	7	0	0	10	0	1	13	1	2	16	3	3	16	5	4	
	14	0	0	3	0	0	6	0	0	9	0	0	12	0	1	15	2	2	15	4	3	15	
	13	0	2	0	0	5	0	0	8	0	0	11	0	0	14	1	1	14	3	2	14	5	
	12	1	0	0	4	0	0	7	0	0	10	0	0	13	0	0	13	2	1	13	4	2	
	11	0	0	3	0	0	6	0	0	9	0	0	12	0	0	12	1	0	12	3	1	12	
	10	0	2	0	0	5	0	0	8	0	0	11	0	0	11	0	0	11	2	0	11	4	
	9	1	0	0	4	0	0	7	0	0	10	0	0	10	0	0	10	1	0	10	3	0	
	8	0	0	3	0	0	6	0	0	9	0	0	9	0	0	9	0	0	9	2	0	9	
	7	0	2	0	0	5	0	0	8	0	0	8	0	0	8	0	0	8	1	0	8	3	
	6	1	0	0	4	0	0	7	0	0	7	0	0	7	0	0	7	0	0	7	2	0	
	5	0	0	3	0	0	6	0	0	6	0	0	6	0	0	6	0	0	6	1	0	6	
	4	0	2	0	0	5	0	0	5	0	0	5	0	0	5	0	0	5	0	0	5	2	
	3	1	0	0	4	0	0	4	0	0	4	0	0	4	0	0	4	0	0	4	1	0	
	2	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	
	1	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	
	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19	19	20	n_1

TABLE 4

CASE $k=2$, $l=3$.

n_2	10	1	2	4	5	5	8	9	11	12	13	17	
	9	1	2	3	4	6	7	8	10	11	14	13	
	8	1	2	3	4	5	6	8	9	11	11	12	
	7	1	2	2	3	5	6	7	8	9	10	11	
	6	1	1	2	4	4	5	7	7	8	8	9	
	5	0	1	3	2	3	6	5	6	6	7	8	
	4	0	2	1	2	5	3	4	5	5	6	5	
	3	1	0	1	4	2	2	4	3	4	4	5	
	2	0	0	3	1	1	3	2	2	3	3	4	
	1	0	2	0	0	2	1	1	2	2	2	2	
	0	1	0	0	1	0	0	1	1	1	1	1	
		0	1	2	3	4	5	6	7	8	9	10	n_1

TABLE 5

CASE $k=2$, $l=5$.

n_2	10	0	5	0	0	13	0	0	22	0	0	31	
	9	2	0	0	9	0	0	17	0	0	26	0	
	8	0	0	6	0	0	13	0	0	21	0	0	
	7	0	3	0	0	10	0	0	16	0	0	22	
	6	1	0	0	7	0	0	13	0	0	17	0	
	5	0	0	4	0	0	10	0	0	13	0	0	
	4	0	2	0	0	7	0	0	10	0	0	13	
	3	1	0	0	4	0	0	7	0	0	9	0	
	2	0	0	3	0	0	4	0	0	6	0	0	
	1	0	2	0	0	2	0	0	3	0	0	5	
	0	1	0	0	1	0	0	1	0	0	2	0	
		0	1	2	3	4	5	6	7	8	9	10	n_1

TABLE 6

CASE $k=2, l=7$.

n_2	10	0	2	2	1	5	4	3	8	5	5	11	
	9	1	1	0	4	3	2	7	4	4	10	5	
	8	0	0	3	2	1	6	3	3	9	4	5	
	7	0	2	1	0	5	2	2	8	3	4	8	
	6	1	0	0	4	1	1	7	2	3	7	3	
	5	0	0	3	0	0	6	1	2	6	2	4	
	4	0	2	0	0	5	0	1	5	1	3	5	
	3	1	0	0	4	0	0	4	0	2	4	1	
	2	0	0	3	0	0	3	0	1	3	0	2	
	1	0	2	0	0	2	0	0	2	0	1	2	
	0	1	0	0	1	0	0	1	0	0	1	0	
		0	1	2	3	4	5	6	7	8	9	10	n_1

TABLE 7

CASE $k=2, l=9$.

n_2	10	0	2	0	2	5	1	3	8	3	4	11	
	9	1	0	1	4	0	2	7	2	3	10	4	
	8	0	0	3	0	1	6	1	2	9	3	3	
	7	0	2	0	0	5	0	1	8	2	2	8	
	6	1	0	0	4	0	0	7	1	1	7	3	
	5	0	0	3	0	0	6	0	0	6	2	1	
	4	0	2	0	0	5	0	0	5	1	0	5	
	3	1	0	0	4	0	0	4	0	0	4	2	
	2	0	0	3	0	0	3	0	0	3	1	0	
	1	0	2	0	0	2	0	0	2	0	0	2	
	0	1	0	0	1	0	0	1	0	0	1	0	
		0	1	2	3	4	5	6	7	8	9	10	n_1

REFERENCES

[1] S. ALOFF AND N. R. WALLACH, *An infinite family of distinct 7-manifolds admitting positive curved Riemannian structures*, Bull. Amer. Math. Soc., 81 (1975), pp. 93–97.

[2] M. BERGER, P. GAUDUCHON AND E. MAZET, *Le spectre d'une variété riemannienne*, Lecture Notes in Mathematics 194, Springer-Verlag, Berlin, New York, Heidelberg, 1971.

[3] A. IKEDA, *On lens spaces which are isospectral but not isometric*, Ann. Scient. Ec. Norm. Sup., 4^e serie, 13 (1980), pp. 303–315.

[4] A. IKEDA AND Y. TANIGUCHI, *Spectra and eigenforms on S^n and $P^n(\mathbb{C})$*, Osaka J. Math., 15 (1978), pp. 515–546.

[5] M. TAKEUCHI, *Modern Theory of Spherical Functions*, Iwanami, Tokyo, 1974.

[6] Y. TANIGUCHI, *Normal homogeneous metrics and their spectra*, Osaka J. Math., 18 (1981), pp. 555–576.

[7] S. YAMAGUCHI, *Spectra of flag manifolds*, Mem. Fac. Sci. Kyushu Univ., Ser. A, 33(1) (1979), pp. 95–112.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics
011

A LOCAL UNCERTAINTY PRINCIPLE*

JOHN J. BENEDETTO†

Abstract. A stationary phase argument is used to characterize the components required to estimate the difference between two Fourier transforms $\hat{f}$ and $\hat{g}$, where the support of f is compact (the Theorem). This characterization allows effective local approximation to $\hat{g}$ by $\hat{f}$ in various norms (the Proposition and the Example). The Heisenberg uncertainty principle asserts poor global approximation; hence, the Theorem, Proposition, and Example demonstrate the compatibility of the uncertainty principle and local determinacy.

Introduction. The Heisenberg uncertainty principle has several mathematical formulations, each of which is a specific theorem, e.g., Fefferman and Phong [4] or Landau, Pollak, and Slepian [6] and Fuchs [5]. Occasionally these theorems are decreed best possible, although one should interpret such optimality in terms of the mathematical *model* which exhibits uncertainty, e.g., the canonical transformations of [4] or the L^2 estimates of [5], [6].

We shall prove that the uncertainty principle, which is valid *globally* for various models, can be transgressed *locally*. This is done by means of a pointwise estimate proved in §2 and the implementation of this estimate in §3 for various special cases. Quantitatively, we are given a *time* interval $[-T,T]$, a *spectral* function V, and a frequency ω. In §2 we shall characterize the ingredients required to estimate $||F|-V|$ in a neighborhood of ω, where F is the Fourier transform of a function f supported by $[-T,T]$. Then, in §3, we shall adapt these ingredients for specific points ω to construct Fourier pairs $f\leftrightarrow F$ such that $||F|-V|$ is small in an ω-neighborhood and where f is supported by $[-T,T]$.

The reason for proving the local results of §2 and §3 is because of the traditional role played by the uncertainty principle in spectrum estimation problems and the limitations attributed to it in this role. We shall describe the spectrum estimation problem in §4 as well as indicating the usefulness of our results to deal with it. Specifically, we shall construct local pointwise approximations to given steeply decaying functions, thereby allowing effective estimation of power spectra having two close peaks.

1. Notation and definitions. $\mathbb{R}$ and $\mathfrak{R}$ will denote the real line, the former indicating the *time* axis and the latter indicating the *frequency* axis. If f is defined on $\mathbb{R}$, then its Fourier transform $\hat{f}(\omega)=F(\omega)$ is the *spectral* function $F(\omega)=\int_{-\infty}^{\infty}f(t)e^{-it\omega}\,dt$ defined on $\mathfrak{R}$. The pairing of f and F is designated by $f\leftrightarrow F$. The support of a function F is $\operatorname{supp}F$ and its supremum norm over $\mathfrak{R}$, resp., over $U\subseteq\mathfrak{R}$, is $\|F\|$, resp., $\|F\|_U$. The L^2 norm of F is $\|F\|_2=((1/2\pi)\int|F(\omega)|^2\,d\omega)^{1/2}=(\int|f(t)|^2\,dt)^{1/2}$. Also, if $U\subseteq\mathfrak{R}$, then $|U|$, resp., $\tilde{U}$, is the Lebesgue measure, resp., complement, of U.

For fixed numbers $\alpha,\Gamma>0$ we take as the prototype of a steeply decaying function the de la Vallée-Poussin kernel

$$V(\omega)=\frac{2\pi}{\alpha}\left(\frac{3\pi}{3\Gamma+\alpha}\right)^{1/2}\chi_{\Gamma+\alpha/2}*\chi_{\alpha/2}(\omega),$$

where χ_α denotes the characteristic function of $[-\alpha,\alpha]$ and $*$ denotes convolution defined as $F*G(\omega)=(1/2\pi)\int F(\omega-\gamma)G(\gamma)\,d\gamma$. V is a trapezoid function supported by

*Received by the editors February 26, 1982, and in revised form November 1, 1982.

†Department of Mathematics, University of Maryland, College Park, Maryland 20742.

$[-(\Gamma+\alpha),\Gamma+\alpha]$ and constant on $[-\Gamma,\Gamma]$; it plays an important role in the study of absolutely convergent Fourier transforms, e.g., Benedetto [1]. An easy calculation shows that $\|V\|_2=1$, $V(0)=(3\pi/(3\Gamma+\alpha))^{1/2}$, and

$$v(t)=\frac{2}{\pi\alpha}\left(\frac{3\pi}{3\Gamma+\alpha}\right)^{1/2}\frac{\sin(\Gamma+\alpha/2)t\sin(\alpha t/2)}{t^2},$$

where $\hat{v}=V$. The value of Γ plays a role in signal processing, e.g., Benedetto [2; 3] and §4; and we shall consider small constant values of α to deal with functions having steep decay or, equivalently, small sidelobes.

2. A local uncertainty theorem. Let T be a fixed positive quantity. In the following result w denotes an infinitely differentiable increasing bijection $w\colon \mathbb{R}\to\hat{\mathbb{R}}$ for which $w\colon [-T,T]\to[-(\Gamma+\alpha),\Gamma+\alpha]$ is also a bijection; and r denotes an absolutely continuous function which is supported by $[-T,T]$, positive on $(-T,T)$, and which satisfies the norm condition $\|r\|_2=\|\hat{r}\|_2=1$.

THEOREM. *Given* α, Γ, *the corresponding de la Vallée-Poussin kernel* V, *and fixed quantities* $T>0$ *and* $\omega\in\mathbb{R}$. *For each* $w\colon \mathbb{R}\to\hat{\mathbb{R}}$ *and each symmetric interval* U_ω *about* t_ω, *where* t_ω *is defined by the condition* $w(t_\omega)=\omega$, *there is a nonnegative function* r, *independent of* U_ω, *such that*

$$\text{(1)}\qquad \big||F(\omega)|-V(\omega)\big|\le|U_\omega|\Big(4r(t_\omega)+\sup_{t\in U_\omega}|r(t)-r(t_\omega)|\Big)$$
$$+\frac{192\|r\|_{U_\omega^{\sim}}}{\min\limits_{+,-}\big|w\big(t_\omega\pm|U_\omega|/2\big)-w(t_\omega)\big|}+r(t_\omega)C(U_\omega,w),$$

where $f(t)=r(t)e^{i\theta(t)}$, θ *is any primitive of* w, *and* $C(U_\omega,w)$ *is the Cornu spiral error term*

$$\text{(2)}\qquad C(U_\omega,w)=\left|\left(\frac{2\pi i}{w'(t_\omega)}\right)^{1/2}-\int_{U_\omega}e^{iw'(t_\omega)(t-t_\omega)^2/2}\,dt\right|.$$

Proof. i. For any $w\colon \mathbb{R}\to\hat{\mathbb{R}}$ we define $r\ge0$ on the interval $[-T,T]$ by the formula

$$\text{(3)}\qquad \frac{1}{2\pi}\int_{-(\Gamma+\alpha)}^{w(t)}V^2(\omega)\,d\omega=\int_{-T}^{t}r^2(u)\,du.$$

Since $w\colon [-T,T]\to[-(\Gamma+\alpha),\Gamma+\alpha]$ is a bijection, we see that r is positive on $(-T,T)$ and that if we set $r=0$ off $[-T,T]$, then

$$\|r\|_2=1.$$

Differentiation of (3) with respect to t yields

$$\text{(4)}\qquad r(t)=\sqrt{\frac{1}{2\pi}w'(t)}\,V(w(t))$$

for each $t\in(-T,T)$; and, by the definitions of V and w, we see that if we define r on $\mathbb{R}$ by (4), then it is an absolutely continuous function on $\mathbb{R}$ supported by $[-T,T]$.

If θ and $\theta+c$ are two primitives of w, then $f(t)=r(t)$ exp $i\theta(t)$ and $f_c(t)=r(t)$ exp $i(\theta(t)+c)$ have the property that $|F|=|F_c|$ on $\hat{\mathbb{R}}$. Also, by the definition of w we see that $t_\omega\in(-T,T)$ if $\omega\in(-(\Gamma+\alpha),\Gamma+\alpha)$ and $t_\omega<-T$ if $\omega<-(\Gamma+\alpha)$.

ii. For $\omega \in \mathfrak{R}$ we let U_ω be an arbitrary open symmetric interval about t_ω. Also, we set $\varphi(t) = \theta(t) - t\omega$. Then we have $\varphi'(t_\omega) = w(t_\omega) - \omega = 0$. Consequently, if we expand φ about t_ω, we obtain

$$\varphi(t) = (\theta(t_\omega) - t_\omega \omega) + \theta''(t_\omega)(t - t_\omega)^2/2 + R(t, \omega).$$

R is the remainder term $(\frac{1}{2})\int_{t_\omega}^{t} \theta^{(3)}(u)(t-u)^2\,du = (\frac{1}{6})\theta^{(3)}(u_t)(t - t_\omega)^3$, some u_t between t_ω and t; and $\lim_{t \to t_\omega} R(t, \omega)/(t - t_\omega)^3 = w''(t_\omega)/3!$.

We write $F(\omega)$ as

$$\text{(5)}\quad F(\omega) = \int_{U_\omega} r(t) e^{i(\theta(t) - t\omega)}\,dt + \int_{U_\omega^\sim} r(t) e^{i(\theta(t) - t\omega)}\,dt$$

$$= \int_{U_\omega} (r(t) - r(t_\omega)) e^{i(\theta(t) - t\omega)}\,dt$$

$$+ \int_{U_\omega} r(t_\omega)\left(e^{i(\theta(t) - t\omega)} - e^{i(\theta(t_\omega) - t_\omega \omega + \theta''(t_\omega)(t - t_\omega)^2/2 + R(t,\omega))}\right) dt$$

$$+ \int_{U_\omega} \left[r(t_\omega) e^{i(\theta(t_\omega) - t_\omega \omega)}\right]\left[e^{i(\theta''(t_\omega)(t - t_\omega)^2/2 + R(t,\omega))} - e^{i(\theta''(t_\omega)(t - t_\omega))^2/2}\right] dt$$

$$+ \int_{U_\omega} r(t_\omega) e^{i(\theta(t_\omega) - t_\omega \omega)} e^{i(\theta''(t_\omega)(t - t_\omega)^2/2)}\,dt$$

$$+ \int_{U_\omega^\sim} r(t) e^{i(\theta(t) - t\omega)}\,dt.$$

iii. It is well known that

$$\int_{-\infty}^{\infty} e^{it^2c^2}\,dt = \sqrt{\pi}\, e^{i\pi/4}/c, \qquad c > 0.$$

Thus for $c^2 = w'(t_\omega)/2$ we obtain

$$\text{(6)}\qquad \int_{-\infty}^{\infty} e^{i\theta''(t_\omega)t^2/2}\,dt = \left(\frac{2\pi i}{w'(t_\omega)}\right)^{1/2}$$

since $e^{i\pi/4} = \sqrt{i}$ and $w' = \theta''$. Note that because of the t^2 in (6) we cannot write $\int_{-r}^{r} e^{iu^2}\,du$ as a real quantity in terms of the cosine.

iv. Equations (4) and (6) and the penultimate term on the right-hand side of (5) suggest adding and subtracting

$$r(t_\omega) e^{i(\theta(t_\omega) - t_\omega \omega)} \left(\frac{2\pi i}{w'(t_\omega)}\right)^{1/2}$$

to F. Also, (4) and the triangle inequality yield

$$\text{(7)}\qquad \big||F(\omega)| - V(\omega)\big| = \left||F(\omega)| - r(t_\omega)\left(\frac{2\pi}{w'(t_\omega)}\right)^{1/2}\right|$$

$$\leq \left|F(\omega) - r(t_\omega) e^{i(\theta(t_\omega) - t_\omega \omega)}\left(\frac{2\pi i}{w'(t_\omega)}\right)^{1/2}\right|.$$

Combining (5) and (7) we obtain

$$
\begin{aligned}
(8)\qquad ||F(\omega)|-V(\omega)| \le\ & 2|U_\omega| r(t_\omega) \\
&+ |U_\omega| \sup_{t\in U_\omega} |r(t)-r(t_\omega)| + r(t_\omega)\int_{U_\omega} |1-e^{iR(t,\omega)}|\,dt \\
&+ r(t_\omega)\left|\left(\frac{2\pi i}{w'(t_\omega)}\right)^{1/2} - \int_{U_\omega} e^{i\theta''(t_\omega)(t-t_\omega)^2/2}\,dt\right| \\
&+ \left|\int_{U_\omega^{\sim}} r(t)e^{i(\theta(t)-t\omega)}\,dt\right|.
\end{aligned}
$$

v. Our result (1) will follow once we estimate the "stationary phase" term $\int_{U_\omega^{\sim}}$ in (8). Since $\operatorname{supp} r \subseteq [-T,T]$, the integral $\int_{U_\omega^{\sim}}$ is really the integral $\int_{[-T,T]\cap U_\omega^{\sim}}$; and the set $[-T,T]\cap U_\omega^{\sim}$ is either a bounded interval or a disjoint union of two bounded intervals. Any such interval $[a,b]$ has the properties that $[a,b]\subseteq[-T,T]$ and $[a,b]\cap U_\omega=\varnothing$.

Since $r\ge 0$ is a function of bounded variation, we shall invoke the mean value theorem for integrals to estimate $\int_{U_\omega^{\sim}}$ in (8). Of course, the integrand is complex and the mean value theorem is only valid for real functions. We circumvent this issue by writing the complex integrand in trigonometric form, and, hence, the integral,

$$\int_a^b r(t)e^{i(\theta(t)-t\omega)}\,dt,$$

can be written as a sum of 12 integrals of the form

$$(9)\qquad \int_c^d e^{i(\theta(t)-t\omega)}\,dt,$$

where each integral is multiplied by some value of r or $r/2$.

We use van der Corput's lemma to estimate the integral (9). The result asserts that

$$(10)\qquad \left|\int_c^d e^{i\varphi(t)}\,dt\right| \le \frac{8}{\rho},$$

where $|\varphi'|>\rho$ on $[c,d]$. In fact, we obtain (10) by means of the mean value theorem for Riemann–Stieltjes integrals, the fact that φ' is monotonic, and the following calculation:

$$\left|\int_c^d e^{i\varphi(t)}\,dt\right| = \left|\frac{1}{i}\int_c^d \frac{1}{\varphi'(t)}\,d\big(e^{i\varphi(t)}\big)\right| \le \frac{4}{|\varphi'(c)|} + \frac{4}{|\varphi'(d)|}.$$

The monotonicity is required to employ the mean value theorem and it is a consequence of the fact that $\varphi'(t)=w(t)-\omega$ is strictly increasing. The factors, 4, are a consequence of the fact that the mean value theorem is only valid for real functions.

If $t\ge t_\omega+|U_\omega|/2$, then $\varphi(t)=\theta'(t)-\omega\ge\theta'(t_\omega+|U_\omega|/2)-\omega=w(t_\omega+|U_\omega|/2)-w(t_\omega)$. Similarly, if $t<t_\omega-|U_\omega|/2$, then $\varphi(t)=\theta'(t)-\omega\le w(t_\omega-|U_\omega|/2)-w(t_\omega)$. Thus, we have $|\varphi'(t)|\ge\min_{+,-}|w(t_\omega\pm|U_\omega|/2)-w(t_\omega)|$ for $t\notin U_\omega$. Taking ρ to be this minimum and combining (10) and the twelve terms of (9), we obtain

$$\left|\int_{U_\omega^{\sim}} r(t)e^{i(\theta(t)-t\omega)}\,dt\right| \le \frac{192\|r\|_{U_\omega^{\sim}}}{\min_{+,-}\big|w\big(t_\omega\pm|U_\omega|/2\big)-w(t_\omega)\big|}.$$

This inequality combines with (8) to give (1). The factor 192 can be lowered. Q.E.D.

3. Consequences of the Theorem.

PROPOSITION. *Given α, Γ, the corresponding de la Vallée-Poussin kernel V, and constants T, $\varepsilon>0$. For any $\omega\notin[-(\Gamma+\alpha),\Gamma+\alpha]$ there is a neighborhood W_ω of ω and an absolutely continuous function f supported by $[-T,T]$ such that $\|F\|_2=\|V\|_2=1$ and*

$$\forall\gamma\in W_\omega,\quad \big||F(\gamma)|-V(\gamma)\big|<\varepsilon. \tag{11}$$

Proof. By continuity it is sufficient to prove (11) for $\gamma=\omega$.

Take $\omega<-(\Gamma+\alpha)$ and note that for any w we shall have $t_\omega<-T$. We shall choose w to be linear on $(-\infty,-T]$. Also, $r(t_\omega)$ vanishes by the construction of r from w in the Theorem. Consequently, the Theorem yields the estimate

$$\big||F(\omega)|-V(\omega)\big|\leq|U_\omega|\sup_{t\in U_\omega}r(t)+\frac{16\|r\|_{U_\omega^\sim}}{w\big(t_\omega+|U_\omega|/2\big)-w(t_\omega)}. \tag{12}$$

For positive L less than 1 and $(T/(\Gamma+\alpha))^{2/3}$ we define

$$w(t)=\frac{1}{L^{3/2}}t+\left(\frac{T}{L^{3/2}}-(\Gamma+\alpha)\right),\qquad t\leq-T.$$

L will be chosen so small that

$$-2\sqrt{\frac{3}{2\Gamma}}\,(\omega+\Gamma+\alpha)L^{1/4}<\frac{\varepsilon}{2},\qquad -16\sqrt{\frac{3}{2T}}\,\frac{L^{1/2}}{\omega+\Gamma+\alpha}<\frac{\varepsilon}{2}. \tag{13}$$

As we shall see, (13) will have the effect of making the right hand side of (12) less than ε.

The length $|U_\omega|$ of U_ω is defined as

$$|U_\omega|=-2(\omega+\Gamma+\alpha)L;$$

and so, since t_ω, defined by $w(t_\omega)=\omega$, is the center of U_ω, we have t_ω increasing to $-T$ as L shrinks. For the sake of a mental picture, keep in mind that $1/L^{3/2}=[-(\omega+\Gamma+\alpha)]/[-2(\omega+\Gamma+\alpha)L^{3/2}]$. Also, we have chosen the slope $1/L^{3/2}$ instead of $1/L$ (resp., $1/L^2$) in order to keep the second term (resp., the first term) on the right hand of (12) small; of course, the first term vanishes in the $1/L$ case.

The slope of the increasing diagonal of the $(\pm T,\pm(\Gamma+\alpha))$ box is $(\Gamma+\alpha)/T$, and thus w can be defined on $[-T,T]$ so that $w'\leq\max(1/L^{3/2},(\Gamma+\alpha)/T)$ there. Further, since $t_\omega-L^{3/2}(\omega+\Gamma+\alpha)=-T$ (by the definition of w) and $L<1$, we obtain $U_\omega\cap(-T,T)=\varnothing$. Consequently, we can choose $w'\leq(\Gamma+\alpha)/T$ on the interval $[t_\omega-L(\omega+\Gamma+\alpha),\infty)$. Of course, $w'=1/L^{3/2}$ on parts of U and we originally chose L small enough so that $1/L^{3/2}>(\Gamma+\alpha)/T$.

With these definitions of w and U_ω, which both depend on L, we evaluate the right hand side of (12). To do this we first define r as in the Theorem to vanish off $[-T,T]$ and by the formula $r(t)=\sqrt{w'(t)/(2\pi)}\;V(w(t))$ for $t\in[-T,T]$. Thus, we have

$$\sup_{t\in U_\omega}r(t)\leq\left(\frac{3\pi}{2\pi(3\Gamma+\alpha)}\right)^{1/2}\left(\frac{1}{L^{3/2}}\right)^{1/2}\leq\left(\frac{3}{2\Gamma}\right)^{1/2}\frac{1}{L^{3/4}}$$

and

$$\|r\|_{U_{\omega}^{\sim}} \leq \left(\frac{3\pi}{2\pi(3\Gamma+\alpha)} \right)^{1/2} \left(\frac{\Gamma+\alpha}{T} \right)^{1/2} \leq \left(\frac{3}{2T} \right)^{1/2}.$$

Substituting these estimates into (12), using the definitions of w and $|U_\omega|$, and setting $f(t)=r(t)e^{i\theta(t)}$, where θ is a primitive of w, we obtain

$$||F(\omega)|-V(\omega)| \leq -2(\omega+\Gamma+\alpha)\sqrt{\frac{3}{2\Gamma}}\, L^{1/4} + \frac{32\sqrt{3/2T}\, L^{3/2}}{|U_\omega|};$$

and so (11) follows by means of (13). Q.E.D.

Example.

a. A construction similar to the Proposition can be made to estimate $V(\omega)$ for $\omega \in [-(\Gamma+\alpha), \Gamma+\alpha]$ by means of functions $f \leftrightarrow F$ for which $\operatorname{supp} f \subseteq [-T, T]$. However, when we deal with the neighborhoods used in the Proposition, the parameter Γ must be chosen large enough to ensure that $r(t_\omega)C(U_\omega, w)$ is small. Quantitatively this comes down to a choice of Γ for which $1/\sqrt{\Gamma} < \varepsilon$; and such bounds are inadequate since the height of V is of the order $1/\sqrt{\Gamma}$. Thus, any effective estimation of V on $[-(\Gamma+\alpha), \Gamma+\alpha]$ must render $r(t_\omega)C(U_\omega, w)$ small independently of taking $\sqrt{\Gamma}$ too large. To this end we give the following calculation in part b, and use this calculation in part c to provide an analogue of the Proposition for the case $\omega \in [-(\Gamma+\alpha), \Gamma+\alpha]$.

b. Since $\int e^{ic^2t^2}\,dt = (1/c)(\sqrt{\pi/2} + i\sqrt{\pi/2})$, $c>0$, we have $\int_0^\infty \cos(ct)^2\,dt = (1/2c)\sqrt{\pi/2}$ Also $\cos(ct)^2=0$ at each $t_n=(1/c)\sqrt{\pi(2n+1)/2}$, $n=0,1,\cdots$. As such we approximate the area beneath or above $\cos(ct)^2$ by $\pm(1/2c)[\sqrt{\pi(2n+3)/2} - \sqrt{\pi(2n+1)/2}]$, respectively. Designating this "triangulation" of $\cos(ct)^2$ by $\cos\Delta(ct)^2$, we compute

$$\int_0^\infty \cos(ct)^2 dt = \frac{1}{s}\int_0^\infty \cos\Delta(ct)^2 dt,$$

where $s=\sqrt{1}-(\sqrt{3}-\sqrt{1})+(\sqrt{5}-\sqrt{3})-(\sqrt{7}-\sqrt{5})+\cdots$. For each N we make the estimate

$$(14)\quad \left| \int_0^{(1/c)\sqrt{(4N-1)\pi/2}} \cos\Delta(ct)^2 dt - \frac{s}{2c}\sqrt{\frac{\pi}{2}} \right|$$
$$= \frac{1}{2c}\sqrt{\frac{\pi}{2}} \sum_{n=N}^{\infty} \left[\left(\sqrt{4n+1}-\sqrt{4n-1}\right) - \left(\sqrt{4n+3}-\sqrt{4n+1}\right) \right] \approx \frac{K}{c\sqrt{N}}$$

by the mean-value theorem. Thus, *we essentially have that $C(U_\omega, w)$ is of order $1/(c\sqrt{N})$ when $|U_\omega|$ is of order $\sqrt{N}/c$ recalling that c is of order $\sqrt{w'(t_\omega)}$.*

c. In order to implement (14) for estimates of the form (11) in the case $\omega \in (-(\Gamma+\alpha), \Gamma+\alpha)$, we proceed as follows. For a given ε and ω we choose N and w so that $1/\sqrt{N} < \varepsilon$, $w(0)=\omega$ and $w' \approx N^a$ in a neighborhood of 0. In particular, because of (14) and the definition of $r(t_\omega)$ we can expect $r(t_\omega)C(U_\omega, w)$ to be less than ε if $|U_\omega| \approx N^{(1-a)/2}$. The remaining terms of (1) can be minimized by more careful treatment of the remainder R in (8) and by a proper choice of a, respectively.

Remark.

a. The terms of the bound (1) in the Theorem correspond to the techniques used in the proof: stationary phase, Fresnel integrals, and Vakman's construction of sophisticated signals, e.g., Vakman [7, §31]. Vakman's construction plays a role in our Proposition and relates time and frequency in a fundamental way just as the Wigner transform and the method of Fefferman and Phong [4] do.

b. The technique in the Proposition, of constructing custom-made bijections w, and the pointwise estimation in the Theorem and the Proposition can be adapted to other situations where different norms might be required, cf., the last part of the Example. For instance, the deconvolution method of [2] sometimes requires that estimators F be flat as well as small in certain neighborhoods. In fact, our results provide one illustration of a general phenomenon: the uncertainty principle, or poor global approximation, does not preclude good local approximation for many norms.

4. The spectrum estimation problem. The spectrum estimation problem is to clarify and quantify the statement: find periodicities in a signal x recorded over the time interval $[-T,T]$.

We assume the following mathematical model, cf. [2].

Assumptions. 1. The signal x is actually defined on the product space $[-T,T]\times P$, where P is a probability space and x is the restriction to $[-T,T]\times P$ of some stationary stochastic process y.

2. The expectation E_f of the periodogram

$$S_f(\omega,\alpha)=\left|\int x(t,\alpha)f(t)e^{-it\omega}\,dt\right|^2$$

is known, where $\operatorname{supp} f\subseteq[-T,T]$.

3. The power spectrum S of x is uniquely determined. This assumption is a theorem in case we make the experimentally reasonable hypothesis that the power spectrum S_y of every stationary extension y of x is compactly supported [2, §III].

Remark. a. Assumption 3 is not universally accepted. In fact, the maximum entropy method of spectrum estimation is essentially opposite the point of view that S is uniquely determined. In maximum entropy the power spectrum is modeled to maximize a certain entropy integral.

b. Besides maximum entropy there are also classical windowing methods of spectrum estimation. For example, the Bartlett–Tukey method produces an asymptotically unbiased estimator E_f by choosing the data window f as an approximate identity.

Our method of spectrum estimation, which is conceptually different than maximum entropy and classical windowing, depends on the Theorem. The discrete part of the measure S reflects periodicities in x, and our task is to formulate an estimate $\tilde{S}$ of S in terms of the incomplete data domain $[-T,T]\times P$.

Method. a. By Assumption 2 and Assumption 3, we have $E_f=S*F^2$ where S is uniquely determined, $f\leftrightarrow F$, and $\operatorname{supp} f\subseteq[-T,T]$.

b. If we estimate E_f by $S*(HF^2)$, where H is the Heaviside function, and sample $S*(HF^2)$ at frequency intervals c, and, finally, deconvolve by means of $(HF^2)^{-1}$, then we are motivated to define the *f-estimator*,

$$S_f=\frac{1}{F(0)^2}\sum_{n=0}^{\infty}a_n\delta'_{nc}*E$$

where

$$a_0=1, \quad a_n=1-\left(\frac{1}{F(0)^2}\right)\sum_0^{n-1} a_m F((n-m)c)^2.$$

The fundamental theoretical source of error between S_f and S arises from considering HF^2 instead of F^2. In this regard and considering the results of §2 and §3, note that $|F'(\omega)|$ is always bounded by $\sqrt{2/3}\, T^{3/2}\|f\|_2$ when $\operatorname{supp} f \subseteq [-T,T]$. On the other hand, the Theorem allows us to construct F so that HF^2 is *as close as possible* to F^2 in specified regions. Thus, if two close peaks in S can be resolved, this algorithm should do it.

Acknowledgment. I would like to thank the referees for several valuable suggestions.

REFERENCES

[1] J. Benedetto, *Spectral Synthesis*, Academic Press, NY., 1975.

[2] ———, *Harmonic analysis and spectral estimation*, J. Math. Anal. Appl., 91 (1983), pp. 444–509.

[3] ———, *Wiener's Tauberian theorem and the uncertainty principle*, Proc. Topics in Modern Harmonic Analysis, Torino-Milano, 1982.

[4] C. Fefferman and D. Phong, *The uncertainty principle and sharp Gårding inequalities*, Comm. Pure and Applied Math., 34 (1981), pp. 285–331.

[5] W. Fuchs, *On the magnitude of Fourier transforms*, International Congress Math., Amsterdam 2 (1954), pp. 106–107.

[6] H. Landau, H. Pollak, and D. Slepian, *Prolate spheroidal wave functions, Fourier analysis, and uncertainty*, I-V, Bell System Tech. J., 40 (1961), pp. 43–64; 40 (1961), pp. 65–84; 41 (1962), pp. 1295–1336; 43 (1964), pp. 3009–3058; 57 (1978), pp. 1371–1430.

[7] D. E. Vakman, *Sophisticated Signals and the Uncertainty Principle in Radar*, Springer-Verlag, New York, 1968.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics
012

DENSE SETS AND FAR FIELD PATTERNS IN ACOUSTIC WAVE PROPAGATION*

DAVID COLTON† AND ANDREAS KIRSCH‡

Abstract. We consider the Dirichlet, Neumann, and transmission boundary value problems corresponding to the scattering of an entire, time harmonic acoustic wave by a bounded obstacle in the plane. We first construct sets of solutions to these problems such that the restrictions of these solutions to the boundary $\partial\Omega$ of the scattering obstacle are dense in $L^2(\partial\Omega)$. These results are then used to determine when the class of far field patterns corresponding to each of these scattering problems is dense or not dense in $L^2[0, 2\pi]$.

1. Introduction. A basic problem in inverse scattering theory for acoustic waves is the classification of far field patterns corresponding to the scattering of a time harmonic incident wave by a bounded, connected obstacle. It is easily verifiable that the class of functions that can be far field patterns is a subset of the class of entire functions [3]. It has also been established that this subset can be further characterized by using the theory of entire functions of exponential type and the corresponding properties of the indicator diagram of such functions [1], [7], [10]. However these results make no use of the boundary conditions satisfied by the total field on the boundary of the scattering obstacle and hence such results can only be characterized as "necessary" rather than "sufficient" conditions for a function to be a far field pattern. In an initial attempt to remedy this defect, Colton [2] has examined the class of far field patterns corresponding to entire incident fields subject to an impedance boundary condition on a bounded, connected obstacle in the plane and has shown that this class is dense in $L^2[0, 2\pi]$. Strangely enough, however, it was shown that such a result is not true in general for the Dirichlet and Neumann problems. In particular it was shown that for the unit disk subject to a Dirichlet or Neumann boundary condition the class of far field patterns is not dense in $L^2[0, 2\pi]$ if the square of the wave number is an eigenvalue of the corresponding interior problem. We use the word "strange" to describe this phenomenon since from physical considerations the behavior of solutions to the interior problem should have nothing to do with the exterior scattering problem, and indeed such considerations have played an important motivating effect on much of the recent work on integral equation methods in scattering theory (cf. [3]). The purpose of this paper is to further examine this phenomena and in particular to answer the following questions:

(1) Is the fact that the far field patterns of the Dirichlet and Neumann problems are not dense at interior eigenvalues true for domains other than the unit disk?

(2) Is the fact that the far field patterns of the impedance boundary value problem are dense true for the transmission boundary value problem (for which the impedance boundary value problem is an approximation)?

The answer to the first question is that the far field patterns for the Dirichlet and Neumann boundary value problems are not dense at the square of an interior eigenvalue if and only if one of the corresponding eigenfunctions is an entire function of a certain type. We then use this result to exhibit a domain for which, in contrast to the case of the unit disk, the far field patterns corresponding to Dirichlet boundary

*Received by the editors September 7, 1982, and in final form April 29, 1983.

† 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.

conditions are dense at an interior eigenvalue. The answer to the second question is that the far field patterns corresponding to the transmission boundary value problem are also not dense if the exterior and interior wave number are related in an appropriate manner, and we shall give sufficient conditions for this to be the case. These results are based on first determining dense sets of solutions in $L^2(\partial\Omega)$ and $L^2(\partial\Omega)\times L^2(\partial\Omega)$ where $\partial\Omega$ is the boundary of the scattering obstacle and then using these results to examine the far field patterns of the boundary value problems under consideration.

Although all of our analysis is done in the plane $\mathbb{R}^2$, our results can easily be generalized to $\mathbb{R}^n$ for an arbitrary integer n. Furthermore, although we shall assume that the boundary of the scattering obstacle is in class C^2, all of our results for the Dirichlet problem remain valid for domains whose boundaries have corners but no cusps. This can be established by using the ideas of Ruland [9] at appropriate points in the proofs of our theorems.

2. Dense sets in $L^2(\partial\Omega)$ and $L^2(\partial\Omega)\times L^2(\partial\Omega)$. Let Ω be a bounded, connected domain in the plane containing the origin with C^2 boundary $\partial\Omega$ having unit outward normal ν, $\Omega_e=\mathbb{R}^2\setminus\overline{\Omega}$, and u^i (the "incident wave") an entire solution of the Helmholtz equation

$$\Delta_2 u+k^2u=0, \tag{2.1}$$

where $k>0$ denotes the wave number. In this paper we shall be considering the following Dirichlet, Neumann, and transmission boundary value problems associated with solutions of (2.1) defined in exterior domains.

Dirichlet problem. Determine a solution $u=u^i+u^s$ of (2.1) in Ω_e such that $u\in C^2(\Omega_e)\cap C(\overline{\Omega}_e)$,

$$u=0 \quad \text{on } \partial\Omega, \tag{2.2}$$

and u^s (the "scattered wave") satisfies the Sommerfeld radiation condition

$$\lim_{r\to\infty} r^{1/2}\left(\frac{\partial u^s}{\partial r}-iku^s\right)=0 \tag{2.3}$$

uniformly with respect to θ where (r,θ) denote polar coordinates.

Neumann problem. Determine a solution $u=u^i+u^s$ of (2.1) in Ω_e such that $u\in C^2(\Omega_e)\cap C^1(\overline{\Omega}_e)$,

$$\frac{\partial u}{\partial\nu}=0 \quad \text{on } \partial\Omega, \tag{2.4}$$

and u^s satisfies the Sommerfeld radiation condition.

Transmission problem. Determine a solution $u_e=u^i+u^s$ of (2.1) in Ω_e and a solution u_i of

$$\Delta_2 u+\kappa^2u=0 \tag{2.5}$$

in Ω such that $u_e\in C^2(\Omega_e)\cap C^1(\overline{\Omega}_e)$, $u_i\in C^2(\Omega)\cap C^1(\overline{\Omega})$,

$$\mu_e u_e-\mu_i u_i=0, \tag{2.6a}$$

$$\text{on } \partial\Omega,$$

$$\frac{\partial u_e}{\partial\nu}-\frac{\partial u_i}{\partial\nu}=0, \tag{2.6b}$$

and u^s satisfies the Sommerfeld radiation condition, where κ,μ_e, and μ_i are positive constants.

The existence and uniqueness of solutions to the above boundary value problems is well known [3]. Our aim in this section is to determine sets of solutions to (2.1) and (2.5) such that the restriction to the boundary of these solutions form dense sets in $L^2(\partial\Omega)$ or $L^2(\partial\Omega)\times L^2(\partial\Omega)$. We begin by defining the following sets:

$$H(k,\Omega_e)=\{u: u\in C^2(\Omega_e)\cap C^1(\overline{\Omega}_e),\ u \text{ satisfies (2.1) and (2.3) in } \Omega_e\},$$

$$A(k,\mathbb{R}^2)=\{u: u(x)=\textstyle\int_0^{2\pi} g(\theta)e^{ikx\cdot\hat{y}}\,d\theta,\ x\in\mathbb{R}^2, \hat{y}=(\cos\theta,\sin\theta),\ g\in L^2[0,2\pi]\},$$

$$H^2(k,\partial\Omega)=\{(u,\partial u/\partial\nu): u\in H(k,\Omega_e),\ x\in\partial\Omega\},$$

$$A^2_{\mu_i,\mu_e}(k,\partial\Omega)=\{(\mu_i u,\mu_e\partial u/\partial\nu): u\in A(k,\mathbb{R}^2),\ x\in\partial\Omega\},$$

$$T_D(k,\Omega_e)=\{u: u=u^i+u^s, u^i\in A(k,\mathbb{R}^2),\ u^s\in H(k,\Omega_e),\ u=0 \text{ on } \partial\Omega\},$$

$$T_N(k,\Omega_e)=\{u: u=u^i+u^s, u^i\in A(k,\mathbb{R}^2),\ u^s\in H(k,\Omega_e),\ \partial u/\partial\nu=0 \text{ on } \partial\Omega\}.$$

We can now state and prove our first theorem.

THEOREM 1. (a) $\partial T_D(k,\Omega_e)/\partial\nu|_{\partial\Omega}$ *is dense in* $L^2(\partial\Omega)$.

(b) $T_N(k,\Omega_e)|_{\partial\Omega}$ *is dense in* $L^2(\partial\Omega)$.

Proof. We first consider part (a). Let $g\in L^2(\partial\Omega)$ be such that

$$\int_{\partial\Omega}\bar{g}\frac{\partial u}{\partial\nu}\,ds=0 \tag{2.7}$$

for every $u\in T_D(k,\Omega_e)$. The result will follow if we can show that this implies that g is identically zero. Let u be an arbitrary element of $T_D(k,\Omega_e)$. Then from Green's formula we have

$$2u(x)=2u^i(x)-\int_{\partial\Omega}\gamma(x,y)\frac{\partial u(y)}{\partial\nu}\,ds(y) \tag{2.8}$$

for $x\in\Omega_e$ where

$$\gamma(x,y)=\frac{i}{2}H_0(k|x-y|) \tag{2.9}$$

with H_0 denoting a Hankel function of the first kind of order zero and $u^i\in A(k,\mathbb{R}^2)$. (2.8) implies that

$$\int_{\partial\Omega}\gamma(x,y)\frac{\partial u(y)}{\partial\nu}\,ds(y)=2u^i(x),\qquad x\in\partial\Omega, \tag{2.10}$$

and

$$\frac{\partial u(x)}{\partial\nu}+\int_{\partial\Omega}\frac{\partial u(y)}{\partial\nu}\frac{\partial}{\partial\nu(x)}\gamma(x,y)\,ds(y)=2\frac{\partial u^i(x)}{\partial\nu},\qquad x\in\partial\Omega. \tag{2.11}$$

We now define the operators $\mathbf{S}, \mathbf{D}$ and $\mathbf{D}^*$ mapping $L^2(\partial\Omega)$ into itself by

$$\begin{aligned}\mathbf{S}\phi(x)&:=\int_{\partial\Omega}\phi(y)\gamma(x,y)\,ds(y),\qquad x\in\partial\Omega,\\ \mathbf{D}\phi(x)&:=\int_{\partial\Omega}\phi(y)\frac{\partial}{\partial\nu(y)}\gamma(x,y)\,ds(y),\qquad x\in\partial\Omega,\\ \mathbf{D}^*\phi(x)&:=\int_{\partial\Omega}\phi(y)\frac{\partial}{\partial\nu(x)}\gamma(x,y)\,ds(y),\qquad x\in\partial\Omega.\end{aligned} \tag{2.12}$$

Note that from potential theoretic arguments it can be easily verified that $\mathbf{I}+\mathbf{D}^*+i\mathbf{S}$ is invertible (cf. [3]) and $\mathbf{D}^*$ is the adjoint of $\mathbf{D}$ with respect to the pairing

$$\langle\phi,\psi\rangle:=\int_{\partial\Omega}\phi\psi\,ds. \tag{2.13}$$

From (2.10), (2.11), we now have that

$$(\mathbf{I}+\mathbf{D}^*+i\mathbf{S})\frac{\partial u(x)}{\partial\nu}=2\left(\frac{\partial u^i}{\partial\nu}+iu(x)\right),\qquad x\in\partial\Omega, \tag{2.14}$$

and hence

$$\frac{\partial u(x)}{\partial\nu}=2(\mathbf{I}+\mathbf{D}^*+i\mathbf{S})^{-1}\left(\frac{\partial u^i(x)}{\partial\nu}+iu^i\right),\qquad x\in\partial\Omega. \tag{2.15}$$

We can now conclude from (2.7) and (2.15) that

$$\begin{aligned}0=\left\langle\bar{g},\frac{\partial u}{\partial\nu}\right\rangle&=2\left\langle\bar{g},(\mathbf{I}+\mathbf{D}^*+i\mathbf{S})^{-1}\left(\frac{\partial u^i}{\partial\nu}+iu^i\right)\right\rangle\\&=2\left\langle(\mathbf{I}+\mathbf{D}+i\mathbf{S})^{-1}\bar{g},\left(\frac{\partial u^i}{\partial\nu}+iu^i\right)\right\rangle.\end{aligned} \tag{2.16}$$

But u^i can be an arbitrary element of $A(k,\mathbb{R}^2)$ and from the Jacobi–Anger expansion

$$e^{ir\cos\theta}=\sum_{n=-\infty}^{\infty}i^nJ_n(r)e^{in\theta} \tag{2.17}$$

where J_n denotes a Bessel function of order n, we can conclude that $J_n(kr)\cos n\theta$ and $J_n(kr)\sin n\theta$ are elements of $A(k,\mathbb{R}^2)$. Hence, from the results of [2], we can conclude that $(\mathbf{I}+\mathbf{D}+i\mathbf{S})^{-1}\bar{g}=0$ and hence $g=0$. The proof of part (a) is now complete.

We now consider part (b) of the theorem. Let $u\in T_N(k,\Omega_e)$. Then, from Green's formula, we have that

$$2u(x)=2u^i(x)+\int_{\partial\Omega}u(y)\frac{\partial}{\partial\nu(y)}\gamma(x,y)\,ds(y),\qquad x\in\Omega_e, \tag{2.18}$$

and hence

$$u(x)-\int_{\partial\Omega}u(y)\frac{\partial}{\partial\nu(y)}\gamma(x,y)\,ds(y)=2u^i(x),\qquad x\in\partial\Omega \tag{2.19}$$

and

$$-\frac{\partial}{\partial\nu}\int_{\partial\Omega}u(y)\frac{\partial}{\partial\nu(y)}\gamma(x,y)\,ds(y)=2\frac{\partial u^i(x)}{\partial\nu},\qquad x\in\partial\Omega. \tag{2.20}$$

We now define the operator $\mathbf{D}_\nu$ on the Sobolev space $H^1(\partial\Omega)$ by

$$\mathbf{D}_\nu\phi(x):=\frac{\partial}{\partial\nu}\int_{\partial\Omega}\phi(y)\frac{\partial}{\partial\nu(y)}\gamma(x,y)\,ds(y),\qquad x\in\partial\Omega \tag{2.21}$$

and note that (2.19), (2.20) imply that

$$(2.22)\qquad (\mathbf{D}_\nu+i\mathbf{D}-i\mathbf{I})u(x)=-2\left(\frac{\partial u^i(x)}{\partial\nu}+iu^i(x)\right),\qquad x\in\partial\Omega.$$

From the results of Kirsch [6] we have that the operator $\mathbf{B}:=\mathbf{D}_\nu+i\mathbf{D}-i\mathbf{I}$ is an isomorphism from $H^1(\partial\Omega)$ onto $L^2(\partial\Omega)$ and hence we can write

$$(2.23)\qquad u(x)=-2\mathbf{B}^{-1}\left(\frac{\partial u^i(x)}{\partial\nu}+iu^i(x)\right).$$

Now suppose $g\in L^2(\partial\Omega)$ is such that $\langle\bar g,u\rangle=0$ for every $u\in T_N(k,\Omega_e)$. Then

$$(2.24)\qquad 0=\langle\bar g,u\rangle=-2\left\langle(\mathbf{B}^{-1})^*\bar g,\frac{\partial u^i}{\partial\nu}+iu^i\right\rangle$$

for every $u^i\in A(k,\mathbb{R}^2)$ where * denotes the adjoint with respect to the pairing given by (2.13). We can now conclude, as in part (a), that $g=0$ and this completes the proof of part (b) of the theorem.

Note. The proofs of [2, Thms. 1 and 2] are incomplete. In these theorems we should begin by assuming only that $g\in L^2(\partial D)$ and then use the results of Kersten [5] to show that $g=\mathbf{K}g$ where $\mathbf{K}$ has a weakly singular kernel. Then since $\mathbf{K}^n$ has a continuous kernel for n sufficiently large, we can conclude that $g\in C(\partial D)$.

We now establish a result for the transmission problem that is analogous to Theorem 1 for the Dirichlet and Neumann problems. To this end we consider the product space $L^2(\partial\Omega)\times L^2(\partial\Omega)$ equipped with the inner product

$$(2.25)\qquad (g_1,g_2):=\int_{\partial\Omega}\bar\phi_1\psi_1\,ds+\int_{\partial\Omega}\bar\phi_2\psi_2\,ds,$$

where $g_1=(\phi_1,\phi_2)$, $g_2=(\psi_1,\psi_2)$.

THEOREM 2. (a) $H^2(k,\partial\Omega)+A^2_{\mu_i,\mu_e}(\kappa,\partial\Omega)$ *is dense in* $L^2(\partial\Omega)\times L^2(\partial\Omega)$.

(b) $H^2(k,\partial\Omega)\cap A^2_{\mu_i,\mu_e}(\kappa,\partial\Omega)=\{(0,0)\}$.

Proof. We first prove part (a). Since $\psi_{1n}(r,\theta):=H_n(kr)\sin n\theta$ and $\psi_{2n}(r,\theta):=H_n(kr)\cos n\theta$, where H_n denotes a Hankel function of the first kind of order n, are both in $H(k,\Omega_e)$, it suffices to show that the relations

$$(2.26\text{a})\qquad \int_{\partial\Omega}\left\{g\psi_{in}+f\frac{\partial\psi_{in}}{\partial\nu}\right\}ds=0,$$

$$(2.26\text{b})\qquad \int_{\partial\Omega}\left\{\mu_i g\gamma_{in}+\mu_e f\frac{\partial\gamma_{in}}{\partial\nu}\right\}ds=0,$$

$i=1,2$, $n=0,1,2,\cdots$, for $f,g\in L^2(\partial\Omega)$ and $\gamma_{1n}(r,\theta):=J_n(\kappa r)\sin n\theta$, $\gamma_{2n}(r,\theta):=J_n(\kappa r)\cos n\theta$, imply that f and g are identically zero (note that we have previously shown that γ_{1n} and γ_{2n} are in $A(\kappa,\mathbb{R}^2)$). To show this we follow the arguments of Colton and Kress [4]. From the addition formula

$$(2.27)\qquad H_0(k|x-\xi|)=H_0(kr_x)J_0(kr_\xi)+2\sum_{n=1}^{\infty}H_n(kr_x)J_n(kr_\xi)\cos n(\theta_x-\theta_\xi),$$

where $r_\xi < r_x$ and (r_x, θ_x), (r_ξ, θ_ξ) are the polar coordinates of x and ξ respectively, we can conclude from (2.26) that the functions w and v defined by

$$w(x) = \int_{\partial\Omega} \left\{ g(y)\gamma_k(x,y) + f(y) \frac{\partial \gamma_k(x,y)}{\partial \nu(y)} \right\} ds(y), \tag{2.28a}$$

$$v(x) = \int_{\partial\Omega} \left\{ \mu_i g(y)\gamma_\kappa(x,y) + \mu_e f(y) \frac{\partial \gamma_\kappa(x,y)}{\partial \nu(y)} \right\} ds(y), \tag{2.28b}$$

are identically zero in Ω and Ω_e respectively, where the subscript on the fundamental solution γ denotes its dependency on the wave number. We note that w is a solution of (2.1) in $\mathbb{R}^2 \setminus \partial\Omega$ and v is a solution of (2.5) in $\mathbb{R}^2 \setminus \partial\Omega$. From the continuity properties of single and double layer potentials [5], [8] we have that

$$w_+ = 2f, \qquad \left(\frac{\partial w}{\partial \nu} \right)_+ = -2g, \tag{2.29}$$

where the plus subscript denotes the limit as x tends to $\partial\Omega$ from Ω_e and

$$v_- = -2\mu_e f, \qquad \left(\frac{\partial v}{\partial \nu} \right)_- = 2\mu_i g, \tag{2.30}$$

where the minus subscript denotes the limit as x tends to $\partial\Omega$ from Ω. By considering appropriate linear combinations of w_-, v_+ and $(\partial w/\partial \nu)_-$, $(\partial v/\partial \nu)_+$ we can conclude that $f, g \in C(\partial D)$ (cf. the note following the proof of Theorem 1). Hence if we define $u := -\mu_i w$, then

$$\mu_e u_+ - \mu_i v_- = 0, \qquad \left(\frac{\partial u}{\partial \nu} \right)_+ - \left(\frac{\partial v}{\partial \nu} \right)_- = 0, \tag{2.31}$$

i.e. u defined in Ω_e and v defined in Ω satisfy homogeneous transmission boundary conditions and hence are both identically zero in Ω_e and Ω, respectively [3]. We can now conclude from either (2.29) or (2.30) that f and g are both zero. This completes the proof of part (a) of the theorem.

To prove part (b), assume (ϕ, ψ) is in $H^2(k, \partial\Omega) \cap A^2_{\mu_i, \mu_e}(\kappa, \partial\Omega)$. Then there exist a function $u \in H^2(k, \partial\Omega)$ and a function $v \in A(\kappa, \mathbb{R}^2)$ such that $\phi = u = \mu_i v$ and $\psi = \partial u / \partial \nu = \mu_e \partial v / \partial \nu$ on $\partial\Omega$. That is u and $\mu_e v$ satisfy homogeneous transmission boundary conditions and hence are both identically zero in Ω_e and Ω respectively, i.e. $(\phi, \psi) = (0, 0)$. The proof of the theorem is now complete.

Note. We have actually proved Theorem 2 for the more general case when, in the definitions of $H^2(k, \partial\Omega)$ and $A^2_{\mu_i, \mu_e}(\kappa, \partial\Omega)$, we replace $H(k, \Omega_e)$ by span $\{\psi_{1n}, \psi_{2n}, n \in \mathbb{N}\}$ and $A(\kappa, \mathbb{R}^2)$ by span $\{\gamma_{1n}, \gamma_{2n}, n \in \mathbb{N}\}$ respectively.

We note that the proof of Theorem 2 provides a method for approximating the solution of the transmission boundary value problem by means of a complete family of solutions. In particular, if u^s is the scattered wave in Ω_e we represent u^s as a finite linear combination of the functions ψ_{1n} and ψ_{2n}, u_i as a finite linear combination of the functions γ_{1n} and γ_{2n}, and consider the corresponding boundary data in the sets $\mu_e H^2(k, \partial\Omega)$ and $-A^2_{\mu_i, \mu_e}(\kappa, \partial\Omega)$ respectively. The unknown coefficients can now be found by determining a best approximation to the boundary data with respect to the norm induced by the inner product defined by (2.25).

3. Far field patterns. We now wish to use the results of the previous section to investigate the far field patterns corresponding to the Dirichlet, Neumann, and transmission boundary value problems. We first define more precisely what we mean by the far field pattern for these problems. From Green's formula we have that for $x \in \Omega_e$

$$u(x)=u^i(x)+\frac{1}{2}\int_{\partial\Omega}\left\{u(y)\frac{\partial}{\partial\nu(y)}\gamma(x,y)-\frac{\partial u(y)}{\partial\nu}\gamma(x,y)\right\}ds(y) \tag{3.1}$$

($u=u^e$ in the case of the transmission boundary value problem) and hence from the asymptotic behavior of Hankel's function we see from (2.9) and (3.1) that

$$u^s(x)=\frac{i}{4}e^{i(kr+\pi/4)}\left(\frac{2}{\pi kr}\right)^{1/2}F(\theta;k)+O\left(\frac{1}{r^{3/2}}\right), \tag{3.2}$$

where

$$F(\theta;k)=\int_{\partial\Omega}\left\{u(y)\frac{\partial}{\partial\nu(y)}e^{-ik\hat{x}\cdot y}-\frac{\partial u(y)}{\partial\nu}e^{-ik\hat{x}\cdot y}\right\}ds(y), \tag{3.3}$$
$$\hat{x}=(\cos\theta,\sin\theta).$$

The function F is known as the far field pattern corresponding to the scattered wave of the boundary value problem under consideration. Our aim in this section is to determine under what conditions the set of far field patterns corresponding to entire incident waves is (or is not) dense in $L^2[0,2\pi]$ where by "entire incident wave" we mean a solution of the Helmholtz equation defined in all of $\mathbb{R}^2$. As pointed out in the Introduction for the impedance boundary value problem this set is dense in $L^2[0,2\pi]$ for any positive value of the wave number [2]. The following examples show that this is not true for the Dirichlet and transmission problems. (The example for the Dirichlet problem can easily be modified to cover the case of the Neumann problem—cf. [2].)

Example 1. Consider the Dirichlet problem when Ω is the unit disk. Then since u^i is an entire solution of the Helmholtz equation we can expand u^i in the form

$$u^i(r,\theta)=\sum_{n=0}^{\infty}J_n(kr)[a_n\cos n\theta+b_n\sin n\theta], \tag{3.4}$$

where the series (3.4) is uniformly convergent on any compact subset of $\mathbb{R}^2$. Then for $r \geq 1$ we can expand u^s in the uniformly convergent series

$$u^s(r,\theta)=-\sum_{n=0}^{\infty}H_n(kr)\frac{J_n(k)}{H_n(k)}[a_n\cos n\theta+b_n\sin n\theta]. \tag{3.5}$$

From (3.5) and the asymptotic behavior of Hankel's function we see that the far field pattern of u^s is given by

$$F(\theta;k)=4i\sum_{n=0}^{\infty}\frac{(-i)^nJ_n(k)}{H_n(k)}[a_n\cos n\theta+b_n\sin n\theta]. \tag{3.6}$$

If k_0^2 is an eigenvalue of the interior Dirichlet problem then $J_n(k_0)=0$ for some integer n_0, and hence in this case $F(\theta;k_0)$ is orthogonal to $\cos n_0\theta$ and $\sin n_0\theta$ for all incident fields u^i. Hence the class of far field patterns for such values of k is not dense in $L^2[0,2\pi]$.

Example 2. Consider the transmission problem when Ω is the unit disk. Then u^i can again be expanded in the form (3.4) and since $u_i \in C^2(\Omega) \cap C^1(\bar{\Omega})$ we can expand u_i in the form

$$u_i(r,\theta) = \sum_{n=0}^{\infty} J_n(\kappa r)[c_n \cos n\theta + d_n \sin n\theta], \tag{3.7}$$

where the series is convergent in $\bar{\Omega}$ and uniformly convergent on compact subsets of Ω. Suppose k and κ are related in such a manner that $J_{n_0}(k) = J_{n_0}(\kappa) = 0$ for some integer n_0, i.e. k and κ are distinct zeros of the Bessel functions $J_{n_0}(r)$. Then representing the scattered wave in the form

$$u^s(r,\theta) = \sum_{n=0}^{\infty} H_n(kr)[f_n \cos n\theta + g_n \sin n\theta], \tag{3.8}$$

we see from the transmission boundary conditions that the unknown coefficients c_n and f_n satisfy the algebraic system

$$\begin{aligned} \mu_e f_n H_n(k) - \mu_i c_n J_n(\kappa) &= -\mu_e a_n J_n(k), \\ k f_n H_n'(k) - \kappa c_n J_n'(\kappa) &= -k a_n J_n'(k), \end{aligned} \tag{3.9}$$

with a similar system being satisfied by d_n and g_n. If $n = n_0$, we can immediately see that $f_{n_0} = 0$ (and $g_{n_0} = 0$). Hence, as in Example 1, we can conclude that in this case the far field pattern is orthogonal to $\cos n_0\theta$ and $\sin n_0\theta$ for all incident fields u^i and hence the class of far field patterns for such values of k and κ is not dense in $L^2[0,2\pi]$.

We shall now establish necessary and sufficient conditions for the far field patterns of the Dirichlet and Neumann problems to be dense in $L^2[0,2\pi]$ for arbitrary domains and sufficient conditions for the far field patterns of the transmission problem not to be dense. To this end we first define the following mappings:

(1) $\mathbf{G}: A(k,\mathbb{R}^2) \to L^2[0,2\pi]$ by $g = \mathbf{G}u$ where $u(x) = \int_0^{2\pi} g(\theta) e^{ikx\cdot\hat{y}} d\theta$, $\hat{y} = (\cos\theta, \sin\theta)$, $x \in \mathbb{R}^2$;

(2) $\mathbf{F}_D: A(k,\mathbb{R}^2) \to L^2[0,2\pi]$ by $F(\theta;k) = \mathbf{F}_D u^i$ where F is the far field pattern of u^s for $u = u^i + u^s \in T_D(k,\Omega_e)$;

(3) $\mathbf{F}_N: A(k,\mathbb{R}^2) \to L^2[0,2\pi]$ by $F(\theta;k) = \mathbf{F}_N u^i$ where F is the far field pattern of u^s for $u = u^i + u^s \in T_N(k,\Omega_e)$;

(4) $\mathbf{F}_T: A(k,\mathbb{R}^2) \to L^2[0,2\pi]$ by $F(\theta;k) = \mathbf{F}_T u^i$ where F is the far field pattern of u^s for $u_e = u^i + u^s$ the solution of the transmission problem.

Let

$$E_D(k,\Omega) = \{u : u \in C^2(\Omega) \cap C(\bar{\Omega}), u \text{ is a solution of (2.1) in } \Omega \text{ and } u = 0 \text{ on } \partial\Omega\},$$

$$E_N(k,\Omega) = \left\{u : u \in C^2(\Omega) \cap C^1(\bar{\Omega}), u \text{ is a solution of (2.1) in } \Omega \text{ and } \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial\Omega\right\},$$

and denote the closure of a set $X \subset L^2[0,2\pi]$ by $\bar{X}$.

THEOREM 3. (a) $L^2[0,2\pi] = \mathbf{G}(E_D(k,\Omega) \cap A(k,\mathbb{R}^2)) \oplus \overline{F_D(A(k,\mathbb{R}^2))}$,

(b) $L^2[0,2\pi] = \mathbf{G}(E_N(k,\Omega) \cap A(k,\mathbb{R}^2)) \oplus \overline{F_N(A(k,\mathbb{R}^2))}$,

and the sums are orthogonal.

Proof. We shall only prove part (a) since the proof of part (b) is essentially the same. We first establish orthogonality. Let $u=u^i+u^s\in T_D(k,\Omega_e)$ and $v\in E_D(k,\Omega)\cap A(k,\mathbb{R}^2)$. By Green's formula we have

$$2u^s(x)=-\int_{\partial\Omega}\gamma(x,y)\frac{\partial u(y)}{\partial\nu}ds(y),\qquad x\in\Omega_e, \tag{3.10}$$

and hence

$$\mathbf{F}_D u^i(\theta;k)=-\int_{\partial\Omega}\frac{\partial u(y)}{\partial\nu}e^{-ik\hat{x}\cdot y}ds(y) \tag{3.11}$$

where $\hat{x}=(\cos\theta,\sin\theta)$. Therefore

$$\begin{aligned}\int_0^{2\pi}\overline{\mathbf{G}v(\theta)}\mathbf{F}_D u^i(\theta;k)\,d\theta&=-\int_{\partial\Omega}\frac{\partial u(y)}{\partial\nu}\int_0^{2\pi}\overline{g(\theta)}e^{-ik\hat{x}\cdot y}d\theta\,ds(y)\\&=-\int_{\partial\Omega}\frac{\partial u(y)}{\partial\nu}\overline{v(y)}\,ds(y)\\&=0.\end{aligned} \tag{3.12}$$

Now let $g\in L^2[0,2\pi]$ be such that

$$\int_0^{2\pi}\overline{g(\theta)}\mathbf{F}_D u^i(\theta;k)\,d\theta=0 \tag{3.13}$$

for every $u^i\in A(k,\mathbb{R}^2)$. Then from (3.12) we see that

$$\int_{\partial\Omega}\frac{\partial u(y)}{\partial\nu}\overline{v(y)}\,ds(y)=0 \tag{3.14}$$

for every $u\in T_D(k,\Omega_e)$ where

$$v(y)=\int_{\partial\Omega}g(\theta)e^{ik\hat{x}\cdot y}d\theta,\qquad y\in\mathbb{R}^2. \tag{3.15}$$

From Theorem 1 we can now conclude that $v=0$ on $\partial\Omega$, i.e. $v\in E_D(k,\Omega)\cap A(k,\mathbb{R}^2)$.

Since, from the Jacobi–Anger expansion (2.17), it follows that the operator $\mathbf{G}$ is invertible, we see from Theorem 3 that a necessary and sufficient condition for the far field patterns of the Dirichlet or Neumann problems to be dense in $L^2[0,2\pi]$ is that $E_D(k,\Omega)\cap A(k,\mathbb{R}^2)=\{0\}$ or $E_N(k,\Omega)\cap A(k,\mathbb{R}^2)=\{0\}$, i.e. the eigenfunctions are not elements of the set $A(k,\mathbb{R}^2)$. We now use this fact to exhibit a domain for which, in contrast to Example 1, the far field patterns of the Dirichlet problem are dense in $L^2[0,2\pi]$ at an eigenvalue. (In this connection see the last paragraph of the Introduction.)

Example 3. Let α and β, $\alpha<\beta$, be the first (real) zeros of the Neumann function $Y_1(r)$ and let

$$\Omega=\{(r,\theta):\alpha<r<\beta,0<\theta<\pi\}. \tag{3.16}$$

Then $k=1$ is an eigenvalue of the interior Dirichlet problem with eigenfunction

$$u(r,\theta)=Y_1(r)\sin\theta. \tag{3.17}$$

We shall show that $E_D(1,\Omega)\cap A(1,\mathbb{R}^2)=\{0\}$. Let $v\in E_D(1,\Omega)\cap A(1,\mathbb{R}^2)$. Then v has the expansion

$$v(r,\theta)=\sum_{n=-\infty}^{\infty} a_n J_n(r)e^{in\theta} \tag{3.18}$$

where the series is uniformly convergent on compact subsets of $\mathbb{R}^2$. Since $v(r,\theta)=0$ for $r=\alpha,\beta,0<\theta<\pi$, we have $a_nJ_n(\alpha)=a_nJ_n(\beta)=0$ for all integers n. But the first zero α of $Y_1(r)$ is less than any positive zero of $J_n(r)$ for all n and hence $a_n=0$ for all integers n, i.e. $v=0$.

We shall now conclude this paper by giving a sufficient condition for the far field patterns of the transmission problem to be not dense in $L^2[0,2\pi]$.

THEOREM 4. *The far field patterns of the transmission boundary value problem corresponding to $u^i\in A(k,\mathbb{R}^2)$ are not dense in $L^2[0,2\pi]$ if there exists a $u\in A(k,\mathbb{R}^2)$, u not identically zero, such that*

$$\begin{pmatrix}\partial u/\partial \nu\\ -u\end{pmatrix}\perp A^2_{\mu_i,\mu_e}(\kappa,\partial\Omega).$$

Proof. From (3.3) we see that

$$\mathbf{F}_T u^i(\theta;k)=\int_{\partial\Omega}\left\{u_e(y)\frac{\partial}{\partial\nu(y)}e^{-ik\hat{x}\cdot y}-\frac{\partial u_e(y)}{\partial\nu}e^{-ik\hat{x}\cdot y}\right\}ds(y), \tag{3.19}$$

where $\hat{x}=(\cos\theta,\sin\theta)$, $u_e=u^i+u^s$. Let $g\in L^2[0,2\pi]$ be such that

$$\int_0^{2\pi}\overline{g(\theta)}\mathbf{F}_T u^i(\theta;k)\,d\theta=0 \tag{3.20}$$

for every $u^i\in A(k,\mathbb{R}^2)$. We want to show that there exists a g that is not identically zero such that (3.20) is valid. To this end we note that (3.20) is equivalent to

(3.21)
$$0=\int_{\partial\Omega}u_e(y)\left[\int_0^{2\pi}\overline{g(\theta)}\frac{\partial}{\partial\nu(y)}e^{-ik\hat{x}\cdot y}d\theta\right]-\frac{\partial u_e(y)}{\partial\nu}\left[\int_0^{2\pi}\overline{g(\theta)}e^{-ik\hat{x}\cdot y}d\theta\right]ds(y)$$

and from the transmission boundary conditions (2.6) we see that (3.21) is equivalent to

$$0=\int_{\partial\Omega}\left\{\mu_i u_i(y)\overline{\frac{\partial u(y)}{\partial\nu}}-\mu_e\frac{\partial u_i(y)}{\partial\nu}\overline{u(y)}\right\}ds(y), \tag{3.22}$$

where

$$u(y)=\int_0^{2\pi}g(\theta)e^{ik\hat{x}\cdot y}d\theta. \tag{3.23}$$

By hypotheses there exists a $g\in L^2[0,2\pi]$, g not identically zero, such that

$$\begin{pmatrix}\partial u/\partial \nu\\ -u\end{pmatrix}\perp A^2_{\mu_i,\mu_e}(\kappa,\partial\Omega),$$

where u has the representaion (3.23). From Theorem 2, in particular the remarks made after the proof of this theorem, we see that for this g, (3.22) and hence (3.20) is valid for all $u^i\in A(k,\mathbb{R}^2)$. Hence the far field patterns are not dense in $L^2[0,2\pi]$.

The reader can easily verify that in the case of Example 2 we can choose u to be

$$u(r,\theta)=J_{n_0}(kr)\cos n_0\theta. \tag{3.24}$$

REFERENCES

[1] D. COLTON, *On the inverse scattering problem for axially symmetric solutions of the Helmholtz equation*, Quart. J. Math., 22 (1971), pp. 125–130.

[2] ______, *Runge's theorem and far field patterns for the impedance boundary value problem in acoustic wave propagation*, this Journal, 13 (1982), pp. 970–977.

[3] D. COLTON AND R. KRESS, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.

[4] ______, *The unique solvability of the null field equations of acoustics*, Quart. J. Mech. Appl. Math., 36 (1983), pp. 87–95.

[5] H. KERSTEN, *Grenz-und Sprungrelationen für Potentiale mit quadrat-summierbarer Flächenbelegung*, Result. d. Math., 3 (1982), pp. 17–24.

[6] A. KIRSCH, *The Robin problem for the Helmholtz equation as a singular perturbation problem*, to appear.

[7] C. MÜLLER, *Radiation patterns and radiation fields*, J. Rat. Mech. Anal., 4 (1955), pp. 235–246.

[8] C. MÜLLER AND H. KERSTEN, *Zwei Klassen vollstandiger Funktionensysteme zur Behandlung der Randwertaufgaben der Schwingungsgleichung* $\Delta U+k^2U=0$, Math. Meth. in the Appl. Sci., 2 (1980), pp. 48–67.

[9] C. RULAND, *Ein Verfahren zur Lösung von* $(\Delta+k^2)u=0$ *in Aussengebieten mit Ecken*, Applicable Analysis, 7 (1978), pp. 69–79.

[10] B. D. SLEEMAN, *The three-dimensional inverse scattering problem for the Helmholtz equation*, Proc. Camb. Phil. Soc., 73 (1973), pp. 477–488.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics

013

CONVERGENCE OF FOURIER SERIES AT A DISCONTINUITY*

RAY REDHEFFER[†]

Abstract. A new proof of convergence of Fourier series, due to Chernoff, is extended so that it applies at points of discontinuity. The argument leading to this extension is very short and leads to an interesting formula for the limit of certain asymmetric partial sums.

Introduction. Many, and perhaps most, of the functions one wants to expand in Fourier series have discontinuities; square waves and sawtooth waves are only two of the more obvious examples. The importance of allowing discontinuities is underlined by the fact that in Fourier analysis a function is considered to be continuous only if its periodic extension is continuous. This requires $f(0)=f(2\pi)$, a hypothesis which is often artificial and irrelevant. Another context in which discontinuities are important is in the summation of series. Some of the most interesting applications to series depend on the fact that the Fourier series for f converges to $[f(c+)+f(c-)]/2$ at points of simple discontinuity.

The purpose of this note is to deduce convergence at a discontinuity by means of a remarkable result that has been recently obtained by Chernoff [1]. Cutting through a tradition of about 150 years, Chernoff obtains pointwise convergence of Fourier series by a very brief argument which makes no use of the Dirichlet theory. The only fact from the traditional approach which is used is the Riemann-Lebesgue lemma, to the effect that the Fourier coefficients of an integrable function tend to zero. For the piecewise smooth functions common in applications this is trivial; just integrate by parts.

By a familiar formula involving the Dirichlet kernel $D_n(x)$, Chernoff extends his result to allow discontinuities. The chief novelty is that, instead of the evaluation

$$D_n(x)=e^{-inx}\sum_{k=0}^{2n} e^{ikx}=e^{-inx}\frac{e^{i(2n+1)x}-1}{e^{ix}-1}$$

which reduces to $\sin(n+\frac{1}{2}x)/\sin\frac{1}{2}x$, he uses only the obvious fact that $D_n(x)$ is even. Although it is of interest that one can avoid summation of a geometric series, the simplification here is perhaps not quite as dramatic as it is in the case of continuity; if one has the Dirichlet formula for partial sums together with

$$D_n(x)=\sin nx\cos\frac{1}{2}x+\cos nx\sin\frac{1}{2}x$$

and the Riemann–Lebesgue lemma, deduction of both Theorems 1 and 2 is no harder than Chernoff's proof of Theorem 1.

In this paper the convergence at a discontinuity is deduced directly from the first theorem of [1], without intervention of any part of the Dirichlet theory. As a dividend, we obtain a necessary and sufficient condition for convergence of the asymmetric partial sums, and a formula for the limit, which seem not to have been noted heretofore. The result is stated in Theorem 3.

*Received by the editors January 14, 1983, and in revised form May 23, 1983. This work was supported in part by the University of Karlsruhe under the auspices of the Deutsche Forschungsgemeinschaft.

[†] Department of Mathematics, University of California, Los Angeles, California 90024.

Throughout the sequel it is assumed that f has period 2π and belongs to $L(-\pi,\pi)$, where $L(a,b)$ denotes the class of functions integrable in the sense of Lebesgue on the interval $[a,b]$.

The theorem of Chernoff. With the notation

$$\hat{f}(k)=\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}dx, \qquad S_{m,n}(x)=\sum_{-m}^{n}\hat{f}(k)e^{ikx},$$

Chernoff's result is as follows:

THEOREM 1. *Let r be a complex number such that the function*

$$\psi(x)=\frac{f(x)-r}{x-c}$$

belongs to $L(c-\delta,c+\delta)$ for some $\delta>0$. Then $S_{m,n}(c)\to r$ when $m,n\to\infty$.

As pointed out in [1], this gives convergence at c if $f'(c)$ exists, or if the right and left hand derivatives exist at c, or if f satisfies a Lipschitz condition at c. But the hypothesis of Theorem 1 is never satisfied at a simple discontinuity.

Discontinuous functions. We shall establish the following:

THEOREM 2. *Suppose there exist complex numbers p and q such that the functions*

$$\phi(x)=\frac{f(x)-p}{x-c}, \qquad \theta(x)=\frac{f(x)-q}{x-c}$$

belong to $L(c-\delta,c)$ and to $L(c,c+\delta)$, respectively, where $\delta>0$. Then $S_{n,n}(c)\to(p+q)/2$ as $n\to\infty$.

For proof let us assume, without loss of generality, that $c=0$. Define a function h by $h(x)=p$ on $[-\pi,0)$ and $h(x)=q$ on $(0,\pi]$. The function $h(x)-(p+q)/2$ is odd; hence the symmetric partial sums of the Fourier series for $h(x)-(p+q)/2$ involve sine terms only and reduce to 0 when $x=0$. This shows that

$$\sum_{-n}^{n}\hat{h}(k)e^{ikc}=\frac{p+q}{2} \qquad (n\geq 1, c=0). \tag{1}$$

In other words, the conclusion of Theorem 2 holds for h, and indeed, in a particularly strong form. Since $f-h$ satisfies the hypothesis of Theorem 1 with $c=r=0$, the equation $f=(f-h)+h$ together with Theorem 1 and (1) gives $S_{n,n}(0)\to 0+(p+q)/2$ as $n\to\infty$. This completes the proof.

Discussion. As pointed out in [1], replacing c by 0 simplifies the calculations but is not essential. In the present setting, if no translation is made to give $c=0$, one requires a function h such that $h(c-)=p$, $h(c+)=q$, and such that the Fourier series for h is easily seen to converge to $(p+q)/2$ at c. Interestingly enough, the obvious choice $h(x)=p$ on $[-\pi,c)$, $h(x)=q$ on $(c,\pi]$ is not suitable. Proof that $S_{n,n}(c)\to(p+q)/2$ for this function is almost as hard as development of the entire Dirichlet theory!

For $c\in(-\pi,\pi)$ one should, instead, choose $h(x)$ to have the values

$$\frac{p+q}{2},\quad p,\quad q,\quad \frac{p+q}{2}$$

on the intervals $[-\pi,c-\delta)$, $(c-\delta,c)$, $(c,c+\delta)$, and $(c+\delta,\pi]$ respectively. Here δ is a positive number so small that the interval $(c-\delta,c+\delta)$ is interior to $(-\pi,\pi)$. By a short calculation

$$\hat{h}(0)=\frac{p+q}{2}, \qquad \hat{h}(k)e^{ikc}=\frac{i}{2\pi k}(p-q)(1-\cos k\delta),$$

and hence

$$\sum_{-n}^{n} \hat{h}(k)e^{ikc} = \frac{p+q}{2}, \qquad n \geq 1. \tag{2}$$

Theorem 2 follows as before, but without the preliminary transformation to make $c=0$.

Equation (2) can be checked by this transformation, however. Namely, extend $h(x)$ to have period 2π, and note that $h(x-c)-(p+q)/2$ is odd. This gives (2) by inspection.

Asymmetric partial sums. As pointed out in [1], at a simple discontinuity the symmetric partial sums $S_{n,n}(c)$ cannot be replaced by the asymmetric sums $S_{m,n}(c)$ used in Theorem 1. The above proof of Theorem 2 allows us to determine the precise degree of asymmetry that is permitted. Let us notice first that if two functions f_1 and f_2 satisfy the hypothesis of Theorem 2, with the same p, q, c, then their difference $f=f_1-f_2$ satisfies the hypothesis of Theorem 1 with $r=0$. This shows that, if any particular sequence of partial sums $S_{m,n}(c)$ for f_1 converges, as $m,n\to\infty$, then the same sequence will converge for f_2, and indeed, to the same value. In other words, the allowable sequences do not depend on the function being considered.

It follows that the sequences giving convergence are the same as for the particular function h used in the proof of Theorem 2. Thus we are led to consider the sum

$$S_{m,n}(c) = \frac{p+q}{2} + \frac{(p-q)i}{2\pi} \sum_{-m}^{n}{}' \left(\frac{1}{k} - \frac{\cos k\delta}{k} \right)$$

where $\delta=\pi$ if $c=0$ and the function leading to (1) is used instead. (Here the $'$ on the sum means that the term for $k=0$ is omitted.) Let $n>m$, as can be assumed without loss of generality. Since the summand is an odd function of k, the sum reduces to

$$E = \sum_{m+1}^{n} \left(\frac{1}{k} - \frac{\cos k\delta}{k} \right).$$

The series with general term $(\cos k\delta)/k$ converges by Abel's test (or it is an alternating series if $c=0$, $\delta=\pi$, as can be attained by translation). Thus we see that

$$E = \log\frac{n}{m} + o(1), \qquad (m,n\to\infty).$$

This gives the following:

THEOREM 3. *Under the hypothesis of Theorem* 2, *with $p\neq q$, the asymmetric partial sums $S_{m,n}(c)$ converge to a finite limit as $m,n\to\infty$ if, and only if, $m,n\to\infty$ in such a way that n/m has a finite nonzero limit. If the latter limit is α, then*

$$\lim S_{m,n}(c) = \frac{p+q}{2} + \frac{(p-q)i}{2\pi}\log\alpha.$$

We get the expected value $(p+q)/2$ if, and only if, $n/m\to 1$ as $m,n\to\infty$.

REFERENCES

[1] PAUL R. CHERNOFF, *Pointwise convergence of Fourier series*, Amer. Math. Monthly, 87 (1980), pp. 399–400.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics

014

ASYMPTOTIC FORMULAS FOR ZERO-BALANCED HYPERGEOMETRIC SERIES*

RONALD J. EVANS[†] AND DENNIS STANTON[‡]

Abstract. A hypergeometric series is called s-balanced if the sum of denominator parameters minus the sum of numerator parameters is s. A nonterminating s-balanced hypergeometric series converges at $x=1$ if s is positive. An asymptotic formula for the partial sums of a zero-balanced ${}_3F_2(1)$ is given. A corollary is the behavior of a zero-balanced ${}_3F_2(x)$ as x approaches 1. Some q-analogues are also given.

1. Introduction. For $0<q<1$, define

$$(1.1)\qquad (a)_k=\prod_{j=0}^{k-1}(1-q^ja),\qquad (a)_\infty=\prod_{j=0}^{\infty}(1-q^ja).$$

In the limiting case $q=1$, define

$$(1.2)\qquad (a)_k=\prod_{j=0}^{k-1}(a+j).$$

Let

$$(1.3)\qquad \lambda(x)=x(x)'_\infty/(x)_\infty,$$

where $(x)'_\infty$ denotes the derivative of $(x)_\infty$ with respect to x.

The following two theorems will be proved in §§3 and 4.

THEOREM 1. *If* $abc=de$ *and* $|c|<1$, *then, in the notation of* (1.1),

$$(1.4)\qquad \sum_{k=0}^{\infty}\left\{\frac{(dq^k)_\infty(eq^k)_\infty(q^{k+1})_\infty}{(aq^k)_\infty(bq^k)_\infty(cq^k)_\infty}-\frac{1}{1-q^{k+1}}\right\}=L_q,$$

where

$$(1.5)\qquad L_q=2\lambda(q)-\lambda(a)-\lambda(b)+\sum_{k=1}^{\infty}\frac{(d/c)_k(e/c)_kc^k}{(a)_k(b)_k(1-q^k)};$$

also, as $m\to\infty$,

$$(1.6)\quad \sum_{k=0}^{m-1}\frac{(a)_k(b)_k(c)_k}{(d)_k(e)_k(q)_k}=\frac{(a)_\infty(b)_\infty(c)_\infty}{(d)_\infty(e)_\infty(q)_\infty}\left\{\sum_{j=0}^{m-1}\frac{1}{1-q^{j+1}}+L_q\right\}+O(q^m),$$

where the implied constant depends on a,b,c,d,e,q *but not on* m.

THEOREM 2. *If* $a+b+c=d+e$ *and* $\mathrm{Re}(c)>0$, *then, in the notation of* (1.2),

$$(1.7)\qquad \sum_{k=0}^{\infty}\left\{\frac{\Gamma(a+k)\Gamma(b+k)\Gamma(c+k)}{\Gamma(d+k)\Gamma(e+k)\Gamma(1+k)}-\frac{1}{k+1}\right\}=L,$$

*Received by the editors February 8, 1983. This research was supported by the National Science Foundation under grants MCS8101860 and MCS8102237.

[†] Department of Mathematics, University of California, San Diego, La Jolla, California, 92093.

[‡] School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

where

$$L=-2\gamma-\frac{\Gamma'(a)}{\Gamma(a)}-\frac{\Gamma'(b)}{\Gamma(b)}+\sum_{k=1}^{\infty}\frac{(d-c)_k(e-c)_k}{(a)_k(b)_k k}, \tag{1.8}$$

where γ *is Euler's constant; also, as* $m\to\infty$,

$$\sum_{k=0}^{m-1}\frac{(a)_k(b)_k(c)_k}{(d)_k(e)_k k!}=\frac{\Gamma(d)\Gamma(e)}{\Gamma(a)\Gamma(b)\Gamma(c)}\{\log m+L+\gamma\}+O\left(\frac{1}{m}\right), \tag{1.9}$$

where the implied constant depends on a,b,c,d,e *but not on* m.

Theorem 2 gives an asymptotic formula as $m\to\infty$ for the mth partial sums of a zero-balanced hypergeometric series ${}_3F_2(\genfrac{}{}{0pt}{}{abc}{de}|1)$. It would be interesting if such a result could be extended to ${}_4F_3$ series. The special case $c=e$ of (1.9) gives the following known asymptotic formula [4, p. 109, (34)] for partial sums of a zero-balanced hypergeometric series ${}_2F_1(\genfrac{}{}{0pt}{}{ab}{d}|1)$:

$$\sum_{k=0}^{m-1}\frac{(a)_k(b)_k}{(d)_k k!}=\frac{\Gamma(d)}{\Gamma(a)\Gamma(b)}\left\{\log m-\gamma-\frac{\Gamma'(a)}{\Gamma(a)}-\frac{\Gamma'(b)}{\Gamma(b)}\right\}+O\left(\frac{1}{m}\right). \tag{1.10}$$

This paper was motivated by the desire to prove the following theorem, stated (in less precise form) without proof by Ramanujan [6, Entry 24, Cor. 2], [2, Entry 24, Cor. 2]. We are grateful to Bruce Berndt for bringing Ramanujan's result to our attention.

THEOREM 3. *If* $a+b+c=d+e$ *and* $\mathrm{Re}(c)>0$, *then as* $u\to 1$ *with* $0<u<1$,

$$\frac{\Gamma(a)\Gamma(b)\Gamma(c)}{\Gamma(d)\Gamma(e)}\,{}_3F_2\left(\begin{matrix}a,b,c\\ d,e\end{matrix}\middle| u\right)=-\log(1-u)+L+O((1-u)\log(1-u)), \tag{1.11}$$

where L *is defined in* (1.8).

In §5, we will deduce Theorem 3 from Theorem 2. It is a mystery to us how Ramanujan found the constant term L in the asymptotic expansion (1.11). Because of the inductive nature of our proofs, this paper unfortunately sheds little light on how Ramanujan might have made this remarkable discovery.

Finally, we mention the q-analogue of Theorem 3. If $abc=de$ and $|c|<1$, then as $u\to 1$

$$\frac{(q)_\infty(d)_\infty(e)_\infty}{(a)_\infty(b)_\infty(c)_\infty}\,{}_3\phi_2\left(\begin{matrix}a,b,c\\ d,e\end{matrix}\middle| u\right)=g_q(u)+L_q+O((1-u)g_q(u)),$$

where $g_q(u)=\sum_{k=0}^{\infty}u^{k+1}/(1-q^{k+1})$, L_q is defined by (1.5), and ${}_3\phi_2$ is defined at the beginning of §2.

2. Preliminary lemmas. We will use the following notation for q-hypergeometric series:

$${}_3\phi_2\left(\begin{matrix}a,b,c\\ d,e\end{matrix}\middle| z\right)=\sum_{k=0}^{\infty}\frac{(a)_k(b)_k(c)_k z^k}{(d)_k(e)_k(q)_k}.$$

Partial sums will be denoted by

$${}_3\phi_2\left(\begin{matrix}a,b,c\\ d,e\end{matrix}\middle| z\right)_m=\sum_{k=0}^{m}\frac{(a)_k(b)_k(c)_k z^k}{(d)_k(e)_k(q)_k}.$$

LEMMA 4. *If* $\mathrm{Re}(C)>0$, $S=D+E-A-B-C$, *and* $\mathrm{Re}(S)>0$, *then*

$$ {}_3F_2\left(\begin{matrix}A,B,C\\D,E\end{matrix}\middle|1\right)=\frac{\Gamma(D)\Gamma(E)\Gamma(S)}{\Gamma(C)\Gamma(A+S)\Gamma(B+S)}\,{}_3F_2\left(\begin{matrix}D-C,E-C,S\\A+S,B+S\end{matrix}\middle|1\right). \tag{2.1}$$

LEMMA 5. *If* $0<q<1$, $|C|<1$, *and* $|DE/ABC|<1$, *then*

$$ {}_3\Phi_2\left(\begin{matrix}A,B,C\\D,E\end{matrix}\middle|\frac{DE}{ABC}\right)=\frac{(DE/AC)_\infty(C)_\infty(DE/BC)_\infty}{(D)_\infty(E)_\infty(DE/ABC)_\infty}\,{}_3\Phi_2\left(\begin{matrix}D/C,E/C,DE/ABC\\DE/AC,DE/BC\end{matrix}\middle|C\right). \tag{2.2}$$

Lemma 4 is proved in [1, p. 14]. Lemma 5 is a q-analogue of Lemma 4 whose proof is completely analogous to the proof for Lemma 4; where Gauss's theorem was invoked, one uses instead the q-analogue of Gauss's theorem given in [1, p. 68, (3)].

LEMMA 6. *If* $0<q<1$ *and* D *and* A *are bounded, then, as* $k\to\infty$,

$$\frac{(Dq^k)_\infty}{(Aq^k)_\infty}=1+O(q^k). \tag{2.3}$$

Proof. This follows easily from the q-binomial theorem [1, p. 66, (4)], namely

$$\sum_{j=0}^{\infty}\frac{(a)_jz^j}{(q)_j}=\frac{(az)_\infty}{(z)_\infty},\qquad |z|<1. \tag{2.4}$$

LEMMA 7. *If* d *and* a *are bounded, then as* $z\to\infty$ *with* $\mathrm{Re}(z)>0$,

$$\frac{\Gamma(a+z)}{\Gamma(d+z)}=z^{a-d}(1+O(z^{-1})). \tag{2.5}$$

Proof. This follows from [4, p. 33, (11)].

LEMMA 8. *Fix* $\varepsilon>0$ *and fix a complex number* E. *Let* $\mathrm{Re}(z)\geq\varepsilon$ *and let* k *be a variable positive integer. Then there exists* $N>0$ *such that*

$$\left(1+\frac{z}{k}\right)^E-1=O\left(\frac{z^N}{k}\right), \tag{2.6}$$

where N *and the implied constant are independent of* z *and* k.

Proof. Let $F=\mathrm{Re}(E)$. If $F\geq 0$, then

$$\left(1+\frac{z}{k}\right)^{-E}-1=-\left(1+\frac{z}{k}\right)^{-E}\left(\left(1+\frac{z}{k}\right)^{E}-1\right)=O\left(\left(1+\frac{z}{k}\right)^{E}-1\right),$$

so it suffices to consider the case $F\geq 0$. Let $N=F+1$. First suppose that $k\leq|z|$. Then

$$\left|1+\frac{z}{k}\right|^F\leq(1+|z|)^F=O(z^F)=O\left(\frac{z^N}{k}\right).$$

Thus

$$\left(1+\frac{z}{k}\right)^E=O\left(\left(1+\frac{z}{k}\right)^F\right)=O\left(\frac{z^N}{k}\right)$$

and (2.6) follows. Finally suppose that $k>|z|$. Then since $F\geq 0$,

$$\left|\left(1+\frac{z}{k}\right)^E-1\right|\leq\sum_{m=1}^{\infty}\left|\binom{E}{m}\right|\left|\frac{z}{k}\right|^m\leq\left|\frac{z}{k}\right|\sum_{m=1}^{\infty}\left|\binom{E}{m}\right|=O\left(\frac{z}{k}\right)=O\left(\frac{z^N}{k}\right).$$

LEMMA 9. *Fix real* $D, D\notin\{0,-1,-2,-3,\cdots\}$. *Let k be a variable positive integer. Let* $\mathrm{Re}(z)\geq 0$. *Then in the notation of* (1.2),

$$\frac{(D-z)_k}{(D)_k}=O(e^{2\pi|z|/3}), \tag{2.7}$$

where the implied constant is independent of z and k.

Proof. For some constant $N>0$ independent of z and k,

$$\left|\frac{(D-z)_k}{(D)_k}\right|=\prod_{j=0}^{k-1}\left|\frac{D+j-z}{D+j}\right|=\prod_{j=0}^{k-1}\left|1-\frac{z}{D+j}\right|\ll(1+|z|)^N\prod_{\substack{j=0\\D+j\geq 1}}^{k-1}\left|1-\frac{z}{D+j}\right|.$$

Thus

$$\begin{aligned}\left|\frac{(D-z)_k}{(D)_k}\right|&\ll(1+|z|)^N\prod_{\substack{j=0\\D+j\geq 1}}^{k-1}\left(1-2\,\mathrm{Re}\left(\frac{z}{D+j}\right)+\left|\frac{z}{D+j}\right|^2\right)^{1/2}\\&\ll(1+|z|)^N\prod_{\substack{j=0\\D+j\geq 1}}^{k-1}\left(1+\left|\frac{z}{D+j}\right|^2\right)^{1/2}\ll(1+|z|)^N\prod_{m=1}^{\infty}\left(1+\frac{|z|^2}{m^2}\right)^{1/2}\\&=(1+|z|)^N\left(\frac{e^{\pi|z|}-e^{-\pi|z|}}{2\pi|z|}\right)^{1/2}\ll(1+|z|)^Ne^{\pi|z|/2}\ll e^{2\pi|z|/3}.\end{aligned}$$

3. Proof of Theorem 1. We begin by proving (1.6) in the case $c=q$. Let $0<t<1$ and let m be a large integer. By the hypothesis $abq=abc=de$,

$$ {}_3\Phi_2\left(\begin{matrix}a,b,q\\d,et\end{matrix}\middle|t\right)_{m-1}=S_1-S_2, \tag{3.1}$$

where

$$S_1={}_3\Phi_2\left(\begin{matrix}a,b,q\\d,et\end{matrix}\middle|t\right) \tag{3.2}$$

and

$$S_2=\frac{(a)_m(b)_m(q)_mt^m}{(d)_m(et)_m(q)_m}\,{}_3\Phi_2\left(\begin{matrix}q,bq^m,aq^m\\dq^m,etq^m\end{matrix}\middle|t\right). \tag{3.3}$$

Apply Lemma 5 with A, B, C, D, E equal to a, b, q, d, et, respectively, to obtain

$$S_1 = \frac{(at)_\infty (bt)_\infty (q)_\infty}{(d)_\infty (et)_\infty (t)_\infty}\, {}_3\phi_2\left(\begin{matrix} d/q, et/q, t \\ at, bt \end{matrix}\,\middle|\, q\right). \tag{3.4}$$

Apply Lemma 5 with A, B, C, D, E equal to $q, bq^m, aq^m, dq^m, etq^m$, respectively, to obtain

$$S_2 = \frac{(a)_\infty (qt)_\infty (b)_m (btq^m)_\infty t^m}{(d)_\infty (et)_\infty (t)_\infty}\, {}_3\phi_2\left(\begin{matrix} d/a, et/a, t \\ qt, btq^m \end{matrix}\,\middle|\, aq^m\right). \tag{3.5}$$

Thus, by (3.1), (3.4), and (3.5),

$${}_3\phi_2\left(\begin{matrix} a, b, q \\ d, et \end{matrix}\,\middle|\, t\right)_{m-1} = R_1(t) + R_2(t) - R_3(t), \tag{3.6}$$

where

$$R_1(t) = \frac{(at)_\infty (bt)_\infty (q)_\infty}{(d)_\infty (et)_\infty (t)_\infty} - \frac{(a)_\infty (qt)_\infty (b)_m (btq^m)_\infty t^m}{(d)_\infty (et)_\infty (t)_\infty}, \tag{3.7}$$

$$R_2(t) = \frac{(at)_\infty (bt)_\infty (q)_\infty}{(d)_\infty (et)_\infty (t)_\infty} \sum_{k=1}^{\infty} \frac{(d/q)_k (et/q)_k (t)_k q^k}{(at)_k (bt)_k (q)_k}, \tag{3.8}$$

and

$$R_3(t) = \frac{(a)_\infty (qt)_\infty (b)_m (btq^m)_\infty t^m}{(d)_\infty (et)_\infty (t)_\infty} \sum_{k=1}^{\infty} \frac{(d/a)_k (et/a)_k (t)_k (aq^m)^k}{(qt)_k (btq^m)_k (q)_k}. \tag{3.9}$$

Taking the limit as $t \to 1$ in (3.6), we obtain

$${}_3\phi_2\left(\begin{matrix} a, b, q \\ d, e \end{matrix}\,\middle|\, 1\right)_{m-1} = R_1 + R_2 - R_3, \tag{3.10}$$

where

$$R_i = \lim_{t \to 1} R_i(t). \tag{3.11}$$

Now,

$$\begin{aligned} R_1 &= \lim_{t \to 1} \frac{1}{(e)_\infty (d)_\infty (1-t)} \left\{ \frac{(at)_\infty (bt)_\infty (q)_\infty}{(qt)_\infty} - (a)_\infty (b)_m (btq^m)_\infty t^m \right\} \\ &= \frac{(a)_\infty (b)_\infty}{(d)_\infty (e)_\infty} \{\lambda(q) - \lambda(a) + \lambda(bq^m) - \lambda(b) + m\}. \end{aligned} \tag{3.12}$$

Since

$$\lambda(x) = \sum_{j=0}^{\infty} \frac{-xq^j}{1 - xq^j}, \tag{3.13}$$

we have

$$(3.14)\quad \lambda(bq^m)-\lambda(b)+m=\sum_{j=0}^{m-1}\frac{1}{1-bq^j}$$
$$=\sum_{j=0}^{m-1}\frac{1}{1-q^{j+1}}+\sum_{j=0}^{m-1}\left\{\frac{-q^{j+1}}{1-q^{j+1}}-\frac{-bq^j}{1-bq^j}\right\}$$
$$=\sum_{j=0}^{m-1}\frac{1}{1-q^{j+1}}+\sum_{j=0}^{\infty}\frac{-q^{j+1}}{1-q^{j+1}}-\sum_{j=0}^{\infty}\frac{-bq^j}{1-bq^j}+O(q^m)$$
$$=\sum_{j=0}^{m-1}\frac{1}{1-q^{j+1}}+\lambda(q)-\lambda(b)+O(q^m).$$

By (3.12) and (3.14),

$$(3.15)\qquad R_1=\frac{(a)_\infty(b)_\infty}{(d)_\infty(e)_\infty}\left\{2\lambda(q)-\lambda(a)-\lambda(b)+\sum_{j=0}^{m-1}\frac{1}{1-q^{j+1}}+O(q^m)\right\}.$$

Since

$$(3.16)\qquad \lim_{t\to 1}\frac{(t)_k}{(t)_\infty}=\frac{(q)_{k-1}}{(q)_\infty},$$

we have

$$(3.17)\qquad R_2=\frac{(a)_\infty(b)_\infty}{(d)_\infty(e)_\infty}\sum_{k=1}^{\infty}\frac{(d/q)_k(e/q)_kq^k}{(a)_k(b)_k(1-q^k)}$$

and

$$(3.18)\qquad R_3=\frac{(a)_\infty(b)_\infty}{(d)_\infty(e)_\infty}\sum_{k=1}^{\infty}\frac{(d/a)_k(e/a)_k(aq^m)^k}{(bq^m)_k(q)_k(1-q^k)}=O(q^m).$$

By (3.10), (3.15), (3.17) and (3.18),

(3.19)
$$ {}_3\phi_2\left(\begin{matrix}a,b,q\\ d,e\end{matrix}\middle|1\right)_{m-1}=\frac{(a)_\infty(b)_\infty}{(d)_\infty(e)_\infty}\left\{\sum_{j=0}^{m-1}\frac{1}{1-q^{j+1}}+2\lambda(q)\right.$$
$$\left.-\lambda(a)-\lambda(b)+\sum_{k=1}^{\infty}\frac{(d/q)_k(e/q)_kq^k}{(a)_k(b)_k(1-q^k)}\right\}+O(q^m).$$

This completes the proof of (1.6) in the case $c=q$.

We next prove that (1.6) holds for $c=q^n$ for all positive integers n. Let $c=q^N$ for an integer $N>1$, and assume as induction hypothesis that (1.6) holds with $c=q^n$ for all n such that $1\le n<N$. Since

$$(3.20)\qquad (a)_k=q^k\left(\frac{a}{q}\right)_k+(1-q^k)(a)_{k-1},$$

we have

$$(3.21)\quad {}_3\phi_2\left(\begin{matrix}a,b,c\\ d,e\end{matrix}\middle|1\right)_{m-1}=\frac{(1-d/q)(1-e/q)}{(1-b/q)(1-c/q)}\,{}_3\phi_2\left(\begin{matrix}a,b/q,c/q\\ d/q,e/q\end{matrix}\middle|1\right)_m$$
$$-\frac{(1-d/q)(1-e/q)}{(1-b/q)(1-c/q)}\,{}_3\phi_2\left(\begin{matrix}a/q,b/q,c/q\\ d/q,e/q\end{matrix}\middle|q\right)_m.$$

Since $a(b/q)(c/q)=(d/q)(e/q)$, the first term on the right of (3.21) can be evaluated by the induction hypothesis.

The last term on the right of (3.21) equals

$$(3.22)\qquad \frac{(1-d/q)(1-e/q)}{(1-b/q)(1-c/q)}\,{}_3\phi_2\left(\begin{matrix}a/q,b/q,c/q\\ d/q,e/q\end{matrix}\middle|q\right)+O(q^m),$$

since the ${}_3\phi_2$ in (3.22) converges; by Lemma 5, the first term in (3.22) in turn equals

$$(3.23)\qquad \frac{(a)_\infty(b)_\infty(c)_\infty}{(1-b/q)(d)_\infty(e)_\infty(q)_\infty}\sum_{k=1}^{\infty}\frac{(d/c)_k(e/c)_k(c/q)^k}{(a)_k(b)_k}.$$

The relations

$$\frac{1}{(b/q)_k(1-q^k)}-\frac{1}{(b/q)_{k+1}}=\frac{q^k}{(b)_k(1-q^k)}$$

and

$$\frac{1}{1-q^{m+1}}-\lambda(b/q)-\frac{1}{1-b/q}=-\lambda(b)+O(q^m)$$

show that (3.21) and (3.23) imply that (1.6) holds for $c=q^N$. This completes the induction, so (1.6) holds for $c=q^N$ for all positive integers N. Taking the limit as m tends to ∞, we see that (1.4) also holds for all c of the form $c=q^N$.

We next prove that (1.4) holds without the restriction $c=q^N$. Since $q^N\to 0$ as $N\to\infty$, it suffices to show that each member of (1.4) is an analytic function of c on the disk $|c|<1$ for each fixed choice of $a,b,d,$ and q.

Fix $t, 0<t<1$. To show that the right member of (1.4) is analytic in c, it suffices to prove that the series

$$(3.24)\qquad \sum_{k=1}^{\infty}\frac{(d/c)_k(ab/d)_k c^k}{(a)_k(b)_k(1-q^k)}$$

converges uniformly in the disk $|c|\le t$. Since $|(d/c)_k c^k|=\Pi_{j=0}^{k-1}|c-dq^j|\ll t_1^k$ for some t_1, $t<t_1<1$, and since $(ab/d)_k/(a)_k(b)_k$ is bounded, the series in (3.24) converges uniformly in the disk $|c|\le t$.

To show that the left member of (1.4) is analytic in c, it suffices to prove that the series

$$(3.25)\qquad \sum_{k=0}^{\infty}\left\{\frac{(dq^k)_\infty(q^{k+1})_\infty(abcq^k/d)_\infty}{(aq^k)_\infty(bq^k)_\infty(cq^k)_\infty}-\frac{1}{1-q^{k+1}}\right\}$$

converges uniformly in the disk $|c|\le t$.By Lemma 6, as $k\to\infty$,

$$\frac{(dq^k)_\infty}{(aq^k)_\infty}=1+O(q^k),\quad \frac{(q^{k+1})_\infty}{(bq^k)_\infty}=1+O(q^k),\quad \frac{(abcq^k/d)_\infty}{(cq^k)_\infty}=1+O(q^k). \tag{3.26}$$

Therefore the summand in (3.25) is $\ll q^k$, so the series in (3.25) converges uniformly in the disk $|c|\le t$. This completes the proof of (1.4).

By (3.26), we see that if the index of summation in (3.25) begins at $k=m$ instead of $k=0$, the resulting series is $O(q^m)$, where the implied constant depends on a,b,c,d,e,q but not on m. Thus (1.6) follows from (1.4).

4. Proof of Theorem 2. By Lemma 7, we see that if the index of summation in (1.7) begins at $k=m$ instead of $k=0$, the resulting series is $O(1/m)$, where the implied constant is independent of m. Since also

$$\sum_{k=0}^{m-1}\frac{1}{k+1}=\log m+\gamma+O\left(\frac{1}{m}\right),$$

(1.9) follows from (1.7). It remains to prove (1.7). If one took $\lim_{q\to 1}$ of each side of (1.4) and then interchanged limits and summations, (1.7) would result. However, since it appears to be a difficult task indeed to justify this interchange of limits and summations, we take a different approach.

The proof in §3 began by showing that (1.4) holds for each c of the form $c=q^n$, where n is a positive integer. Mimicking this proof with $q=1$, we can deduce that (1.9) holds for $c=1$, as follows. In place of (3.1), write, for $\varepsilon>0$,

$$_3F_2\left(\begin{matrix}a,b,1\\ d,e+\varepsilon\end{matrix}\middle|1\right)_{m-1}=H_1-H_2,$$

where

$$H_1={}_3F_2\left(\begin{matrix}a,b,1\\ d,e+\varepsilon\end{matrix}\middle|1\right)$$

and

$$H_2=\frac{(a)_m(b)_m}{(d)_m(e+\varepsilon)_m}\,{}_3F_2\left(\begin{matrix}1,b+m,a+m\\ d+m,e+\varepsilon+m\end{matrix}\middle|1\right).$$

Apply Lemma 4 to get analogues of (3.4) and (3.5) for H_1 and H_2. Let $\varepsilon\to 0$ to obtain the analogue of (3.10) of the form

$$_3F_2\left(\begin{matrix}a,b,1\\ d,e\end{matrix}\middle|1\right)_{m-1}=G_1+G_2-G_3. \tag{4.1}$$

The analogue of (3.12) is

$$G_1=\frac{\Gamma(d)\Gamma(e)}{\Gamma(a)\Gamma(b)}\left(\frac{\Gamma'(1)}{\Gamma(1)}-\frac{\Gamma'(a)}{\Gamma(a)}-\frac{\Gamma'(b)}{\Gamma(b)}+\frac{\Gamma'(b+m)}{\Gamma(b+m)}\right).$$

Since [4, p. 33, (8)]

$$\frac{\Gamma'(b+m)}{\Gamma(b+m)}=\log m+O\left(\frac{1}{m}\right),$$

we obtain the following analogue of (3.15):

$$G_1=\frac{\Gamma(d)\Gamma(e)}{\Gamma(a)\Gamma(b)}\left(-\gamma-\frac{\Gamma'(a)}{\Gamma(a)}-\frac{\Gamma'(b)}{\Gamma(b)}+\log m\right)+O\left(\frac{1}{m}\right). \tag{4.2}$$

Apply Lemma 7 to obtain the following analogues of (3.17) and (3.18):

$$G_2=\frac{\Gamma(d)\Gamma(e)}{\Gamma(a)\Gamma(b)}\sum_{k=1}^{\infty}\frac{(d-1)_k(e-1)_k}{(a)_k(b)_k k} \tag{4.3}$$

and

$$G_3=\frac{\Gamma(d)\Gamma(e)}{\Gamma(a)\Gamma(b)}\sum_{k=1}^{\infty}\frac{(d-a)_k(e-a)_k}{(b+m)_k(1)_k k}=O\left(\frac{1}{m}\right). \tag{4.4}$$

Combining (4.1)–(4.4), we deduce that (1.9) holds for $c=1$.

An induction argument analogous to that following (3.19) shows that (1.9) holds for each positive integer c. Taking the limit as m tends to ∞, we see that (1.7) also holds for each positive integer c.

To prove that (1.7) holds for all c with $\operatorname{Re}(c)>0$, it suffices by Carlson's theorem [1, p. 39] to prove that, *for fixed a, b, d and fixed* $\varepsilon>0$, both sides of (1.7) are analytic in c and equal to $O(e^{2\pi|c|/3})$ for $\operatorname{Re}(c)\geq\varepsilon$.

Write $D=\operatorname{Re}(d-\varepsilon)$, adjusting ε if necessary so that $D\notin\{0,-1,-2,-3,\cdots\}$. Write $z=c+D-d$, so in the notation of (1.2),

$$S:=\sum_{k=1}^{\infty}\frac{(d-c)_k(e-c)_k}{(a)_k(b)_k k}=\sum_{k=1}^{\infty}A_k\frac{(D-z)_k}{(D)_k}$$

with

$$A_k=\frac{(a+b-d)_k(D)_k}{(a)_k(b)_k k}.$$

By Lemma 7, $A_k=O(k^{-1-\varepsilon})$. By Lemma 9, $(D-z)_k/(D)_k=O(e^{2\pi|z|/3})$. Thus S is analytic in z and equals $O(e^{2\pi|z|/3})$ for $\operatorname{Re}(z)\geq 0$. It follows that S is analytic in c and equal to $O(e^{2\pi|c|/3})$ for $\operatorname{Re}(c)\geq\varepsilon$.

It remains to prove that

$$T:=\sum_{k=1}^{\infty}\left\{\frac{\Gamma(a+k)\Gamma(b+k)\Gamma(c+k)}{\Gamma(1+k)\Gamma(d+k)\Gamma(a+b-d+c+k)}-\frac{1}{k+1}\right\}$$

is analytic in c and equal to $O(e^{2\pi|c|/3})$ for $\operatorname{Re}(c)\geq\varepsilon$. Let $E=d-a-b$. By Lemma 7,

$$\begin{aligned}T&=\sum_{k=1}^{\infty}\left\{k^{-E-1}(c+k)^{E}(1+k^{-1}O(1))-\frac{1}{k+1}\right\}\\&=O(1)+\sum_{k=1}^{\infty}k^{-1}\left\{\left(1+\frac{c}{k}\right)^{E}-1\right\}\{1+k^{-1}O(1)\},\end{aligned}$$

where the expressions $O(1)$ are bounded analytic functions of c for $\mathrm{Re}(c)\geq\varepsilon$. By Lemma 8, $(1+c/k)^{E}-1=O(c^{N}/k)$, so T is analytic in c and equals $O(c^{N})$ for $\mathrm{Re}(c)\geq\varepsilon$.

5. Proof of Theorem 3. Define

$$f(k)=\frac{\Gamma(a+k)\Gamma(b+k)\Gamma(c+k)}{\Gamma(d+k)\Gamma(e+k)\Gamma(1+k)},$$

and $V=\Sigma_{k=0}^{\infty}f(k)u^{k}+\log(1-u)-L$, where L is defined in (1.8). We must show that as $u\to 1$,

$$V=O((1-u)\log(1-u)).$$

By (1.7),

$$V=\sum_{k=0}^{\infty}\left(f(k)-\frac{1}{k+1}\right)(u^{k}-1)+\sum_{k=0}^{\infty}\frac{u^{k}-u^{k+1}}{k+1}.$$

The last sum is $(u-1)/u\log(1-u)=O((1-u)\log(1-u))$ as $u\to 1$. Finally, by Lemma 7,

$$\sum_{k=1}^{\infty}\left|\left(f(k)-\frac{1}{k+1}\right)(u^{k}-1)\right|\ll\sum_{k=1}^{\infty}\frac{1-u^{k}}{k^{2}}$$

$$=(1-u)\sum_{k=1}^{\infty}k^{-2}\sum_{n=0}^{k-1}u^{n}=(1-u)\sum_{n=0}^{\infty}u^{n}\sum_{k=n+1}^{\infty}k^{-2}$$

$$<(1-u)\left\{\frac{\pi^{2}}{6}+\sum_{n=1}^{\infty}\frac{u^{n}}{n}\right\}=O((1-u)\log(1-u)).$$

6. Concluding remarks. The series

$${}_3F_2\left(\begin{matrix}a,b,c\\d,e\end{matrix}\middle|1\right)$$

converges for $\mathrm{Re}(e+d-a-b-c)>0$. Theorem 2 gives information of the divergence at the boundary $a+b+c=d+e$. We have not investigated related problems, such as $a+b+c=d+e+1$.

Bailey and Darling have given transformations for truncated 1-balanced ${}_3F_2$'s [1, p. 94–95]. We were unable to use similar techniques to derive Theorem 2. There may be similar results for special truncated very well poised ${}_6F_5$'s.

The special case $c=e$ of Theorem 3 gives an asymptotic expansion of a zero-balanced ${}_2F_1(x)$ as $x\to 1$. This is equivalent to (1.10). This result is easy to obtain in the following way. The point $x=1$ is a regular singular point of the differential equation for ${}_2F_1(x)$. There are two independent solutions (u_1 and u_2) near $x=1$. If the ${}_2F_1$ is zero-balanced, one solution is logarithmic. The precise definitions of u_1 and u_2 and the constants c_1 and c_2 such that ${}_2F_1(x)=c_1u_1+c_2u_2$, are given in [3, eq. 2.10 (14)]. The asymptotic formula follows immediately.

For the ${}_3F_2(x)$ case, Norlund [5] has explicitly given three independent solutions ($u_1, u_2,$ and u_3) near $x=1$. (The authors would like to thank Dennis Hejhal for pointing this out.) Again the zero-balancing condition gives a logarithmic solution. So an expansion of the form of Theorem 3 is guaranteed. However, the constant L is not given. One would need to find the constants c_1, c_2, and c_3 such that ${}_3F_2(x)=c_1u_1+c_2u_2+c_3u_3$. This is not an easy task.

REFERENCES

[1] W. N. Bailey, *Generalized Hypergeometric Series*, Stechert–Hafner, New York, 1964.

[2] B. C. Berndt, *Chapter* 11 *of Ramanujan's second notebook*, to appear.

[3] A. Erdélyi et al., *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[4] Y. L. Luke, *The Special Functions and Their Approximations*, Vol. 1, Academic Press, New York, 1969.

[5] N. Norlund, *Hypergeometric functions*, Acta Math., 94 (1955), pp. 289–349.

[6] S. Ramanujan, *Notebooks*, 2 volumes, Tata Institute of Fundamental Research, Bombay, 1957.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics

015

CHARACTERIZATION OF QUADRATURE FORMULA II*

FRANZ PEHERSTORFER[†]

Abstract. In a recent paper (SIAM J. Math. Anal., 12 (1981), pp. 935–942) we have described positive quadrature formulas (qf). The purpose of this note is to complete our results on positive qf and to extend them to qf which have a given number of positive and negative weights. Furthermore we display the connection between our results and the results of Sottas and Wanner (BIT, 22 (1982), pp. 339–352).

1. Introduction. Let w be a nonnegative weight function on $[-1,+1]$. We consider interpolatory quadrature formulas of the type

$$\int_{-1}^{+1} f(x)w(x)\,dx=\sum_{i=1}^{n}\lambda_i f(x_i)+R_n(f), \tag{1}$$

where $-1<x_1<x_2<\cdots<x_n<1$. If $R_n(f)=0$ for all $f\in\mathbb{P}_{2n-1-m}$ ($\mathbb{P}_{2n-m-1}$ denotes the set of polynomials of degree at most $2n-1-m$), we say that (1) is a $(2n-1-m,n,w)$ quadrature formula (qf).

In [7] we have given a full description of positive $(2n-1-m,n,w)$ qf. Recently, Sottas and Wanner [10] have given an independent characterization of $(2n-1-m,n,w)$ qf which have a given number of positive and negative weights. In this paper we extend our investigations of [7] and display the connection between our results and the results of Sottas and Wanner.

2. Characterizations. Henceforth let p_n denote that polynomial of degree n with leading coefficient one which is orthogonal to $\mathbb{P}_{n-1}$ on $[-1,+1]$ with respect to the weight function w. Thus the polynomials (p_n) satisfy a recurrence relation of the form

$$p_n(x)=(x-\alpha_n)p_{n-1}(x)-\beta_n p_{n-2}(x),$$

where $p_{-1}=0, p_0=1$. Note that $\beta_n>0$ and $|\alpha_n|<1$.

The following lemma can be derived from [10, Thm. 1]. However we shall give a proof which is independent of this result.

LEMMA 1. *Let $n,k\in\mathbb{N}_0$ be such that $n\ge 2k$. Suppose that the polynomial $q_{n-k+1}(x)=\prod_{i=1}^{n-k+1}(x-y_i)$, $y_i\in\mathbb{R}$, $y_1<y_2<\cdots<y_{n-k+1}$, has no common zero with p_{n-k} and that $q_{n-k+1}\perp P_{n-k-2}$. Then each polynomial t_n of degree n with $t_n\perp\mathbb{P}_{n-2k-1}$ has a unique representation of the form*

$$t_n(x)=r_k(x)p_{n-k}(x)+s_{k-1}(x)q_{n-k+1}(x),$$

where $r_k\in\mathbb{P}_k$ and $s_{k-1}\in\mathbb{P}_{k-1}$.

Proof. Let $x_1<x_2<\cdots<x_{n-k}$ denote the zeros of p_{n-k}. Construct r_k and s_{k-1} such that

$$r_k(y_i)=t_n(y_i)/p_{n-k}(y_i)\quad \text{for } i=1,\cdots,k+1, \tag{2}$$

and

$$s_{k-1}(x_i)=t_n(x_i)/q_{n-k+1}(x_i)\quad \text{for } i=1,\cdots,k. \tag{3}$$

* Received by the editors July 7, 1982.

[†] Institut für Mathematik, Universität Linz, A-4040 Linz, Austria.

Then for $n=2k$ the lemma is proved. For $n>2k$ let us consider the polynomial

$$u(x):=t_n(x)-(r_k(x)p_{n-k}(x)+s_{k-1}(x)q_{n-k+1}(x))\in\mathbb{P}_n.$$

We claim that $u=0$. Suppose to the contrary that $u\neq 0$. Since $u\perp\mathbb{P}_{n-2k-1}$, u can be represented in the form

$$u(x)=\prod_{i=1}^{l}(x-t_i)v(x),$$

where $t_1<t_2<\cdots t_l$, $l\geq n-2k$ and $v\in\mathbb{P}_{n-l}$ is nonnegative or nonpositive on $\mathbb{R}$. Furthermore there exist $\lambda_i\in\mathbb{R}$, $i=1,\cdots,l$, such that

$$\sum_{i=1}^{l}\lambda_i p(t_i)=\int_{-1}^{+1}p(x)v(x)w(x)\,dx \quad \text{for all } p\in\mathbb{P}_{l+n-2k-1}.$$

Now let $x_{v_1},\cdots,x_{v_{k^*}}$ $(y_{\mu_1},\cdots,y_{\mu_{m^*}})$ denote those x_i's, $i\in\{1,\cdots,k\}$, (y_i's, $i\in\{1,\cdots,k+1\}$) at which u does not change sign. Note that in view of (2) and (3) $u(x_i)=0$ for $i=1,\cdots,k$ and $u(y_i)=0$ for $i=1,\cdots,k+1$. Defining

$$\tilde{p}(x):=p_{n-k}(x)/\prod_{j=1}^{k^*}(x-x_{v_j}),\qquad \tilde{q}(x):=q_{n-k+1}(x)/\prod_{j=1}^{m^*}(x-y_{\mu_j}),$$

we obtain

$$\text{(4)}\qquad \sum_{i=1}^{l}\lambda_i\tilde{p}(t_i)p(t_i)=\int_{-1}^{+1}p_{n-k}(x)p(x)\tilde{v}(x)w(x)\,dx=0$$

for all $p\in\mathbb{P}_{1-(k-k^*)-1}$, where $\tilde{v}(x)=v(x)/\prod_{j=1}^{k^*}(x-x_{v_j})\in\mathbb{P}_{n-l-k^*}$; resp.

$$\text{(5)}\qquad \sum_{i=1}^{l}\lambda_i\tilde{q}(t_i)q(t_i)=\int_{-1}^{+1}q_{n-k+1}(x)q(x)\tilde{\tilde{v}}(x)w(x)\,dx=0$$

for all $q\in\mathbb{P}_{l-(k+1-m^*)-1}$, where $\tilde{\tilde{v}}(x)=v(x)/\prod_{j=1}^{m^*}(x-y_{\mu_j})\in\mathbb{P}_{n-l-m^*}$. Since at most $l-(k-k^*)(l-(k+1-m^*))$ of the l points $\tilde{p}(t_i)(\tilde{q}(t_i))$ are not zero, it follows from (4) ((5)) that

$$\lambda_i\tilde{p}(t_i)=0 \quad \text{for } i=1,\cdots,l$$

and

$$\lambda_i\tilde{q}(t_i)=0 \quad \text{for } i=1,\cdots,l.$$

Using the fact that $\tilde{p}$ and $\tilde{q}$ have no common zero we get that $\lambda_i=0$ for $i=1,\cdots,l$; this contradiction proves the lemma.

Remark 1. Lemma 1 remains true for $n=2k-1$, if the leading coefficients of r_k and s_{k-1} are fixed.

Notation. Let $p_n^{(1-x^2)}$ denote that polynomial of degree n with leading coefficient one which is orthogonal to $\mathbb{P}_{n-1}$ on $[-1,+1]$ with respect to the weight function $(1-x^2)w$.

$P_n(z)=z^n+\cdots$ denotes that polynomial which is orthogonal on the unit circle with respect to the weight function $f(\varphi):=w(\cos\varphi)|\sin\varphi|$ for $\varphi\in[0,2\pi)$. Furthermore let U be the open unit disk $\{z\in\mathbb{C}\,|\,|z|<1\}$.

It is well known (see e.g.[4] and [11]) that the polynomials $\{P_n\}$ satisfy a recurrence relation of the type

$$P_{n+1}(z)=zP_n(z)-a_nP_n^*(z),$$

where $P_n^*(z)=z^n\bar{P}_n(z^{-1})$. Note that $|a_n|<1$. Concerning the determination of the so-called parameters a_n we refer to [4]. For example, if $w^{(\lambda)}(x)=(1-x^2)^{\lambda-1/2}$ for $\lambda\in(-\frac{1}{2},\infty)$ one obtains from [4, Thm. 31] that $a_{2n}^{(\lambda)}=0$ and $a_{2n+1}^{(\lambda)}=-\lambda/(n+1+\lambda)$ for $n\in\mathbb{N}_0$. Let us note that the formula given in [7, p. 938] for $\lambda=\frac{1}{2}$ is incorrect.

LEMMA 2. *Let t_n be a polynomial of degree n with leading coefficient one, such that $t_n\perp\mathbb{P}_{n-m-1}$, $0\le m\le n$. Then there exists a real unique polynomial q_m of degree m with leading coefficient one, such that*

$$t_n(x)=2^{-n+1}\mathrm{Re}\{z^{-n+1}q_m(z)P_{2n-1-m}(z)\},$$

$x=\frac{1}{2}(z+1/z)$, $z=e^{i\varphi}$, $\varphi\in[0,\pi]$.

Proof. First let us note (see [4, p. 65] and [11, p. 294]) that for $v\in\mathbb{N}_0$, $x=\frac{1}{2}(z+1/z)$, $z=e^{i\varphi}$, $\varphi\in[0,\pi]$

$$2^{v-1}p_v(x)=\mathrm{Re}\{z^{-v+1}P_{2v-1}(z)\}=\frac{\mathrm{Re}\{z^{-v}P_{2v}(z)\}}{(1-a_{2v-1})},$$

$$2^{v-1}p_{v-1}^{(1-x^2)}(x)=\frac{\mathrm{Im}\{z^{-v+1}P_{2v-1}(z)\}}{\sin\varphi}=\frac{\mathrm{Im}\{z^{-v}P_{2v}(z)\}}{(1+a_{2v-1})\sin\varphi}.$$

Suppose that $2k-1\le m\le 2k$. It follows from Lemma 1 that there exist polynomials $r_k\in\mathbb{P}_k$ and $s_{k-1}\in\mathbb{P}_{k-1}$, such that

$$t_n(x)=r_k(x)p_{n-k}(x)-s_{k-1}(x)(1-x^2)p_{n-k-1}^{(1-x^2)}(x).$$

Case 1. $m=2k$. Setting

$$q_{2k}(z)=(2z)^k\left\{r_k\left(\frac{1}{2}\left(z+\frac{1}{z}\right)\right)+\frac{z^2-1}{2z}s_{k-1}\left(\frac{1}{2}\left(z+\frac{1}{z}\right)\right)\right\},\tag{6}$$

the assertion follows.

Case 2. $m=2k-1$. Let c,d be the leading coefficients of r_k (resp. s_{k-1}). Since the leading coefficient of t_n is one, we have that $c+d=1$. From $\int_{-1}^{+1}x^{n-2k}t_nw=0$ it follows that

$$c\int_{-1}^{+1}p_{n-k}^2w-d\int_{-1}^{+1}[p_{n-k-1}^{(1-x^2)}]^2(1-x^2)w=0.$$

Using the fact (see [4, p. 68]) that

$$\int_{-1}^{+1}p_{n-k}^2w=\frac{D_{n-k}}{1-a_{2(n-k)-1}}$$

and

$$\int_{-1}^{+1}[p_{n-k-1}^{(1-x^2)}]^2(1-x^2)w=\frac{D_{n-k}}{1+a_{2(n-k)-1}},$$

where $D_{n-k}\in\mathbb{R}^+$, simple calculation gives

$$c=(1-a_{2(n-k)-1})/2 \quad\text{and}\quad d=(1+a_{2(n-k)-1})/2.$$

Setting

$$(7)\qquad q_{2k-1}(z)=2(2z)^{k-1}\left\{\frac{r_k(1/2(z+1/z))}{(1-a_{2(n-k)-1})}+\frac{z^2-1}{2z}\frac{s_{k-1}(1/2(z+1/z))}{(1+a_{2(n-k)-1})}\right\}$$

the lemma is proved.

THEOREM 1. *Let $n,m\in\mathbb{N}_0$, $n\geq m$. A $(2n-1-m,n,w)$ qf based on the nodes $x_1,\cdots,x_n\in\mathbb{R}$, $-1<x_1<x_2<\cdots<x_n<1$ has $n-l$ positive weights and l negative weights if and only if there exists a polynomial q_m of degree m with real coefficients and leading coefficient one, such that*

$$2^{-n+1}\mathrm{Re}\{z^{-n+1}q_m(z)P_{2n-1-m}(z)\}=\prod_{i=1}^{n}(x-x_i),$$

$x=\frac{1}{2}(z+1/z)$, $z=e^{i\varphi}$, $\varphi\in[0,\pi]$, *and q_m has $m-2l$ zeros in the unit circle, no zeros on this circle and $2l$ zeros outside of this circle.*

Proof. Necessity. In view of Lemma 2 there exists a real polynomial q_m of degree m with leading coefficient one, such that

$$t_n(x)=2^{-n+1}\mathrm{Re}\{z^{-n+1}q_m(z)P_{2n-1-m}(z)\},$$

$x=\cos\varphi$, $z=e^{i\varphi}$, $\varphi\in[0,\pi]$. On the other hand we have that

$$t_n(x)=\prod_{j=1}^{n}(\cos\varphi-\cos\varphi_j)=(2z)^{-n}\prod_{j=1}^{n}(1-2z\cos\varphi_j+z^2),$$

where $\varphi_j=\arccos x_j$ for $j=1,\cdots,n$, $z=e^{i\varphi}$, $\varphi\in[0,\pi]$.

Hence

$$zq_m(z)P_{2n-1-m}(z)+q_m^*(z)P_{2n-1-m}^*(z)=\prod_{j=1}^{n}(1-2z\cos\varphi_j+z^2).$$

Now let Ω_ν, $\nu\in\mathbb{N}_0$, denote the polynomial of second kind with respect to the weight function $w(\cos\varphi)|\sin\varphi|$. Then it is well known (see [4, p. 7]) that the following relation holds:

$$(8)\qquad P_\nu^*(z)\Omega_\nu(z)+P_\nu(z)\Omega_\nu^*(z)=K_\nu z^\nu\qquad(\nu\in\mathbb{N}_0,K_\nu\in\mathbb{R}^+).$$

Let us put

$$\tilde{P}_{2n,m}(z)=zq_m(z)P_{2n-1-m}(z)\quad\text{and}\quad\tilde{\Omega}_{2n,m}(z)=zq_m(z)\Omega_{2n-1-m}(z).$$

Then we have that

$$\tilde{P}_{2n,m}^*(z)=q_m^*(z)P_{2n-1-m}^*(z)\quad\text{and}\quad\tilde{\Omega}_{2n,m}^*(z)=q_m^*(z)\Omega_{2n-1-m}^*(z).$$

With the help of (8) one deduces (compare [4] and [7]) that

$$(9)\qquad 1+\sum_{k=1}^{2n-1-m}c_kz^k+O(z^{2n-m})=\frac{-\tilde{\Omega}_{2n,m}(z)+\tilde{\Omega}_{2n,m}^*(z)}{\tilde{P}_{2n,m}(z)+\tilde{P}_{2n,m}^*(z)}$$

$$=\frac{1}{2}\sum_{j=1}^{n}\lambda_j\frac{1-z^2}{1-2z\cos\varphi_j+z^2}\qquad\text{for } z\in U,$$

where

$$c_k=2\int_{-1}^{+1}T_kw\Big/\int_{-1}^{+1}w=\sum_{j=1}^{n}\lambda_j\cos k\varphi_j \quad \text{for } k=0,\cdots,2n-1-m,$$

T_k denotes the Chebyshev polynomial of first kind, and

$$\lambda_j=\frac{-2(-\tilde{\Omega}_{2n,m}+\tilde{\Omega}^*_{2n,m})(z_j)}{z_j(d/dz)(\tilde{P}_{2n,m}+\tilde{P}^*_{2n,m})(z_j)} \tag{10}$$
$$=\frac{-2e^{-in\varphi_j}(-\tilde{\Omega}_{2n,m}+\tilde{\Omega}^*_{2n,m})(e^{i\varphi_j})}{i(d/d\varphi)e^{-in\varphi_j}(\tilde{P}_{2n,m}+\tilde{P}^*_{2n,m})(e^{i\varphi_j})},$$

for $j=1,\cdots,n$, $z_j=e^{i\varphi_j}$. The second equality in (9) follows by partial fraction expansion.

Now we claim that q_m has no zero on the circumference. Since $t_n(\pm 1)\neq 0$ implies that $q_m(\pm 1)\neq 0$, we have to show only that q_m has no zero of the form $e^{i\psi}$, $\psi\in(0,\pi)$. Suppose to the contrary that q_m has such a zero $e^{i\psi}$, $\psi\in(0,\pi)$. Then q_m (resp. q^*_m) can be represented in the form

$$q_m(z)=(z-e^{i\psi})(z-e^{-i\psi})q_{m-2}(z)$$
$$\left(\text{resp. } q^*_m(z)=(z-e^{i\psi})(z-e^{-i\psi})q^*_{m-2}(z)\right),$$

from which it follows that $e^{i\psi}$ and $e^{-i\psi}$ are also zeros of $-\tilde{\Omega}_{2n,m}+\tilde{\Omega}^*_{2n,m}$ and $\tilde{P}_{2n,m}+\tilde{P}^*_{2n,m}$. Thus we get by (10) that there is a $j^*\in\{1,\cdots,n\}$, such that $\lambda_{j^*}=0$, which is a contradiction to $\lambda_j\neq 0$ for $j=1,\cdots,n$.

Using relation (8) we find that at the zeros $z_j=e^{i\varphi_j}$ of $\tilde{P}_{2n,m}+\tilde{P}^*_{2n,m}$ the following relation holds:

$$(\tilde{P}_{2n,m}-\tilde{P}^*_{2n,m})(z_j)(-\tilde{\Omega}_{2n,m}+\tilde{\Omega}^*_{2n,m})(z_j)=2K_{2n-1-m}z_j^{2n}|q_m(z_j)|^2.$$

Thus we obtain by setting

$$R_n(\cos\varphi)=\operatorname{Re}\{z^{-n}\tilde{P}_{2n,m}(z)\}$$

and

$$S_{n-1}(\cos\varphi)=\frac{\operatorname{Im}\{z^{-n}\tilde{P}_{2n,m}(z)\}}{\sin\varphi}$$

that

$$\lambda_j=\frac{K_{2n-1-m}|q_m(e^{i\varphi_j})|^2}{\sin\varphi_j S_{n-1}(\cos\varphi_j)\,(d/d\varphi)R_n(\cos\varphi_j)}. \tag{11}$$

Denoting by $\Delta_0^{2\pi}\arg f(e^{i\varphi})$ the net change in $\arg f(e^{i\varphi})$ as φ varies from 0 to 2π we get with the aid of (11) that

$$\Delta_0^{2\pi}\arg e^{-in\varphi}\tilde{P}_{2n,m}(e^{i\varphi})=2\pi\left(\sum_{j=1}^{n}\operatorname{sgn}\lambda_j\right). \tag{12}$$

Using the facts that the *qf* has l negative weights and that P_{2n-1-m} has all zeros in U we obtain

$$2\pi(n-2l)=\Delta_0^{2\pi}\arg q_m(e^{i\varphi})+2\pi(n-m);$$

hence

$$\Delta_0^{2\pi}\arg q_m(e^{i\varphi})=2\pi(m-2l). \tag{13}$$

Sufficiency. The assertion follows from (9), (13) and (12).

Remark 2. Let us note that the polynomial $\operatorname{Re}\{z^{-n+1}q_m(z)P_{2n-1-m}(z)\}$, $z=e^{i\varphi}$, $\varphi\in[0,\pi]$, has n simple zeros in $(-1,+1)$, if q_m has all zeros in U.

THEOREM 1′. *Let $n,k\in\mathbb{N}_0$, $k\le[(n+1)/2]$. $x_1,\cdots,x_n\in\mathbb{R}$, $-1<x_1<x_2<\cdots<x_n<1$, are the nodes of a $(2n-1-2k,n,w)$ $((2n-2k,n,w))$qf with $(n-l)$ positive weights and l negative weights if and only if $t_n(x):=\prod_{i=1}^n(x-x_i)$ has a unique representation of the form*

$$t_n(x)=r_k(x)p_{n-k}(x)-s_{k-1}(x)(1-x^2)p_{n-k-1}^{(1-x^2)}(x),$$

where $r_k\in\mathbb{P}_k$ and $s_{k-1}\in\mathbb{P}_{k-1}$ have (leading coefficient $(1-a_{2(n-k)-1})/2$ resp. $(1+a_{2(n-k)-1})/2$ and) no common zero and

$$I_{-1}^{+1}\frac{s_{k-1}}{r_k}=k-2l.$$

As usual $I_{-1}^{+1}s_{k-1}/r_k$ denotes the Cauchy index of s_{k-1}/r_k between -1 and $+1$; see e.g. [2].

Proof. Necessity. Let q_m be the unique polynomial of Theorem 1. Putting:

for $m=2k$

$$r_k(x)=2^{-k}\operatorname{Re}\{e^{-ik\varphi}q_m(e^{i\varphi})\}\quad\text{and}\quad s_{k-1}(x)=2^{-k}\frac{\operatorname{Im}\{e^{-ik\varphi}q_m(e^{i\varphi})\}}{\sin\varphi};$$

for $m=2k-1$

$$r_k(x)=(1-a_{2(n-k)-1})2^{-k}\operatorname{Re}\{e^{-i(k-1)\varphi}q_m(e^{i\varphi})\}$$

and

$$s_{k-1}(x)=(1+a_{2(n-k)-1})2^{-k}\frac{\operatorname{Im}\{e^{-i(k-1)\varphi}q_m(e^{i\varphi})\}}{\sin\varphi}$$

the assertion follows from the facts that q_m has no zero on the circumference and that

$$\begin{matrix}\Delta_0^{2\pi}\arg e^{-ik\varphi}q_{2k}(e^{i\varphi})\\ \Delta_0^{2\pi}\arg e^{-i(k-1)\varphi}q_{2k-1}(e^{i\varphi})\end{matrix}=2\pi I_{-1}^{+1}\frac{s_{k-1}}{r_k}. \tag{14}$$

Sufficiency. Setting q_m as in (6) and (7) and using relation (14) the sufficiency part is proved.

Remark 3. Using the fact that $(x^2-1)p_{n-k-1}^{(1-x^2)}$ has a unique representation of the form $(x-\alpha)p_{n-k}+\mu p_{n-k-1}$ (see [11, Thm. 2.5]) one obtains immediately (see (15) and (16)) the connection between the polynomials r_k, s_{k-1} and the polynomials f,g of Sottas and Wanner (see [10, Thm. 1]).

As a simple consequence of Theorem 1′ we obtain

COROLLARY 1. *Let* $n \geq 2k$. $x_1, \cdots, x_n \in \mathbb{R}$ *are the nodes of a positive* $(2n-1-2k, n, w)$*qf if and only if* $x_1, \cdots, x_n$ *are the zeros of a polynomial of the form*

$$r_k p_{n-k} - s_{k-1}(1-x^2)p_{n-k-1}^{(1-x^2)},$$

where r_k (s_{k-1}) *is a polynomial of degree* k $(k-1)$ *with positive leading coefficient, which has* k $(k-1)$ *simple zeros in* $(-1,+1)$ *and the zeros of* r_k *and* s_{k-1} *separate each other.*

Proof. Follows immediately from Theorem 1′ and the fact that $r_k p_{n-k} - s_{k-1}(1-x^2)p_{n-k-1}^{(1-x^2)}$ has n simple zeros in $(-1,+1)$, if r_k, s_{k-1} satisfy the above conditions.

For positive quadrature formulas we obtain additionally the following characterizations (compare also [10, Corollary to Thm. 2]).

THEOREM 2. *Let* $n \geq 2k$. *Then the following conditions are equivalent*:

(a) x_i *are the nodes of a positive* $(2n-1-2k, n, w)$ *qf.*

(b) $t_n(x) := \prod_{i=1}^{n}(x-x_i)$ *can be generated by a recurrence relation of the form*

$$t_j(x) = (x-\alpha_j')t_{j-1}(x) - \beta_j' t_{j-2}(x)$$

where α_j', $\beta_j' \in \mathbb{R}$ *satisfy the following conditions*:

$$\alpha_j' = \alpha_j \quad \beta_j' = \beta_j \quad \textit{for } j = 1, \cdots, n-k;$$

$|\alpha_j'| < 1$, $\beta_j' > 0$, $(1-\alpha_j') - \beta_j' Q_{j-1}(1) > 0$ *and* $(1+\alpha_j') + \beta_j' Q_{j-1}(-1) > 0$ *for* $j = n-k+1, \cdots, n$, *where* $Q_{j-1}(\pm 1) := t_{j-2}(\pm 1)/t_{j-1}(\pm 1)$.

(c) x_i *are the zeros of a polynomial of the form*

$$f_k(x)p_{n-k}(x) - g_{k-1}(x)p_{n-k-1}(x),$$

where $f_k(g_{k-1})$ *is a polynomial of degree* k $(k-1)$ *with positive leading coefficient, which has* k $(k-1)$ *simple zeros in* $(-1,+1)$, *the zeros of* f_k *and* g_{k-1} *separate each other,* $f_k(1) - g_{k-1}(1)Q_{n-k}(1) > 0$ *and* $\operatorname{sgn}(f_k(-1) - g_{k-1}(-1)Q_{n-k}(-1)) = (-1)^k$; $Q_{n-k}(\pm 1) = p_{n-k-1}(\pm 1)/p_{n-k}(\pm 1)$.

Proof. (a)⇒(b). Since the *qf* is positive, i.e. $\lambda_j > 0$ for $j = 1, \cdots, n$, it follows that the sequence $\{c_v\}_0^{2n-1}$ defined by

$$c_0 = 2, \qquad c_v = \sum_{j=1}^{n} \lambda_j \cos v\varphi_j \quad \text{for } v = 1, \cdots, 2n-1,$$

where $\varphi_j = \arccos x_j$, is positive definite (see [1] and [4]). Now let us denote by $\tilde{P}_{2n-1}$ that polynomial which is orthogonal with respect to the sequence $\{c_v\}_0^{2n-1}$. Since $\{c_v\}_0^{2n-1}$ is positive definite, it follows (see [4, pp. 4–5]) that $\tilde{P}_{2n-1}$ can be generated by a recurrence relation of the type

$$\tilde{P}_{v+1}(z) = z\tilde{P}_v(z) - \tilde{a}_v \tilde{P}_v^*(z), \qquad v = 0, \cdots, 2n-2,$$

with $|\tilde{a}_v| < 1$ for $v = 0, \cdots, 2n-2$. Using the fact that

$$c_v = \sum_{j=1}^{n} \lambda_j \cos v\varphi_j = 2\int_{-1}^{+1} T_v w \Big/ \int_{-1}^{+1} w$$

for $v = 0, \cdots, 2n-2k-1$, it follows that

$$\tilde{a}_v = a_v \quad \text{for } v = 0, \cdots, 2n-2k-2.$$

From [4, Thm. 31.1, Thm. 31.2] it follows that $t_n(x)=2^{-n+1}\operatorname{Re} z^{-n+1}\tilde{P}_{2n-1}(z)$ can be generated by a recurrence relation of the above type.

(b)⇒(c). Let $\tilde{g}_0(x)=1$, $\tilde{g}_1(x)=x-\alpha'_n$ and

$$\tilde{g}_j(x)=(x-\alpha'_{n+1-j})\tilde{g}_{j-1}(x)-\beta'_{n+2-j}\tilde{g}_{j-2}(x)$$

for $j=2,\cdots,k$. Putting $f_k=\tilde{g}_k$ it follows by induction that

$$t_n=f_k p_{n-k}-\beta'_{n+1-k}\tilde{g}_{k-1}p_{n-k-1}.$$

Furthermore we get from the recurrence relation in (b) that $\tilde{g}_j(1)>0$ and $\operatorname{sgn}\tilde{g}_j(-1)=(-1)^j$. The implication follows now from the recurrence relation of $\tilde{g}_j$.

(c)⇒(a). Let $y_1,\cdots,y_{k-1}$ denote the zeros of g_{k-1}. Putting

(15) $$2s_{k-1}(x)=(Q_{n-k}(1)-Q_{n-k}(-1))g_{k-1}(x)$$

and

(16)

$$2r_k(x)=2f_k(x)-[(Q_{n-k}(1)-Q_{n-k}(-1))x+(Q_{n-k}(1)+Q_{n-k}(-1))]g_{k-1}(x),$$

we obtain with the aid of [11, Thm. 2.5], that

$$r_k p_{n-k}+s_{k-1}(x^2-1)p_{n-k-1}^{(1-x^2)}=f_k p_{n-k}-g_{k-1}p_{n-k-1}.$$

Taking into consideration the facts that

$$\operatorname{sgn} r_k(y_i)=\operatorname{sgn} f_k(y_i)=(-1)^{k-i} \quad \text{for } i=1,\cdots,k-1$$

and

$$\operatorname{sgn} r_k(-1)=(-1)^k \quad \text{and} \quad r_k(+1)>0,$$

it follows that r_k has k simple zeros in $(-1,+1)$, and that the zeros of r_k and s_{k-1} separate each other. In view of Corollary 1 the implication is proved.

The following simple lemma is often useful.

LEMMA 3. *Suppose $n,m\in\mathbb{N}_0$, $n>m$. Let v_m be a real polynomial which is positive on $[-1,+1]$. If the polynomial t_n is orthogonal with respect to the weight function w/v_m, then the zeros of t_n are the nodes of a positive $(2n-1-m,n,w)$ qf.*

Proof. Since there exist positive weights $\mu_1,\cdots,\mu_n$, such that

$$\sum_{i=1}^{n}\mu_i v_m(x_i)p(x_i)=\int_{-1}^{+1}v_m p\frac{w}{v_m}$$

for all $p\in\mathbb{P}_{2n-1-m}$, the assertion follows by setting $\lambda_i=\mu_i v_m(x_i)$ for $i=1,\cdots,n$.

As a consequence of Lemma 3 we obtain (compare [7, Cor. 3])

COROLLARY 2. *Let $n,m\in\mathbb{N}_0$, $n\geq m+1$ and let $w(x)=(1-x)^\alpha(1+x)^\beta$, $\alpha,\beta\in\{\pm\frac{1}{2}\}$. If the polynomial $\sum_{j=0}^m d_j z^j$, $(d_0,\cdots,d_{m-1})\in\mathbb{R}^m$, $d_m=1$, has all zeros in $|z|<\frac{1}{2}$, then the qf based on nodes which are the zeros of the polynomial $\sum_{j=0}^m d_j p_{n-m+j}^w$ is a positive $(2n-1-m,n,w)$ qf.*

Proof. Let $q_m(z)=\sum_{j=0}^m d_j 2^{-j}z^j$. Then q_m has all zeros in the open unit disk and thus (see [11, p. 31] and [3]) $\sum_{j=0}^m d_j p_{n-m+j}^w$ is orthogonal to $\mathbb{P}_{n-1}$ with respect to the weight function $w(x)/|2^m q_m(e^{i\varphi})|^2$.

Polynomials which are orthogonal with respect to a weight function of the type w/v, where v is a polynomial which is positive on $[-1,+1]$, were studied in [8].

Next let us consider quadrature formulas with preassigned nodes (see e.g. [5, pp. 402–412]). From Corollary 1 we obtain immediately

COROLLARY 3. *Let $n,k\in\mathbb{N}_0$, $n\geq 2k$ and let $y_1,\cdots,y_{2k}\in\mathbb{R}$, $(-1<)y_1<y_2<\cdots<y_{2k}(<1)$ be given. There exists a positive $(2n-1-2k,n,w)$ qf with nodes at the given points y_i if and only if the system of linear equations*

$$\left(\sum_{j=0}^{k} A_j y_i^j\right) p_{n-k}(y_i)-\left(\sum_{j=0}^{k-1} B_j y_i^j\right)(1-y_i^2)p_{n-k-1}^{(1-x^2)}(y_i)=0, \qquad i=1,\cdots,2k,$$

$A_k+B_{k-1}=1$, has a unique solution $(A_0,\cdots,A_k,B_0,\cdots,B_{k-1})\in\mathbb{R}^{2k+1}$ such that $\sum_{j=0}^{k}A_jx^j$, $\sum_{j=0}^{k-1}B_jx^j$ satisfy the conditions of Corollary 1.

Quadrature formulas which have preassigned nodes at the zeros of p_n are of special interest (see [6] and [9]). By Theorem 2 we are able to give a full characterization of such quadrature formulas. We need the following:

LEMMA 4. *Let $n,k,j\in\mathbb{N}_0$, $n\geq k+1\geq j+2$.*

$$p_n(x)=\tilde{p}_{k,n}(x)p_{n-k}(x)-\beta_{n-k+1}\tilde{p}_{k-1,n}(x)p_{n-k-1}(x),$$

where $\tilde{p}_{k,n}$ is defined by the recurrence relation

$$\tilde{p}_{j,n}(x)=(x-\alpha_{n+1-j})\tilde{p}_{j-1,n}(x)-\beta_{n+2-j}\tilde{p}_{j-2,n}(x),$$

$\tilde{p}_{0,n}(x)=1$, $\tilde{p}_{1,n}(x)=x-\alpha_n$.

Furthermore the following relation holds:

$$\tilde{p}_{k,n}=\tilde{p}_{j,n}(x)\tilde{p}_{k-j,n-j}(x)-\beta_{n-j+1}\tilde{p}_{j-1,n}(x)\tilde{p}_{k-j-1,n-j-1}(x).$$

Proof. The first relation follows immediately from the recurrence relation of p_n.

Next let us show by induction that

$$\tilde{p}_{k,n}(x)=(x-\alpha_n)\tilde{p}_{k-1,n-1}(x)-\beta_n\tilde{p}_{k-2,n-2}(x). \tag{17}$$

For $k=2$ the assertion follows immediately. Let us assume that (17) is true for $3\leq i\leq k-1$. Using the relation

$$\tilde{p}_{k,n}(x)=(x-\alpha_{n+1-k})\tilde{p}_{k-1,n}(x)-\beta_{n+2-k}\tilde{p}_{k-2,n}(x)$$

and the inductive hypothesis for $\tilde{p}_{k-1,n}$ and $\tilde{p}_{k-2,n}$ it follows that (17) is also true for $i=k$.

Thus we have shown that the relation

$$\tilde{p}_{k,n}(x)=\tilde{p}_{j,n}(x)\tilde{p}_{k-j,n-j}(x)-\beta_{n-j+1}\tilde{p}_{j-1,n}(x)\tilde{p}_{k-j-1,n-j-1}(x) \tag{18}$$

is true for $j=1$. Assume that (18) is true for $2\leq i\leq j$. Replacing $\tilde{p}_{k-j,n-j}$ in (18) by its value from (17) we obtain that (18) is also valid for $i=j+1$.

COROLLARY 4. *Let $n,k\in\mathbb{N}_0$, $n\geq 2k$. There exists a positive $(4n+1-2k,2n+1,w)$ qf which has n nodes at the zeros of p_n if and only if there exist polynomials f_k, g_{k-1}, such that*

$$f_k\tilde{p}_{n-k,2n+1-k}-g_{k-1}\tilde{p}_{n-k-1,2n-k}=p_n$$

and f_k, g_{k-1} satisfy the conditions of Theorem 2(c).

Proof. Necessity. Let x_i, $i=1,\cdots,2n+1$, be the nodes of the positive *qf* and set $t_{2n+1}(x)=\Pi_{i=1}^{2n+1}(x-x_i)$. Then it follows in view of Theorem 2(c) that

$$t_{2n+1}=f_kp_{2n+1-k}-g_{k-1}p_{2n-k}.$$

In view of Lemma 4 we obtain that

$$\begin{aligned} t_{2n+1} = (f_k \tilde{p}_{n-k,2n+1-k} - g_{k-1}\tilde{p}_{n-k-1,2n-k})p_{n+1} \\ -\beta_{n+2}(f_k \tilde{p}_{n-k-1,2n+1-k} - g_{k-1}\tilde{p}_{n-k-2,2n-k})p_n. \end{aligned}$$

Since t_{2n+1} vanishes at the zeros of p_n, it follows that $f_k \tilde{p}_{n-k,2n+1-k} - g_{k-1}\tilde{p}_{n-k-1,2n-k}$ vanishes at the zeros of p_n.

Sufficiency. Putting $t_{2n+1} = f_k p_{2n+1-k} - g_{k-1} p_{2n-k}$ the assertion follows immediately.

Acknowledgment. I would like to thank Professor Wanner whose enquiry on the connection between the results of [10] and our paper [7] stimulated our investigations. The statement of Theorem 1 above was also suggested by him.

REFERENCES

[1] N. Ahiezer and M. Krein, *The L-problem of moments*, in Some Questions in the Theory of Moments, N. Ahiezer and M. Krein, eds., Translations of Mathematical Monographs, Vol. 2, American Mathematical Society, Providence, RI, 1962.

[2] F. R. Gantmacher, *Matrizenrechnung* II, VEB Deutscher Verlag der Wissenschaften, Berlin, 1959.

[3] Ja. L. Geronimus, *On a problem of S. N. Bernstein*, Dokl. Akad. Nauk SSSR, 229 (1976), pp. 538–541; Soviet Math. Dokl., 17 (1976), pp. 1051–1054.

[4] ———, *Polynomials orthogonal on a circle and their applications*, Zapiski Naukno-Issled. Inst. Mat. Meh. Har'kov Mat. Obsc., (4) 19 (1948), pp. 35–120, Amer. Math. Soc. Transl., (1) 3 (1962), pp. 1–78.

[5] F. B. Hildebrand, *Introduction to Numerical Analysis*, 2nd ed., McGraw-Hill, New York, 1974.

[6] G. Monegato, *An overview of results and questions related to Kronrod schemes*, in Numerische Integration, G. Hämmerlin, ed., Proc., Conf. Math. Res. Inst. Oberwolfach, 1978, ISNM 45, Birkhäuser, Basel, 1979, pp. 231–240.

[7] F. Peherstorfer, *Characterization of positive quadrature formulas*, this Journal, 12 (1981), pp. 935–942.

[8] T. E. Price, Jr., *Orthogonal polynomials for nonclassical weight functions*, SIAM J. Numer. Anal., 16 (1979), pp. 999–1006.

[9] P. Rabinowitz, *The exact degree of precision of generalized Gauss-Kronrod integration rules*, Math. Comp., 35 (1980), pp. 1275–1283.

[10] G. Sottas and G. Wanner, *The number of positive weights of a quadrature formula*, BIT 22 (1982), pp. 339–352.

[11] G. Szegö, *Orthogonal Polynomials*, 4th ed., AMS Colloquium Publications, American Mathematical Society, Providence, RI, 1967.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics
016

ON OSCILLATION PROPERTIES AND THE INTERVAL OF ORTHOGONALITY OF ORTHOGONAL POLYNOMIALS*

ERIK A. VAN DOORN[†]

Abstract. This paper is mainly concerned with the true interval of orthogonality for a sequence of orthogonal polynomials, which is the smallest closed interval containing the limit points of the set of zeros of the polynomials. We give bounds for the endpoints of this interval in terms of the coefficients in the three term recurrence formula and show them to be generalizations of most existing results. Similar findings are reported for the limit interval of orthogonality, which is defined as the smallest closed interval containing the derived set of the set of limit points. Our bounds are based upon an oscillation theorem for orthogonal polynomials which is of independent interest.

AMS-MOS subject classification (1980). Primary 42 C 05

1. Introduction. Let $\{c_n\}_{n=1}^{\infty}$ and $\{\lambda_n\}_{n=2}^{\infty}$ be sequences of real numbers and assume that λ_n is positive. Then it is a classical result that the polynomials $P_n(x)$, $n=0,1,\cdots$, defined by the recurrence formula

$$(1)\qquad \begin{aligned} &P_n(x)=(x-c_n)P_{n-1}(x)-\lambda_nP_{n-2}(x), \qquad n=1,2,\cdots,\\ &P_{-1}(x)=0, \qquad P_0(x)=1, \end{aligned}$$

where it is convenient for us to define $\lambda_1=0$, are orthogonal with respect to a (not necessarily unique) mass distribution $d\psi(x)$ on the real line. That is, there is a bounded, nondecreasing function ψ with an infinite spectrum (= support of $d\psi$) such that

$$(2)\qquad \int_{-\infty}^{\infty} P_m(x)P_n(x)\,d\psi(x)=k_n\delta_{nm} \qquad (k_n>0).$$

$P_n(x)$ has n real, distinct zeros $x_{n1}<x_{n2}<\cdots<x_{nn}$ with the property

$$(3)\qquad x_{n+1,i}<x_{ni}<x_{n+1,i+1}, \qquad i=1,2,\cdots,n,$$

so that

$$(4)\qquad \xi_i=\lim_{n\to\infty} x_{ni} \quad\text{and}\quad \eta_j=\lim_{n\to\infty} x_{n,n-j+1}$$

both exist in the extended real number system (see, e.g., [6, §I.5]). The interval $[\xi_1,\eta_1]$ is called the *true interval of orthogonality* since it is the smallest closed interval in which the support of a distribution corresponding to $\{P_n\}$ is concentrated. The *spread* of the true interval of orthogonality is defined as $\eta_1-\xi_1$, while its *centre*, defined only when $\xi_1>-\infty$ or $\eta_1<\infty$, is given by $\frac{1}{2}(\xi_1+\eta_1)$.

Regarding the finiteness of ξ_1, we will have use for a criterion which is essentially due to Stieltjes [20] and elaborated by Chihara [1]. Namely, in order that $\xi_1\geq A>-\infty$, it is necessary and sufficient that there exist numbers γ_n such that

$$(5)\qquad c_n-A=\gamma_{2n-2}+\gamma_{2n-1} \quad\text{and}\quad \lambda_{n+1}=\gamma_{2n-1}\gamma_{2n}, \qquad n>0,$$

*Received by the editors September 24, 1982, and in revised form March 24, 1983.

[†] Netherlands Postal and Telecommunications Services, Dr. Neher Laboratories, P. O. Box 421, 2260 AK Leidschendam, the Netherlands. Present address, Stichting Mathematisch Centrum, Kruislaan 413, 1098 SJ Amsterdam, the Netherlands.

where $\gamma_0 \geq 0$ and $\gamma_n > 0$ for $n > 0$. Here $\gamma_0 \geq 0$ may be replaced by $\gamma_0 = 0$, since the existence of a sequence $\{\gamma_n\}$ satisfying (5) and $\gamma_0 > 0$ implies the existence of a sequence $\{\gamma_n'\}$ satisfying (5) and $\gamma_0' = 0$ (or, in fact, any number between 0 and γ_0). When (5) holds one also has $\eta_1 = \infty$ if and only if $\{\gamma_n\}$ is unbounded.

From (3) and (4) we obviously have $\xi_i \leq \xi_{i+1} < \eta_{j+1} \leq \eta_j$, so that

$$(6) \qquad \sigma = \lim_{i \to \infty} \xi_i \quad \text{and} \quad \tau = \lim_{j \to \infty} \eta_j$$

exist, again allowing for $\pm\infty$. It is important to note at this point that

$$(7) \qquad \xi_{i+1} = \xi_i \Rightarrow \sigma = \xi_i, \qquad i = 0, 1, \cdots$$

and

$$(8) \qquad \eta_{j+1} = \eta_j \Rightarrow \tau = \eta_j, \qquad j = 0, 1, \cdots,$$

where $\xi_0 \equiv -\infty$, $\eta_0 \equiv \infty$ (see, e.g., [6, Thm. II.4.6]).

It can be shown [6, Thm. III.4.2] that the sets of orthogonal polynomials $\{P_n^{(k)}(x)\}_n$, $k = 0, 1, \cdots$, which are determined through the recurrence formula (1) by the sequences $\{c_n^{(k)} = c_{n+k}\}_{n=1}^{\infty}$ and $\{\lambda_n^{(k)} = \lambda_{n+k}\}_{n=2}^{\infty}$, have true intervals of orthogonality $[\xi_1^{(k)}, \eta_1^{(k)}]$ with the properties

$$(9) \qquad \xi_1^{(k)} \leq \xi_1^{(k+1)} \leq \sigma \quad \text{and} \quad \tau \leq \eta_1^{(k+1)} \leq \eta_1^{(k)}, \qquad k = 0, 1, \cdots.$$

Further, the next theorem is easily seen to hold as a consequence of [6, Thms. IV.2.1 and IV.3.2].

THEOREM 1.

$$\xi_1^{(k)} \to \sigma \quad \textit{and} \quad \eta_1^{(k)} \to \tau, \qquad k \to \infty.$$

We emphasize that σ and τ are determined only by the limiting behaviour of the parameter sequences $\{c_n\}$ and $\{\lambda_n\}$, so that any finite number of changes in the parameter values has no influence on the values of σ and τ. In view of this fact, we are justified in calling $[\sigma, \tau]$ the *limit interval of orthogonality*. The *spread* and the *centre* of the limit interval of orthogonality are defined as $\tau - \sigma$ and $\frac{1}{2}(\sigma + \tau)$, respectively, provided these quantities are meaningful.

It is the purpose of this paper to give bounds on the true and limit intervals of orthogonality in terms of the parameters c_n and λ_n. Our main tool will be the oscillation theorem for orthogonal polynomials given in §2, which is of independent interest. An extension of this result will be derived in the Appendix.

We note that any result on ξ_1 (or σ), e.g., Stieltjes' criterion (5), may be transformed into a result on η_1 (or τ) and vice versa by considering the polynomials $\bar{P}_n(x) = (-1)^n P_n(-x)$, which satisfy the recurrence relation (1) with parameter sequences $\{\bar{c}_n = -c_n\}$ and $\{\bar{\lambda}_n = \lambda_n\}$. Therefore, as far as the endpoints are concerned, we shall concentrate only on one side of the intervals of orthogonality. In fact, upper bounds on ξ_1 and σ will be given in §3 and lower bounds in §4. Several known results will appear as corollaries to our theorems. We remark that some of these known results are given in the literature under the condition that the distribution $d\psi$ with respect to which the polynomials P_n are orthogonal is unique. This is because they are stated (or derived) in terms of supporting points of $d\psi$ instead of limit points of zeros of the polynomials P_n, while both points of view are equivalent only if $d\psi$ is unique (cf. [3] and [6, Chap. II]).

In the final section, some bounds will be derived on spread and centre of the true and limit intervals of orthogonality and these will be compared with existing results.

2. The basic oscillation theorem. We need some preliminary results and notation first. Let $\mathbf{u}=\{u_0,u_1,\cdots,u_n,\cdots\}$ be an infinite sequence of real numbers. The finite sequence consisting of the first $n+1$ elements of $\mathbf{u}$ will be denoted by $\mathbf{u}_{(n)}$, i.e., $\mathbf{u}_{(n)}=\{u_0,u_1,\cdots,u_n\}$. By $S(\mathbf{u}_{(n)})$, we denote the number of sign changes in the sequence $\mathbf{u}_{(n)}$ by deleting all zero terms, with the special convention $S(\mathbf{0}_{(n)})=-1$, $\mathbf{0}_{(n)}$ denoting the sequence consisting of $n+1$ zeros. We let $S(\mathbf{u})=\lim_{n\to\infty}S(\mathbf{u}_{(n)})$, which exists but, of course, may be infinite.

Our next prerequisite concerns Sturmian sequences of polynomials. We recall the definition (see [17, pp. 7–8]).

DEFINITION 1. A sequence of $n+1$ polynomials $\{R_0,R_1,\cdots,R_n\}$, $n>0$, is called a *Sturmian sequence* on the interval (a,b) if these four conditions are satisfied:

(i) $R_n(x)\neq 0$ for $x=a,\ b$,

(ii) $R_0(x)\neq 0$ for all $x\in[a,b]$,

(iii) $R_i(x)=0\ (0<i<n)\,\&\,x\in[a,b]\Rightarrow R_{i-1}(x)R_{i+1}(x)<0$,

(iv) $R_n(x)=0\,\&\,x\in[a,b]\Rightarrow R_{n-1}(x)R'_n(x)>0$.

This definition is justified by the following theorem [17, Satz 7].

THEOREM 2 (Sturm's theorem). *If the sequence of polynomials* $\{R_0,R_1,\cdots,R_n\}$ *is a Sturmian sequence on the interval* (a,b), *then the number of zeros of* R_n *in the interval* (a,b) *equals* $S(\mathbf{R}(a))-S(\mathbf{R}(b))$, *where* $\mathbf{R}(x)=\{R_0(x),R_1(x),\cdots,R_n(x)\}$.

The relevance of this theorem for this paper resides in the next lemma, which concerns the sequence of orthogonal polynomials $\{P_0,P_1,\cdots,P_n,\cdots\}$ defined by the recurrence relation (1).

LEMMA 1. *The sequence* $\mathbf{P}_{(n)}=\{P_0,P_1,\cdots,P_n\}$, *where* $n>0$, *is a Sturmian sequence on any interval* (a,b) *where* $P_n(a)\neq 0$ *and* $P_n(b)\neq 0$.

Proof. See [21, p. 45].

We are now in a position to state our basic result.

THEOREM 3 (basic oscillation theorem). *For the polynomials* $\{P_n\}_{n=0}^{\infty}$ *defined by the recurrence relation* (1) *one has*:

(i) $S(\mathbf{P}(x))=k\Leftrightarrow\eta_{k+1}\leq x<\eta_k,\ k=0,1,\cdots,$

(ii) $S(\mathbf{P}(x))=\infty\Leftrightarrow x<\tau$ *or* $x=\tau<\eta_j$ *for all* j,

(iii) $S(\tilde{\mathbf{P}}(x))=k\Leftrightarrow\xi_k<x\leq\xi_{k+1},\ k=0,1,\cdots,$

(iv) $S(\tilde{\mathbf{P}}(x))=\infty\Leftrightarrow x>\sigma$ *or* $x=\sigma>\xi_i$ *for all* i,

where $\mathbf{P}(x)=\{P_0(x),P_1(x),\cdots\}$, $\tilde{\mathbf{P}}(x)=\{\tilde{P}_0(x),\tilde{P}_1(x),\cdots\}$ *and* $\tilde{P}_n(x)=(-1)^nP_n(x)$.

Proof. It is evident that (ii) and (iv) are implied by (i) and (iii), respectively, while (iii) readily follows from (i) by considering the polynomials $\tilde{P}_n(x)=(-1)^nP_n(-x)$ mentioned in the introduction. So it remains to prove (i).

To this end, let x and n be such that $P_n(x)\neq 0$. Choose η such that $\max(x,x_{nn})<\eta<\eta_0\equiv\infty$. By (3) we then have $\eta>x_{ii}$ $(i=1,2,\cdots,n)$, and (1) subsequently implies $P_i(\eta)>0$ for $i=0,1,\cdots,n$, whence $S(\mathbf{P}_{(n)}(\eta))=0$. Now applying Sturm's theorem to $\mathbf{P}_{(n)}$ in the interval (x,η), we get $S(\mathbf{P}_{(n)}(x))-S(\mathbf{P}_{(n)}(\eta))=$number of zeros of P_n in (x,η), i.e.,

$$S\big(\mathbf{P}_{(n)}(x)\big)=\text{number of zeros of } P_n \text{ in } (x,\infty). \tag{10}$$

Letting n tend to infinity in (10), (i) emerges as a consequence of (3) and (4). □

Aspects of the basic oscillation theorem may be found in the literature under various guises. Thus a special case of it was employed by Stieltjes [20, p. 564] in the context of continued fractions, while parts (ii) and (iv) of the theorem are essentially contained in [23, Thm. 8(a)] in the context of difference equations. Further, by making the identification

$$P_n(x)=\det(A_n-xI_n), \tag{11}$$

where I_n is the $n\times n$ identity matrix and

$$A_n=\begin{bmatrix} c_1 & \sqrt{\lambda_2} & & 0 \\ \sqrt{\lambda_2} & \ddots & \ddots & \\ & \ddots & \ddots & \sqrt{\lambda_{n-1}} \\ 0 & & \sqrt{\lambda_{n-1}} & c_n \end{bmatrix}, \tag{12}$$

our questions regarding (essentially) the zeros x_{nk} may be put in terms of eigenvalues of symmetric tridiagonal matrices for which the Sturmian approach is well known (see, e.g., [16, Chap. 7]). Indeed, we shall repeatedly make use of this identification to obtain new results or point out alternative proofs.

In closing this section, we remark that Chihara ([1], [4], see also [6]) has obtained characterizations for ξ_1, η_1, σ and τ which are in appearance quite different from the basic oscillation theorem. A third characterization, which may be conceived as a consequence of Chihara's results, has been stated and given an independent proof by Whitehurst [22, Chap. 4]. It is not very difficult to prove directly the equivalence of Chihara's or Whitehurst's results and the basic oscillation theorem.

3. Upper bounds on ξ_1 and σ. Our starting point in this section will be a lemma concerning the system of equations

$$z_n+a_nz_{n-1}+b_nz_{n-2}=0, \qquad n=1,2,\cdots. \tag{13}$$

LEMMA 2. *If the system of equations* (13), *where* $b_n>0$, *possesses a solution* $z_{-1},z_0,z_1,\cdots$ *satisfying* $z_nz_{n+1}<0$ *for* $n\geq N\geq 0$, *then*

$$a_M+\sum_{m=M+1}^{M+k}\left(a_m-2\sqrt{b_m}\right)>0 \tag{14}$$

for any two integers $k\geq 0$ *and* $M>N+1$ ($M\geq N+1$ *if* $z_{N-1}=0$).

Proof. Assuming that a given solution has $z_m\neq 0$ for $m=M-1,M,\cdots,M+k-1$, we can write down the equalities

$$a_M=-\frac{z_M}{z_{M-1}}-b_M\frac{z_{M-2}}{z_{M-1}},$$

and, for $m=M,M+1,\cdots,M+k-1$,

$$a_{m+1}-2\sqrt{b_{m+1}}=\frac{z_m}{z_{m-1}}-\frac{z_{m+1}}{z_m}-\frac{\left(z_m+z_{m-1}\sqrt{b_{m+1}}\right)^2}{z_{m-1}z_m}.$$

Summing these $k+1$ equalities yields

$$a_M+\sum_{m=M+1}^{M+k}\left(a_m-2\sqrt{b_m}\right)=-\left\{\frac{z_{M+k}}{z_{M+k-1}}+b_M\frac{z_{M-2}}{z_{M-1}}+\sum_{m=M}^{M+k-1}\frac{\left(z_m+z_{m-1}\sqrt{b_{m+1}}\right)^2}{z_{m-1}z_m}\right\},$$

from which the lemma follows at once. □

Returning to the recurrence formula (1), we let x be any real number,

$$(15)\qquad y_n=P_n(x),\qquad n=-1,0,1,\cdots,$$

and $\mathrm{y}=\{y_0,y_1,\cdots\}$. Further, let $\{\chi_1,\chi_2,\cdots\}$ be any sequence of positive numbers and define

$$(16)\qquad z_{-1}=0,\quad z_0=1\quad\text{and}\quad z_n=(\chi_1\chi_2\cdots\chi_n)^{-1}y_n,\, n>0.$$

If we let b_1 be positive but otherwise arbitrary,

$$(17)\qquad a_n=(c_n-x)/\chi_n\quad\text{and}\quad b_{n+1}=\lambda_{n+1}/(\chi_n\chi_{n+1}),\qquad n>0,$$

then $\{z_n\}_{n=-1}^{\infty}$ satisfies the recurrence relation (13) with $b_n>0$, so that Lemma 2 applies. Translating this result in terms of y_n, c_n, λ_n, χ_n and x yields

$$(18)\qquad \frac{c_M}{\chi_M}+\sum_{m=M+1}^{M+k}\left(\frac{c_m}{\chi_m}-2\left(\frac{\lambda_m}{\chi_{m-1}\chi_m}\right)^{1/2}\right)>x\sum_{m=M}^{M+k}\frac{1}{\chi_m}$$

for $k\geq 0$ and $M>N+1$ ($M\geq N+1$ if $y_{N-1}=0$), whenever $y_ny_{n+1}<0$ for $n\geq N\geq 0$.

By the basic oscillation theorem one has $x\leq\xi_1$ if and only if $S(\tilde{\mathrm{y}})=0$. That is, $x\leq\xi_1$ if and only if $y_ny_{n+1}<0$ for $n\geq 0$, since $y_n=0$ is clearly impossible when $x\leq\xi_1$. Further noting that $y_{-1}=0$, we conclude that the inequality $x\leq\xi_1$ implies the inequalities (18) for all $k\geq 0$ and $M>0$. From this result one easily deduces the following theorem.

THEOREM 4. *For any sequence of positive numbers* $\{\chi_1,\chi_2,\cdots\}$ *and integers* $k\geq 0$ *and* $M>0$ *one has*

$$(19)\qquad \xi_1<\left(\frac{c_M}{\chi_M}+\sum_{m=M+1}^{M+k}\left(\frac{c_m}{\chi_m}-2\left(\frac{\lambda_m}{\chi_{m-1}\chi_m}\right)^{1/2}\right)\right)\left(\sum_{m=M}^{M+k}\frac{1}{\chi_m}\right)^{-1}.$$

Taking $k=0$ and $\chi_n=1$ for all n, we obtain Corollary 4.1, which is also a direct consequence of Stieltjes' criterion (5) and therefore well known (see, e.g., [6, p. 109]).

COROLLARY 4.1.

$$\xi_1<c_n,\qquad n=1,2,\cdots.$$

Letting $k=1$ and $\chi_n=1$ for all n, a result emerges which was first given (with an error) by Maki [11] and later improved by Chihara [5].

COROLLARY 4.2.

$$\xi_1<\frac{1}{2}(c_n+c_{n+1})-\sqrt{\lambda_{n+1}},\qquad n=1,2,\cdots.$$

We remark that the other part of the Maki–Chihara result to the effect that $\frac{1}{2}(c_n+c_{n+1})-\sqrt{\lambda_{n+1}}$ is unbounded when $\xi_1>-\infty$ and $\eta_1=\infty$, can also be generalized in the spirit of Theorem 4, at least when $\chi_n=1$ for all n. One should simply use Maki's argument on the basis of which lies the result of Stieltjes mentioned in the introduction.

Assuming that $\inf\{c_n\} > -\infty$, we can choose $k=1$ and $\chi_n = c_n - c$ in (19), where c is any number smaller than c_n for all n. After some rearranging, we then get

$$(20)\quad \xi_1 < c + 2\frac{(c_n-c)(c_{n+1}-c)-(\lambda_{n+1}(c_n-c)(c_{n+1}-c))^{1/2}}{c_n+c_{n+1}-2c}, \qquad n=1,2,\cdots.$$

In combination with Corollary 4.1, this result yields a useful third corollary. Namely, if there are values of $\zeta_n \equiv \frac{1}{2}(c_n+c_{n+1}-((c_n-c_{n+1})^2+4\lambda_{n+1})^{1/2})$, $n=1,2,\cdots$, with the property $\zeta_n < c_m$ for all m, we can choose c equal to any of those ζ_n, ζ_1 say, after which the choice $n=1$ yields that $\xi_1 < \zeta_1$. Hence, in this case, $\xi_1 < \zeta_n$ for all n. If, on the other hand, $\zeta_n > c_m$ for some m and all n, Corollary 4.1 implies that the same conclusion holds. Thus, we have the following result, which is sharper than Corollary 4.2, while involving the same parameters.

COROLLARY 4.3.

$$\xi_1 < \frac{1}{2}\left(c_n + c_{n+1} - \left((c_{n+1}-c_n)^2 + 4\lambda_{n+1}\right)^{1/2}\right), \qquad n=1,2,\cdots.$$

We note that upper bounds for ξ_1 can be obtained on the basis of the interpretation (11) for $P_n(x)$. Namely, considering that the eigenvalues of A_n equal those of $K_nA_nK_n$, where K_n is the $n\times n$ matrix consisting of elements $k_{ij}=1$ when $i+j=n+1$ $(i,j=1,2,\cdots,n)$ and 0 elsewhere, one also has

$$(21)\qquad P_n(x) = \det(K_nA_nK_n - xI_n).$$

Hence, we can identify $P_n(x)$ with the nth polynomial in an orthogonal sequence $\{\hat{P}_m(x)\}$ determined by the recurrence formula (1) through the parameters $\hat{c}_m = c_{n+1-m}$ $(m\le n)$, $\hat{c}_m = c_m$ $(m>n)$, $\hat{\lambda}_m = \lambda_{n+2-m}$ $(m \le n+1)$ and $\hat{\lambda}_m = \lambda_m$ $(m>n+1)$. It now follows from (3) and (4) that

$$(22)\qquad \xi_1 < x_{n1} = \hat{x}_{n1} < \hat{x}_{k1}, \qquad k=1,2,\cdots,n-1,$$

where $\hat{x}_{m1}$ denotes the smallest zero of $\hat{P}_m(x)$. However, the only practical bounds obtained by this approach are $\xi_1 < \hat{x}_{11}$, but this gives Corollary 4.1, and $\xi_1 < \hat{x}_{21}$, which amounts to Corollary 4.3.

Remark. A third proof of Corollary 4.3 may be given on the basis of Chihara's characterization for ξ_1 (cf. [6, Thm. IV.2.1]).

The arguments leading to Theorem 4 need only slight modification to obtain results on the limit interval of orthogonality. For by the basic oscillation theorem we have $x<\sigma$ only if $S(\bar{y})$ is finite; that is, only if $y_ny_{n+1}<0$ for n sufficiently large (by definition of σ, $y_n=0$ occurs for at most finitely many n if $x<\sigma$). Hence the inequality $x<\sigma$ implies the inequality (18) for M sufficiently large and all $k\ge 0$. From this it is easy to derive Theorem 5, which, however, also derives directly from the Theorems 1 and 4.

THEOREM 5. *For any sequence of positive numbers* $\{\chi_1,\chi_2,\cdots\}$ *and integer* $k\ge 0$, *one has*

$$(23)\quad \sigma \le \liminf_{M\to\infty}\left\{\left(\frac{c_M}{\chi_M} + \sum_{m=M+1}^{M+k}\left(\frac{c_m}{\chi_m} - 2\left(\frac{\lambda_m}{\chi_{m-1}\chi_m}\right)^{1/2}\right)\right)\left(\sum_{m=M}^{M+k}\frac{1}{\chi_m}\right)^{-1}\right\}.$$

Taking $k=0$ and χ_n arbitrary, we get the analogue of Corollary 4.1, which has been obtained previously by Wouk [23, last inequality of Thm. 8(e)] and Chihara [1, Thm. 6]; see also [6, Thm. IV.3.1].

COROLLARY 5.1.

$$\sigma \leq \liminf_{n\to\infty} \{c_n\}.$$

We also state as a corollary the analogue of Corollary 4.3, although its proof is most conveniently given via Theorem 1 and Corollary 4.3.

COROLLARY 5.2.

$$\sigma \leq \liminf_{n\to\infty} \frac{1}{2}\left\{c_n + c_{n+1} - \left((c_n - c_{n+1})^2 + 4\lambda_{n+1}\right)^{1/2}\right\}.$$

An interesting case arises when we let k tend to infinity in Theorem 5. However, we had better do this not in (23), but at at earlier stage in the reasoning leading to Theorem 5. Namely, from Theorem 4 we see that for all $M>0$

$$\xi_1 \leq \lim_{k\to\infty} \inf\{f(M,k)\},$$

where $f(M,k)$ denotes the expression between braces in (23). Hence, by Theorem 1,

$$\sigma \leq \lim_{M\to\infty} \inf\left\{ \lim_{k\to\infty} \inf\{f(M,k)\}\right\}. \tag{24}$$

Now let us assume that $\Sigma\chi_n^{-1} = \infty$. Then, evidently, $\liminf_{k\to\infty}\{f(M,k)\} = \liminf_{k\to\infty}\{f(1,k)\}$, so that we obtain the next theorem.

THEOREM 6. *For any sequence of positive numbers* $\{\chi_0, \chi_1, \cdots\}$ *such that* $\Sigma\chi_n^{-1} = \infty$, *one has*

$$\sigma \leq \lim_{k\to\infty} \inf\left\{\left(\sum_{m=1}^{k}\left(\frac{c_m}{\chi_m} - 2\left(\frac{\lambda_m}{\chi_{m-1}\chi_m}\right)^{1/2}\right)\right)\left(\sum_{m=1}^{k}\frac{1}{\chi_m}\right)^{-1}\right\}. \tag{25}$$

Taking $\chi_n = 1$ for all n, we obtain the important Corollary 6.1, which has been given previously by Wouk [23, Thm. 8(g)].

COROLLARY 6.1.

$$\sigma \leq \lim_{k\to\infty} \inf\left\{\frac{1}{k}\sum_{m=1}^{k}\left(c_m - 2\sqrt{\lambda_m}\right)\right\}.$$

4. Lower bounds on ξ_1 and σ. As in the previous section we start our discussion by considering the system of equations (13). If we plot a solution $z_{-1}, z_0, z_1, \cdots$ of this system by joining successive coordinates (i, z_i) by straight line segments, then the points where such a line segment meets the x-axis will be called a node of the solution. We can now cite the following classical result [14].

LEMMA 3 (Sturm's separation theorem for difference equations). *For any system of equations* (13) *where* $b_n>0$, *the nodes of any two linearly independent solutions separate each other.*

Suppose $a_n + 1 < -b_n < 0$ for $n > N \geq 0$ and let two arbitrary numbers $\hat{z}_N > \hat{z}_{N-1} \geq 0$ determine a solution $\{\hat{z}_n\}_{-1}^{\infty}$ of (13). Then we have by induction

$$\hat{z}_n - \hat{z}_{n-1} = -(a_n+1)(\hat{z}_{n-1} - \hat{z}_{n-2}) - (a_n + b_n + 1)\hat{z}_{n-2} > 0$$

for $n > N$. Lemma 3 now implies that any solution $\{z_n\}$ of (13) has at most one node in the interval $[N-1, \infty)$. Hence, also noting that $z_n z_{n-2} \leq 0$ if $z_{n-1} = 0$, we can state the following lemma, which is also essentially contained in [9].

LEMMA 4. *If $a_n+b_n+1<0$ and $b_n>0$ for $n>N$, then any nontrivial solution $\{z_n\}$ of* (13) *for which $z_{m-1}z_m\le 0$ for some $m\ge N$ has the property that* $\operatorname{sign}(z_{m+k})=\operatorname{sign}(z_m)$ *if $z_m\ne 0$, and* $=\operatorname{sign}(-z_{m-1})$ *if $z_m=0$, for all $k>0$.*

Back to our orthogonal system (1) we let x be any real number and define the quantities y_n as in (15). Further, we let $\{\chi_0,\chi_1,\cdots\}$ be any sequence of positive numbers and define

$$(26)\qquad z_{-1}=0,\ z_0=1\quad\text{and}\quad z_n=(-1)^n(\chi_1\chi_2\cdots\chi_n)^{-1}y_n,\qquad n>0.$$

Finally, we let b_1 be positive,

$$(27)\qquad a_n=-(c_n-x)/\chi_n\quad\text{and}\quad b_{n+1}=\lambda_{n+1}/(\chi_n\chi_{n+1}),\qquad n>0.$$

Then $\{z_n\}$ satisfies the recurrence relation (13) with $b_n>0$, so that the second condition in Lemma 4 is satisfied for $n>0$. In terms of c_n, λ_n, χ_n and x, the first condition in this lemma reads

$$(28)\qquad c_n-\frac{\lambda_n}{\chi_{n-1}}-\chi_n>x,$$

provided $n>1$. Supposing (28) to be valid for $n>0$, we can choose $b_1>0$ so small that $a_n+b_n+1<0$ for $n>0$. Hence, Lemma 4 applies and we have $\operatorname{sign}((-1)^k y_k)=\operatorname{sign}(z_k)=\operatorname{sign}(z_0)=1$, since $z_{-1}z_0=0$. Thus, by the basic oscillation theorem, $x\le\xi_1$. A trivial argument subsequently leads to our next theorem.

THEOREM 7. *For any sequence of positive numbers $\{\chi_0,\chi_1,\cdots\}$, one has*

$$(29)\qquad \inf_{n\ge 1}\left\{c_n-\frac{\lambda_n}{\chi_{n-1}}-\chi_n\right\}\le\xi_1.$$

Remark. This theorem may also be obtained via the identification (11) for $P_n(x)$. Namely, the zeros $x_{n1},x_{n2},\cdots,x_{nn}$ of $P_n(x)$ are the eigenvalues of A_n; and therefore, also of the matrix $\Phi_n^{-1}A_n\Phi_n$, where $\Phi_n=\operatorname{diag}(\phi_1,\phi_2,\cdots,\phi_n)$ and $\phi_i>0$. With Gershgorin's theorem (see [12, p. 146]), one may subsequently prove that

$$(30)\qquad x_{n1}\ge\min_{i\le n}\left\{c_i-\frac{\phi_{i-1}}{\phi_i}\sqrt{\lambda_i}-\frac{\phi_{i+1}}{\phi_i}\sqrt{\lambda_{i+1}}\right\},$$

where $\phi_0=1$, say. Taking $\{\phi_i\}$ such that $\phi_{i+1}=\chi_i\phi_i/\sqrt{\lambda_{i+1}}$ and letting n tend to infinity yields (29).

Various consequences of Theorem 7 suggest themselves; e.g., one could take $\chi_n=1$ for all n, or, $\chi_0=1$ and $\chi_n=\lambda_{n+1}$ $(n>0)$, the latter result being implicit in Maki [11]. We will explicitly state as a corollary the case $\chi_0=1$ and $\chi_n=\sqrt{\lambda_{n+1}}$ $(n>0)$, since this result improves directly upon Lemma 3 of Nevai [15, p. 21].

COROLLARY 7.1.

$$\inf_{n\ge 1}\left\{c_n-\sqrt{\lambda_n}-\sqrt{\lambda_{n+1}}\right\}\le\xi_1.$$

By choosing $\chi_0=1$ and $\chi_n=\lambda_{n+1}/(c_{n+1}-\phi_{n+1})$ $(n>0)$, where $\phi_n<c_n$ $(n>1)$, we obtain the following useful, alternative formulation of Theorem 7.

THEOREM 7′. *For any sequence $\{\phi_1,\phi_2,\cdots\}$, with $\phi_1\le c_1$ and $\phi_n<c_n$ $(n>1)$, one has*

$$(31)\qquad \inf_{n\ge 1}\{\phi_n-\lambda_{n+1}/(c_{n+1}-\phi_{n+1})\}\le\xi_1.$$

Thus formulated, Theorem 7 is seen to improve upon a result of Léopold [10], specified for the present context, which amounts to (31) with a fixed value ϕ ($\le c_n$ for all n) for all ϕ_n.

As a final lower bound for ξ_1, we mention a theorem of Chihara. Actually, Chihara gives the corresponding result for σ, but his argument applies equally well here (cf. [2], [4] and [6, Thm. IV.3.3]).

THEOREM 8 (Chihara). *For any chain sequence* $\{\beta_n\}_{n=1}^{\infty}$, *one has*

$$\inf_{n\ge 1}\frac{1}{2}\left\{c_n+c_{n+1}-\left((c_{n+1}-c_n)^2+4\lambda_{n+1}/\beta_n\right)^{1/2}\right\}\le\xi_1. \tag{32}$$

Remark. $\{\beta_n\}_{n=1}^{\infty}$ is a *chain sequence* if there exists a sequence $\{g_k\}_{k=0}^{\infty}$ with $0\le g_0<1$ and $0<g_k<1$ ($k>0$), such that $\beta_n=(1-g_{n-1})g_n$; $\{g_k\}$ is called a *parameter sequence* for $\{\beta_n\}$. For instance, $\{\frac{1}{4}\}$ is a chain sequence for which $\{\frac{1}{2}\}$ is a parameter sequence.

Remark. Theorems 7 and 8 are in a sense best possible since equality may be obtained in (29) and (32). To this end, one should take $\beta_n=\alpha_n(\xi_1)\equiv\lambda_{n+1}/((c_{n+1}-\xi_1)(c_n-\xi_1))$ (which is a chain sequence according to [6, Thm. IV.2.1]) in (32) and $\chi_n=(c_n-\xi_1)(1-g_{n-1})$, with $\{g_k\}$ a parameter sequence for $\{\alpha_n(\xi_1)\}$, in (29). Thus we have actually obtained new characterizations for the true interval of orthogonality.

Using an argument similar to that for Theorem 7 or, alternatively, exploiting Theorems 1 and 7, one easily produces the following general lower bound for σ.

THEOREM 9. *For any sequence of positive numbers* $\{\chi_0,\chi_1,\cdots\}$, *one has*

$$\liminf_{n\to\infty}\left\{c_n-\frac{\lambda_n}{\chi_{n-1}}-\chi_n\right\}\le\sigma. \tag{33}$$

We will explicitly state as a corollary of Theorem 9 the case where $\chi_n=\sqrt{\lambda_{n+1}}$ for $n>0$.

COROLLARY 9.1.

$$\liminf_{n\to\infty}\left\{c_n-\sqrt{\lambda_n}-\sqrt{\lambda_{n+1}}\right\}\le\sigma.$$

The latter result has been given by Wouk [23, Thm. 8(f)], while it is a slight generalization of a result of Chihara [2, p. 704]; see also Nevai [15, p. 22].

In this context we remark that the proof and subsequent formulation of another one of Wouk's results [23, Thm. 8(h)] contains an error. The corrected version of this theorem is an easy consequence of the above corollary.

For completeness' sake we finally mention the analogue to Theorem 8, Chihara's lower bound for σ.

THEOREM 10 (Chihara [2], [4], see also [6, Thm. IV.3.3]). *For any chain sequence* $\{\beta_n\}$

$$\liminf_{n\to\infty}\frac{1}{2}\left\{c_n+c_{n+1}-\left((c_{n+1}-c_n)^2+4\lambda_{n+1}/\beta_n\right)^{1/2}\right\}\le\sigma. \tag{34}$$

Remark. It can be shown that the left-hand sides of (33) and (34) can be made arbitrarily close to σ by a suitable choice of $\{\chi_n\}$ and $\{\beta_n\}$, respectively.

5. Bounds on spread and centre. As mentioned in the introduction, we can straightforwardly produce lower (upper) bounds for η_1 (or τ) on the basis of upper (lower) bounds for ξ_1 (or σ) by considering the polynomials $\bar{P}_n(x)=(-1)^nP_n(-x)$

which are determined by the recurrence formula (1) via the parameters $\bar{c}_n = -c_n$ and $\bar{\lambda}_n = \lambda_n$, and thus have $[-\eta_1, -\xi_1]$ ($[-\tau, -\sigma]$) as their true (limit) interval of orthogonality. Then various upper (lower) bounds on the spread of the true (or limit) interval of orthogonality may be obtained by combining upper (lower) bounds for ξ_1 (or σ) with lower (upper) bounds for η_1 (or τ). Similarly, we should combine upper (lower) bounds for ξ_1 (or σ) with upper (lower) bounds for η_1 (or τ) to obtain upper (lower) bounds on the centre of the true (or limit) interval of orthogonality. We will not pursue this approach in any detail except that we show how known results on the spread of the true interval of orthogonality may be reproduced in this way. Also, we show that additional information on the centre of the true (or limit) interval of orthogonality may be obtained by exploiting Stieltjes' criterion (5).

Let us first note that as a consequence of Corollary 4.3 and its dual result for η_1, we have the following theorem, which is essentially due to Mirsky [13], who states it in a finite eigenvalue context (the term *spread* is taken from Mirsky).

THEOREM 11.

$$\eta_1 - \xi_1 > \left((c_{n+1} - c_n)^2 + 4\lambda_{n+1} \right)^{1/2}, \qquad n = 1, 2, \cdots .$$

This is the simplest result combining parameters c_n and λ_n. A bound involving only c_n's, which is not necessarily worse than Theorem 11, is

$$(35) \qquad \eta_1 - \xi_1 > c_n - c_m, \qquad n, m = 1, 2, \cdots,$$

which follows from Corollary 4.1. However, Theorem 11 does improve upon a result involving only λ_n's which, together with (35), was given already by Shohat [18], [19], viz.,

$$(36) \qquad \eta_1 - \xi_1 > 2\sqrt{\lambda_n}, \qquad n = 2, 3, \cdots .$$

But then, the latter inequality can be sharpened in another direction on the basis of (19) (with $\chi_n \equiv 1$) as follows.

THEOREM 12. *For any two integers $k > 0$ and $M \geq 0$, one has*

$$(37) \qquad \eta_1 - \xi_1 > \frac{4}{k+1} \sum_{m=M+1}^{M+k} \sqrt{\lambda_m} .$$

In particular, it follows that $\eta_1 - \xi_1 \geq 4\sqrt{\lambda}$ when $\lambda_m \to \lambda$ as $m \to \infty$.

So much for the spread.

Regarding the centre of the true interval of orthogonality, let us assume $\eta_1 < \infty$. Then, by Stieltjes' criterion (in dual form), we have

$$-c_n = -\eta_1 + \gamma_{2n-2} + \gamma_{2n-1}, \qquad \lambda_{n+1} = \gamma_{2n-1}\gamma_{2n}$$

for $n > 0$, where $\gamma_0 = 0$ and $\gamma_n > 0$ for $n > 0$. For convenience, we define $\gamma_{-1} = 1$. By (29) we then get

$$(38) \qquad \inf_{n \geq 1} \left\{ \eta_1 - \gamma_{2n-2} - \gamma_{2n-1} - \frac{\gamma_{2n-3}\gamma_{2n-2}}{\chi_{n-1}} - \chi_n \right\} \leq \xi_1 .$$

Subsequently, substituting $\chi_n = \gamma_{2n-1}$ for $n \geq 0$ yields

$$(39) \qquad -\eta_1 + 2\inf\{c_n\} \leq \xi_1 .$$

Combining this inequality and its dual result, we obtain the next theorem.

THEOREM 13. *If* $\xi_1 > -\infty$ *or* $\eta_1 < \infty$, *then*

$$\inf\{c_n\} \le \frac{1}{2}(\xi_1 + \eta_1) \le \sup\{c_n\}. \tag{40}$$

Similarly, we obtain the corresponding result for the centre of the limit interval of orthogonality.

THEOREM 14. *If* $\sigma > -\infty$ *or* $\tau < \infty$, *then*

$$\lim_{n\to\infty} \inf\{c_n\} \le \frac{1}{2}(\sigma + \tau) \le \lim_{n\to\infty} \sup\{c_n\}. \tag{41}$$

Appendix. A second order oscillation theorem. In this appendix, we shall assume $\xi_1 > -\infty$. We define

$$Q_n(x) = P_n(x)/P_n(\xi_1), \qquad n = 0, 1, \cdots, \tag{A1}$$

where $\{P_n\}$ is given by (1), and wish to study the behaviour of the sequence $\mathbf{Q}(x) = \{Q_0(x), Q_1(x), \cdots\}$. To this end, we define the polynomials $P_n^*(x)$, $n = 0, 1, \cdots$, by

$$P_n^*(x) = P_{n+1}(\xi_1)(Q_{n+1}(x) - Q_n(x))/(x - \xi_1), \tag{A2}$$

i.e., $\{P_n^*\}$ is the set of *kernel polynomials* with parameter ξ_1 which is associated with our original system $\{P_n\}$ (see [6, §I.7]). These kernel polynomials form an orthogonal system. The zeros of $P_n^*(x)$ will be denoted by x_{nk}^*, $k = 1, 2, \cdots, n$, and in an obvious manner we define the numbers ξ_k^* and η_k^*, $k = 0, 1, \cdots$. The following lemma holds.

LEMMA A1. *For all* $k > 0$, *one has* $\xi_k^* = \xi_{k+1}$ *and* $\eta_k^* = \eta_k$.

Proof. There is a separation theorem saying that

$$x_{nk} < x_{nk}^* < x_{n+1,k+1} \tag{A3}$$

[6, Thm. I.7.2], whence the second statement holds.

Regarding ξ_k^* we can only conclude from (A3) that

$$\xi_k \le \xi_k^* \le \xi_{k+1}, \qquad k = 1, 2, \cdots. \tag{A4}$$

However, there exists a distribution $d\psi(x)$ with respect to which the polynomials P_n are orthogonal whose support contains the points ξ_k, $k = 1, 2, \cdots$, but no other points smaller than σ [6, Thm. II.4.5]. The polynomials P_n^* are then orthogonal with respect to the distribution $d\psi^*(x) = (x - \xi_1)d\psi(x)$ [21, Thm. 3.1.4]. Assuming that $d\psi^*$ is the only distribution with respect to which the P_n^* are orthogonal, we subsequently obtain from [6, Thm. II.4.5] that $\xi_k^* = \xi_{k+1}$ $(k = 1, 2, \cdots)$.

Now suppose that $d\psi^*$ is not uniquely determined by $\{P_n^*\}$. We see from (A4) that $\xi_1^* \le \xi_2$. But $\xi_1^* < \xi_2$ would be contradictory to the fact that the support of $d\psi^*$ contains at least one point in $(-\infty, \xi_1^*]$ (see [6, Thm. II.4.4(i)]). Consequently, $\xi_1^* = \xi_2$. Invoking [3, Thm. 5], we conclude that $d\psi^*$ is the unique distribution corresponding to $\{P_n^*\}$ whose support is contained in $[\xi_2, \infty)$, and that $\xi_k^* = \xi_{k+1}$ for $k > 1$ too. □

The following second order oscillation theorem is the main result of this appendix.

THEOREM A1. *The polynomials* Q_n *defined by* (A1) *and* (1) *satisfy*

$$S(\mathbf{Q}(x)) = S(\Delta_Q(x)) = k \tag{A5}$$

iff $\xi_k < x \le \xi_{k+1}$ $(k = 0, 1, \cdots)$. *Here* $\mathbf{Q}(x) = \{Q_0(x), Q_1(x), \cdots\}$ *and* $\Delta_Q(x) = \{Q_0(x), Q_1(x) - Q_0(x), Q_2(x) - Q_1(x), \cdots\}$.

Proof. The fact that $S(\mathrm{Q}(x))=k$ iff $\xi_k<x\leq\xi_{k+1}$ is a restatement of the basic oscillation theorem. The second part follows by application of the basic oscillation theorem to the polynomials P_n^* and observing that, by Corollary 4.1, $Q_0(x)(Q_1(x)-Q_0(x))<0$ when $x>\xi_1$. □

When $\eta_1<\infty$ a similar theorem may be obtained for the polynomials

$$R_n(x)=P_n(x)/P_n(\eta_1), \qquad n=0,1,\cdots. \tag{A6}$$

In closing, we remark that a finite version of Theorem A1 is stated in [7] in the context of birth-death processes. Indeed, the results of this paper apply to these stochastic processes as is shown in [8].

REFERENCES

[1] T. S. CHIHARA, *Chain sequences and orthogonal polynomials*, Trans. Amer. Math. Soc., 104 (1962), pp. 1–16.

[2] ______, *On recursively defined orthogonal polynomials*, Proc. Amer. Math. Soc., 16 (1965), pp. 702–710.

[3] ______, *On indeterminate Hamburger moment problems*, Pacific J. Math., 27 (1968), pp. 475–484.

[4] ______, *Orthogonal polynomials whose zeros are dense in intervals*, J. Math. Anal. Appl., 24 (1968), pp. 362–371.

[5] ______, *On the true interval of orthogonality*, Quart. J. Math., Oxford, 22 (1971), pp. 605–607.

[6] ______, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[7] E. A. VAN DOORN, *On the time dependent behaviour of the truncated birth-death process*, Stochastic Processes Appl., 11 (1981), pp. 261–271.

[8] ______, *Conditions for exponential ergodicity and bounds for the decay parameter of a birth-death process*, submitted for publication.

[9] P. HARTMAN AND A. WINTNER, *On linear difference equations of second order*, Amer. J. Math., 72 (1950), pp. 124–128.

[10] E. LÉOPOLD, *Location of the zeros of polynomials satisfying the three terms recurrence relation*. III. *Positive coefficients case*, Centre de Physique Théorique, CNRS Marseille, 1982.

[11] D. P. MAKI, *On the true interval of orthogonality for orthogonal polynomials*, Quart. J. Math., Oxford, 21 (1970), pp. 61–65.

[12] M. MARCUS AND H. MINC, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.

[13] L. MIRSKY, *Inequalities for normal and Hermitian matrices*, Duke Math. J., 24 (1957), pp. 591–599.

[14] E. J. MOULTON, *A theorem in difference equations on the alternation of nodes of linearly independent solutions*, Ann. Math., 13 (1912), pp. 137–139.

[15] P. G. NEVAI, *Orthogonal Polynomials*, Mem. Amer. Math. Soc. 18, no. 213, 1979.

[16] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[17] O. PERRON, *Algebra*, *Vol.* II, Walter de Gruyter, Berlin, 1927.

[18] J. SHOHAT, *Théorie générale des polynômes orthogonaux de Tchebichef*, Mémorial des Sciences Mathématiques, Fasc. LXVI, Gauthier-Villars, Paris, 1934.

[19] ______, *The relation of the classical orthogonal polynomials to the polynomials of Appell*, Amer. J. Math., 58 (1936), pp. 453–464.

[20] T. J. STIELTJES, *Recherches sur les fractions continues*, Oeuvres, Tome II, P. Noordhoff, Groningen, 1918, pp. 398–566.

[21] G. SZEGÖ, *Orthogonal Polynomials*, 4th ed., AMS Colloquium Publications 23, American Mathematical Society, Providence, RI, 1975.

[22] T. A. WHITEHURST, *On random walks and orthogonal polynomials*, Ph.D. Thesis, Indiana University, Bloomington, 1978.

[23] A. WOUK, *Difference equations and J-matrices*, Duke Math. J., 20 (1953), pp. 141–159.

SIAM J. MATH. ANAL.
Vol. 15, No. 5, September 1984

© 1984 Society for Industrial and Applied Mathematics

017

A PROPERTY OF ORTHOGONAL POLYNOMIAL FAMILIES WITH POLYNOMIAL DUALS*

MARCI PERLSTADT†

Abstract. We show that for those discrete orthogonal polynomial families, $\{p_i(\mu(x))\}$, that have polynomial duals, the "finite convolution-type integral" operator, $\Sigma_{y=0}^{M} w(y)\Sigma_{i=0}^{L} p_i(\mu(x))p_i(\mu(y))/h_i$, commutes with a second order difference operator.

1. Introduction. Let $p_i(\mu(x))$, $(i,x=0,1,\cdots,N$, where N is finite or infinite), be a discrete family of orthogonal polynomials with weight function, $w(x)$, and normalization factor, h_i, i.e.

$$(1.1)\qquad \sum_{x=0}^{N} p_i(\mu(x))p_j(\mu(x))w(x)=\delta_{ij}h_i.$$

Expanding a square integrable function $f(\mu(x))$ in terms of the $p_i(\mu(x))$'s yields

$$f(\mu(x))=\sum_{i=0}^{N}\hat{f}(i)\frac{p_i(\mu(x))}{\sqrt{h_i}}\quad \text{where}\,\hat{f}(i)=\sum_{x=0}^{N} f(\mu(x))w(x)\frac{p_i(\mu(x))}{\sqrt{h_i}}.$$

The $\hat{f}(i)$'s can be interpreted as "Fourier coefficients" for the expansion and as a reminder of this interpretation we write

$$F(f)=\hat{f}\quad\text{and}\quad F^{-1}(\hat{f})=f,$$

for the "Fourier" and "inverse Fourier" transforms of f.

We wish to consider the analogue of "timelimiting" and "bandlimiting" for the standard Fourier and inverse Fourier transform. By "timelimiting" $f(x)$ to M, we mean multiplying f by the characteristic function of the set $\{0,1,\cdots,M\}$ and by bandlimiting $f(x)$ to L, we mean multiplying $\hat{f}$ by the characteristic function of $\{0,1,\cdots,L\}$. Here it is assumed $L,M<N$. We will abuse our notation slightly and denote these operations as L and M, i.e.

$$Mf=f\cdot\chi_{\{0,1,\cdots,M\}},\qquad L\hat{f}=\hat{f}\cdot\chi_{\{0,1,\cdots,L\}}.$$

The operator we wish to study is $E=LFM$, namely the operator that timelimits, inverts, and then bandlimits a function. We remark that $E^*=MF^{-1}L$ and thus we can consider the self-adjoint operator $E^*E=MF^{-1}LFM$. In particular

$$E^*Ef(\mu(x))=\sum_{y=0}^{M} w(y)f(\mu(y))\sum_{i=0}^{L}\frac{p_i(\mu(x))p_i(\mu(y))}{h_i},\qquad x\in\{0,1,\cdots,M\}.$$

It should be noted that any time a timelimited function f is to be recovered from knowledge of $\hat{f}$ only on $\{0,1,\cdots,L\}$, we are faced with the need to study E^*E. Namely we are given

$$Mf=f\quad\text{and}\quad LFf=g\quad(\text{known}).$$

*Received by the editors December 7, 1982, and in revised form March 15, 1983.

†Department of Mathematics, Drexel University, Philadelphia, Pennsylvania 19104. This research was partially supported by the National Science Foundation under grant MCS-8302526.

Combining these two equations yields

$$LFMf=Ef=g.$$

To solve we multiply both sides of the equation by E^*

$$E^*Ef=E^*g.$$

Thus we would like to determine the eigenstructure of E^*E. Ordinarily this can pose a somewhat forbidding problem. Our approach, however, will be to find a simple means of determining the eigenfunctions by extending the methods of Slepian, Landau, and Pollak to this situation.

2. The Slepian–Landau–Pollak approach. In [1], [2], [3], Slepian, Landau, and Pollak consider the operator E^*E for the case of $f\in L^2(\mathbb{R}^1)$ and F and F^{-1}, the standard Fourier and inverse Fourier transforms on the real line. In this instance operator L represents $f\cdot\chi_L$ where $L=[-T/2,T/2]$ and M represents $\hat{f}\cdot\chi_M$ where $M=[-W,W]$. The eigenfunctions of the finite convolution integral operator E^*E are found by producing a second order differential operator, $\tilde{D}$, with simple spectrum such that $\tilde{D}$ and E^*E commute. Thus $\tilde{D}$ and E^*E share their eigenfunctions and the determination of these eigenfunctions is considerably simpler using $\tilde{D}$. In [4] Slepian extends these results to the standard Fourier transform on $\mathbb{R}^n$ and in [5] to Fourier series. In both cases a commuting $\tilde{D}$ is found.

The existence of such a commuting $\tilde{D}$ is not to be expected in general. In [6] Morrison shows that for the case of the standard Fourier transform on $\mathbb{R}^1$, the commuting $\tilde{D}$ does not exist unless L, M are symmetric intervals about the origin. There are, however, other directions one may turn in attempting to generalize these results. Thus in [7] a number of cases including Gegenbauer polynomials are considered and, for appropriate choices of L and M, a commuting $\tilde{D}$ is found. In [8] Grünbaum extends these results to expansions in the classical orthogonal polynomials: Jacobi, Hermite, Laguerre and Bessel. In [9] these results are extended to certain discrete orthogonal polynomial families, namely the Poisson–Charlier, Meixner, Krawtchouk, and Hahn polynomials. In the discrete cases a second-order difference operator $\tilde{D}$ is found that commutes with E^*E.

As mentioned earlier, the existence of such a $\tilde{D}$ is not the usual case. In fact, there seems to be two basic properties of the polynomial families mentioned above that are necessary for the construction of $\tilde{D}$. They are:

(i) the existence of a second-order difference equation of the form

$$Dp_i(x)=\frac{1}{w(x)}\Delta[w(x-1)\delta(x-1)\nabla p_i(x)]=\lambda_i p_i(x) \tag{2.1}$$

(here $\Delta f(x)=f(x+1)-f(x)$ and $\nabla f(x)=f(x)-f(x-1)$), and

(ii) the existence of a first-order difference equation of the form

$$r(x)\Delta p_i(x)=S(x,i)p_i(x)+t(i)p_{i-1}(x). \tag{2.2}$$

Note that, of course, for the continuous (classical) cases, the difference equations are replaced by differential equations.

Bochner [10] has shown that in the continuous case the only polynomial families satisfying a second-order differential equation corresponding to the form of (2.1) are the classical cases[1]. Lesky [11] has shown that the only families satisfying a second-order difference equation of the form (2.1) are the Poisson–Charlier, Meixner, Krawtchouk, and Hahn polynomials[1]. If, however, one is willing to loosen the restrictions on the form of (2.1) and consider $p_i(\mu(x))$ instead of just $p_i(x)$, then some new families arise. In particular, we will consider here those polynomial families with polynomial duals.

3. Polynomial families with polynomial duals. Let $p_i(\mu(k))$ and $R_k(\lambda(i))$ be two families of orthogonal polynomials normalized so that $p_i(\mu(0))=1$ and $R_k(\lambda(0))=1$. These two families are dual[2] if

$$p_i(\mu(k))=R_k(\lambda(i)).$$

Our interest in polynomials with polynomial duals stems from the fact that they guarantee a three-term recurrence (in k) for the $R_k(\lambda(i))$'s and thus a second-order difference operator of the form (2.1) for the $p_i(\mu(k))$'s (and for the $R_k(\lambda(i))$'s). Thus any orthogonal polynomial family with a polynomial dual is a prime candidate for an operator $\tilde{D}$ to commute with E^*E. Recent work [13], [14], [15], [16], [17], [20] in the area of orthogonal polynomials makes it possible for us to show here that such a $\tilde{D}$ can be found for all such families.

It is now known that in addition to the Poisson–Charlier, Meixner, and Krawtchouk (all of which are self-dual), as well as the Hahn polynomials and their duals [12], there is Wilson's discovery [13], [16] of the Racah polynomials (which are clearly self-dual). In addition, where appropriate, there are "q-generalizations" of these polynomials: q-Krawtchouk [18], q-Hahn and dual q-Hahn [19] and, more generally, the recently discovered Askey–Wilson q-analogue of the Racah polynomials [14], [15]. In [20] Leonard shows that these are the only families with polynomial duals.

Before proceeding we further remark that the difference formulas (2.1), (2.2) can in fact be rewritten as divided difference formulas. Thus, for example, (2.1) can be recast as

$$\alpha(x_k)p_i[x_{k-1},x_k,x_{k+1}]+\beta(x_k)p_i[x_{k-1},x_k]=\lambda_i p_i(x_k).$$

Here $x_k=\mu(k)$ and

$$f[x_{k-1},x_k]=\frac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}},\qquad f[x_{k-1},x_k,x_{k+1}]=\frac{f[x_k,x_{k+1}]-f[x_{k-1},x_k]}{x_{k+1}-x_{k-1}}.$$

This is useful to remember not only because of the analogy with the differential equations of the continuous case, but also because it provides an important clue in the construction of $\tilde{D}$. For the families studied in [8] and [9], the $\delta(x)$ in (2.1) and $r(x)$ in (2.2) were identical. In the cases studied here we will have

$$[\mu(x+1)-\mu(x-1)]\delta(x)=r(x) \tag{3.1}$$

since $r(x)$, $\delta(x)$ come respectively from first- and second-order divided differences. This observation will help in constructing $\tilde{D}$ in §4. In particular the $\varepsilon(x)$ term (see (4.2)) was taken as $\varepsilon(x)=x-M$ for the cases studied in [8] and [9]. Here we will take $\varepsilon(x)$ so as to compensate for the extra factor in (3.1), i.e.

$$\Delta\varepsilon(x)=\mu(x+1)-\mu(x-1).$$

[1] Note that these cases also satisfy equations of the form (2.2).

[2] Note that for any orthogonal polynomial family $\{p_i(\mu(k))\}$, one can consider its dual: $R_k(\lambda(i))=p_i(\mu(k))$. The dual is always orthogonal but need not be polynomial.

4. The operator $\tilde{D}$. Returning to the problem of finding a commuting second-order difference operator $\tilde{D}$ for E^*E, we recall the form of the second-order difference operator for $p_i(\mu(x))$

$$D=\frac{1}{w(x)}\Delta[w(x-1)\delta(x-1)\nabla]. \tag{4.1}$$

We will show that $\tilde{D}$ can be constructed to have the form

$$\tilde{D}=\frac{1}{w(x)}\Delta[w(x-1)\delta(x-1)\varepsilon(x-1)\nabla]+G(L)c(x), \tag{4.2}$$

where $\varepsilon(x-1)$, $G(L)$, $c(x)$ will be explicitly determined in §5 and 6.

Note that it will suffice to choose these unknowns in such a manner that $\varepsilon(M)=0$ and that

$$\tilde{D}_xK_L(x,y)=\tilde{D}_yK_L(x,y) \quad \text{where } K_L(x,y)=\sum_{i=0}^{L}\frac{p_i(\mu(x))p_i(\mu(y))}{h_i}.$$

This follows by essentially the same argument as in [9] but we outline it here for completeness. By repeated application of the summation by parts formula we have

$$\begin{aligned}E^*E\tilde{D}_xf(\mu(x))&=\sum_{y=0}^{M}K_L(x,y)\big[\tilde{D}_yf(\mu(y))\big]w(y)\\ &=A-B+C|_{y=-1}^{y=M}+\sum_{y=0}^{M}f(\mu(y))\big[\tilde{D}_yK_L(x,y)\big]w(y),\end{aligned}$$

where

$$\begin{aligned}&A=K_L(x,y+1)\cdot\varepsilon(y)w(y)\cdot\delta(y)\cdot\Delta f(\mu(y)),\\ &B=\varepsilon(y+1)\cdot w(y+1)\cdot\delta(y+1)\cdot f(\mu(y+1))\Delta K_L(x,y+1),\\ &C=f(\mu(y+1))\Delta[\varepsilon(y)\cdot w(y)\cdot\delta(y)\nabla K_L(x,y+1)].\end{aligned}$$

Then by expanding the Δ term in C and by combining terms, we get $A-B+C|_{y=-1}^{y=M}=0$ if $\varepsilon(M)=0$. Thus since

$$\tilde{D}_xE^*Ef(\mu(x))=\sum_{y=0}^{M}f(\mu(y))[\tilde{D}_xK_L(x,y)]w(y),$$

it suffices to show that $\tilde{D}_xK_L(x,y)=\tilde{D}_yK_L(x,y)$ (provided $\varepsilon(M)=0$).

We note that the real work here is in carrying out the Racah and q-Racah cases [14], [16], as the rest of the cases are simply special limiting cases of these polynomials. In fact, the Racah polynomials can be obtained from the q-Racah polynomials, but, as the notation used for the two is quite different, both derivations are given. Furthermore, the proof for each case follows the same general outline in [9].

Before continuing we remark that this particular form (4.2) for $\tilde{D}$ is strongly reminiscent of the commuting differential operator Slepian, Landau, and Pollak find for the case where $f,\hat{f}\in L^2(\mathbb{R}^1)$, $L=[-T,T]$, $M=[-W,W]$. There

$$E^*Ef(x)=\int_{-T}^{T}\frac{\sin(W(x-y))}{x-y}f(y)\,dy, \qquad x\in A.$$

The commuting $\tilde{D}$ can be chosen to have the form

$$\tilde{D}f(x)=((T^2-x^2)f'(x))'-W^2x^2f(x), \tag{4.3}$$

and the eigenfunctions of (4.3) are the prolate spheroidal wave functions. Note that (T^2-x^2) and $-W^2x^2$ in (4.3) play roles corresponding respectively to the $\varepsilon(x)$ and $G(L)c(x)$ terms in (4.2).

5. Wilson's Racah polynomials. All of the polynomials we have mentioned can be expressed as hypergeometric series or basic hypergeometric series:

$${}_rF_s\left(\begin{matrix}a_1,\cdots,a_r\\ b_1,\cdots,b_s\end{matrix};x\right)=\sum_{n=0}^{\infty}\frac{(a_1)_n\cdots(a_r)_nx^n}{(b_1)_n\cdots(b_s)_nn!},$$

$${}_{r+1}\phi_r\left(\begin{matrix}a_1,\cdots,a_{r+1}\\ b_1,\cdots,b_r\end{matrix};q,x\right)=\sum_{n=0}^{\infty}\frac{(a_1;q)_n\cdots(a_{r+1};q)_nx^n}{(b_1;q)_n\cdots(b_r;q)_n(q;q)_n},$$

where

$$(a)_n=\begin{cases}(a)(a+1)\cdots(a+n-1), & n=1,2,\cdots,\\ 1, & n=0,\end{cases}$$

and

$$(a;q)_n=\begin{cases}(1-a)(1-aq)\cdots(1-aq^{n-1}), & n=1,2,\cdots,\\ 1, & n=0.\end{cases}$$

The Racah polynomials in particular can be defined [16] as

$$p_n[(x+a)^2]={}_4F_3\left(\begin{matrix}-n,a+b+c+d+n-1,-x,x+2a\\ a+b,a+c,a+d\end{matrix};1\right). \tag{5.0}$$

The orthogonality relationship (1.1) has

$$\mu(x)=(x+a)^2,$$
$$w(x)=\frac{(2a)_x(a+1)_x(a+b)_x(a+c)_x(a+d)_x}{(1)_x(a)_x(a-b+1)_x(a-c+1)_x(a-d+1)_x}=w(x-1)\frac{\beta(x)}{\gamma(x)},$$

where

$$\beta(x)=(2a+x-1)(a+x)(a+b+x-1)(a+c+x-1)(a+d+x-1),$$
$$\gamma(x)=(x)(a+x-1)(a-b+x)(a-c+x)(a-d+x),$$

and

$$h_i=\frac{i!(a+b+c+d-1)}{(a+b+c+d-1)_i(2i+a+b+c+d-1)}\cdot\frac{(c+d)_i(b+d)_i(b+c)_i}{(a+b)_i(a+c)_i(a+d)_i}\cdot H=h_{i-1}\frac{B_i}{D_i},$$

where

$$B_i=(i)(2i-3+a+b+c+d)(b+c+i-1)(b+d+i-1)(c+d+i-1),$$
$$D_i=(a+b+c+d+2i-1)(a+b+i-1)$$
$$\cdot(a+c+i-1)(a+d+i-1)(a+b+c+d+i-2),$$
$$H=\frac{(2a+1)_N(1-c-d)_N}{(a-c+1)_N(a-d+1)_N}.$$

It is required that $a+b$, $a+c$, or $a+d=-N$. The above orthogonality relationship and the difference formulas that follow are derived in [13], [16]. In addition to the difference formulas of the form (2.1) and (2.2) we will need the Christoffel–Darboux formula. These are given below using the notation $x_a=(x+a)^2$.

(i) *Second-order difference equation.*

$$(5.1)\qquad \frac{1}{w(x)}\Delta[w(x-1)\delta(x-1)\nabla p_i(x_a)]=i(a+b+c+d+i-1)p_i(x_a)$$

where

$$\delta(x)=\frac{(2a+x)(a+b+x)(a+c+x)(a+d+x)}{2(2a+2x+1)(x+a)}.$$

(ii) *First-order difference equation.*

(5.2)

$$2(x+a)\delta(x)\Delta\left[\frac{p_i(x_a)}{\sqrt{h_i}}\right]=i(x+2a)(x+a+b+c+d+i-1)\left[\frac{p_i(x_a)}{\sqrt{h_i}}\right]$$
$$+\frac{B_i\left[\left(p_i(x_a)/\sqrt{h_i}\right)-\left(p_{i-1}(x_a)/\sqrt{h_{i-1}}\right)\left(\sqrt{D_i}/\sqrt{B_i}\right)\right]}{(2i-3+a+b+c+d)(2i-2+a+b+c+d)}$$

(iii) *Christoffel–Darboux.*

$$(5.3)\qquad (x_a-y_a)\sum_{n=0}^{k}[p_n(x_a)p_n(y_a)/h_n]$$
$$=C_{k+1}[p_{k+1}(x_a)p_k(y_a)-p_k(x_a)p_{k+1}(y_a)]/\left(\sqrt{h_{k+1}}\sqrt{h_k}\right);$$

here $C_k=\sqrt{B_k}\sqrt{D_k}/(a+b+c+d+2k-1)(a+b+c+d+2k-2)(a+b+c+d+2k-3)$.

Claim. *The commuting $\tilde{D}$ can be chosen as*

$$\tilde{D}=\frac{1}{w(x)}\Delta[w(x-1)\delta(x-1)\varepsilon(x-1)\nabla]+G(L)x_a$$

where $\varepsilon(x)=x(x+2a+1)-M(M+2a+1)=(x-M)(x+M+2a+1)$ *and* $G(L)=-L(a+b+c+d+L)$.

Proof. Clearly $\varepsilon(M)=0$ and thus we need only show that

$$\tilde{D}_xK_L(x,y)=\tilde{D}_yK_L(x,y)\quad\text{where } K_L(x,y)=\sum_{i=0}^{L}p_i(x_a)p_i(y_a)/h_i.$$

We begin by noting that using (5.1) and (5.2) gives

$$\begin{aligned}\tilde{D}_x\left[\frac{p_i(x_a)}{\sqrt{h_i}}\right]&=\varepsilon(x-1)i(a+b+c+d+i-1)\left[\frac{p_i(x_a)}{\sqrt{h_i}}\right]+2(x+a)\delta(x)\Delta\left[\frac{p_i(x_a)}{\sqrt{h_i}}\right]\\&\qquad+G(L)x_a\left[\frac{p_i(x_a)}{\sqrt{h_i}}\right]\\&=[(x-1)(x+2a)(i)(a+b+c+d+i)\\&\qquad-(i)(a+b+c+d+i-1)M(M+2a+1)\\&\qquad\qquad+i(x+2a)(a+b+c+d+i)]\frac{p_i(x_a)}{\sqrt{h_i}}\\&\quad+\frac{B_i\left(p_i(x_a)/\sqrt{h_i}\right)}{(2i-3+a+b+c+d)(2i-2+a+b+c+d)}\\&\quad-\frac{\sqrt{B_i}\sqrt{D_i}\left(p_{i-1}(x_a)/\sqrt{h_{i-1}}\right)}{(2i-3+a+b+c+d)(2i-2+a+b+c+d)}+G(L)x_a\frac{p_i(x_a)}{\sqrt{h_i}}.\end{aligned}$$

Thus

(5.4)

$$\begin{aligned}\left(\tilde{D}_x-\tilde{D}_y\right)K_L(x,y)&=[(x-1)(x+2a)-(y-1)(y+2a)]\\&\qquad\cdot\sum_{i=0}^{L}(i)(a+b+c+d+i)\frac{p_i(x_a)p_i(y_a)}{h_i}\\&\quad+(x-y)\sum_{i=0}^{L}(i)(a+b+c+d+i)\frac{p_i(x_a)p_i(y_a)}{h_i}\\&\quad+\sum_{i=0}^{L}\frac{\sqrt{B_i}\sqrt{D_i}\,[p_i(x_a)p_{i-1}(y_a)-p_{i-1}(x_a)p_i(y_a)]}{(2i-3+a+b+c+d)(2i-2+a+b+c+d)\sqrt{h_i}\sqrt{h_{i-1}}}\\&\quad+(x_a-y_a)G(L)\sum_{i=0}^{L}\frac{p_i(x_a)p_i(y_a)}{h_i}.\end{aligned}$$

Noting that by (5.3)

$$\begin{aligned}&\sum_{i=0}^{L}\frac{\sqrt{B_i}\sqrt{D_i}\,[p_i(x_a)p_{i-1}(y_a)-p_{i-1}(x_a)p_i(y_a)]}{(2i-3+a+b+c+d)(2i-2+a+b+c+d)\sqrt{h_i}\sqrt{h_{i-1}}}\\&\quad=\sum_{i=0}^{L}(2i-1+a+b+c+d)(x_a-y_a)\sum_{n=0}^{i-1}[p_n(x_a)p_n(y_a)/h_n]\\&\quad=(x_a-y_a)\sum_{n=0}^{L-1}[p_n(x_a)p_n(y_a)/h_n]\sum_{i=n+1}^{L}(2i-1+a+b+c+d)\\&\quad=(x_a-y_a)\sum_{n=0}^{L}[L(L+a+b+c+d)-n(n+a+b+c+d)][p_n(x_a)p_n(y_a)/h_n],\end{aligned}$$

and applying this to (5.4), yields

$$\begin{aligned}(\tilde{D}_x-\tilde{D}_y)K_L(x,y)&=(x-y)(x+y+2a)\sum_{i=0}^{L}(i)(a+b+c+d+i)[p_i(x_a)p_i(y_a)/h_i]\\&\quad+(x_a-y_a)G(L)\sum_{i=0}^{L}\frac{p_i(x_a)p_i(y_a)}{h_i}\\&\quad+(x_a-y_a)\sum_{i=0}^{L}(L(L+a+b+c+d)\\&\qquad\qquad-n(n+a+b+c+d))\left[\frac{p_n(x_a)p_n(y_a)}{h_n}\right]=0\end{aligned}$$

(since $G(L)=-L(a+b+c+d+L)$).

6. Askey and Wilson's q-Racah polynomials. In [14], [15] Askey and Wilson generalize the Racah polynomials (5.0) and consider the q-Racahs:

$$p_n(\mu(x))={}_4\phi_3\left(\begin{matrix}q^{-n},q^{n+1}ab,q^{-x},q^{x+1}cd\\aq,bdq,cq\end{matrix};q,q\right)\tag{6.0}$$

where $\mu(x)=q^{-x}+q^{x+1}cd$ and aq, cq, or $bdq=q^{-N}$. The orthogonality relation (1.1) in this case has

$$w(x)=\frac{(cdq;q)_x(1-cdq^{2x+1})(aq;q)_x(bdq;q)_x(cq;q)_x(abq)^{-x}}{(q;q)_x(1-cdq)(cdq/a;q)_x(cq/b;q)_x(dq;q)_x},$$

$$h_n=\frac{(q;q)_n(1-abq)(bq;q)_n(aq/d;q)_n(abq/c;q)_n(cdq)^n}{(abq;q)_n(1-abq^{2n+1})(aq;q)_n(bdq;q)_n(cq;q)_n}h_0,$$

where h_0 is a constant depending on a,b,c,d, and q. This relationship and the second-order difference formula are derived in [14]. The first-order difference is derived in [17]. We state these below using the notation $\hat{x}=\mu(x)$.

(i) *Second-order difference equation.*

$$\frac{1}{w(x)}\Delta[w(x-1)\delta(x-1)\nabla p_n(\hat{x})]=-(1-q^{-n})(1-q^{n+1}ab)p_n(\hat{x}),\tag{6.1}$$

where $\delta(x)=(1-cdq^{x+1})(1-cq^{x+1})(1-bdq^{x+1})(1-aq^{x+1})/(1-cdq^{2x+1})(1-cdq^{2x+2})$.

(ii) *First-order difference equation.*

$$\begin{aligned}&q^{-x}(1-cdq^{2x+1})\delta(x)\left[\Delta[p_n(\hat{x})]/\sqrt{h_n}\right]\\&\quad=q^{-n}(1-abq^{2n+1})C_n\left[\left(p_n(\hat{x})/\sqrt{h_n}\right)-\left(p_{n-1}(\hat{x})/\sqrt{h_{n-1}}\right)\left(\sqrt{h_{n-1}}/\sqrt{h_n}\right)\right]\\&\qquad+(1-q^{-n})(q^{n+1}ab-q^{-x})(1-q^{x+1}cd)\left[p_n(\hat{x})/\sqrt{h_n}\right],\end{aligned}\tag{6.2}$$

where $C_n=q(1-q^n)(1-bq^n)(c-abq^n)(d-aq^n)/(1-abq^{2n})(1-abq^{2n+1})$.

(iii) *Christoffel–Darboux.*

$$(6.3)\qquad C_{n+1}\sqrt{h_n/h_{n+1}}\,[p_{n+1}(\hat{x})p_n(\hat{y})-p_n(\hat{x})p_{n+1}(\hat{y})]/\sqrt{h_n}\sqrt{h_{n+1}}$$

$$=[(q^{-x}-q^{-y})+(q^{x+1}-q^{y+1})cd]\sum_{i=0}^{n}[p_i(\hat{x})p_i(\hat{y})/h_i].$$

CLAIM. *$\tilde{D}=(1/w(x))\Delta[w(x-1)\delta(x-1)\varepsilon(x-1)\nabla]+\hat{x}G(L)$ where $\varepsilon(x)=(q^{-x}+cdq^{x+2}-q^{-M}-cdq^{M+2})/(1-q)=(q^{-x}-q^{-M})(1-cdq^{M+x+2})/(1-q)$ and $G(L)=[(q^{-L}+abq^{L+2})/(q-1)]-[(1+abq^2)/(q-1)]=(q^L-1)(abq^2-q^{-L})/(q-1)$.*

Proof. The proof follows the same general outline as the proof for the Racahs. We sketch it briefly noting that $\varepsilon(M)=0$ and that using (6.1) and (6.2) we have

$$\begin{aligned}&\tilde{D}_x\left(p_i(\hat{x})/\sqrt{h_i}\right)\\&\quad=-\varepsilon(x-1)(1-q^{-i})(1-q^{i+1}ab)\left[p_i(\hat{x})/\sqrt{h_i}\right]\\&\qquad+q^{-i}(1-abq^{2i+1})C_i\left[\left(p_i(\hat{x})/\sqrt{h_i}\right)-\left(p_{i-1}(\hat{x})/\sqrt{h_{i-1}}\right)\left(\sqrt{h_{i-1}}/\sqrt{h_i}\right)\right]\\&\qquad+(1-q^{-i})(q^{i+1}ab-q^{-x})(1-q^{x+1}cd)\left[p_i(\hat{x})/\sqrt{h_i}\right]+G(L)\hat{x}\left[p_i(\hat{x})/\sqrt{h_i}\right].\end{aligned}$$

Thus $\left(\tilde{D}_x-\tilde{D}_y\right)K_L(x,y)=\mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV}$, where

$$\mathrm{I}=-[\varepsilon(x-1)-\varepsilon(y-1)]\sum_{i=0}^{L}\left[(1-q^{-i})(1-q^{i+1}ab)\left[\frac{p_i(\hat{x})p_i(\hat{y})}{h_i}\right]\right],$$

$$\begin{aligned}\mathrm{II}=\sum_{i=0}^{L}\Bigg[&(1-q^{-i})[(q^{i+1}ab-q^{-x})(1-q^{x+1}cd)\\&-(q^{i+1}ab-q^{-y})(1-q^{y+1}cd)]\left[\frac{(p_i(\hat{x})p_i(\hat{y}))}{h_i}\right]\Bigg],\end{aligned}$$

$$\begin{aligned}\mathrm{III}=\sum_{i=0}^{L}q^{-i}(1-abq^{2i+1})C_i&\Big[(p_i(\hat{x})p_{i-1}(\hat{y})\\&-p_{i-1}(\hat{x})p_i(\hat{y}))/\left(\sqrt{h_i}\sqrt{h_{i-1}}\right)\Big]\left[\sqrt{h_{i-1}}/\sqrt{h_i}\right],\end{aligned}$$

$$\mathrm{IV}=G(L)(\hat{x}-\hat{y})\sum_{i=0}^{L}\frac{p_i(\hat{x})p_i(\hat{y})}{h_i}.$$

Therefore, in a manner similar to §5, we obtain

$$\begin{aligned}\mathrm{I}+\mathrm{II}&=\left[(q^{-x}-q^{-y})+(cdq^{x+1}-cdq^{y+1})\right]\\&\quad\cdot\sum_{i=0}^{L}\left[\left[\frac{(1-q^{-i})(1-abq^{i+2})}{(q-1)}\right]\left[\frac{p_i(\hat{x})p_i(\hat{y})}{h_i}\right]\right]\\&=(\hat{x}-\hat{y})\sum_{i=0}^{L}\left[\left[\frac{(q^{-i}+abq^{i+2})}{(1-q)}\right]\left[\frac{p_i(\hat{x})p_i(\hat{y})}{h_i}\right]\right]\\&\quad+(\hat{x}-\hat{y})\left[\frac{(1+abq^2)}{(q-1)}\right]\sum_{i=0}^{L}\frac{p_i(\hat{x})p_i(\hat{y})}{h_i},\end{aligned}$$

and

$$\mathrm{III}=\sum_{i=0}^{L} q^{-i}(1-abq^{2i+1})\sum_{k=0}^{i-1}\left[\left[(q^{-x}-q^{-y})+(cdq^{x+1}-cdq^{y+1})\right]\left[\frac{p_k(\hat{x})p_k(\hat{y})}{h_k}\right]\right]$$

$$=\left[(q^{-x}-q^{-y})+(cdq^{x+1}-cdq^{y+1})\right]\sum_{k=0}^{L-1}\left(\frac{p_k(\hat{x})p_k(\hat{y})}{h_k}\right)\sum_{i=k+1}^{L} q^{-i}(1-abq^{2i+1})$$

$$=\left[(q^{-x}-q^{-y})+(cdq^{x+1}-cdq^{y+1})\right]$$

$$\cdot\sum_{k=0}^{L}\left[\left(\left[\frac{(q^{-k}+abq^{k+2})}{(q-1)}\right]-\left[\frac{(q^{-L}+abq^{L+2})}{(q-1)}\right]\right)\left[\frac{p_k(\hat{x})p_k(\hat{y})}{h_k}\right]\right].$$

It follows that $(\tilde{D}_x-\tilde{D}_y)K_L(x,y)=0$.

7. Some special cases. We remark again that all of the polynomials with polynomial duals arise as special cases of Askey and Wilson's q-Racah polynomials [20]. Thus a particular $\tilde{D}$ can be found by taking appropriate limits of the $\tilde{D}$ in §6. The results in §5 can be obtained by taking limits as $q\to 1$ and reparameterizing the Racahs as:

$$r_n(\lambda_x)={}_4F_3\left(\begin{matrix}-n,n+\alpha+\beta+1,-x,x+\gamma+\delta+1\\ \alpha+1,\beta+\delta+1,\gamma+1\end{matrix};1\right)\tag{7.1}$$

where $\lambda_x=x(x+\gamma+\delta+1)$.

One can further obtain $\tilde{D}$ for the limiting case where $q\to -1$. The orthogonality of these polynomials was given by Leonard [20]. All of the "classical" polynomials (Poisson–Charlier, Meixner, Hahn, Krawtchouk) for which $\tilde{D}$ was constructed in [9] can also be obtained by appropriate limits [15]. Letting $\beta\to\infty$ and $\alpha+1=-N$ in (7.1) gives the Hahn duals and $\tilde{D}$ for this case is given by:

$$\tilde{D}_x=\frac{1}{w(x)}\Delta[w(x-1)\delta(x-1)\varepsilon(x-1)\nabla]\left[p_n(\lambda_x)/\sqrt{h_n}\right]+G(L)\lambda_x\left[p_n(\lambda_x)/\sqrt{h_n}\right],$$

where

$$w(x)=\frac{(x+\alpha+\beta)(x+\alpha)(N-x)(2x+\alpha+\beta+1)}{(2x+\alpha+\beta-1)(x)(x+\beta)(x+\alpha+\beta+N)}w(x-1),$$

$$\delta(x)=\frac{(x+\alpha+\beta+1)(x+\alpha+1)(N-1-x)}{(2x+\alpha+\beta+1)(2x+\alpha+\beta+2)},$$

$$\varepsilon(x)=(x)(x+\alpha+\beta+2)-M(M+\alpha+\beta+2),$$

$$G(L)=L,$$

$$h_i=[(N-n)(\alpha+n)/(n)(N+\beta-n)]h_{i-1}.$$

Some other limiting cases of interest include the q-Hahn polynomials [14], [19] which result from taking $d=0$ and $cq=q^{-N}$ in(6.0),the dual q-Hahns [14], [19] with $b=0$ and $aq=q^{-N}$ in(6.0),and Stanton's q-analogue of the Krawtchouk polynomials [18], [14]. With appropriate limits the corresponding $\tilde{D}$ can be constructed.

8. $EE^* = LFMF^{-1}L$. In [7] it is noted that one can equally well study the operator $EE^* = LFMF^{-1}L$. Namely if f is an eigenfunction of $MF^{-1}LFM$ with eigenvalue $\lambda \neq 0$, then LFf is an eigenfunction of $LFMF^{-1}L$ with eigenvalue λ. We can represent EE^* by the $(L+1)x(L+1)$ matrix with entries

$$(EE^*)_{i,j} = \sum_{x=0}^{M} p_i(\mu(x))p_j(\mu(x))w(x), \qquad 0 \leq i,j \leq L.$$

A tridiagonal matrix T that commutes with EE^* can be found by applying $\tilde{D}$ to $p_i(x)$ and obtaining the resulting three-term recurrence in terms of $p_{i-1}(x)$, $p_i(x)$, and $p_{i+1}(x)$. From this, the matrix T can be read off. For a detailed example of this sort, see [7]. Note that the reduction of work in using T rather than EE^* to compute the eigenfunctions is equivalent to the reduction of work in computing the eigenfunctions of a tridiagonal matrix rather than of a full matrix.

If we represent T in the form:

$$T = \begin{pmatrix} a_1 & b_1 & & & \\ b_1 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & & & b_{L-1} \\ & & & b_{L-1} & a_L \end{pmatrix},$$

then T has simple spectrum if $b_i \neq 0$, $i = 0, 1, \cdots, L-1$. Thus by computing the entries of T, we can be guaranteed simplicity of spectrum for our commuting operators.

We further remark that for all cases studies so far, the existence of a commuting difference operator $\tilde{D}$ for E^*E has been equivalent to the existence of a commuting tridiagonal matrix T for EE^*. Thus by forming EE^* for any given case and checking to see if an appropriate T exists, one has some indication as to the possible existence of $\tilde{D}$. We note numerical tests indicate that T will not exist, in general, if the operators M and L of §1 are taken as

$$Mf = f \cdot \chi_{\{M_1, M_1+1, \cdots, M_2\}}, \qquad Lf = f \cdot \chi_{\{L_1, l_1+1, \cdots, L_2\}},$$

unless $M_1 = 0$ or $M_2 = N$, and $L_1 = 0$ or $L_2 = N$. This is reminiscent of the results of Morrison [6] for the standard Fourier transform.

Acknowledgment. The author would like to thanks Professors Richard Askey and James Wilson for their help and for access to their preprints.

REFERENCES

[1] D. Slepian and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: I, Bell System Tech. J., 40 (1961), pp. 43–64.

[2] H. J. Landau and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: II, Bell System Tech. J., 40 (1961), pp. 65–84.

[3] ———, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: III, Bell System Tech. J., 41 (1962), pp. 1295–1336.

[4] D. Slepian, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: IV, Bell System Tech. J., 43 (1964), pp. 3009–3058.

[5] ———, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: V, Bell System Tech. J., 57 (1978), pp. 1371–1430.

[6] J. Morrison, *On the commutation of finite integral operators with difference kernels, and linear self-adjoint differential operators*, Abstract, Notices AMS, 9 (1962), p. 119.

[7] F. A. Grünbaum, L. Longhi, and M. Perlstadt, *Differential operators commuting with finite convolution integral operators: Some nonabelian examples*, SIAM J. Appl. Math, 42 (1982), pp. 941–955.

[8] F. A. Grünbaum, *A new property of a reproducing kernel for classical orthogonal polynomials*, J. Math. Anal. Appl. 95 (1983), pp. 491–500.

[9] M. Perlstadt, *Chopped orthogonal polynomial expansions—Some discrete cases*, SIAM J. Alg. Disc. Meth., 4 (1983), pp. 94–100.

[10] S. Bochner, *Über Sturm–Liouvillesche Polynomsyteme*, Math. Z., 29 (1929), pp. 730–736.

[11] P. Lesky, *Orthogonale Polynomsysteme als Losungen Sturm–Liouvillescher Differenzengleichungen*, Monat. Math., 66 (1962), pp. 203–214.

[12] S. Karlin and J. L. McGregor, *The Hahn polynomial formulas and an application*, Scripta Math., 26 (1961), pp. 33–46.

[13] J. A. Wilson, *Hypergeometric series recurrence relation and some new orthogonal functions*, Ph.D. thesis, Univ. Wisconsin, Madison, 1978.

[14] R. Askey and J. Wilson, *A set of orthogonal polynomials that generalize the Racah coefficients or 6-j symbols*, this Journal, 10 (1979), pp. 1008–1016.

[15] ———, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. AMS, to appear.

[16] J. A. Wilson, *Some hypergeometric orthogonal polynomials*, this Journal, 11 (1980), pp. 690–701.

[17] ———, *Hypergeometric series recurrence relations and properties of some orthogonal functions*, to appear.

[18] D. Stanton, *Some q-Krawtchouk polynomials on Chevalley groups*, Amer. J. Math., 102 (1980), pp. 625–662.

[19] ———, *Product formulas for q-Hahn polynomials*, this Journal, 11 (1980), pp. 100–107.

[20] D. Leonard, *Orthogonal polynomials, duality, and association schemes*, this Journal, 13 (1982), pp. 656–663.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics
001

BIFURCATION TO QUASI-PERIODIC TORI IN THE INTERACTION OF STEADY STATE AND HOPF BIFURCATIONS*

JÜRGEN SCHEURLE[†] AND JERROLD MARSDEN[‡]

Abstract. Bifurcations to quasi-periodic tori in a two parameter family of vector fields are studied. At criticality, the vector field has an equilibrium point with a zero eigenvalue and a pair of complex conjugate eigenvalues. This situation has been studied by Langford, Iooss, Holmes and Guckenheimer. Here we provide explicitly computed conditions under which the stability of the secondary branch of tori, and whether the flow on them is quasiperiodic, can be determined. The results are applied to "Brusselator" system of reaction diffusion equations.

Introduction. Consider a smooth vector field on $\mathbb{R}^3$ which has a singular point at the origin. Suppose that the linearized vector field at the origin has an eigenvalue zero and a pair of pure imaginary eigenvalues $\pm i\gamma$, $\gamma>0$. The aim of this paper is to show that in a two parameter unfolding of this singularity satisfying explicitly computed nondegeneracy conditions, there are continuous curves, emanating as a tertiary bifurcation from a secondary curve of Hopf periodic orbits, along which one has invariant 2-tori carrying quasi-periodic flow. In particular, this shows that within a structurally stable situation one has an abundance of bifurcations that are the first part of a Landau sequence:

trivial equilibrium → primary branch of equilibria → secondary branch of Hopf orbits → tertiary branch of quasi-periodic 2-tori.

Explicit exchange of stability results are established for each bifurcation. Because of the concrete nature of the formulas for the nondegeneracy conditions that are derived, the results can be applied to specific problems for specific choices of parameters. We work out these conditions for a reaction diffusion problem as an example. We expect that the method will also apply to certain plasma instability problems; see Crawford [1983].

We shall work within the class of C^k vector fields with a zero at the origin. No other symmetry conditions are imposed. The eigenvalues are assumed to cross the imaginary axis "with nonzero speed" with respect to the unfolding parameters and some nondegeneracy conditions are imposed on the second and third order terms of the unfolding. These assumptions will imply that the trivial solution undergoes transcritical and Hopf bifurcations. Langford [1979] showed that the interaction between this steady state and Hopf bifurcation leads to invariant tori under some generic-type assumptions. Although Langford's paper forms the basis for the present work, our approach is more in the spirit of singularity theory and the work of Holmes [1980] and Guckenheimer [1981], [1982] in that it uses normal forms and the ideas of unfolding. Of course there are now many papers in bifurcation theory using this approach, such as Golubitsky and Schaeffer [1979], Schaeffer and Golubitsky [1981] and Golubitsky and Langford [1981]. Some partial results similar in spirit to ours have been given by Broer [1982] (see also

*Received by the editors December 13, 1982, and in revised form September 26, 1983.

[†] Universität Stuttgart, Pfaffenwaldring 57, D-7000 Stuttgart 80, West Germany. The research of the author was supported by Deutsche Forschungsgemeinschaft contract 13 Sche 233/2-1 while he was visiting the University of California, Berkeley.

[‡] Department of Mathematics, University of California, Berkeley, California 94720. The research of this author was partially supported by the U. S. Department of Energy under contract DE-AT03-82ER12097.

Broer [1981b], Braaksma, and Broer [1981], and Chow and Hale [1982]). As we have mentioned, we impose no symmetry conditions, but also do not forbid them. In particular, we begin with a normal form somewhat more general than that considered by Guckenheimer [1981], [1982].

To construct the bifurcating 2-tori, we use a theorem of Sacker [1965]. To locate the tori carrying quasi-periodic flow, we use KAM theory and methods of Scheurle [1982]. The results will be local, and will be robust against higher order perturbations, so we have a form of structural stability. (For global results, the work of Chenciner [1982] may be relevant.) Despite the fact that individual quasi-periodic flows are structurally unstable, their occurrence in this bifurcation is stable and in fact their occurrence along appropriate arcs in parameter space is an open condition. To make the explicit computation for the example considered in §3, we use Poincaré–Birkhoff normal forms and center manifold theory, a technique of Ruelle and Takens [1971] that proved effective for explicit calculations in the Hopf bifurcation (see Marsden and McCracken [1976] Hassard and Wan [1978], and Hassard, Kazarinoff and Wan [1981]).

There are other singularities and corresponding unfoldings where our method should be applicable to yield invariant tori with quasi-periodic flow. If, for example, the spectrum of the linearized vector field is as above and the second order terms vanish identically, then we have a more degenerate singularity, and a two-parameter unfolding is reasonable only within the class of $\mathbb{Z}_2$-symmetric vector fields. In this context, one has an interaction between a pitchfork and a Hopf bifurcation. For this case, Langford and Iooss [1980] succeeded in showing the existence of invariant 2-tori provided the 5-jet satisfies a certain (implicitly given) nondegeneracy condition. In the case of a vector field in $\mathbb{R}^4$ with two pairs of purely imaginary eigenvalues of the linearization, one expects the existence of invariant 3-tori under suitable conditions (see Iooss and Langford [1980], and Guckenheimer [1980]). However, as far as we know, only in some symmetric cases has it been shown that the flow on some of these tori is actually quasi-periodic. In particular, Guckenheimer [1980] assumes a type of axial symmetry and Broer [1981b] and Braaksma and Broer [1981] deal with divergence-free vector fields.

Although quasi-periodic motions are chaotic in some sense, one should also mention that even much more complicated dynamical behavior has been discovered in these problems. In the case considered in the present paper, Guckenheimer [1981], [1982] showed that a generic perturbation of a certain truncation of the system possesses transversal homoclinic orbits and hence horseshoes. He uses a geometric argument based on an argument of Silnikov. The precise hypotheses can be expected to be difficult to check in specific examples of perturbations. P. Holmes [1980] applied Melnikov's method to prove the existence of transversal homoclinic orbits for a very particular unfolding of this singularity. Exact verifiable results are again difficult to obtain since the Melnikov function is exponentially small and is not seen at finite orders in perturbation theory (cf. Holmes and Marsden [1982]). These results together with our result, however, strongly suggests the verifiable coexistence of both quasi-periodic motions and horseshoes, even for equal parameter values. We plan to address our attention to this question in a forthcoming paper.

The structure of the paper is as follows: In §1 we discuss the normal form of the unfolding and we prove that there is a curve in the parameter space along which one has Neimark–Sacker bifurcations. We construct the bifurcating invariant tori near this curve and show that they are asymptotically stable if they bifurcate to the right, and unstable, if they bifurcate to the left. In §2 we show that there are continuous curves in parameter space emanating from the critical curve, along which one has invariant tori with quasi-periodic flow. Finally in §3 we apply our theory to a model system of reaction diffusion equations, the so-called Brusselator. Although this is a system of partial differential equations, center manifold theory and the method of Birkhoff–Poincaré normal forms is used to reduce it to the normal form discussed before.

1. Bifurcation of periodic solutions into invariant 2-tori. We consider the following unfolding of a three-dimensional codimension-two singularity (cf. Guckenheimer [1981], [1982]):

$$
\begin{aligned}
&\dot{r}=(\lambda-\sigma)r+arz+dr^3+erz^2+\cdots+O(l),\\
&\dot{z}=\lambda z+bz^2+cr^2+fr^2z+gz^3+\cdots+O(l),\\
&\dot{\theta}=\gamma+h_1z+h_2r^2+h_3z^2+\cdots+O(l)/r.
\end{aligned}
\tag{1.1}
$$

Here (r,z,θ) are cylindrical coordinates in $\mathbb{R}^3$. It is assumed that the vector field on the right is sufficiently smooth and has been written in Birkhoff–Poincaré normal form up to terms of order l, i.e. the coefficients $a,b,c,\cdots$ are real numbers, whereas $O(l)$ stands for functions which are of order l in r and z uniformly in θ. γ is a given positive constant, while λ and σ are (real) unfolding parameters. We shall refer to λ as the *bifurcation parameter* and to σ as the *splitting parameter* (cf. Langford [1979]). Note that we cannot further simplify the θ-equation as done by Guckenheimer [1981], [1982] or Broer [1982], since we are mainly interested in quasi-periodic solutions. Although invariant tori are preserved by these simplifications, Poincaré rotation numbers are in general changed.

If $abc\neq 0$ in (1.1), then the equation is classified by Langford [1979] into six qualitatively distinct cases according to the values of a,b, and c. In each case, if $\sigma\neq 0$, one has transcritical steady state bifurcations as well as Hopf bifurcations as λ varies. These are the *primary* and *secondary bifurcations*. In one case the possibility of a so-called tertiary bifurcation is allowed, i.e. the bifurcation of periodic solutions into invariant 2-tori (through the Neimark–Sacker bifurcation). This case arises when

$$
ab<0 \quad\text{and}\quad c(b-a)>0. \tag{H1}
$$

Under an additional nondegeneracy condition on d,e,f, and g given explicitly below in (1.14), we shall show in this section that such a bifurcation indeed occurs.

As far as primary and secondary bifurcations are concerned, there are three curves in the (λ,σ)-plane, along which such bifurcations occur, near the origin. These curves are given by the asymptotic formulas

$$
\begin{aligned}
&\mathcal{C}_1:\lambda=O(\sigma^2),\\
&\mathcal{C}_2:\lambda=\frac{b}{b-a}\sigma+O(\sigma^2),\\
&\mathcal{C}_3:\lambda=\sigma+O(\sigma^2),
\end{aligned}
\tag{1.2}
$$

respectively. Along $\mathcal{C}_1$ we have a transcritical stationary bifurcation from the trivial solution. For $\sigma>0$ the trivial solution loses its stability to the supercritical stationary solution branching from it. Along $\mathcal{C}_2$ we have a Hopf bifurcation from the stable stationary solutions that branched out along $\mathcal{C}_1$. For $\sigma>0$ this Hopf bifurcation is supercritical, and so the periodic solutions acquire the stability for a small parameter range. Finally, along $\mathcal{C}_3$ these periodic solutions run back into the trivial solution again in a subcritical Hopf bifurcation. Near this bifurcation point the periodic solutions are unstable. There is an intermediate neutral stability curve for the periodic solutions asymptotically given by

$$\mathcal{C}_4: \lambda = \frac{2b}{2b-a}\sigma + O(\sigma^2). \tag{1.3}$$

Here the Floquet exponents of the periodic solutions are purely imaginary. This gives rise to a tertiary bifurcation into invariant 2-tori (cf. Marsden and McCracken [1976] and Iooss [1979] for expositions). See Fig. 1. For $\sigma<0$ the Hopf bifurcation from the trivial solution occurs first as in Fig. 1(c).

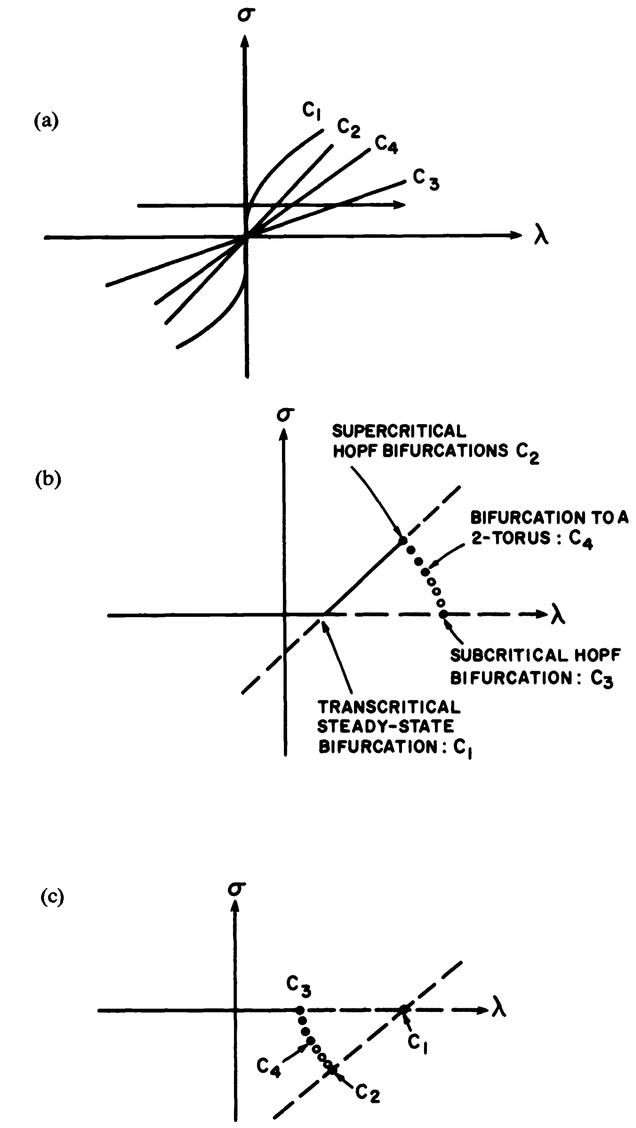

(a)
σ
C1
C2
C4
C3
λ
(b)
σ
SUPERCRITICAL HOPF BIFURCATIONS C2
BIFURCATION TO A 2-TORUS : C4
λ
SUBCRITICAL HOPF BIFURCATION: C3
TRANSCRITICAL STEADY-STATE BIFURCATION : C1
(c)
σ
C3
λ
C1
C4
C2

FIG. 1. (a) *The curves* $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$ *and* $\mathcal{C}_4$ *in the* (λ, σ)*-plane* (b) *The bifurcation diagram corresponding to the horizontal section shown in* (a); $\sigma>0$. (c) *The case* $\sigma<0$.

In order to construct these tori one proceeds as follows (see Guckenheimer [1982] for details): Truncate the equations in (1.1) after terms of order $l-1$. The truncated system is axisymmetric with respect to the z-axis and so the θ-equation decouples from the (r,z)-part, and corresponding to the periodic solutions discussed above for $r\neq 0$ are the zeros (with $r\neq 0$) of the vector field $f=(f_1,f_2)$ given by

$$\begin{aligned}(1.4)\qquad & f_1(\sigma,\lambda,r,z)=(\lambda-\sigma)r+arz+\cdots+(\text{monomial of order } l-1),\\ & f_2(\sigma,\lambda,r,z)=\lambda z+bz^2+cr^2+\cdots+(\text{monomial of order } l-1).\end{aligned}$$

Their "Floquet exponents" are given by the formula

$$(1.5)\qquad \mu_\pm=\frac{1}{2}\operatorname{tr}Df\pm\left(\frac{1}{4}\operatorname{tr}^2Df-\det Df\right)^{1/2}.$$

Here Df is the Jacobian matrix of f with respect to r and z. Hence, for the reduced system, the neutral stability curve for these solutions determined by the condition $\operatorname{Re}\mu_\pm=0$, is given by

$$(1.6)\qquad f=0,\qquad \operatorname{tr}Df=0,$$

assuming $\det Df>0$. In order to solve (1.6), we introduce rescaled variables as follows:

$$(1.7)\qquad \lambda=\sigma\tilde{\lambda},\quad r=|\sigma|\tilde{r},\quad z=\sigma\tilde{z},\quad \tilde{r}>0,$$

and consider the corresponding rescaled (or "blown-up") vector field

$$(1.8)\qquad \tilde{f}=\sigma^{-2}f.$$

Note that there is a reflectional symmetry in our problem which guarantees that for given λ and σ, solutions appear in pairs $(\theta,\pm r,z)$. Hence we can replace $|\sigma|$ by σ in (1.7). For small $|\sigma|$, the implicit function theorem applies to yield a solution

$$(1.9)\qquad \tilde{\lambda}=\lambda_0(\sigma)=\frac{2b}{2b-a}+O(\sigma)$$

of (1.6). Moreover, the implicit function theorem applies to yield the periodic orbits

$$(1.10)\qquad \tilde{r}=r_0(\sigma,\tilde{\lambda}),\qquad \tilde{z}=z_0(\sigma,\tilde{\lambda})$$

of $\tilde{f}$ for parameter values near the curve given by (1.9). Their Floquet exponents $\mu_\pm=\alpha\pm i\beta$ have the properties

$$(1.11)\quad \alpha(\sigma,\lambda_0(\sigma))\equiv 0,\quad \frac{\partial\alpha}{\partial\tilde{\lambda}}(0,\lambda_0(0))=\alpha_0=\frac{1}{2}-\frac{b}{a}>0,\quad \beta(0,\lambda_0(0))=\beta_0>0.$$

All these functions are smooth (in fact locally analytic). Note that for practical reasons it suffices to compute them up to terms of order $l-1$ in σ.

Next we introduce local variables near the periodic orbits (1.10)

$$(1.12)\qquad \tilde{\lambda}=\lambda_0(\sigma)+\Lambda,\quad \tilde{r}=r_0(\sigma,\tilde{\lambda})+R,\quad \tilde{z}=z_0(\sigma,\tilde{\lambda})+Z,$$

and transform the linearized vector field $D\tilde{f}$ to Jordan normal form. This can be achieved by a similarity transformation which is analytic in σ and Λ. Furthermore, the transformed vector field is brought into Birkhoff–Poincaré normal form up to terms of order four. Again, this can be done via a coordinate transformation which is analytic in

σ and $\tilde{\Lambda}$. We end up with a vector field, the components of which have the following form in polar coordinates $(\tilde{\rho},\phi)$, $\tilde{\rho}\geq 0$, in the $(\tilde{R},\tilde{Z})$ plane:

$$\begin{aligned} \tilde{f}_{\tilde{\rho}} &= \alpha\tilde{\rho} = A\tilde{\rho}^3 + O(\tilde{\rho}^5), \\ \tilde{f}_{\phi} &= \beta + B\tilde{\rho}^2 + O(\tilde{\rho}^4). \end{aligned} \tag{1.13}$$

Here α and β are as in (1.11), A and B are functions of σ and $\tilde{\Lambda}$, and the symbols $O(\tilde{\rho}^5)$ and $O(\tilde{\rho}^4)$ stand for functions which are of order five and four, respectively, uniformly in ϕ,σ and $\tilde{\Lambda}$ and which are 2π-periodic in ϕ. All functions are smooth (in fact analytic). We have the following asymptotic formula for A:

$$A(\sigma,\tilde{\Lambda}) = \frac{b}{(2b-a)^2}\left[4b\left(\frac{b}{a}-1\right)d - \frac{4bc}{a}e - 2bf + 6cg\right]\sigma + O(|\tilde{\Lambda}|+\sigma^2). \tag{1.14}$$

This formula is obtained by straightforward calculations as outlined above. We also have $A(0,0)=0$. This is a consequence of the fact that the vector field in (1.8) is Hamiltonian and integrable for $\sigma=0$ and $\tilde{\lambda}=\lambda_0(0)$ (see Fig. 2, Guckenheimer [1982] and Langford [1982]). We shall make the following hypothesis:

$$\frac{\partial}{\partial\sigma}A(0,0) = \Omega \neq 0. \tag{H2}$$

For given values of a,b,c, and h_1, this is a nondegeneracy assumption for the third order terms in (1.1). Formula (1.14) tells us that this assumption is fulfilled for all values of the coefficients h_2 and h_3, and for (d,e,f,g) outside a hyperplane in $\mathbb{R}^4$. Thus it is generic within the class of vector fields considered here.

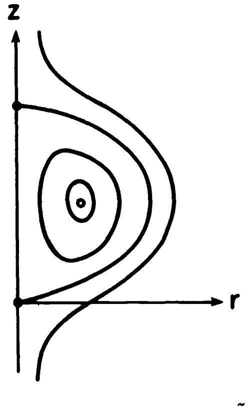


FIG. 2. *The homoclinic orbit and phase portrait for $\tilde{f}$ with $\sigma=0$ and $\tilde{\lambda}=\lambda_0(0)$.*

If we neglect the higher order terms in (1.13) for a moment, zeros of the equation

$$\alpha\tilde{\rho} + A\tilde{\rho}^3 = 0 \tag{1.15}$$

correspond to limit cycles of that system. Moreover, if we add the equation $\dot{\theta}=\gamma$, they correspond to invariant 2-tori of the $(\theta,\tilde{r},\tilde{z})$-system. In fact, we shall use these tori as a zeroth approximation for the invariant tori of the complete system.

Because of (1.11) and (H2) we can write

$$\alpha(\sigma,\tilde{\Lambda}) = \tilde{\Lambda}\tilde{\alpha}(\tilde{\sigma},\tilde{\Lambda}), \qquad A(\sigma,\tilde{\Lambda}) = \sigma A_1(\sigma,\tilde{\Lambda}) + \tilde{\Lambda}A_2(\sigma,\tilde{\Lambda}), \tag{1.16}$$

where the functions $\tilde{\alpha}, A_1$ and A_2 are smooth (even analytic), and

$$\tilde{\alpha}(0,0)=\alpha_0>0, \qquad A_1(0,0)=\Omega\neq 0. \tag{1.17}$$

Hence, if $|\tilde{\Lambda}|$ is sufficiently small depending on σ, then (1.15) has a solution

$$\rho_0=\sqrt{\frac{-\tilde{\Lambda}\tilde{\alpha}}{\sigma A_1+\tilde{\Lambda}A_2}}\,, \tag{1.18}$$

either for $\tilde{\Lambda}\geq 0$ if $\sigma\Omega<0$, or for $\tilde{\Lambda}\leq 0$ if $\sigma\Omega>0$. Thus, a vertical bifurcation of tori is excluded by our assumptions. Observe that $\rho_0\to 0$ as $\tilde{\Lambda}\to 0$.

In order to continue these tori to the complete system, we now introduce the rescaled parameter Λ via

$$\tilde{\Lambda}=\sigma^m\Lambda, \quad \delta\leq|\Lambda|\leq 1, \quad \Lambda\Omega<0. \tag{1.19}$$

where $m\geq 1$ is an odd integer to be determined later and δ is some given positive number. The restriction of Λ to such a domain is quite natural, because $\tilde{\Lambda}=0$ corresponds to the neutral stability curve of the truncated system and differs from $\mathcal{C}_4$ to a higher order in σ. Moreover, we restrict the $\tilde{\rho}$-variable to a neighborhood of the tori given by (1.18), i.e. we introduce a local action variable ρ via

$$\tilde{\rho}=\rho_0+\sigma^n|\Lambda|^{1/2}\rho, \qquad |\rho|\leq 1, \tag{1.20}$$

where the integer $n\geq m$ will be chosen later.

After making the above substitutions of variables in (1.1) (the θ-variable is left unchanged so far), we end up with a system of the following type:

$$\begin{aligned}\dot{\rho}&=\sigma\{2\rho_0^2A\rho+g_1(\sigma,\Lambda,\phi,\rho)+h_1(\sigma,\Lambda,\theta,\phi,\rho)\},\\ \dot{\phi}&=\sigma\{\beta+B\rho_0^2+g_2(\sigma,\Lambda,\phi,\rho)+h_2(\sigma,\Lambda,\theta,\phi,\rho)\},\\ \dot{\theta}&=\gamma+\sigma\{f_0(\sigma,\Lambda,\phi)+g_3(\sigma,\Lambda,\phi,\rho)+h_3(\sigma,\Lambda,\theta,\phi,\rho)\}.\end{aligned} \tag{1.21}$$

Here the functions ρ_0, β, A, B, f_0 and g_k $(k=1,2,3)$ are analytic in the variables σ, $|\Lambda|^{1/2}$, ϕ,ρ restricted to the indicated regions. Furthermore, f_0 and g_k are 2π-periodic trigonometric polynomials of finite degree in ϕ. f_0 can in fact be chosen so that it has degree two by absorbing the higher order terms in g_3. Moreover, these functions have the following order of magnitude with respect to σ, uniformly in the other variables as $|\sigma|\to 0$:

$$\begin{aligned}&\rho=O\big(|\sigma|^{(m-1)/2}\big), \qquad \beta=O(1),\\ &A=O(|\sigma|), \qquad B=O(1),\\ &f_0=O(1), \qquad g_k=O\big(|\sigma|^n+|\sigma|^{5(m-1)/2-n}\big) \qquad (k=1,2,3).\end{aligned} \tag{1.22}$$

The functions h_k $(k=1,2,3)$ are as smooth as the original vector field with respect to θ,ϕ,ρ and are at least continuous in the parameters, provided that n is not too large compared with l. They are 2π-periodic in both the variables ϕ and θ, and

$$h_k=O\big(|\sigma|^{l-2}+|\sigma|^{l-n-1}\big) \quad \text{as } |\sigma|\to 0. \tag{1.23}$$

Observe that without loss of generality we can assume that f_0 does not explicitly depend on ϕ. For otherwise, we transform the θ-variable via

$$(1.24) \qquad \theta=\tilde{\theta}+\Psi(\sigma,\Lambda,\phi),$$

where Ψ is the solution of the equation

$$(1.25) \qquad \frac{\partial\Psi}{\partial\phi}=\frac{f_0-[f_0]}{\beta+B\rho_0^2}$$

Here $[f_0]$ denotes the constant term of the trigonometric polynomial f_0. Clearly (1.25) has a unique solution, and this is again a 2π-periodic trigonometric polynomial in ϕ with coefficients depending smoothly (even analytically) on σ and $|\Lambda|^{1/2}$. The new θ-equation reads as follows:

$$(1.26) \qquad \dot{\tilde{\theta}}=\gamma+\sigma[f_0]+\sigma\{g_3+h_3\}-\sigma\frac{\partial\Psi}{\partial\phi}\{g_2+h_2\}.$$

The other equations in (1.21) are not affected by this transformation. Subsequently we shall simply assume that f_0 does not depend on ϕ in (1.21).

Now we are ready to apply Sacker's theorem (Sacker [1965, Thm. 1]) on invariant submanifolds to prove the following result:

THEOREM 1.1. *Let the vector field in* (1.1) *be of class C^r with $r\geq 21$. Moreover, let the coefficients $a,b,c,\cdots$ satisfy the hypotheses* (H1) *and* (H2). *Let $\delta\in(0,1)$ be some given number and denote by $S_\delta^\pm$ the regions of the (λ,σ)-plane which are bounded by the two curves $\lambda=\sigma\lambda_0(\sigma)\pm\delta\sigma^{10}$ and $\lambda=\sigma\lambda_0(\sigma)\pm\sigma^{10}$, where $\lambda_0(\sigma)$ is as in* (1.9). *Then, there is a positive σ_0 such that, for all parameter values either in $S_\delta^+\cap\{(\lambda,\sigma)||\sigma|\leq\sigma_0\}$ or in $S_\delta^-\cap\{(\lambda,\sigma)||\sigma|\leq\sigma_0\}$ depending on the sign of Ω, the system* (1.1) *possesses an invariant 2-torus of class C^{r-1}. This torus depends continuously on the parameters. If $\Omega>0$ it exists in S_δ^+ and is locally attractive. If $\Omega<0$ it exists in S_δ^- and is locally repelling.* (*Thus we have the usual exchange of stability phenomenon.*)

Proof. We apply Sacker's theorem to the transformed system (1.21). To this end we set $l=21$, $m=9$, $n=10$ for the integers in (1.1), (1.19) and (1.20). Then, in Sacker's notation we have $u=\rho$, $x_1=\phi$, $x_2=\theta$, $\mu=\sigma^{10}$, $\omega_1=\sigma\beta+\sigma B\rho_0^2$, $\omega_2=\gamma+\sigma f_0$, $P_1=\sigma^2\rho_0^2A/\sigma^{10}$, $\hat{A}_1=\{g_2+h_2\}/\sigma^9$, $\hat{A}_2=\{g_3+h_3\}/\sigma^9$, and $\hat{G}=\{g_1+\mathrm{h}_1\}/\sigma^9$. So we are in his degenerate case i). Because of (H2), P_1 is strictly bounded away from zero. The sign of P_1 is equal to that of Ω. Furthermore, (1.22) and (1.23) imply that the functions $\hat{A}_k$ and $\hat{G}$ together with all their derivatives with respect to θ,ϕ,ρ, up to order r tend to zero uniformly as $\sigma\to 0$. Hence, for $\delta\leq|\Lambda|\leq 1$ and $|\Lambda|$ sufficiently small Sacker's theorem applies to yield an invariant manifold of (1.21) given by a function

$$(1.27) \qquad \rho=\tau(\sigma,\Lambda,\theta,\phi),$$

which is C^{r-1} and 2π-periodic in θ and ϕ and together with its derivatives continuous in Λ and σ. Note that the estimates in Sacker's proof are independent of $\omega=(\omega_1,\omega_2)$ in bounded regions. Hence it does not matter that ω depends on σ in our case. Now, in view of (1.19), Λ has to be chosen positive if $\Omega<0$ and negative if $\Omega>0$. This proves the two alternatives in the existence part of the theorem. In order to prove the stability assertion, note that the constructed torus is locally stable if $-\mu P_1$ is positive i.e., if $\Omega<0$ (cf. (1.16)) and unstable if this matrix is negative, i.e. if $\Omega>0$. □

Remark 1.2. The regularity assumption in Theorem 1.1 is obviously not optimal. Moreover, the existence domain of the tori in the (λ,σ)-plane is only estimated very roughly. It cannot be expected in general that the tori reach the curve $\lambda=\sigma\lambda_0(\sigma)$, i.e. $\delta=0$, for this curve is only an approximate neutral stability curve for the periodic solutions of (1.1). However, we still aim to show that the tori actually emanate along curves from the exact neutral stability curve $\mathcal{C}_4$.

To prove this, one has to localize the vector field around the exact periodic solutions of (1.1) rather than around approximate solutions as in (1.12). Moreover, the complete 2-jet of the localized vector field has to be taken into consideration for the construction of a zeroth approximation of the tori. To this end, one has to transform away all nonresonant terms of the 2-jet. In fact, there is an almost identical transformations of the variables $\tilde{\Lambda}, R, Z, \theta$ of the form

$$(1.28)\qquad \begin{aligned} \tilde{\Lambda}&=\tilde{\tilde{\Lambda}}+\Lambda_0(\sigma),\\ R&=\tilde{R}+R_0(\sigma,\tilde{\Lambda},\tilde{\theta},\tilde{R},\tilde{Z}),\\ Z&=\tilde{Z}+Z_0(\sigma,\tilde{\Lambda},\tilde{\theta},\tilde{R},\tilde{Z}),\\ \theta&=\tilde{\theta}+\psi_0(\sigma,\tilde{\Lambda},\tilde{\theta},\tilde{R},\tilde{Z}) \end{aligned}$$

such that, in the new variables, our method applies uniformly for $|\tilde{\tilde{\Lambda}}|\leq 1$ and $|\sigma|$ sufficiently small. The periodic solutions are given by $\tilde{R}=0$, $\tilde{Z}+=0$, and $\tilde{\tilde{\Lambda}}=0$ corresponds to the curve $\mathcal{C}_4$. The remaining resonant terms in the 2-jet (with respect to $\tilde{R}$ and $\tilde{Z}$) of the transformed vector field are independent of $\tilde{\theta}$ and differ from the old ones only by terms of order greater than or equal to l in σ. The same is true for higher order terms. In (1.28) all functions are smooth, 2π-periodic in $\tilde{\theta}$, and polynomial of order two in $\tilde{R}$ and $\tilde{Z}$. Such a transformation is easily constructed using the classical implicit function theorem, for there are no small divisors involved.

COROLLARY 1.3. *Let the assumptions of Theorem* 1.1 *hold. Then, for each fixed* σ *sufficiently small, there is a continuous branch of invariant* 2-*tori of* (1.1) *which emanates at the curve* $\mathcal{C}_4$ *from periodic solutions. It can be parametrized by* λ. *For each parameter value, the torus contains the corresponding periodic orbit in its interior.* (*The torus collapses to the periodic orbit if the curve* $\mathcal{C}_4$ *is approached.*) *It is locally stable if it bifurcates supercritically, and unstable if it bifurcates subcritically.*

Remark 1.4. A result is similar to Corollary 1.3 has been proved by Langford [1979]. However, his condition is rather implicit, and it seems to be hard to check it in applications.

Remark 1.5. In many formulations for the bifurcation to tori, a nonresonance condition up to order four is needed. This does not occur here since the characteristic exponents of the periodic solutions tend to zero as $\sigma\to 0$ while their period tends to $\gamma>0$.

2. The existence of bifurcating tori with quasi-periodic flow. It is well known (cf. Iooss [1979]), that the flow on a bifurcating 2-torus is qualitatively described by Poincaré's rotation number p. If p is rational, then the flow is periodic and otherwise it is quasi-periodic, so ergodic. Also it is known that there is a set of parameter values of positive Lebesgue measure for which the flow is ergodic, provided that p varies effectively with a parameter. The measure of this set tends to 1 when the corresponding family of flows approaches a family of parallel flows (see Arnol'd [1965] and Herman [1977]). The purpose of this section is to show that this actually happens in the present situation for generic paths through the (λ,σ)-plane close to the origin. In particular we shall see that there are specific continuous curves emanating from $\mathcal{C}_4$ along which all the tori are quasi-periodic. In between these curves one expects the phenomenon of phase locking to occur which means that p takes constant rational values in open regions of the parameter space (see Arnol'd [1965]).

Let $\rho=\tau(\sigma,\Lambda,\theta,\phi)$ be the invariant 2-manifold of (1.21) constructed in the previous section. Then the flow on this submanifold induced by (1.21) is given by the equations

$$\dot\phi=\sigma\{\beta+B\rho_0^2+g_2+h_2\},\qquad \dot\theta=\gamma+\sigma\{f_0+g_3+h_3\} \tag{2.1}$$

where the ρ-argument in the functions g_k and h_k (in (1.21)) is replaced by τ. Rescaling the time by a factor $\kappa\sim 1$ we get

$$\dot\phi=\frac{\sigma}{\kappa}\{\beta+B\rho_0^2+g_2+h_2\},\qquad \dot\theta=\frac{\gamma}{\kappa}+\frac{\sigma}{\kappa}\{f_0+g_3+h_3\}. \tag{2.2}$$

Now, because $\beta(0,\lambda_0(0))=\beta_0>0$ and $\gamma>0$, we can replace σ and κ in (2.2) by new parameters

$$\omega_1=\frac{\sigma}{\kappa}(\beta+B\rho_0^2),\qquad \omega_2=\frac{\gamma}{\kappa}+\frac{\sigma}{\kappa}f_0. \tag{2.3}$$

Recall that ρ_0^2 is of order $O(|\sigma|^{m-1})$ as $\sigma\to 0$ and $f_0=O(1)$ can be made independent of ϕ. Hence, when $m\geq 3$ the implicit function theorem gives a unique smooth solution

$$\sigma=\sigma(\Lambda,\omega_1,\omega_2),\qquad \kappa=\kappa(\Lambda,\omega_1,\omega_2) \tag{2.4}$$

of equations (2.3) for ω_1 near 0 and ω_2 near γ. In particular, we have

$$\sigma(\Lambda,0,\gamma)=0,\qquad \kappa(\Lambda,0,\gamma)=1,\quad \sigma=O(|\omega_1|)\quad \text{as } \omega_1\to 0. \tag{2.5}$$

We shall look for curves in (ω_1,ω_2)-space parametrized by Λ, where the flow of (2.2) is quasi-periodic with two given basic frequencies $|\omega_1^0|$ and ω_2^0 near 0 and γ, respectively (ω_1^0 is allowed to be negative). To this end we first consider the modified system

$$\dot\phi=\omega_1^0+\Delta\omega_1+\frac{\sigma}{\kappa}\{g_2+h_3\},\qquad \dot\theta=\omega_2^0+\Delta\omega_2+\frac{\sigma}{\kappa}\{g_3+h_3\}. \tag{2.6}$$

Note that the right-hand sides of these equations are smooth in ϕ and θ and, together with their derivatives, continuous in Λ, ω_1 and ω_2 (recall that we have only proved that τ is continuous in σ and Λ).

LEMMA 2.1. *Let the functions g_k and h_k be of class C^r, $r\geq 8$, in θ and ϕ and let them satisfy* (1.22) *and* (1.23) *with respect to $\sigma(l,m,n$ as in* §1). *Moreover, let q be a given integer with $q\leq\min(n,5(m-1)/2-n,\ l-2,l-n-1)$. Then, there is a positive constant c with the following property. If the vector $\omega^0=(\omega_1^0,\omega_2^0)$ is contained in the set*

$$I_\varepsilon=\left\{\omega^0\in\mathbb{R}\,\middle|\,|j_1\omega_1^0+j_2\omega_2^0|\geq\varepsilon|j|^2\ \textit{for all}\ j\neq 0\in\mathbb{Z}^2\right\} \tag{2.7}$$

for some $\varepsilon\in(0,1)$, then there is a continuous function $\Delta\omega=\Delta\omega(\Lambda,\omega)$ defined for

$$|\omega_1|<c\varepsilon^{2/\{\min(n,5(m-1)/2-n,l-2,l-n-1)-q+1\}} \tag{2.8}$$

and a 2π-periodic coordinate transformation of class C^{r-3}

$$\phi=\tilde\phi+U(\Lambda,\omega,\tilde\phi,\tilde\theta),\qquad \theta=\tilde\theta+V(\Lambda,\omega,\tilde\phi,\tilde\theta) \tag{2.9}$$

continuous in Λ and ω, such that (2.6) *is transformed into*

$$\dot{\tilde\phi}=\omega_1^0,\qquad \dot{\tilde\theta}=\omega_2^0. \tag{2.10}$$

In particular, we have

$$(2.11) \qquad \Delta\omega = O\left(|\omega_1|^q\right) \quad \textit{as } \omega_1 \to 0.$$

Proof. The existence of such a transformation is proved, for instance, by Zehnder [1975, Thm. 4.1] (cf. also Moser [1966] and Hörmander [1977]). The estimates (2.8) and (2.11) are due to the fact, that the functions $(\sigma/\kappa)\{g_2+h_2\}/\varepsilon^2\omega_1^q$ and $(\sigma/\kappa)\{g_3+h_3\}/\varepsilon^2\omega_1^q$, together with their derivatives, have to be sufficiently small. Here the factor ε^2 in the denominator stems from the nonresonance condition (2.7) and the second factor ω_1^q guarantees (2.11). In view of (1.22), (1.23) and (2.5), these functions can be made arbitrarily small by choosing an appropriate c in (2.8). □

In particular, this lemma implies that the flow of the modified system (2.6) is quasi-periodic with the two basic frequencies ω_1^0 and ω_2^0:

$$(2.12) \qquad \phi = \omega_1^0 t + U\left(\Lambda, \omega, \omega_1^0 t, \omega_2^0 t\right), \qquad \theta = \omega_2^0 t + V\left(\Lambda, \omega, \omega_1^0 t, \omega_2^0 t\right).$$

Hence, it remains to be shown that there is a curve in (ω_1, ω_2)-space, on which the systems (2.2) and (2.6) agree. Such a curve is determined by solutions of the equations

$$(2.13) \qquad \omega_1 = \omega_1^0 + \Delta\omega_1(\Lambda, \omega_1, \omega_2), \qquad \omega_2 = \omega_2^0 + \Delta\omega_2(\Lambda, \omega_1, \omega_2).$$

We shall solve (2.13) for ω_1 and ω_2 using Brouwer's fixed point theorem.

Let $\eta = \eta(\varepsilon)$ denote the right-hand side of the inequality in (2.8), and set $\omega_0 = (0, \gamma)$. Then, with respect to ω, the right-hand sides of the equations in (2.13) define a continuous map from the ball $B_\eta(\omega_0)$ with center ω_0 and radius η, into $\mathbb{R}^2$. This map depends continuously on Λ. In order to make sure that it maps $B_\eta(\omega_0)$ into itself, we require

$$(2.14) \qquad \left|\omega^0 - \omega_0\right| \le \eta - \tilde{c}\eta^q,$$

where the constant $\tilde{c}$ stems from (2.11). Here a difficulty arises. Namely, it is not clear whether or not there are elements ω^0 in I_ε satisfying (2.14). But a measure-theoretical result helps. It is "well known" (see Siegel and Moser [1971] and Rüssmann [1979]) that the Lebesgue measure of the complement set of I_ε in any ball is of order $O(|\varepsilon|)$ as $\varepsilon \to 0$, uniformly for balls with small radius, i.e. we have

$$(2.15) \qquad \mu\left(B_{\eta - \tilde{c}\eta^q}(\omega_0) \setminus I_\varepsilon\right) \le \tilde{\tilde{c}}\varepsilon$$

with some constant $\tilde{c}$, where μ denotes Lebesgue measure. Thus

$$(2.16) \qquad \pi\left(\eta - \tilde{c}\eta^q\right)^2 - \tilde{\tilde{c}}\varepsilon > 0$$

implies

$$(2.17) \qquad B_{\eta - \tilde{c}\eta^q}(\omega_0) \cap I_\varepsilon \neq \emptyset.$$

Hence, if $q > 1$ and η is of order less than $\sqrt{\varepsilon}$ as $\varepsilon \to 0$, there are many frequency vectors ω^0 such that Brouwer's fixed point theorem applies to yield a solution

$$(2.18) \qquad \omega = \omega(\Lambda)$$

of (2.13) which depends continuously on Λ (see Chow, Mallet-Paret and York [1978]). In fact, the relative measure of the set $B_{\eta - \tilde{c}\eta^q}(\omega_0) \cap I_\varepsilon$ in this ball tends to 1 as $\varepsilon \to 0$.

THEOREM 2.2. *Let the assumptions of Theorem* 1.1 *hold. Let $S_\delta^\pm$ be the regions of the (λ,σ)-plane as defined there, and let β_0 be as in* (1.11). *Then each interval around* 0 *contains a subset I with the following properties: The relative measure of I tends to* 1 *as the interval shrinks to the point* 0. *For all $\sigma_0 \in I$, near the line $\sigma=\sigma_0$ there is a continuous curve in $S_\delta^\pm$*

$$\lambda=s_\pm^1(\sigma_0,\Lambda),\qquad \sigma=s_\pm^2(\sigma_0,\Lambda)\qquad (\delta\leq\Lambda\leq 1), \tag{2.19}$$

along which the flow on the tori constructed in the previous section is quasi-periodic with two basic frequencies $\nu_1\sim\sigma_0\beta_0$ and $\nu_2\sim\gamma$. The frequency ratio ν_1/ν_2 is a constant along each curve. These curves join the two boundaries of $S_\delta^\pm$ and form a set of positive measure.

Proof. Let the integers l,m,n be chosen as in the proof of Theorem 1.1 and set $q=2$. The right-hand side of the inequality in (2.8) is of order less than $\sqrt{\varepsilon}$ as $\varepsilon\to 0$. Thus, according to the above discussion and (2.4), for almost all frequency vectors ω^0 near $\omega_0=(0,\gamma)$, there are values $\sigma=\sigma(\Lambda,\omega^0)$, $\kappa=\kappa(\Lambda,\omega^0)$ depending continuously on Λ, such that the flow of (2.2) is quasi-periodic and given by (2.12). In view of (2.3) and (2.11), we have

$$\left|\frac{\sigma\beta_0}{\kappa}-\omega_1^0\right|=O\left(|\sigma|^2\right),\qquad \left|\frac{\gamma}{\kappa}-\omega_2^0\right|=O\left(|\sigma|\right)\quad \text{as } \sigma\to 0. \tag{2.20}$$

Now, consider the curves $\sigma=\sigma(\Lambda,\omega^0)$ in parameter space. In the original time scaling, the corresponding frequencies are $\nu_1=\kappa\omega_1^0$ and $\nu_2=\kappa\omega_2^0$. Hence, for different ratios ω_1^0/ω_2^0 these curves cannot intersect. However, if the ratios for two different ω^0 are coincident, the corresponding curves may coincide. Therefore, it suffices to consider this subspace of ω^0's where $\omega_2^0=\gamma$. In this case, (2.3) implies

$$\kappa=1+O\left(|\sigma|\right). \tag{2.21}$$

Thus, with

$$\sigma_0=\omega_1^0/\beta_0, \tag{2.22}$$

we conclude from (2.20) that

$$|\sigma-\sigma_0|-O\left(|\sigma_0|^2\right),\quad |\nu_1-\omega_1^0|=O\left(|\sigma_0|^2\right),\quad |\nu_2-\omega_2^0|=O\left(|\sigma_0|\right) \tag{2.23}$$

as $\sigma_0\to 0$. Now set

$$s_\pm^1(\sigma_0,\cdot)=\sigma(\omega^0,\cdot),s_\pm^2=s_\pm^1\lambda_0\left(s_\pm^1\right)+\left(s_\pm^1\right)^{10}\Lambda \tag{2.24}$$

with λ_0 from (1.9), and note that the measure theoretical result mentioned above, carries over to the subspace of frequency vectors ω^0 with fixed second component $\omega_2^0=\gamma$. Thus, the theorem follows. □

Remark 2.3. As in the case of Corollary 1.3 it can be shown that the curves in Theorem 2.2 actually emanate at the neutral stability curve $\mathcal{C}_4$ of the periodic solutions of (1.1). This has an interesting consequence. Passing through the parameter space on a path which runs into one of these curves, one has a bifurcation from periodic orbits into quasi-periodic solutions with two basic frequencies. See Fig. 3. These are stable if they are supercritical. Although individual quasi-periodic flows are nongeneric, one generically sees this bifurcation in the unfolding in (1.1) which thus exhibits the first part of the Landau sequence of transitions from trivial stationary solutions to quasi-periodic solutions.

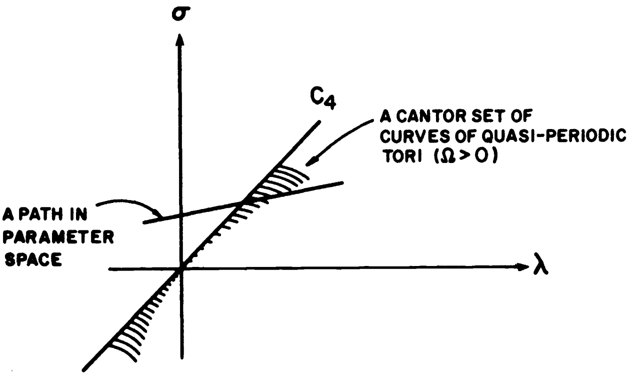

σ
C_4
A CANTOR SET OF CURVES OF QUASI-PERIODIC TORI (Ω>O)
A PATH IN PARAMETER SPACE
λ

FIG. 3. *A path in parameter space generally meets a set of positive measure of quasi-periodic tori.*

Remark 2.4. Note, that for an analytic vector field in (1.1) the tori with quasi-periodic flow and frequency vectors contained in some set I_ε are analytic 2-manifolds. This can, for instance, be proved by the method in Scheurle [1982], which combines the construction of the torus, with that of the flow on it and thus avoids a separate reduction step as in Theorem 1.1 (cf. also Bogoliubov, Mitropolskii and Samoilenko [1976]). In general, even C^∞-smoothness is lost by such a reduction process (see Sacker [1965]).

Remark 2.5. Consider the differential equation

$$x'=f(\sigma,\lambda,x), \tag{2.25}$$

where f: $\mathbb{R}\times\mathbb{R}\times\mathbb{R}^3\to\mathbb{R}^3$ is sufficiently smooth. Assume the existence of a trivial solution $x\equiv 0$, i.e. $f(\sigma,\lambda,0)\equiv 0$. Moreover, let $D_x f(\sigma,\lambda,0)$ have the eigenvalues $\alpha(\sigma,\lambda)$, and $\beta(\sigma,\lambda)\pm i\gamma(\sigma,\lambda)$ where $\alpha(0,0)=0$, $\beta(0,0)=0$, $\gamma(0,0)\neq 0$ and

$$\frac{\partial(\alpha,\beta)}{\partial(\sigma,\lambda)}\neq 0 \quad \text{at } (\sigma,\lambda)=(0,0). \tag{2.26}$$

This is a generalized Hopf condition due to Langford [1979]. Then, introducing α and $\alpha-\beta$ as new parameters, (2.25) can be transformed to a system of form (1.1). Of course, the coefficients $a,b,c,\cdots$, and γ will depend on the parameters in general. However, our theory still applies to this situation, if in (H1), (1.2), (1.3), (1.9), (1.11) and (1.14) the values of the coefficients at $\sigma=0$, $\lambda=0$ are inserted.

3. The Brusselator. We apply our theory to a model system of reaction diffusion equations, namely to the so-called Brusselator. Although this system originally consists of partial differential equations, there are parameter values where the solutions we are interested in, lie in a three-dimensional, locally invariant center manifold, and the restriction to this submanifold has a singularity of the type discussed in the previous sections. In particular, the generalized Hopf condition (2.26) is fulfilled, so that (1.1) is a reasonable unfolding. This example has been studied in detail by Guckenheimer [1982] (see also Keener [1976]). (See Schaeffer and Golubitsky [1981] for a discussion of the steady state bifurcations near a double zero eigenvalue.) We shall give explicit formulas for the relevant coefficients and in particular we compute the third order coefficients. We shall use a complex representation for convenience.

The following reaction schema is considered:

$$A\to X,\quad B+X\to Y+D,\quad 2X+Y\to 3X,\quad X\to E.$$

Here A, B, D, E are reactants whose concentrations are assumed to be fixed throughout the reaction. It is the dynamics of the intermediates X and Y which is examined. In addition, the reaction is assumed to take place in a one-dimensional medium (with position variable ξ) and that X and Y diffuse with diffusion constants D_1 and D_2. This yields the following system of reaction diffusion equations

$$\begin{aligned} &(3.1) && \frac{\partial X}{\partial t}=D_1\frac{\partial^2 X}{\partial \xi^2}+X^2Y-(B+1)X+A,\\ & && \frac{\partial Y}{\partial t}=D_2\frac{\partial^2 Y}{\partial \xi^2}-X^2Y+BX. \end{aligned}$$

It is further assumed that the reaction is at equilibrium at the end points of the interval $[0,\pi]$ so that $X(0)=X(\pi)=A$ and $Y(0)=Y(\pi)=B/A$ for all $t\geq 0$.

Obviously, the Brusselator problem has the trivial equilibrium solution $X(\xi,t)=A$, $Y(\xi,t)=B/A$. Let us introduce the relative coordinates

$$(3.2) \qquad u=X-A, \qquad v=Y-B/A.$$

Then the equations (3.1) become

$$\begin{aligned} &(3.3) && \frac{\partial u}{\partial t}=D_1\frac{\partial^2 u}{\partial \xi^2}+(B-1)u+A^2v+\left(\frac{B}{A}u^2+2Auv+u^2v\right),\\ & && \frac{\partial v}{\partial t}=D_2\frac{\partial^2 v}{\partial \xi^2}-Bu-A^2v-\left(\frac{B}{A}u^2+2Auv+u^2v\right) \end{aligned}$$

and the corresponding boundary conditions are $u(0,t)=u(\pi,t)=0$ and $v(0,t)=v(\pi,t)=0$. If $w=(u,v)$, we write $w_t=Lw+N(w)$, where Lw is the linear part of the right-hand side of (3.3). Let us consider L as a linear operator in the Banach space $C^0[0,\pi]$ with domain of a definition equal to the subspace $C_0^2[0,\pi]$ of C^2-functions which are zero at the boundary of $[0,\pi]$. Representing w as a Fourier series $w(\xi,t)=\sum_{n=1}^{\infty}w_n(t)\sin n\xi$, we find that the two-dimensional spaces spanned by the vector valued functions $w_n \sin n\xi$ are invariant for L with spectrum given by the eigenvalues of

$$E_n=\begin{pmatrix} B-1-n^2D_1 & A^2 \\ B & -A^2-n^2D_2 \end{pmatrix}.$$

In particular, L generates a holomorphic semi-group e^{Lt} in C^0.

Guckenheimer [1982] showed that there are parameter values for which, at the trivial equilibrium, there is a simple zero eigenvalue of E_k, pure imaginary eigenvalues $\pm i\gamma$ of E_1, and all other eigenvalues have negative real parts. This is a consequence of the fact that $\operatorname{tr} E_n$ is a monotonically decreasing function of n^2 and $\det E_n$ is strictly convex. If one regards (A^2,B) as being experimental parameters with the diffusion rates (D_1,D_2) fixed, then the corresponding conditions are

$$(3.4) \qquad A^2=D_2k^2\frac{D_1+D_2-D_1k^2}{1+D_1k^2-D_2k^2}, \qquad B=1+A^2+D_1+D_2$$

subject to the following inequalities on the diffusion rates

$$(3.5) \qquad \det E_{k\pm 1}>0.$$

There are solutions to (3.4) and (3.5) with A, B, D_1 and D_2 all positive.

Now let A^2 and B vary in a neighborhood of such a critical point. Then a straightforward computation shows that Langford's condition (2.26) is satisfied. Moreover, the center manifold theorem applies to yield a three-dimensional submanifold of C_0^2 near the origin (depending smoothly on the parameters) which is locally invariant and attractive for the flow induced by (3.3), and which contains all "small" solutions which are bounded for all $t \in \mathbb{R}$, in particular equilibria, periodic orbits and invariant tori near the origin. Hence, the restriction of (3.3) to this submanifold gives a complete description of the dynamics we are interested in. The corresponding vector field V can be transformed to the form (1.1). According to Remark 2.5 it suffices to compute the coefficients $a, b, c, \cdots$ at the critical parameter values, in order to check the hypotheses of §1 and 2.

Let E be the three-dimensional eigenspace for L corresponding to the eigenvalues with zero real part and let P: $C_0^2[0,\pi] \to E$ be the projection onto E. Moreover, let F be the complementary eigenspace, and set $Q = P - \mathrm{id}$, $Pw = x$, and $Qw = y$. Since the center manifold is tangent to E at $w = 0$, the Taylor expansions of the vector field V and the restriction of $P(L+N)$ to E agree up to terms of order two. Let us choose a basis for E such that the linear part is given in (complex) Jordan normal form

$$L|_E = \begin{pmatrix} -i\gamma & 0 & 0 \\ 0 & i\gamma & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \gamma(\det E_1)^{1/2}. \tag{3.6}$$

Then the second order terms of $N|_E$ are given by

$$\begin{aligned} N_2(x) = \big[& (x_1^2\alpha_{11} + x_2^2\alpha_{22} + x_1x_2\alpha_{12})\sin^2\xi + x_3^2\alpha_{33}\sin^2 k\xi \\ & + (x_1x_3\alpha_{13} + x_2x_3\alpha_{23})\sin\xi \sin k\xi\big] \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad x_j \in \mathbb{C}, \end{aligned} \tag{3.7}$$

where

(3.8)

$$\alpha_{11} = \frac{1}{A}\left(\frac{B}{2}(1 - D_2^2\gamma^{-2}) - A^2(1 - \gamma^{-2}D_2(1 + D_1))\right) + \frac{i}{A}(BD_2\gamma^{-1} - A^2\gamma^{-1}(1 + D_1 + D_2)),$$

$$\alpha_{12} = \frac{1}{A}(B(1 + \gamma^{-2}D_2^2) - 2A^2(1 + \gamma^{-2}D_2(1 + D_1))),$$

$$\alpha_{33} = \frac{A}{B}k^4D_2^2 - A^3B,$$

$$\alpha_{13} = \frac{\sqrt{2}}{A}(BD_2k^2 - A^2(A^2 + k^2D_2)) - \frac{i\sqrt{2}}{A}(A^2\gamma^{-1}(1 + D_1)(A^2 + k^2D_2) - B\gamma^{-1}k^2D_2^2),$$

$$\alpha_{22} = \bar{\alpha}_{11}, \qquad \alpha_{23} = \bar{\alpha}_{13}.$$

Now the coefficients a, b, c in (1.1) can be read off from $PN_2(x)$. We list the result of this computation:

$$a = \operatorname{Re}\alpha_{13}\frac{1+i}{\pi}\sqrt{2}\int_0^\pi \sin^2\xi \sin k\xi \, d\xi, \tag{3.9}$$

$$b = \alpha_{33}\frac{2\mu}{\pi}\int_0^\pi \sin^3 k\xi \, d\xi,$$

$$c = \alpha_{12}\cdot\frac{2\mu}{\pi}\int_0^\pi \sin^2\xi \sin k\xi \, d\xi,$$

$$\mu = \left((A^2 + k^2D_2)^2 - A^2B\right)^{-1}k^2D_2.$$

To compute third order coefficients requires considerably more effort. Here not only $PN_3(x)$ has to be considered but also the second order terms of the center manifold representation and of the Poincaré–Birkhoff transformation are involved. (The techniques used are analogous to those in Marsden and McCracken [1976] and Hassard and Wan [1978].) To be more precise, if the center manifold is given by

$$y=C(x), \qquad x\in E, \tag{3.10}$$

and if

$$x=\tilde{x}+T(\tilde{x}) \tag{3.11}$$

is the Poincaré–Birkhoff transformation, then we have to compute the (resonant) third order terms of the vector field

$$\tilde{X}(\tilde{x})=[\mathrm{id}+DT(\tilde{x})]^{-1}[PL\tilde{x}+PLT(\tilde{x})+PN(\tilde{x}+T(\tilde{x})+C(\tilde{x}+T(\tilde{x})))]. \tag{3.12}$$

Obviously, we have

$$\begin{aligned}\tilde{X}_3(\tilde{x})=&PN_3(\tilde{x})+2PN_2(\tilde{x},T_2(\tilde{x}))+2PN_2(\tilde{x},C_2(\tilde{x}))\\ &-DT_2(\tilde{x})(PLT_2(\tilde{x})+PN_2(\tilde{x})-DT_2(\tilde{x})PL\tilde{x}).\end{aligned} \tag{3.13}$$

Hence, $T_2(\tilde{x})$ and $C_2(\tilde{x})$ have to computed.

The contribution of $PN_3(\tilde{x})$ to the resonant terms in (1.1) is

(3.14)

$$\begin{aligned}d_1=&\frac{1}{2\pi}\left(-3-\gamma^{-2}D_2^2+2\gamma^{-1}D_2+(1+D_1)(-2\gamma^{-2}D_2+\gamma^{-1}+3\gamma^{-2}D_2^2)\right)\int_0^\pi \sin^4\xi\,d\xi,\\ e_1=&\frac{1}{\pi}\Big(-2B(A^2+k^2D_2)(1-\gamma^{-1}D_2)\\ &\qquad-(A^2+k^2D_2)^2(1-\gamma^{-1}(1+D_1))\Big)\int_0^\pi \sin^2\xi\sin^2 k\xi\,d\xi,\\ f_1=&\frac{2\mu}{\pi}\left(-2(A^2+k^2D_2)(1+\gamma^{-2}D_2(1+D_1))-B(1+\gamma^{-2}D_2^2)\right)\int_0^\pi \sin^2\xi\sin^2 k\xi\,d\xi,\\ g_1=&-\frac{2\mu}{\pi}B(A^2+k^2D_2)^2\int_0^\pi \sin^4 k\xi\,d\xi.\end{aligned}$$

Next we compute the second order terms $C_2(\tilde{x})$ in the center manifold representation (3.10). These are determined by the solution of the linear equation

$$DC_2(\tilde{x})L\tilde{x}=LC_2(\tilde{x})+QN_2(\tilde{x}), \qquad \tilde{x}\in E. \tag{3.15}$$

Setting

$$C_2(\tilde{x})=\sum_{i,j=1}^{3}\beta_{ij}\tilde{x}_i\tilde{x}_j, \qquad \beta_{ij}=\beta_{ji} \tag{3.16}$$

and using (3.7), we get the following expressions for the 2-vectors $\beta_{ij}=(\beta_{ij}^1,\beta_{ij}^2)$:

(3.17)
$$\begin{aligned}
\beta_{11}&=-\alpha_{11}\left(L_{1F}^{-1}+2i\gamma\right)^{-1}Q\begin{pmatrix}1\\-1\end{pmatrix}\sin^2\xi,\\
\beta_{12}&=-\frac{1}{2}\alpha_{12}L_{1F}^{-1}Q\begin{pmatrix}1\\-1\end{pmatrix}\sin^2\xi,\\
\beta_{22}&=-\alpha_{22}\left(L_{1F}^{-1}-2i\gamma\right)^{-1}Q\begin{pmatrix}1\\-1\end{pmatrix}\sin^2\xi,\\
\beta_{33}&=-\alpha_{33}L_{1F}^{-1}Q\begin{pmatrix}1\\-1\end{pmatrix}\sin^2 k\xi,\\
\beta_{13}&=-\frac{1}{2}\alpha_{13}\left(L_{1F}+i\gamma\right)^{-1}Q\begin{pmatrix}1\\-1\end{pmatrix}\sin\xi\sin k\xi,\\
\beta_{23}&=-\frac{1}{2}\alpha_{23}\left(L_{1F}-i\gamma\right)^{-1}Q\begin{pmatrix}1\\-1\end{pmatrix}\sin\xi\sin k\xi.
\end{aligned}$$

Note that

(3.18)
$$\beta_{12}=\bar{\beta}_{12},\quad \beta_{33}=\bar{\beta}_{33},\quad \beta_{11}=\bar{\beta}_{22},\quad \beta_{13}=\bar{\beta}_{23}.$$

Inserting (3.16) into $2PN_2(\tilde{x},C_2(\tilde{x}))$ yields the following contribution to d,e,f,g in (1.1):

(3.19)
$$\begin{aligned}
d_2=\operatorname{Re}\frac{2(1+i)}{\pi}\Bigg[&\left(\frac{B}{A}\left(1-i\gamma^{-1}D_2\right)-A\left(1-i\gamma^{-1}(1+D_1)\right)\right)\int_0^\pi\beta_{11}^1\sin^2\xi\,d\xi\\
&+\left(\frac{B}{A}\left(1+i\gamma^{-1}D_2\right)-A\left(1+i\gamma^{-1}(1+D_1)\right)\right)\int_0^\pi\beta_{12}^1\sin^2\xi\,d\xi\\
&+A\left(1-i\gamma^{-1}D_2\right)\int_0^\pi\beta_{11}^2\sin^2\xi\,d\xi+A\left(1+i\gamma^{-1}D_2\right)\int_0^\pi\beta_{12}^2\sin^2\xi\,d\xi\Bigg],
\end{aligned}$$

$$\begin{aligned}
e_2=\operatorname{Re}\frac{2\sqrt{2}\,(1+i)}{\pi}\Bigg[&\left(\frac{B}{\sqrt{2}\,A}\left(1+i\gamma^{-1}D_2\right)-\frac{A}{\sqrt{2}}\left(1+i\gamma^{-1}(1+D_1)\right)\right)\int_0^\pi\beta_{11}^1\sin^2\xi\,d\xi\\
&+\left(\frac{B}{A}\left(A^2+k^2D_2\right)-AB\right)\int_0^\pi\beta_{13}^1\sin k\xi\sin\xi\,d\xi\\
&+\frac{A}{\sqrt{2}}\left(1+i\gamma^{-1}D_2\right)\int_0^\pi\beta_{33}^2\sin^2\xi\,d\xi+A\left(A^2+k^2D_2\right)\int_0^\pi\beta_{13}^2\sin k\xi\sin\xi\,d\xi\Bigg],
\end{aligned}$$

$$\begin{aligned}
f_2=\operatorname{Re}\frac{4\sqrt{2}\,u}{\pi}\Bigg[&\left(\frac{B}{A}\left(1-i\gamma^{-1}D_2\right)-A\left(1-i\gamma^{-1}(1+D_1)\right)\right)\int_0^\pi\beta_{13}^1\sin\xi\sin k\xi\,d\xi\\
&+\left(\frac{B}{A}\left(A^2+k^2D_2\right)-AB\right)\frac{1}{\sqrt{2}}\int_0^\pi\beta_{12}^1\sin^2 k\xi\,d\xi\\
&+A\left(1-i\gamma^{-1}D_2\right)\int_0^\pi\beta_{13}^2\sin\xi\sin k\xi\,d\xi+A\left(A^2+k^2D_2\right)\frac{1}{\sqrt{2}}\int_0^\pi\beta_{12}^2\sin^2 k\xi\,d\xi\Bigg],
\end{aligned}$$

$$g_2=\frac{4\mu}{\pi}\left[\left(\frac{B}{A}\left(A^2+k^2D_2\right)-AB\right)\int_0^\pi\beta_{33}^1\sin^2 k\xi\,d\xi+A\left(A^2+k^2D_2\right)\int_0^\pi\beta_{33}^2\sin^2 k\xi\,d\xi\right].$$

Now we compute $T_2(\tilde{x})$. This bilinear form is again determined by a linear equation

$$DT_2(\tilde{x})PL\tilde{x}-PLT_2(\tilde{x})=\Pi PN_2(\tilde{x}), \qquad \tilde{x}\in E, \tag{3.20}$$

where Π denotes projection onto the nonresonant terms. Here the ansatz

$$T_2(\tilde{x})=\sum_{i,j=1}^{3}\gamma_{ij}\tilde{x}_i\tilde{x}_j, \qquad \gamma_{ij}=\gamma_{ji}, \tag{3.21}$$

leads to the following values for the components of $\gamma_{ij}\in\mathbf{C}^3$.

$$\begin{aligned}
&\gamma_{11}^1=\frac{1-i}{\gamma\pi}\sqrt{2}\,\alpha_{11}\int_0^\pi\sin^3\xi\,d\xi, && \gamma_{22}^2=\overline{\gamma_{11}^1},\\
&\gamma_{22}^1=\frac{i-1}{3\gamma\pi}\sqrt{2}\,\alpha_{22}\int_0^\pi\sin^3\xi\,d\xi, && \gamma_{11}^2=\overline{\gamma_{22}^1},\\
&\gamma_{12}^1=\frac{i-1}{\sqrt{2}\,\gamma\pi}\alpha_{12}\int_0^\pi\sin^3\xi\,d\xi, && \gamma_{12}^2=\overline{\gamma_{12}^1},\\
&\gamma_{33}^1=\frac{i-1}{\gamma\pi}\sqrt{2}\,\alpha_{33}\int_0^\pi\sin^2 k\xi\sin\xi\,d\xi, && \gamma_{33}^2=\overline{\gamma_{33}^1},\\
&\gamma_{23}^1=\frac{i-1}{\gamma\pi 2\sqrt{2}}\alpha_{23}\int_0^\pi\sin^2\xi\sin k\xi\,d\xi, && \gamma_{13}^2=\overline{\gamma_{23}^1},\\
&\gamma_{13}^1=0, && \gamma_{23}^2=0,\\
&\gamma_{11}^3=\frac{\mu i}{\pi\gamma}\alpha_{11}\int_0^\pi\sin^2\xi\sin k\xi\,d\xi, && \gamma_{22}^3=\overline{\gamma_{11}^3},\\
&\gamma_{13}^3=\frac{\mu i}{\pi\gamma}\alpha_{13}\int_0^\pi\sin\xi\sin^2 k\xi\,d\xi, && \gamma_{23}^3=\overline{\gamma_{13}^3},\\
&\gamma_{12}^3=\gamma_{33}^3=0.
\end{aligned} \tag{3.22}$$

This leads to the following contribution of $2PN_2(\tilde{x},T_2(\tilde{x}))$ in (1.1):

$$\begin{aligned}
d_3&=\operatorname{Re}\frac{(1+i)\sqrt{2}}{\pi}\Bigg[\left(4\alpha_{11}\gamma_{12}^1+2\alpha_{22}\gamma_{11}^2+\alpha_{12}\left(\gamma_{11}^1+2\gamma_{12}^2\right)\right)\int_0^\pi\sin^3\xi\,d\xi\\
&\qquad\qquad+\alpha_{23}\gamma_{11}^3\int_0^\pi\sin^2\xi\sin k\xi\,d\xi\Bigg],\\
e_3&=\operatorname{Re}\frac{(1+i)\sqrt{2}}{\pi}\Bigg[\left(2\alpha_{11}\gamma_{33}^1+\alpha_{12}\gamma_{33}^2\right)\int_0^\pi\sin^3\xi\,d\xi+4\alpha_{33}\gamma_{13}^3\int_0^\pi\sin^2 k\xi\sin\xi\,d\xi\\
&\qquad\qquad+2\alpha_{23}\gamma_{13}^2\int_0^\pi\sin^2\xi\sin k\xi\,d\xi\Bigg],\\
f_3&=\frac{2\mu}{\pi}\Bigg[4\left(\alpha_{11}\gamma_{23}^1+\alpha_{22}\gamma_{13}^2\right)\int_0^\pi\sin^2\xi\sin k\xi\,d\xi\\
&\qquad\qquad+2\left(\alpha_{13}\left(\gamma_{23}^3+\gamma_{12}^1\right)+\alpha_{23}\left(\gamma_{13}^3+\gamma_{12}^2\right)\right)\int_0^\pi\sin\xi\sin^2 k\xi\,d\xi\Bigg],\\
g_3&=\frac{2\mu}{\pi}\left(\alpha_{13}\gamma_{33}^1+\alpha_{23}\gamma_{33}^2\right)\int_0^\pi\sin\xi\sin^2 k\xi\,d\xi.
\end{aligned} \tag{3.23}$$

Thus, putting these computations together, we end up with the following formulas for the third order coefficients in the normal form (1.1):

$$d=\sum_{n=1}^{3} d_n,\quad e=\sum_{n=1}^{3} e_n,\quad f=\sum_{n=1}^{3} f_n,\quad g=\sum_{n=1}^{3} g_n \tag{3.24}$$

with d_n, e_n, f_n, g_n ($n=1,2,3$) given by (3.14), (3.19) and (3.23), respectively. Here we have used the facts that, according to (3.20), the argument of the bracket in (3.13) is just $[\mathrm{id}-\Pi]PN_2(\tilde{x})$ and

$$\Pi DT_2(\tilde{x})[\mathrm{id}-\Pi]PN_2(\tilde{x})=0. \tag{3.25}$$

In conclusion, we remark that there are solutions of (3.4) and (3.5), such that hypothesis (H1) of §1 is fulfilled for a,b and c in (3.11), e.g. for $k=5$, $D_1=0.02$ and $D_2=0.09$ (see Guckenheimer [1982]). Moreover, if one considers the coefficients in (3.9) and (3.24) as functions of D_1 and D_2 (A^2 and B as in (3.4)), then it is plausible that Ω in (H2) (see 1.14) does not identically vanish. (The precise verification of this is awkward because the constraints on the variables prohibit an asymptotic analysis. However, the numerical computation of Ω for any particular parameter values by, for example, appropriate truncation of the Fourier series, seems to be straightforward, but is not undertaken here.) Hence, apart from isolated exceptional values of the diffusion rates D_1 and D_2 in this set, both hypotheses (H1) and (H2) are fulfilled. Thus, for these values, one has, in particular, bifurcation to quasi-periodic orbits lying on invariant tori. If $\Omega>0$, then these quasi-periodic orbits form an asymptotically stable 2-torus.

Acknowledgments. We thank John David Crawford, Marty Golubitsky, John Guckenheimer and Gerard Iooss for useful discussions. We also thank Henk Broer for making his preprints available to us.

REFERENCES

V. I. Arnol'd [1965], *Small denominators* I. *Mappings of the circumference onto itself*, AMS Translations, Series 2, 46, Providence, RI, pp. 213–284.

N. N. Bogoliubov, J. A. Mitropolskii and A. M. Samoilenko [1976], *Method of Accelerated Convergence in Nonlinear Mechanics*, Springer-Verlag, New York.

B. L. J. Braaksma and H. W. Broer [1981], *Quasi-periodic flow near a codimension one singularity of a divergence free vector field in dimension four*, preprint.

H. Broer [1981a], *Formal normal form theorems for vector fields and some consequences for bifurcations in the volume preserving case*, Lecture Notes in Mathematics, 898, Springer-Verlag, New York, pp. 54–74.

______ [1981b], *Quasi-periodic flow near a codimension one singularity of a divergence free vector field in dimension three*, Lecture Notes in Mathematics, 898, Springer-Verlag, New York, pp. 75–89.

______ [1982], *Quasi-periodicity in local bifurcation theory*, preprint.

A. Chenciner [1982], *Courbes fermées invariantes non normalement hyperboliques au voisinage d'une bifurcation de Hopf degenérée pour difféomorphismes de* ($\mathbb{R}^2$, 0), C. R. Acad. Sci., 294, pp. 269–272.

S. N. Chow and J. K. Hale [1982], *Methods of Bifurcation Theory*, Springer-Verlag, New York.

S. N. Chow, J. Mallet-Paret and J. Yorke [1978], *Finding zeros of maps*: *homotopy methods that are constructive with probability one*, Math. Comp., 32, pp. 887–899.

J. D. Crawford [1983], *The Hopf bifurcation and plasma instabilities*, Ph. D. thesis, Univ. of California, Berkeley.

M. Golubitsky and W. F. Langford [1981], *Classification and unfoldings of degenerate Hopf bifurcations*, J. Differential Equations, 41, pp. 375–415.

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

J. GUCKENHEIMER [1980], *On quasi-periodic flow with three independent frequencies*, preprint.

______ [1981], *On a codimension two bifurcation*, Lecture Notes in Mathematics, 898, Springer-Verlag, New York, 99–142.

______ [1982], *Multiple bifurcation problems of codimension two*, preprint.

B. D. HASSARD, N. D. KAZARINOFF AND Y. H. WAN [1981], *Theory and applications of Hopf bifurcation*, London Math. Soc. Lect. Notes # 41, Camb. Univ. Press, London.

B. HASSARD AND Y. H. WAN [1978], *Bifurcation formulae derived from center manifold theory*, J. Math. Anal. Appl., 63, pp. 297–312.

M. R. HERMAN [1977], *Mesure de Lebesgue et nombre de rotation*, Lecture Notes in Mathematics, 597, Springer-Verlag, New York, pp. 271–293.

P. HOLMES [1980], *Unfolding a degenerate nonlinear oscillator: a codimension two bifurcation*, Ann. NY Acad. Sci., 357, pp. 473–488.

P. J. HOLMES AND J. MARSDEN [1982], *Horseshoes in perturbations of Hamiltonian systems with two degrees of freedom*, Comm. Math. Phys., 82, pp. 523–544.

L. HÖRMANDER [1977], *Implicit function theorems*, Stanford Lectures, Stanford Univ., Stanford, CA.

J. P. KEENER [1976], *Secondary bifurcation in nonlinear diffusion reaction equations*, Studies in Applied Mathematics 55, pp. 187–211.

G. IOOSS [1979], *Bifurcations of Maps and Applications*, North-Holland Mathematics Studies 36. North-Holland, Amsterdam.

G. IOOSS AND W. F. LANGFORD [1980], *Conjectures on routes to turbulence via bifurcations*, Ann. NY Acad. Sci., 357, pp. 489–505.

W. F. LANGFORD [1979], *Periodic and steady-state mode interactions lead to tori*, SIAM J. Appl. Math., 37, pp. 22–48.

______ [1981], *A review of interactions of Hopf and steady-state bifurcations* in Nonlinear Dynamics and Turbulence, G. Iooss and D. Joseph, eds., Pitman, London.

______ [1982], *Chaotic dynamics in the unfoldings of degenerate bifurcations*, Proc. International Symposium on Applied Mathematics and Information Sciencs, Kyoto Univ.

W. F. LANGFORD AND G. IOOSS [1980], *Interactions of Hopf and pitchfork bifurcations*, Bifurcation Problems and Their Numerical Solution, Birkhäuser, Berlin, 103–134.

J. E. MARSDEN AND M. MCCRACKEN [1976], *The Hopf Bifurcation and Its Applications*, Applied Mathematical Science, Vol. 19, Springer-Verlag, New York.

J. MOSER [1966], *A rapidly convergent iteration method and nonlinear partial differential equations*, I *and* II, Ann. Scuola Norm. Sup. Pisa, 20, pp. 265–315, 499–535.

D. RUELLE AND F. TAKENS [1971], *On the nature of turbulence*, Comm. Math. Phys., 20, pp. 167–192.

H. RÜSSMANN [1979], *Konvergente Reihenentwicklungen in der Störungs theorie der Himmelsmechanik* Selecta Mathematics V. Springer-Verlag, New York, pp. 92–260.

R. J. SACKER [1965], *A new approach to the perturbation theory of invariant surfaces*, Comm. Pure Appl. Math., 18, pp. 717–732.

D. G. SCHAEFFER AND M. A. GOLUBITSKY [1981], *Bifurcation analysis near a double eigenvalue of a model chemical reaction*, Arch. Rat. Mech. Anal., 75, pp. 315–347.

J. SCHEURLE [1982], *Bifurcation of quasi-periodic solutions from equilibrium points of reversible dynamical systems*, Arch. Rat. Mech. Anal., to appear.

C. L. SIEGEL AND J. K. MOSER [1971], Lectures on Celestial Mechanics, Springer-Verlag, New York.

E. ZEHNDER [1975], *Generalized implicit function theorems with applications to some small divisor problems* I, Comm. Pure Appl. Math., 28, pp. 91–140.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics
002

RESONANCE ZONES IN TWO-PARAMETER FAMILIES OF CIRCLE HOMEOMORPHISMS*

GLEN RICHARD HALL[†]

Abstract. We consider a two-parameter family of diffeomorphisms of the circle where one of the parameters controls the amount of rigid rotation while the second controls the nonlinearity. In particular, we show that the regions in the parameter plane for which the map has a periodic orbit of a particular rotation number (resonance zones) increase in size linearly as the second parameter is increased from zero. This is a discretization of the phenomenon known as "phase locking" for ordinary differential equations. Using this, we obtain some results on the smoothness of the curves between the resonance zones.

AMS-MOS subject classification (1980). Primary 58F22, 58F14

Key words. periodic orbits, resonance, phase locking

Introduction. Let $\mathbb{T}=\mathbb{R}/\mathbb{Z}$ be the circle with unit circumference and consider the two parameter family of maps from $\mathbb{T}$ onto $\mathbb{T}$ given by

$$(1)\qquad \theta\to\langle\theta+\phi+\alpha\gamma(\theta)\rangle$$

where $\langle\cdot\rangle$ denotes fractional part, ϕ and α are parameters and γ is a smooth function, periodic with period one. When $\alpha=0$, this map is merely "rigid rotation" by ϕ: hence it will have periodic orbits if and only if ϕ is rational. When $\alpha>0$, the set of parameter values where a periodic orbit exists of a particular period and rotation number opens into a region in the (ϕ,α) plane (see Brunovsky [3]). In this paper we study the rate at which these resonance zones open near $\alpha=0$. This can be considered a discretization of the phenomenon of 'phase locking' in O.D.E.'s which has been extensively studied (see, for example, Loud [8], Bushard [4], [5]).

When α is small, the resonance zones corresponding to periodic orbits of different period or rotation number will remain disjoint. Between these zones there will be arcs in the parameter plane where the above map has no periodic orbits and all orbits are dense. Herman [6] has shown that when $\gamma\in C^3$ for $\alpha_0>0$, α_0 small the set of ϕ such that at parameter (ϕ,α_0) the map (1) has no periodic orbits will have positive measure. Herman [6] also showed that this set tends to one with full measure as α tends to zero. Arnol'd [1] showed that certain of these arcs of nonresonance will be smooth depending on a number theoretic condition on the rotation number. In the final section of this paper we point out that although all the nonresonance arcs will have a first derivative at $\alpha=0$, for certain rotation numbers (depending on γ) they will not have second derivatives at $\alpha=0$. The precise nature of the nonresonance arcs for arbitrary irrational rotation number remains open.

1. Definitions and notation.

We let $\gamma:\mathbb{R}\to\mathbb{R}$ be a function satisfying

(1) $\gamma\in C^1$ and $|d\gamma/d\theta|\leq 1$,
(2) $\forall\theta\in\mathbb{R}$, $\gamma(\theta+1)=\gamma(\theta)$,
(3) $\int_0^1\gamma(\theta)\,d\theta=0$,

*Received by the editors May 18, 1983, and in revised form November 16, 1983. This research was sponsored by the U.S. Army Research Office under contract DAAG 29-80-C-0041, and by a National Science Foundation postdoctoral fellowship.

[†] Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706.

and we let $\Sigma_{n=-\infty}^{\infty} a_n e^{in2\pi\theta}=\gamma(\theta)$ be the Fourier series of γ. Given such a γ, we may define a two-parameter family of maps $f\colon \mathbb{R}^3\to\mathbb{R}$ by

$$f\colon(\theta,\phi,\alpha)\to\theta+\phi+\alpha\gamma(\theta).$$

The parameter ϕ controls the rotation or "twist" while α controls the nonlinearity of f. Note that for fixed (ϕ,α) we have

$$\forall\theta\in\mathbb{R},\quad f(\theta+1,\phi,\alpha)=f(\theta,\phi,\alpha)+1;$$

hence f is the left of a two-parameter family of degree one maps of the circle. By condition (1) on γ, f is a homeomorphism when $\alpha\in[0,1)$.

Notation. For fixed (ϕ,α) we let $f^n(\theta,\phi,\alpha)$ denote the nth iterate of $f(\cdot,\phi,\alpha)$, i.e. $f^n(\theta,\phi,\alpha)=f(f^{n-1}(\theta,\phi,\alpha),\phi,\alpha)$.

DEFINITION. For $\phi\in\mathbb{R}$, $\alpha\in[0,1)$ the rotation number of $f(\cdot,\phi,\alpha)$ is defined to be

$$\rho(\phi,\alpha)=\lim_{n\to\infty}\frac{f^n(\theta,\phi,\alpha)-\theta}{n}.$$

We will use the following facts about the rotation number due to Poincare.

THEOREM A. *For $\phi\in\mathbb{R}$, $\alpha\in[0,1)$*

1) *$\rho(\phi,\alpha)$ exists and is independent of the θ in the definition,*

2) *$\rho(\phi,\alpha)$ is continuous in (ϕ,α) and increasing in ϕ,*

3) *if $\rho(\phi,\alpha)=p/q\in Q$, then there exists $\theta\in[0,1)$ such that $f^q(\theta,\phi,\alpha)=\theta+p$.*

Proof. See Herman [7]. □

Remark. Since $f(\cdot,\phi,\alpha)$ is the lift of a homeomorphism of a circle when $\alpha\in[0,1)$, we may reinterpret part (3) of Theorem A as saying that this circle map has a periodic orbit with period q and rotation number p/q.

DEFINITION. For $\beta\in\mathbb{R}$ we let

$$A_\beta=\{(\phi,\alpha)\colon\phi\in\mathbb{R},\alpha\in[0,1),\rho(\phi,\alpha)=\beta\}.$$

Remark. When $\beta=p/q$ is rational, then the set $A_{p/q}$ is called the p/q resonance horn or the p/q Arnol'd tongue.

Notation. We say $\gamma\in C^{r+\varepsilon}$ for $r\geqq 1$ an integer and $\varepsilon\in(0,1]$ if $\gamma\in C^r$ and there exists a constant $c>0$ such that

$$\forall\theta_1,\theta_2\in\mathbb{R},\quad\left|\frac{d^r\gamma}{d\theta^r}(\theta_1)-\frac{d^r\gamma}{d\theta^r}(\theta_2)\right|<C|\theta_1-\theta_2|^\varepsilon.$$

If $\gamma\in C^{r+\varepsilon}$ and is given by the Fourier series $\gamma(\theta)=\Sigma_{n=-\infty}^{\infty}a_n e^{in2\pi\theta}$, then $|a_n|=O(|n|^{-r-\varepsilon})$ (see [1]).

2. Resonance horns. For γ as in §1 and rational p/q in lowest terms, we wish to consider the set $A_{p/q}$ for α near zero.

THEOREM 1 (Herman [7], Boyland [2]). *For each rational p/q there exist Lipschitz functions $\phi_1,\phi_2\colon[0,1)\to\mathbb{R}$*

1) $\forall\alpha\in[0,1)$, $\phi_1(\alpha)\leqq\phi_2(\alpha)$,

2) $\phi_1(0)=p/q=\phi_2(0)$,

3) $(\phi,\alpha)\in A_{p/q}$ *if and only if* $\phi_1(\alpha)\leqq\phi\leqq\phi_2(\alpha)$.

Moreover, the Lipschitz constant of ϕ_1, ϕ_2 is independent of p/q (i.e., it depends only on γ).

We will include a proof of Theorem 1 since it allows us to set up notation for

THEOREM 2. *With $\gamma(\theta)=\Sigma_{n=-\infty}^{\infty}a_n e^{in2\pi\theta}$, for each rational p/q, the functions ϕ_1, ϕ_2 are differentiable at $\alpha=0$, $(d\phi_1/d\alpha)(0)\leqq 0\leqq(d\phi_2/d\alpha)(0)$ and there exists an $\varepsilon>0$ such* *that if* $\alpha \in [0, \varepsilon)$, *then*

$$(*) \qquad \phi_2(\alpha)-\phi_1(\alpha) \geqq \alpha\left(\sum_{n=-\infty}^{\infty} |a_{nq}|^2\right)^{1/2} \Big/ 2.$$

Remark. Generically, all the Fourier coefficients of γ are nonzero, so generically all the horns "open" about the vertical ray at a positive rate. (See Fig. 1.) This is a discretization of Bushard [4, Thm. 1].

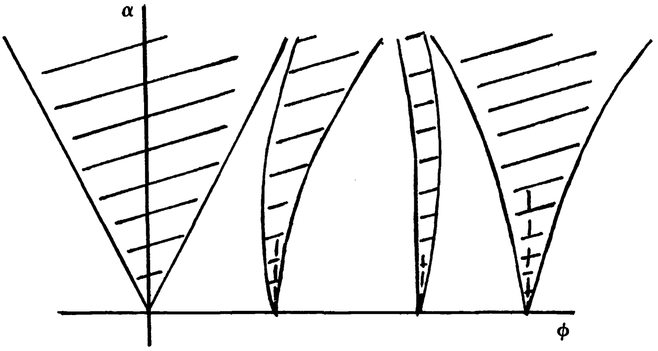


FIG. 1

Proof of Theorem 1. Since, by Theorem A, the function $\rho(\phi, \alpha)$ is continuous and increasing in ϕ, we may define functions $\phi_1, \phi_2 : [0,1) \to \mathbb{R}$ satisfying (1,2,3) of Theorem 1. It remains to show that ϕ_1, ϕ_2 are Lipschitz.

For $\zeta \in [0,1)$ we note that

$$f^q(\zeta, \phi_1(\alpha), \alpha) \leqq \zeta + p, \qquad f^q(\zeta, \phi_2(\alpha), \alpha) \geqq \zeta + p.$$

Since $(\partial f^q/\partial\phi)(\theta, \phi, \alpha) \geqq 1$ for all (θ, ϕ, α) with $\alpha \in [0,1)$, we may use the implicit function theorem to obtain a curve $\phi_\zeta : [0,1) \to \mathbb{R}$ such that $f^q(\zeta, \phi, \alpha) = \zeta + p$ if and only if $\phi = \phi_\zeta(\alpha)$. Moreover $\phi_\zeta(0) = p/q$ and

$$\frac{d\phi_\zeta}{d\alpha}(\alpha) = -\frac{\partial f^q}{\partial \alpha}(\zeta, \phi_\zeta(\alpha), \alpha) \Big/ \frac{\partial f^q}{\partial \phi}(\zeta, \phi_\zeta(\alpha), \alpha).$$

Using the chain rule we obtain

$$\left|\frac{d\phi_\zeta}{d\alpha}(\alpha)\right| < \sup_{\theta \in [0,1)} |\gamma(\theta)|.$$

By Theorem A part (3) we see that

$$\phi_1(\alpha) = \inf_{\zeta \in [0,1)} \phi_\zeta(\alpha) \quad \text{and} \quad \phi_2(\alpha) = \sup_{\zeta \in [0,1)} \phi_\zeta(\alpha).$$

Hence ϕ_1, ϕ_2 are Lipschitz with constant independent of p/q. □

Proof of Theorem 2. Noting that for any q,

$$f^q(\theta, \phi, \alpha) = \theta + q\phi + \alpha \sum_{j=0}^{q-1} \gamma(\theta + j\phi) + \alpha h(\theta, \phi, \alpha)$$

where h is C^1 and $h \to 0$ as $\alpha \to 0$, we see that, using the notation above,

$$\frac{d\phi_\zeta}{d\alpha}(0) = -\sum_{j=0}^{q-1} \gamma(\zeta + jp/q)/q.$$

Since γ is given in Fourier series by $\gamma(\theta) = \sum_{n=-\infty}^{\infty} a_n e^{in2\pi\theta}$, we see that

$$\frac{d\phi_\zeta}{d\alpha}(0) = -\sum_{n=-\infty}^{\infty} a_{nq} e^{inq2\pi\zeta}.$$

The L^2-norm of $(d\phi_\zeta/d\alpha)(0)$ as a function of ζ is then equal to $(\sum_{n=-\infty}^{\infty} |a_{nq}|^2)^{1/2}$. But $\int_0^1 (d\phi_\zeta/d\alpha)(0)\,d\zeta = a_0 = 0$ so $(d\phi_\zeta/d\alpha)(0)$ must attain in absolute value the value of its L^2-norm and assume both positive and negative values. For each $\zeta \in [0,1)$, $\alpha \in [0,1)$, we have

$$\phi_1(\alpha) \leqq \phi_\zeta(\alpha) \leqq \phi_2(\alpha)$$

and this shows the inequality (*).

Let $\beta = \inf_{\zeta \in [0,1)} (d\phi_\zeta/d\alpha)(0)$. Then $\beta = (d\phi_{\zeta_0}/d\alpha)(0)$ for some $\zeta_0 \in [0,1)$ and

$$\limsup_{\alpha \to 0} \frac{\phi_1(\alpha) - p/q}{\alpha} \leqq \beta.$$

If there exists $\delta > 0$ such that

$$\liminf_{\alpha \to 0} \frac{\phi_1(\alpha) - p/q}{\alpha} < \beta - \delta,$$

then there exists $\alpha_n \to 0$ such that

$$\frac{\phi_1(\alpha_n) - p/q}{\alpha_n} < \beta - \delta$$

and hence there exist $\zeta_n \in [0,1)$ with

$$\frac{\phi_{\zeta_n}(\alpha_n) - p/q}{\alpha_n} < \beta - \delta$$

since $\phi_1(\alpha_n) = \inf_{\zeta \in [0,1)} \phi_\zeta(\alpha_n)$. By taking a convergent subsequence of the ζ_n's and noting that $(d\phi_\zeta/d\alpha)(\alpha)$ converges to $(d\phi_\zeta/d\alpha)(0)$ uniformly in ζ as $\alpha \to 0$, we see that this contradicts the choice of β. Hence $d\phi_1/d\alpha$ exists at $\alpha = 0$ and equals β. That $d\phi_2/d\alpha$ exists at $\alpha = 0$ and equals $\sup_{\zeta \in [0,1)} (d\phi_\zeta/d\alpha)(0)$ follows in precisely the same manner, which completes the proof of the theorem. □

Remarks. 1) The curves ϕ_1, ϕ_2 are also characterized by the existence of a "node" orbit with rotation number p/q. If we let $F: \mathbb{R}^3 \to \mathbb{R}^2$ be defined by

$$F(\theta,\phi,\alpha) = \left(f^q(\theta,\phi,\alpha) - \theta - p,\ \frac{\partial f^q}{\partial \theta}(\theta,\phi,\alpha) - 1 \right)$$

and if γ is C^2, then the curves ϕ_1, ϕ_2 may be obtained by applying the implicit function theorem to the equation $F = 0$. Hence ϕ_1 and ϕ_2 are, generically, piecewise as smooth as γ on $\alpha \in (0,1)$.

2) The rate at which the Fourier coefficients of γ decrease is controlled by the smoothness of γ. In particular, if γ is $C^{r+\delta}$, then $|a_n| \leqq c/n^{r+\delta}$ where c is a constant

independent of n. From this we easily obtain that

$$\left(\sum_{n=-\infty}^{\infty}|a_{nq}|^2\right)^{1/2} \leqq c/q^{r+\delta}$$

which may be used in $(*)$ when γ is $C^{r+\delta}$.

3) Let $B_\alpha=\{\phi:\rho(\phi,\alpha)\notin Q\}$ and let λ be Lebesgue measure on $\mathbb{R}$. As noted in the introduction, Herman [6] has shown that for any bounded interval $I\subseteq\mathbb{R}$ and $0\leqq\alpha<1$, $\lambda(B_\alpha\cap I)>0$ whenever $B_\alpha\cap I\neq\varnothing$ and $\gamma\in C^3$ and $\lambda(B_\alpha\cap I)\to\lambda(I)$ as $\alpha\to 0$. An immediate consequence of Theorem 2 is

COROLLARY. *Let $\gamma(\theta)=\sum_{n=-\infty}^{\infty}a_n e^{in\theta}$ be as above with $a_n\neq 0$ for infinitely many n. Then for any bounded interval $I\subseteq\mathbb{R}$ there exist constants $\varepsilon_I, c_I>0$ such that*

$$\lambda(B_\alpha\cap I)<\lambda(I)(1-c_I\alpha)$$

whenever $\alpha\in[0,\varepsilon_I]$.

3. Nonresonance. Fix γ as in §1. For a given irrational η there exists a Lipschitz curve $\psi_\eta:[0,1)\to\mathbb{R}$ such that $\rho(\phi,\alpha)=\eta$ if and only if $\phi=\psi_\eta(\alpha)$ (see Herman [7]).

THEOREM 3. *For any irrational η, the derivative of $\psi_\eta(\alpha)$ exists when $\alpha=0$ and equals* 0.

Proof. Recall that since $\rho(\psi_\eta(\alpha),\alpha)=\eta$, there is a unique $f(\cdot,\psi_\eta(\alpha),\alpha)$ invariant probability measure μ_α and

$$\begin{aligned}\eta&=\rho(\psi_\eta(\alpha),\alpha)\\&=\lim_{n\to\infty}\frac{f^n(\theta,\psi_\eta(\alpha),\alpha)-\theta}{n}\\&=\lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}\left(f^{j+1}(\theta,\psi_\eta(\alpha),\alpha)-f^j(\theta,\psi_\eta(\alpha),\alpha)\right)\\&=\lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}\left(f(\cdot,\psi_\eta(\alpha),\alpha)-\text{identity}\right)\left(f^j(\theta,\psi_\eta(\alpha),\alpha)\right)\\&=\int_0^1\left(f(\theta,\psi_\eta(\alpha),\alpha)-\theta\right)d\mu_\alpha(\theta)\\&=\psi_\eta(\alpha)+\alpha\int_0^1\gamma(\theta)\,d\mu_\alpha(\theta)\end{aligned}$$

(see Herman [7] for details). But noting that $\psi_\eta(0)=\eta$ we see that

$$\frac{1}{\alpha}\left(\psi_\eta(\alpha)-\psi_\eta(0)\right)=\int_0^1\gamma(\theta)\,d\mu_\alpha(\theta).$$

It follows from the fact that the measure $d\mu_\alpha$ is determined, up to an error small with n, by the first n iterates of $f(\cdot,\psi_\eta(\alpha),\alpha)$ that $\int_0^1\gamma(\theta)\,d\mu_\alpha\to\int_0^1\gamma(\theta)\,d\theta=0$ as $\alpha\to 0$ (see Herman [7, p. 73]). So the derivative of $\psi_\eta(\alpha)$ exists at $\alpha=0$ and equals zero. □

If, for example, γ is analytic and η is sufficiently poorly approximable by rationals, then the curve ψ_η will also be analytic (see Arnol'd [1]). However, this will not hold for all irrationals.

THEOREM 4. *For $\gamma(\theta)=\sum_{n=-\infty}^{\infty} a_n e^{in2\pi\theta}$ with infinitely many $a_n \neq 0$ and $g:(0,1)\to\mathbb{R}$ a strictly positive, continuous function with $g(\alpha)\to 0$ as $\alpha\to 0$, there exists a residual set of irrationals η such that*

$$(**)\qquad \limsup_{\alpha\to 0}\left|\frac{\psi_\eta(\alpha)-\eta}{\alpha g(\alpha)}\right|=\infty.$$

Proof of Theorem 4. Fix a rational p/q such that $a_{nq}\neq 0$ for some n. Then for any $\delta>0$ there exists an open interval $J(\delta,p/q)\subseteq\mathbb{R}$ with either right or left end point at p/q such that for each $\eta\in J(\delta,p/q)\sim Q$ there exists $\alpha\in[0,\delta]$ such that

$$|\psi_\eta(\alpha)-\eta|\geqq\frac{|a_{nq}|\cdot\alpha}{4}$$

(see Fig. 2). Hence for each $M>0$ there exists $\delta_M>0$ depending on p/q such that for each $\eta\in J(\delta_M,p/q)\sim Q$ there exists $\alpha\in[0,\delta_M)$ with

$$\left|\frac{\psi_\eta(\alpha)-\eta}{\alpha g(\alpha)}\right|\geqq M,$$

and $\delta_M\to 0$ as $M\to\infty$. Since $a_n\neq 0$ for infinitely many n, it follows that

$$K_M\equiv\bigcup_{r/s\in Q}\{J(\delta_M,r/s):a_{ns}\neq 0 \text{ for some } n\}$$

is open and dense. But then $(**)$ holds for every $\eta\in(\bigcap_{M\geqq 1}K_M)\sim Q$ which is a residual set and the proof is complete. □

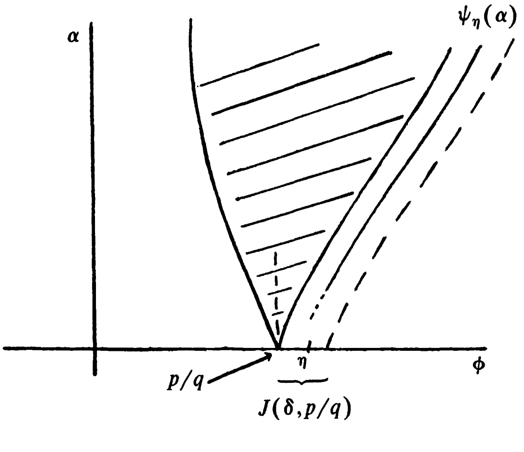


FIG. 2

Remark. Since the first derivative of ψ_η at zero is zero, the above implies in particular that ψ_η has no second derivative at zero, i.e. $\lim_{\alpha\to 0}((\psi_\eta(\alpha)-\eta)/\alpha^2)$ does not exist. We conjecture that ψ_η is in fact C^1 on $[0,1)$ for all irrationals η. It is possible that ψ_η is even smoother on $(0,1)$ for arbitrary irrationals η.

Acknowledgments. The author would like to thank Richard McGehee for suggesting this problem and for his encouragement. Also, thanks are due the referee for several important suggestions and improvements.

REFERENCES

[1] V. I. ARNOL'D, *Small denominators* I. *Mappings of the circumference onto itself*, AMS Translations, Series 2, 46, Providence, RI, 1965, pp. 213–284.

[2] P. L. BOYLAND, *Bifurcation of circle maps: Arnol'd tongues, bistability and rotation intervals*, Thesis, Univ. of Iowa, Ames, 1983.

[3] P. BRUNOVSKY, *Generic properties of the rotation number of one-parameter diffeomorphisms of the circle*, Czech. Math. J., 24 (99) (1974), pp. 74–90.

[4] L. B. BUSHARD, *Periodic solutions and locking in on the periodic surface*, J. Nonlinear Mechanics, 8 (1973), pp. 129–141.

[5] ———, *Behavior of the periodic surface for a periodically perturbed autonomous system and periodic solutions*, J. Differential Equations, 12 (1972), pp. 487–503.

[6] M. R. HERMAN, *Mesure de Lebesgue et nombre de rotation*, Lecture Notes in Mathematics, 597, Springer, Berlin, 1977, pp. 271–293.

[7] ———, *Sur la conjugaison differentiable des diffeomorphismes du cercle à des rotations*, Publ. Math. I.H.E.S., 49 (1979), pp. 5–234.

[8] W. S. LOUD, *Phase shift and locking-in regions*, Quart. J. Appl. Math., 25 (1967), pp. 222–227.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics

003

OSCILLATION THEOREMS FOR NONLINEAR SECOND ORDER DIFFERENTIAL EQUATIONS WITH A NONLINEAR DAMPING TERM*

S. R. GRACE[†], B. S. LALLI[†] AND C. C. YEH[‡]

Abstract. Sufficient conditions for the oscillation of the nonlinear second order differential equation

$$(a(t)\psi(x(t))\dot{x}(t))^{\cdot}+p(t)k(t,x(t),\dot{x}(t))\dot{x}(t)+q(t)f(x(t))=0,$$

are established. A systematic study is attempted which extends and correlates a number of existing results.

1. Introduction. Our main objective in this paper is the study of the oscillatory behavior of the differential equation

$$(1)\quad (a(t)\psi(x(t))\dot{x}(t))^{\cdot}+p(t)k(t,x(t),\dot{x}(t))\dot{x}(t)+q(t)f(x(t))=0 \qquad \left(\cdot=\frac{d}{dt}\right),$$

where the functions a,p,q: $[t_0,\infty)\rightarrow R=(-\infty,\infty)$, k: $[t_0,\infty)\times R^2\rightarrow[0,\infty)$, ψ,f: $R\rightarrow R$ are continuous, $a(t)>0$, $\psi(x)>0$ for all x and $xf(x)>0$ for $x\neq 0$.

The functions appearing in equation (1) will be assumed to be sufficiently smooth for a local existence and uniqüeness theorem to hold for equation (1) on $0\leq t_0\leq t<\infty$.

In what follows, we consider only solutions of equation (1) which are defined for all large t. A solution of equation (1) is called oscillatory if it has no last zero, otherwise it is called nonoscillatory.

A well-known sufficient condition for oscillation of the linear equation

$$(2)\qquad \ddot{x}(t)+q(t)x(t)=0,$$

where q: $[t_0,\infty)\rightarrow R$ is continuous, is that

$$(3)\qquad \int^{\infty}q(s)\,ds=\infty.$$

This result has been extended in [1] to the nonlinear equation

$$(4)\qquad \ddot{x}(t)+q(t)f(x(t))=0,$$

where f is nondecreasing, continuously differentiable and $xf(x)>0$ for $x\neq 0$. Coles [3] and Kamenev [7] obtained oscillation criteria for (4) and/or related equations by using weighted integral conditions which include (3) as a special case. Other oscillation criteria which involve the behavior of the integral of q are established by Wintner [24] who showed that the condition

$$\lim_{t\rightarrow\infty}\frac{1}{t}\int_{t_0}^{t}du\int_{t_0}^{u}q(s)\,ds=\infty$$

is sufficient for (2) to be oscillatory. In [6], Kamenev improved Wintner's result by using the nth primitive

$$A_n(t)=\frac{1}{(n-1)!}\int_{t_0}^{t}(t-u)^{n-1}q(u)\,du$$

* Received by the editors August 25, 1982, and in revised form April 15, 1983.

[†] Department of Mathematics, University of Saskatchewan, Saskatoon, Canada S7N OWO.

[‡] Department of Mathematics, Central University, Chung-Li, Taiwan, Republic of China.

of the coefficient $q(t)$ for some integer $n \geq 3$. Yeh [25], [26], extended Kamenev's result in [6] to a larger class of equations which include equations (2) and (4). Extensions to the above mentioned criteria as well as other criteria are the subject of many studies. For general interest, we refer the reader to the papers [1]–[26].

The main results of this paper are presented in the form of seven theorems. In Theorems 1–3, we discuss the oscillatory behavior of (1) when this equation is either superlinear, i.e., $\int_{\pm\varepsilon}^{\pm\infty} \psi(u)/f(u)\,du < \infty$, or sublinear, i.e., $\int_{\pm 0}^{\pm\varepsilon} \psi(u)/f(u)\,du < \infty$, for every $\varepsilon > 0$. Theorems 4 and 5 concern the oscillation of equation (1) where weighted averaging procedure is used. These theorems are given in a form which is useful in investigating the oscillatory and nonoscillatory solutions of equations of the form $\ddot{x}(t) + k^2 t^{-2} x(t) = 0$, according to different values of k. Theorems 6 and 7 ensure the oscillation of equation (1) when $p(t) = 0$ and $q(t)$ is of varying sign. Examples are inserted in the text to illustrate the relevance of the theorems.

The results obtained here are presented in a form which is essentially new. Our results of this paper extend and unify some of the results in [1], [3], [6]–[10], [13]–[16], [18], [22]–[26].

To obtain our results we need the following lemma.

LEMMA. *Let $p(t) \geq 0$ and $q(t)$ be nonnegative and not identically zero on any ray of the form $[t^*, \infty)$, $t^* \geq t_0$, and assume that*

$$k(t,x,y) \leq |y|^{\alpha}, \quad -\infty < x,y < \infty, \quad t \geq t_0 \text{ and some constant } \alpha \geq 0, \tag{5}$$

$$\psi(x) \geq c > 0 \quad \text{for all } x, \tag{6}$$

and

$$\begin{aligned} &\frac{1}{a(t)}\left(1 + \int_{t_0}^{t} \frac{p(s)}{a^{\alpha+1}(s)}\,ds\right)^{-1/\alpha} \notin \mathfrak{L}(t_0, \infty) && \text{if } \alpha > 0, \\ &\int_{t_0}^{\infty} \frac{1}{a(s)} \exp\left(\int_{t_0}^{s} \frac{-p(\tau)}{ca(\tau)}\,d\tau\right) ds = \infty && \text{if } \alpha = 0. \end{aligned} \tag{7}$$

Then if $x(t)$ is a nonoscillatory solution of (1), *we must have*

$$x(t)\dot{x}(t) > 0 \quad \textit{for all large } t.$$

Proof. Let $x(t)$ be a nonoscillatory solution of (1) and assume $x(t) > 0$ for $t \geq t_0 \geq 0$. If $\dot{x}(t_1) = 0$ and $q(t_1) > 0$ for some $t_1 \geq t_0$, then

$$\left(a(t)\psi(x(t))\dot{x}(t)\right)^{\cdot}\big|_{t=t_1} = -q(t_1)f(x(t_1)) < 0,$$

from which we can prove that $\dot{x}(t)$ cannot have another zero after it vanishes once. Thus $\dot{x}(t)$ has a fixed sign for all sufficiently large t. Let $\dot{x}(t) < 0$ for $t \geq t_2 \geq t_1$. Then

$$\dot{u}(t) + \gamma \frac{p(t)}{a^{\alpha+1}(t)} u^{\alpha+1}(t) \geq 0 \quad \text{for } t \geq t_2, \tag{8}$$

where $u(t) = -a(t)\psi(x(t))\dot{x}(t)$ and $\gamma = c^{-(\alpha+1)}$.

Integrating (8) from t_2 to t, we obtain

$$\psi(x(t))\dot{x}(t) \leq -\frac{1}{a(t)}\left[u^{-\alpha}(t_2) + \alpha\gamma \int_{t_2}^{t} \frac{p(s)}{a^{\alpha+1}(s)}\,ds\right]^{-1/\alpha} \quad \text{if } \alpha > 0, \tag{9}$$

and

$$(10)\qquad \psi(x(t))\dot{x}(t)\leq -\frac{1}{a(t)}u(t_2)\exp\left(-\int_{t_2}^{t}\frac{p(\tau)}{ca(\tau)}d\tau\right)\quad \text{if } \alpha=0.$$

Now integrate (9) and (10) and use (7), and obtain a contradiction. This proves our lemma.

Note. If $p(t)=0$, then condition (5) can be disregarded and condition (7) takes the form

$$\int^{\infty}\frac{1}{a(s)}ds=\infty.$$

2. Main results.

THEOREM 1. *Let conditions* (5)–(7) *hold,*

$$(11)\qquad p(t)\geq 0,\ q(t)\geq 0 \textit{ for } t\geq t_0 \textit{ and } q(t) \textit{ is not eventually zero on } [t_0,\infty);$$

$$(12)\qquad f'(x)\geq 0 \textit{ for all } x,\ ('=\tfrac{d}{dx}).$$

Suppose that there exists a differentiable function $\rho\colon [t_0,\infty)\to(0,\infty)$ *such that*

$$(13)\qquad \int^{\infty}\rho(s)q(s)\,ds=\infty.$$

Then each of the following conditions ensures the oscillation of each of the continuable solutions of (1):

(I) $\dot{\rho}(t)\leq 0$ *for* $t\geq t_0$;

(II) $\dot{\rho}(t)\geq 0$, $(a(t)\dot{\rho}(t))^{\cdot}\leq 0$ *for* $t\geq t_0$, *and*

$$(14)\qquad \int_{\varepsilon}^{\infty}\frac{\psi(u)}{f(u)}du<\infty \quad \textit{and} \quad \int_{-\varepsilon}^{-\infty}\frac{\psi(u)}{f(u)}du<\infty \quad \textit{for every } \varepsilon>0;$$

(III) $\int^{\infty}|(a(s)\dot{\rho}(s))^{\cdot}|\,ds<\infty$ *and* (14) *holds.*

Proof. Let $x(t)$ be a nonoscillatory solution of (1), say $x(t)>0$ for $t\geq t_1\geq t_0$. By the lemma, there exists $t_2\geq t$ such that $\dot{x}(t)>0$ for $t\geq t_2$. Let

$$w(t)=\frac{a(t)\psi(x(t))\dot{x}(t)}{f(x(t))}\rho(t).$$

By differentiation, we obtain that for every $t\geq t_2$

$$\begin{aligned}(15)\qquad \dot{w}(t)&=\frac{(a(t)\psi(x(t))\dot{x}(t))^{\cdot}}{f(x(t))}\rho(t)+\frac{a(t)\psi(x(t))\dot{x}(t)}{f(x(t))}\dot{\rho}(t)\\&\quad-\frac{a(t)\psi(x(t))(\dot{x})^2}{f^2(x(t))}f'(x(t))\rho(t)\\&=-\rho(t)q(t)-\rho(t)p(t)k(t,x(t),\dot{x}(t))\frac{\dot{x}(t)}{f(x(t))}\\&\quad+\frac{a(t)\psi(x(t))\dot{x}(t)}{f(x(t))}\dot{\rho}(t)-\frac{a(t)\psi(x(t))\dot{x}^2}{f^2(x(t))}f'(x(t))\rho(t).\end{aligned}$$

Thus

$$\dot{w}(t)\leq -\rho(t)q(t)+a(t)\dot{\rho}(t)\frac{\psi(x(t))\dot{x}(t)}{f(x(t))}. \tag{16}$$

Integrating (16) from t_2 to t, we obtain

$$w(t)\leq w(t_2)-\int_{t_2}^{t}\rho(s)q(s)\,ds+\int_{t_2}^{t}a(s)\dot{\rho}(s)\frac{\psi(x(s))\dot{x}(s)}{f(x(s))}\,ds. \tag{17}$$

We consider the following cases.

Case 1. Let (I) hold. Then (17) becomes

$$w(t)\leq w(t_2)-\int_{t_2}^{t}\rho(s)q(s)\,ds\to -\infty \quad \text{as } t\to\infty,$$

which contradicts the fact that $w(t)>0$ for $t\geq t_2$.

Case 2. Let (II) hold. Then (17) becomes

$$\begin{aligned} w(t)\leq w(t_2)&-\int_{t_2}^{t}\rho(s)q(s)\,ds-a(t)\dot{\rho}(t)\int_{x(t)}^{\infty}\frac{\psi(u)}{f(u)}\,du \\ &+a(t_2)\dot{\rho}(t_2)\int_{x(t_2)}^{\infty}\frac{\psi(u)}{f(u)}\,du+\int_{t_2}^{t}(a(s)\dot{\rho}(s))^{\cdot}\left(\int_{x(s)}^{\infty}\frac{\psi(u)}{f(u)}\,du\right)ds, \end{aligned} \tag{18}$$

which implies that

$$w(t)\leq w(t_2)+a(t_2)\dot{\rho}(t_2)\int_{x(t_2)}^{\infty}\frac{\psi(u)}{f(u)}\,du-\int_{t_2}^{t}\rho(s)q(s)\,ds\to -\infty \quad \text{as } t\to\infty,$$

contradicting the fact that $w(t)>0$ for $t\geq t_2$.

Case 3. Let (III) hold. From (14), it follows that

$$0\leq\int_{x(t_2)}^{\infty}\frac{\psi(u)}{f(u)}\,du\leq M \quad \text{for some constant } M.$$

Note that $\int^{\infty}|(a(s)\dot{\rho}(s))^{\cdot}|\,ds<\infty$ implies $|a(t)\dot{\rho}(t)|\leq M_1$ for all t, where M_1 is a positive constant. Thus (18) becomes

$$\begin{aligned} w(t)\leq w(t_2)+a(t_2)\dot{\rho}(t_2)&\int_{x(t_2)}\frac{\psi(u)}{f(u)}\,du-\int_{t_2}^{t}\rho(s)q(s)\,ds \\ &+M\int_{t_2}^{\infty}|(a(s)\dot{\rho}(s))^{\cdot}|\,ds\to -\infty \quad \text{as } t\to\infty, \end{aligned}$$

which is again a contradiction.

The following theorem concerns the case when (13) fails.

THEOREM 2. *Let conditions* (5)–(7), (11), (12) *and* (14) *hold. Suppose that there exists a differentiable function* ρ: $[t_0,\infty)\to(0,\infty)$ *such that* $\dot{\rho}(t)\leq 0$ *for* $t\geq t_0$ *and*

$$\int^{\infty}\frac{1}{a(s)\rho(s)}\int_{s}^{\infty}\rho(\tau)q(\tau)\,d\tau\,ds=\infty. \tag{19}$$

Then (1) *is oscillatory.*

Proof. Let $x(t)$ be a nonoscillatory solution of (1). Assume that $x(t)>0$ for $t\geq t_1\geq t_0$. Following the same reasoning as in the proof of Theorem 1, we get (17). Since $\dot{x}(t)\geq 0$ for $t\geq t_2$, we have

$$0\leq\frac{a(t_2)\psi(x(t_2))\dot{x}(t_2)}{f(x(t_2))}\rho(t_2)-\int_{t_2}^{\infty}\rho(s)q(s)\,ds.$$

Hence, for all $t\geq t_2$,

$$\int_t^{\infty}\rho(s)q(s)\,ds\leq a(t)\rho(t)\frac{\psi(x(t))\dot{x}(t)}{f(x(t))},$$

and integrating we obtain

$$\int_{t_2}^{t}\frac{1}{a(s)\rho(s)}\int_s^{\infty}\rho(\tau)q(\tau)\,d\tau\,ds\leq\int_{t_2}^{t}\frac{\psi(x(s))\dot{x}(s)}{f(x(s))}\,ds.$$

This contradicts (14), since the integral on the left diverges.

THEOREM 3. *In Theorem* 1 (III), *let the condition* (14) *be replaced by*

$$\int_0^{+\varepsilon}\frac{\psi(u)}{f(u)}du<\infty \quad \textit{and} \quad \int_0^{-\varepsilon}\frac{\psi(u)}{f(u)}du<\infty \quad \textit{for every } \varepsilon>0. \tag{20}$$

Then every bounded solution of (1) *is oscillatory.*

Proof. Let $x(t)$ be a bounded nonoscillatory solution of (1), say $x(t)>0$ for $t\geq t_1\geq t_0$. As in the proof of Theorem 1, we obtain (17). Thus,

$$\begin{aligned} w(t)\leq w(t_2)&-\int_{t_2}^{t}\rho(s)q(s)\,ds+a(t)\dot{\rho}(t)\int_0^{x(t)}\frac{\psi(u)}{f(u)}du\\ &-a(t_2)\dot{\rho}(t_2)\int_0^{x(t_2)}\frac{\psi(u)}{f(u)}du-\int_{t_2}^{t}(a(s)\dot{\rho}(s))^{\cdot}\left(\int_0^{x(s)}\frac{\psi(u)}{f(u)}du\right)ds. \end{aligned}$$

The rest of the proof is similar to that of Theorem 1, Case 3, and is omitted.

THEOREM 4. *Let conditions* (5)–(7) *and* (11) *hold, and*

$$f'(x)>0 \quad \textit{and} \quad \frac{\psi(x)}{f'(x)}\leq\alpha_1 \quad \textit{for } x\neq 0. \tag{21}$$

Suppose that there exists a differentiable function $\rho\colon[t_0,\infty)\to(0,\infty)$ *such that*

$$\limsup_{t\to\infty}\int_{t_0}^{t}\left[\rho(s)q(s)-\frac{\alpha_1}{4}\frac{a(s)\dot{\rho}^2(s)}{\rho(s)}\right]ds=\infty. \tag{22}$$

Then (1) *is oscillatory.*

Proof. Let $x(t)$ be a nonoscillatory solution of (1), say $x(t)>0$ for $t\geq t_1\geq t_0$. Using arguments similar to those in the proof of Theorem 1 we get (15). Thus,

$$\begin{aligned}\dot{w}(t)\leq&-\rho(t)q(t)+\frac{a(t)\dot{\rho}^2(t)}{4\rho(t)}\frac{\psi(x(t))}{f'(x(t))}\\&-a(t)\psi(x(t))\left[\sqrt{\rho(t)f'(x(t))}\,\frac{\dot{x}(t)}{f(x(t))}-\frac{\dot{\rho}(t)}{2\sqrt{\rho(t)f'(x(t))}}\right]^2\\\leq&-\rho(t)q(t)+\frac{\alpha_1}{4}\frac{a(t)\dot{\rho}^2(t)}{\rho(t)}.\end{aligned}$$

Integrating the above inequality from t_2 to t, we get

$$\int_{t_2}^{t}\left[\rho(s)q(s)-\frac{\alpha_1 a(s)\dot{\rho}^2(s)}{4\rho(s)}\right]ds\leq w(t_2)-w(t)\leq w(t_2)<\infty,\qquad t\geq t_2.$$

This contradicts (22). The case $x(t)<0$ for $t\geq t_1$ is similar.

THEOREM 5. *Let condition* (22) *in Theorem* 4 *be replaced by*

$$\begin{aligned}(23)\qquad\limsup_{t\to\infty}\frac{1}{t^{n-1}}\int_{t_0}^{t}(t-u)^{n-3}\bigg[&(t-u)^2\rho(u)q(u)\\&-\frac{\alpha_1 a(u)[(t-u)\dot{\rho}(u)-(n-1)\rho(u)]^2}{4\rho(u)}\bigg]du=\infty,\end{aligned}$$

for some integer $n\geq 3$. *Then* (1) *is oscillatory.*

Proof. Let $x(t)$ be a nonoscillatory solution of (1). Assume $x(t)>0$ for $t\geq t_1\geq t_0\geq 0$. Following the same way as in the proof of Theorem 1, we get (15). Thus

$$\begin{aligned}\int_{t_2}^{t}(t-u)^{n-1}\dot{w}(u)\,du\leq&-\int_{t_2}^{t}(t-u)^{n-1}\rho(u)q(u)\,du\\&+\int_{t_2}^{t}(t-u)^{n-1}\dot{\rho}(u)\frac{a(u)\psi(x(u))\dot{x}(u)}{f(x(u))}du\\&-\int_{t_2}^{t}(t-u)^{n-1}\rho(u)\frac{a(u)\psi(x(u))\dot{x}^2(u)f'(x(u))}{f^2(x(u))}du.\end{aligned}$$

Since

$$\begin{aligned}\int_{t_2}^{t}(t-u)^{n-1}\dot{w}(u)\,du=&-(t-t_2)^{n-1}w(t_2)\\&+\int_{t_2}^{t}(n-1)(t-u)^{n-2}\frac{a(u)\rho(u)\psi(x(u))\dot{x}(u)}{f(x(u))}du,\end{aligned}$$

we get

$$t^{1-n}\int_{t_2}^{t}(t-u)^{n-3}\left[(t-u)^2\rho(u)q(u)-\frac{\alpha_1}{4}\frac{a(u)[(t-u)\dot\rho(u)-(n-1)\rho(u)]^2}{\rho(u)}\right]du$$

$$\le\left(1-\frac{t_2}{t}\right)^{n-1}w(t_2)-t^{1-n}\int_{t_2}^{t}a(u)\psi(x(u))\left([(t-u)^{n-1}\rho(u)f'(x(u))]^{1/2}\frac{\dot x(u)}{f(x(u))}\right.$$

$$\left.-\frac{(t-u)^{(n-3)/2}[\dot\rho(u)(t-u)-(n-1)\rho(u)]}{2[\rho(u)f'(x(u))]^{1/2}}\right)^2du$$

$$\le\left(1-\frac{t_2}{t}\right)^{n-1}w(t_2)\to w(t_2)\quad \text{as } t\to\infty,$$

which contradicts (23). Thus our proof is complete.

COROLLARY. *Let condition* (23) *in Theorem* 5 *be replaced by*

(24) $$\limsup_{t\to\infty} t^{1-n}\int_{t_0}^{t}(t-u)^{n-1}\rho(u)q(u)\,du=\infty,$$

and

(25) $$\lim_{t\to\infty} t^{1-n}\int_{t_0}^{t}(t-u)^{n-3}\left[\frac{a(u)}{\rho(u)}[(t-u)\dot\rho(u)-(n-1)\rho(u)]^2\right]du<\infty,$$

for some integer $n\ge 3$. Then the conclusion of Theorem 5 *holds.*

Remarks.

1. If $p=0$ (or $k=1=\psi$), then in Theorem 5, q (or q and p) need not be of fixed sign to ensure the oscillation of (1). In that case, Theorem 5 includes [25, Thm. 2], [26, Thms. 1,2], [6, Thm.], [24, Thm.].

2. We can easily verify that the equation $\ddot x+t^{-2}x=0$ is oscillatory by Theorem 5 for $\rho(t)=t$ and $n=3$, while none of the results mentioned in Remark 1 can be applied to this equation. Hence, we conclude that our Theorem 5 is stronger and more general than these results.

For illustration, we consider the following examples.

Example 1. Consider the equations

(E1) $$\left(\frac{t}{1+\sin^2\log t}(1+x^2)\dot x\right)^{\cdot}+\frac{1}{t}x=0,\qquad t>0,$$

and

(E2) $$\left(\frac{1}{1+\sin^2\log t}(1+x^2)\dot x\right)^{\cdot}+\frac{1}{t}\dot x+\frac{1}{t^2}x=0,\qquad t>0.$$

Equation (E1) is oscillatory by Theorem 1 (I) for $\rho(t)=1$ and all bounded solutions of (E2) are oscillatory by Theorem 3 for $\rho(t)=t$. Equations (E1) and (E2) admit the oscillatory solution $x(t)=\sin\log t$.

Example 2. The equations

(E3) $$\left(\frac{1}{t}\dot x\right)^{\cdot}+\frac{2}{t^3}x=0,\qquad t>0,$$

and

$$\text{(E4)}\qquad \left(\frac{1}{t}(1+x^2)\dot{x}\right)^{\cdot}+\frac{2}{t^3}(x+x^3)=0,\qquad t>0,$$

are oscillatory by Theorem 4 for $\rho(t)=t^2$. One such solution of (E3) is $x=t\sin\log t$.

Example 3. Consider the equations

$$\text{(E5)}\qquad \left(\frac{1}{t}(1+\sin^2 x)\dot{x}\right)^{\cdot}+\frac{1}{t^3}x^3=0,\qquad t>0,$$

and

$$\text{(E6)}\qquad \left(\sqrt{t}\,(1+\sin^2 x)\dot{x}\right)^{\cdot}+t\dot{x}^3+\frac{1}{t}x^3=0,\qquad t>0.$$

One can easily check that equation (E5) is oscillatory by Theorem 1 (II) and (III) for $\rho(t)=t^2$. Equation (E6) is oscillatory by Theorem 1 (I)–(III) for $\rho(t)=1$. We may note that the oscillatory character of these equations is not discernible from previously known oscillation criteria.

Example 4. Consider the equations

$$\text{(E7)}\qquad \left(\frac{1}{t}e^x\dot{x}\right)^{\cdot}+\frac{1}{2t^2}e^{3\dot{x}^3}+\frac{1}{2t^2\ln t}x=0,\qquad t\geq e^e,$$

and

$$\text{(E8)}\qquad \left(\frac{1}{t}e^x\dot{x}\right)^{\cdot}+\frac{1}{t^2\ln t}x=0,\qquad t\geq e^e.$$

One can easily check that all bounded solutions of (E7) and (E8) are oscillatory by Theorem 3 for $\rho(t)=t$. These equations have the nonoscillatory unbounded solution $x(t)=\ln t$.

The remainder of the theorems in this paper concern the oscillation of all solutions of (1) when $q(t)$ is of varying sign and

$$p(t)=0.$$

THEOREM 6. *If, in addition to conditions* (12) *and* (20), *we assume that there exists a differentiable function*

$$\rho\colon [t_0,\infty)\to(0,\infty),$$

such that

$$\text{(26)}\qquad \int^{\infty}\rho(s)q(s)\,ds=\infty$$

and

$$\text{(27)}\qquad \int^{\infty}\frac{1}{a(s)\rho(s)}\int_{t_0}^{s}\rho(\tau)q(\tau)\,d\tau\,ds=\infty,$$

then each of the following conditions ensures the oscillation of each of the continuable solutions of (1):

(A_1) $\dot{\rho}(t)\leq 0$ *and* $(a(t)\dot{\rho}(t))^{\cdot}\geq 0$ *for* $t\geq t_0$;

(A_2) *condition* (14) *and* $\int^{\infty}|(a(s)\dot{\rho}(s))^{\cdot}|\,ds<\infty$;

(A_3) *condition* (21) (*or condition* (20)) *and* $\int^\infty a(s)\dot\rho^2(s)/\rho(s)\,ds<\infty$;

(A_4) $f'(x)>0$, $0<\psi(x)\leq c$ *for all* x, $\dot\rho(t)>0$ *for* $t\geq t_0$ *and there exists a constant* $C_1>0$, *such that*

$$\frac{G''(x)G(x)}{G'^2(x)}\leq -\frac{1}{C_1} \quad \textit{for } x\neq 0$$

and

$$\frac{\rho(t)(a(t)\dot\rho(t))^{\cdot}}{a(t)\dot\rho^2(t)}\leq -C_1 \quad \textit{for } t\geq t_0,$$

where

$$G(x)=\int_0^x \frac{\psi(u)}{f(u)}\,du.$$

Proof. The proof of Theorem 6 can be modelled on that of [4, Thm. 2.2] and hence is omitted.

THEOREM 7. *Let conditions* (12) *and* (14) *hold. Suppose that there exists a differentiable function* ρ: $[t_0,\infty)\to(0,\infty)$, *such that*

$$\int^\infty \frac{1}{a(s)\rho(s)}\,ds=\infty, \tag{28}$$

and either

$$\dot\rho(t)\geq 0, \qquad (a(t)\dot\rho(t))^{\cdot}\leq 0 \quad \textit{for } t\geq t_0 \tag{29}$$

and

$$\int^\infty \rho(s)q(s)\,ds=\infty, \tag{30}$$

or

$$\dot\rho(t)>0, \qquad (a(t)\dot\rho(t))^{\cdot}\geq 0 \quad \textit{for } t\geq t_0 \tag{31}$$

and

$$\lim_{t\to\infty}\frac{1}{a(t)\dot\rho(t)}\int^t \rho(s)q(s)\,ds=\infty. \tag{32}$$

Then (1) *is oscillatory.*

Proof. Let $x(t)$ be a nonoscillatory solution of (1), say $x(t)>0$ for $t\geq t_1\geq t_0$. Let

$$w(t)=\frac{a(t)\psi(x(t))\dot x(t)}{f(x(t))}\rho(t).$$

Then $w(t)$ satisfies

$$\dot w(t)=-\rho(t)q(t)+\frac{a(t)\dot\rho(t)\psi(x(t))\dot x(t)}{f(x(t))}-\frac{a(t)\rho(t)\psi(x(t))f'(x(t))\dot x^2(t)}{f^2(x(t))}.$$

Thus,

$$w(t)\le w(t_1)-\int_{t_1}^{t}\rho(s)q(s)\,ds+\int_{t_1}^{t}\frac{a(s)\dot\rho(s)\psi(x(s))\dot x(s)}{f^2(x(s))}\,ds$$
$$-\int_{t_1}^{t}a(s)\rho(s)f'(x(s))\psi(x(s))\left(\frac{\dot x(s)}{f(x(s))}\right)^2 ds.$$

Consider the following two cases.

Case 1. Let (29) hold. Bonnet's form of the second mean-value theorem implies

$$\int_{t_1}^{t}\frac{a(s)\dot\rho(s)\psi(x(s))\dot x(s)}{f(x(s))}\,ds\le c_1\int_{x(t_1)}^{x(t)}\frac{\psi(u)}{f(u)}\,du\le c_2$$

for some constants $c_1, c_2>0$. Thus, by (30), there exists a $t_2\ge t_1$, such that

(33)
$$\frac{a(t)\rho(t)\psi(x(t))\dot x(t)}{f(x(t))}+\int_{t_1}^{t}a(s)\rho(s)f'(x(s))\psi(x(s))\left(\frac{\dot x(s)}{f(x(s))}\right)^2 ds\le -1,\qquad t\ge t_2.$$

Case 2. If (31) holds, then, by using again the Bonnet theorem, for some $c_3>0$, $t\ge t_1$, we have

$$\int_{t_1}^{t}\frac{a(s)\dot\rho(s)\psi(x(s))\dot x(s)}{f(x(s))}\,ds\le c_3a(t)\dot\rho(t),$$

and consequently we obtain

$$\frac{1}{a(t)\dot\rho(t)}\left[a(t)\rho(t)\psi(x(t))\,\frac{\dot x(t)}{f(x(t))}+\int_{t_1}^{t}a(s)\rho(s)f'(x(s))\psi(x(s))\left(\frac{\dot x(s)}{f(x(s))}\right)^2 ds\right]$$
$$\le c_3+\frac{w(t_1)}{a(t)\dot\rho(t)}-\frac{1}{a(t)\dot\rho(t)}\int_{t_1}^{t}\rho(s)q(s)\,ds.$$

Using (32), it follows that for some $t_2\ge t_1$ and for every $t\ge t_2$ (33) holds. Since, from (33), we have $\dot x(t)<0$ for $t\ge t_2$, (33) can be rewritten as follows:

$$(34)\quad \frac{a(t)\rho(t)\psi(x(t))(-\dot x(t))}{f(x(t))}\ge 1+\int_{t_2}^{t}a(s)\rho(s)f'(x(s))\psi(x(s))\left(\frac{\dot x(s)}{f(x(s))}\right)^2 ds.$$

Multiplying (34) by

$$\frac{-f'(x(t))\dot x(t)}{f(x(t))}\left[1+\int_{t_2}^{t}a(s)\rho(s)\psi(x(s))f'(x(s))\left(\frac{\dot x(s)}{f(x(s))}\right)^2 ds\right]^{-1}\ge 0,$$

and integrating, we obtain

$$(35)\quad \ln\frac{f(x(t_2))}{f(x(t))}\le\ln\left[1+\int_{t_2}^{t}a(s)\rho(s)\psi(x(s))f'(x(s))\left(\frac{\dot x(s)}{f(x(s))}\right)^2 ds\right],\qquad t\ge t_2.$$

By (35) and (33) we conclude that

$$\frac{\psi(x(t))\dot{x}(t)}{f(x(t))} \le \frac{\psi(x(t))\dot{x}(t)}{f(x(t_2))} \le -\frac{1}{a(t)\rho(t)}, \qquad t \ge t_2,$$

which, by (28), leads to a contradiction. The proof is now complete.

For illustration we consider the following example.

Example 5. The equations

$$\text{(E9)} \qquad \left(2\cosh(\sin t)\,\frac{1}{1+x^2}\,\dot{x}\right)^{\cdot} + (\sin t)e^{-\sin t}x = 0$$

and

$$\text{(E10)} \qquad (2\,\mathrm{sech}(\sin t)(1+x^2)\dot{x})^{\cdot} + (\sin t)e^{-5\sin t}x^5 = 0$$

have the nonoscillatory solution $x(t) = e^{\sin t}$; the only assumption that fails is condition (30) (or (32)).

Remarks.

1. Some of the results in this paper are extendable to inequalities of the form $x(t)\{(a(t)\psi(x(t))\dot{x}(t))^{\cdot} + p(t)k(t,x(t),\dot{x}(t))\dot{x}(t) + Q(t,x(t))\} \le 0$, where Q: $[t_0,\infty) \times R \to R$ is continuous and

$$Q(t,x) \ge q(t)f(x) \quad \text{for } x \ne 0,$$

as well as the functional inequalities of the form

$$x(t)\{(a(t)\psi(x(t))\dot{x}(t))^{\cdot} + p(t)k(t,x(t),\dot{x}(t))\dot{x}(t) + Q(t,x[g(t)])\} \le 0$$

where Q is above and g: $[t_0,\infty) \to R$ is continuous and $\lim_{t\to\infty} g(t) = \infty$.

2. It is obvious that condition (27) used in Theorem 6 is weaker than condition (28) given in Theorem 7, and consequently our Theorems 6 and 7 generalize and improve the corresponding ones in [2], [3] and [7].

3. In Theorems 4 and 5, if we replace all conditions on f by

$$\text{(36)} \qquad \frac{f(x)}{x} \ge c > 0 \quad \text{for } x \ne 0,$$

then the conclusions of these theorems remain valid. It is obvious that the function f in condition (36) need not be a monotone function, e.g., $f(x) = xe^{\sin x}$.

4. One can easily check that in view of Theorems 4 and 5, Kamenev's results in [8] can be easily extended to equation (1). Here we omit the detail.

5. Our results include some of the results of Wong [22], [23], Kartsatos [10], Yeh [25], [26], Naito [14], Graef, Rankin and Spikes [6], Grace and Lalli [4], Legatos and Kartsatos [13], Staikos and Sficas [16] and Opial [15].

Finally, it remains an open question to the authors if the results of this paper can be extended for (1) where q: $[t_0,\infty) \to R$ is a continuous function and is of varying sign.

REFERENCES

[1] N. P. Bhatia, *Some oscillation theorems for second order differential equations*, J. Math. Anal. Appl., 15 (1966), pp. 442–446.

[2] W. J. Coles, *Oscillation criteria for nonlinear second order equations*, Ann. Mat. Pura Appl., (4), 82 (1969), pp. 123–134.

[3] W. J. COLES, *A nonlinear oscillation theorem*, in International Conference on Differential Equations, Academic Press, New York, 1975, pp. 193–202.

[4] S. R. GRACE AND B. S. LALLI, *Oscillation theorems for perturbed nonlinear differential equations*, J. Math. Anal. Appl., 77 (1980), pp. 205–214.

[5] J. R. GRAEF, S. M. RANKIN AND P. W. SPIKES, *Oscillation theorems for perturbed nonlinear differential equations*, J. Math. Anal. Appl., 65 (1978), pp. 375–390.

[6] I. V. KAMENEV, *Integral criterion for oscillation of linear differential equations of second order*, Mat. Zametki, 23 (1978), pp. 249–251. (In Russian.)

[7] ______, *Oscillation of solutions of second order nonlinear equations with sign-variable coefficients*, Differencial'nye Uravnenija, 6 (1970), pp. 1718–1721. (In Russian.)

[8] ______, *Oscillation of solutions of a second order differential equation with "integrally small" coefficient*, Differencial'nye Uravnenija, 13 (1977), pp. 1491–1497. (In Russian.)

[9] A. G. KARTSATOS, *Properties of bounded solutions of nonlinear equations of second order*, Proc. Amer. Math. Soc., 19 (1968), pp. 1057–1059.

[10] A. G. KARTSATOS AND H. ONOSE, *Remarks on oscillation of second order differential equations*, Bull. Fac. Sci. Ibaraki Univ., Math., 5 (1973), pp. 23–31.

[11] A. KROOPNICK, *Oscillation properties of* $(M(t)\dot{x})^{\cdot}+a(t)b(x)=0$, J. Math. Anal. Appl., 63 (1978), pp. 141–144.

[12] A. C. LAZER, *A stability result for the differential equation* $\ddot{y}+p(t)y=0$, Michigan Math. J., 12 (1965), pp. 193–196.

[13] G. G. LEGATOS AND A. G. KARTSATOS, *Further results on the oscillation of solutions of second order equations*, Math. Japan, 14 (1968), pp. 67–73.

[14] M. NAITO, *Oscillation criteria for a second order differential equation with a damping term*, Hiroshima Math. J., 4 (1974), pp. 347–354.

[15] Z. OPIAL, *Sur une critère d'oscillation de l'équation différentielle* $(Q(t)\dot{x})^{\cdot}+f(t)x=0$, Ann. Polon. Math., 6 (1959), pp. 99–104.

[16] V. A. STAIKOS AND Y. G. SFICAS, *Oscillation for forced second order nonlinear differential equations*, Atli Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur., 8 (55) (1973), pp. 20–30.

[17] C. A. SWANSON, *Comparison and Oscillation Theory of Linear Differential Equations*, Academic Press, New York, 1968.

[18] C. C. TRAVIS, *A note on second order nonlinear oscillation*, Math. Japan, 18 (1973), pp. 261–264.

[19] D. WILLET, *Classification of second order linear differential equations with respect to oscillation*, Adv. Math. 3 (1969), pp. 594–623.

[20] J. S. W. WONG, *Some stability conditions for* $\ddot{x}+a(t)x^{2n-1}=0$, SIAM J. Appl. Math., 15 (1967), pp. 889–892.

[21] ______, *Remarks on stability conditions for the differential equation* $\ddot{x}+a(t)f(x)=0$, J. Austral. Math. Soc., 9 (1969), pp. 496–502.

[22] ______, *On second order nonlinear oscillation*, Funkcial. Ekvac., 11 (1968), pp. 207–234.

[23] ______, *Oscillation theorems for second order nonlinear differential equations*, Bull. Inst. Math. Acad. Sinica, 3 (1975), pp. 283–309.

[24] A. WINTNER, *A criterion for oscillatory stability*, Quart. Appl. Math., 7 (1949), pp. 115–117.

[25] C. C. YEH, *An oscillation criterion for second order nonlinear differential equations with functional arguments*, J. Math. Anal. Appl., 76 (1980), pp. 72–76.

[26] ______, *Oscillation theorms for nonlinear second order differential equations with damped term*, Proc. Amer. Math. Soc., 84 (1982), 397–402.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics
004

ON A COMPARISON THEOREM OF HILLE FOR SELF-ADJOINT SECOND ORDER LINEAR DIFFERENTIAL SYSTEMS*

D. F. ST. MARY[†]

Abstract. Hille–Wintner type comparison theorems are developed in which two second order self-adjoint linear systems are compared. In particular, classes of $n\times n$ Hermitian matrix functions are defined, using Opial-type inequalities, which can be used for comparison purposes in the Hille theorems. Additional theorems are presented in which a system and a scalar differential equation are compared.

1. Let $p(t)$, $q(t)$ be real valued continuous functions defined on $[a,\infty)$, and

$$P(t)\equiv\lim_{T\to\infty}\int_t^T p(s)\,ds=\int_t^\infty p(s)\,ds,\qquad Q(t)\equiv\int_t^\infty q(s)\,ds$$

exist (finitely).

THEOREM. *Let $P(t)\geq|Q(t)|$ on $[a,\infty)$. If $u''+p(t)u=0$ is nonoscillatory on $[a,\infty)$, then $u''+q(t)u=0$ is nonoscillatory on $[a,\infty)$.*

The previous theorem is the basic statement of the comparison principle and was first stated in a weaker form by Hille [11]; it has evolved through the efforts of Wintner [18], Taam [16], Hartman [10], Willett [17], and Wong [19]. No *general* analogue of this result for systems appears in the literature, although much recent activity has developed around it. Erbe [6], [7] has considered the problem for scalar differential equations of order three and four, and Butler [3] has taken another look at the scalar second order problem. Jones [12], and Etgen and Lewis [8] consider the case of systems (the latter in B^*-algebras) but relate a system to a scalar equation.

In this paper we present several versions of the theorem for the corresponding matrix systems. In particular, we show that for certain classes of nonoscillatory comparison functions $p(t)$, for which incidentally $P(t)$ need not be nonnegative, Hille's theorem holds for a pair of *systems*. In §3 we revert to comparing matrix systems to scalar equations and derive the Taam and Hartman generalizations. In the latter case we present a theorem which also encompasses the Sturm comparison theorem.

2. Let $p(t)$, $q(t)$ be continuous $n\times n$ Hermitian matrix valued functions defined on $[a,\infty)$. The differential system

$$E[p]\qquad\qquad U''+p(t)U=0$$

is said to be *nonoscillatory* on $[a,\infty)$, if there exists a conjoined (i.e. $U'^*U-U^*U'\equiv0$) $n\times n$ matrix solution $U(t)$ of $E[p]$ which is nonsingular on $[\alpha,\infty)$ for some α. Throughout the discussion $P(t)$, $Q(t)$ will denote $n\times n$ Hermitian matrix functions such that $P'(t)=-p(t)$, $Q'(t)=-q(t)$. In particular, if we assume that $p(t)$ satisfies

$$(t-a)^{-1}\int_a^t\int_a^s p(\sigma)\,d\sigma\,ds\geq\lambda E \tag{2.1}$$

for all t large, λ some real number, E the $n\times n$ identity matrix, and that $E[p]$ is nonoscillatory, then the existence of the limit as $t\to+\infty$ of the left side of (2.1) follows,

* Received by the editors January 25, 1982, and in final revised form April 27, 1983. This research was partially supported by the National Science Foundation under grant PRM-8115410.

[†] Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003.

see [5]. Further, if the limit is denoted by C and $P(t)$ is defined by

$$P(t)=C-\int_a^t p(s)\,ds, \tag{2.2}$$

then there exists an $n\times n$ Hermitian matrix function $V(t)$ such that

$$V(t)=P(t)+\int_t^\infty V^2(s)\,ds \quad \text{on } [\alpha,\infty). \tag{2.3}$$

(Note that in this case if $\int_a^\infty p(s)\,ds$ exists then $P(t)=\int_t^\infty p$.) Finally, we remark that a well-known criterion for nonoscillation of $E[p]$ is the existence on $[\alpha,\infty)$, for some $\alpha\geq a$, of an $n\times n$ Hermitian matrix function $W(t)$ for which the Riccati inequality $\mathfrak{R}[W]\equiv W'+W^2+p\leq 0$ holds on $[\alpha,\infty)$. The corresponding result for systems of the form $E[p;r]$ to be considered later also holds. Matrix inequalities are considered in the positive (nonnegative) definite sense.

We now present a sequence of Hille type theorems which successively define classes of comparison functions. The comparison functions satisfy Opial type inequalities [13]. At the conclusion of this presentation we shall discuss an example demonstrating a relationship among the theorems.

THEOREM 2.1. *Let c, C be $n\times n$ Hermitian matrix valued functions satisfying $C'(t)=-c(t)$ on $[a,\infty)$ and let* (i) $(P(t)+C(t))^2\leq c(t)$ *on* $[a,\infty)$. *If* (ii) $(Q(t)+C(t))^2\leq(P(t)+C(t))^2$ *on* $[a,\infty)$, *then $E[q]$ is nonoscillatory. In particular, if $P^2(t)\leq p(t)/4$, then $Q^2(t)\leq P^2(t)$ implies $E[q]$ is nonoscillatory.*

Proof. It follows from [5, Lemma 4.1] that (i) implies $E[p]$ is nonoscillatory, but then (ii) in conjunction with (i) yields, in the same manner, the nonoscillation of $E[q]$. The corollary is obtained on taking $C(t)\equiv P(t)$. Hypothesis (i) is clear. We show that (ii) is satisfied by using the matrix inequality

$$AB^*+BA^*\leq AA^*+BB^*,$$

$$(Q+C)^2=Q^2+QP+PQ+P^2\leq 4P^2=(P+C)^2.$$

We remark that for Hermitian matrices A,B, $0\leq A\leq B$ does not imply $A^2\leq B^2$. In order to begin to observe the differences between the various hypotheses we shall be considering, we note that the function $P(t)=((1+\sin t)/kt)E$ does not satisfy $P^2\leq p/4$ for any k, nor does it satisfy $\int_t^\infty P^2\leq P(t)/4$ (to be commented upon later) for any k, but if $C(t)=(2/kt)E$, then (i) holds for $k\geq 8$.

THEOREM 2.2. *Let $\mathcal{P}(t),\mathcal{Q}(t)$ be $n\times n$ Hermitian matrix functions satisfying $\mathcal{P}'(t)=-P^2(t)$, $\mathcal{Q}'(t)=-Q^2(t)$ and let $\mathcal{P}^2(t)\leq P^2(t)/16$ on $[a,\infty)$. If* (i) $Q^2(t)\leq P^2(t)$ *on* $[a,\infty)$, *or* (ii) $\mathcal{Q}^2(t)\leq\mathcal{P}^2(t)$ *on* $[a,\infty)$, *then $E[q]$ is nonoscillatory.*

Proof. Let $p_1(t)\equiv 4P^2(t)$; then for .$P_1(t)\equiv 4\mathcal{P}(t)$, $4P_1^2(t)\leq p_1(t)$. Thus $p_1(t)$ is a comparison function for Theorem 2.1. Using (ii) Theorem 2.1 implies that $E[4Q]^2$ is nonoscillatory from which the conclusion follows [5, Lemma 4.2]. In the case of (i), the nonoscillation of $E[p_1]$ implies that of $E[4Q^2]$ by the Sturm comparison theorem and hence the conclusion. We remark that implicit in his proof is the conclusion that $E[p]$ is nonoscillatory, see also [5, Thm. 4.3].

It is not difficult to present examples of functions $p(t)$ such that $P(t)=\int_t^\infty p$ exists and $(\int_t^\infty P^2)^2\leq P^2(t)/16$ but which do not satisfy $P^2(t)\leq p(t)/4$ for all large t.

Let $Y_p(t;s)$ be the solution of the initial value problem $Y'=P(t)Y$, $Y(s)=E$, E the $n\times n$ identity matrix. Define, when the integral exists,

$$\bar{P}(t)\equiv\int_t^\infty Y_p^*(s;t)P^2(s)Y_p(s;t)\,ds.$$

Analogously define $\bar{Q}(t)$. It was shown in [5, Thm. 3.2] that (2.1) in conjunction with the nonoscillation of $E[p]$ yields the existence of $\bar{P}(t)$, where $P(t)$ is defined by (2.2).

THEOREM 2.3. *Let $\bar{P}(t)$ exist and*

$$\left(\int_t^\infty Y_p^*(s;t)\bar{P}^2(s)Y_p(s;t)\,ds\right)^2 \leq \bar{P}^2(t)/16.$$

If $\bar{Q}(t)$ exists, $\bar{Q}^2(t)\leq\bar{P}^2(t)$, and, for

$$A(t)=\left(\left(\bar{Q}(t)+Q(t)\right)-\left(\bar{P}(t)+P(t)\right)\right),\qquad B(t)=\int_t^\infty Y_p^*4\bar{P}^2Y_p\,ds,$$

$A(t)B(t)+B(t)A(t)\leq 0$, then $E[q]$ is nonoscillatory.

Proof. [5, Thm. 4.4] implies that $E[p]$ is nonoscillatory. Let $W(t)=Q(t)+\bar{Q}(t)+B(t)$; then W is Hermitian and

$$W'(t)=-q-Q^2(t)-\bar{Q}(t)Q(t)-Q(t)\bar{Q}(t)-4\bar{P}^2(t)-B(t)P(t)-P(t)B(t),$$

so

$$W'+W^2+q=AB+BA-(\bar{P}-B)^2+2(B^2-\bar{P}^2)+\bar{Q}^2-\bar{P}^2\leq 0.$$

Thus $E[q]$ is nonoscillatory.

The conditions relating P,Q and $\bar{P},\bar{Q}$ in this theorem reduce in the scalar case to those of Willett [17] and Wong [19] which yield a generalization of the classical comparison theorem. The proof of the final theorem in the series is completely analogous to that of Theorem 2.3; one uses $W(t)=Q(t)+\bar{Q}(t)+\tilde{Q}(t)+B(t)$.

THEOREM 2.4. *Let $\bar{P}(t)$ and $\tilde{P}(t)$ exist, where*

$$\tilde{P}(t)\equiv\int_t^\infty Z_p^*(s;t)\bar{P}^2(s)Z_p(s;t)\,ds,$$

$Z_p(t;T)$ is the solution of the initial value problem $Z'=(\bar{P}(t)+P(t))Z$, $Z(T)=E$, and let

$$\left(\int_t^\infty Z_p^*(s;t)\tilde{P}^2(s)Z_p(s;t)\,ds\right)^2\leq\tilde{P}^2(t)/16.$$

If $\bar{Q}(t)$ and $\tilde{Q}(t)$ exist, $\tilde{Q}^2(t)\leq\tilde{P}^2(t)$, and, for $A(t)=(\tilde{Q}+\bar{Q}+Q)-(\tilde{P}+\bar{P}+P)$, $B(t)=\int_t^\infty Z_p^*4\tilde{P}^2Z_p\,ds$, $A(t)B(t)+B(t)A(t)\leq 0$, then $E[q]$ is nonoscillatory.

We shall now present an example in which Theorems 2.1, 2.3, and 2.4 all apply and upon the successive application of each we observe an improved result concerning the nonoscillation of $E[q]$. We shall take for $q(t)$ the 2×2 matrix which has $(\alpha\sin\beta t)/t^\gamma$, $\gamma>0$, on its main diagonal and has $(a\cos bt)/t^\eta$, $\eta>0$, in the other positions, and for $p(t)$ the 2×2 matrix $(1/4t^2)E$. It follows that $P(t)=(1/4t)E$; $Q(t)$ has $((\alpha\cos\beta t)/\beta t^\gamma)+O(1/t^{\gamma+1})$ on its main diagonal and $((-a\sin bt)/bt^\eta)+O(1/t^{\eta+1})$ in the off-diagonal positions; and $P^2\leq p/4$. Thus, applying Theorem 2.1, $E[q]$ is nonoscillatory whenever $Q^2(t)\leq P^2(t)$. In all cases we choose to analyze such inequalities by showing that the maximum eigenvalue $\Lambda(Q^2(t))$, of $Q^2(t)$ is less than or equal to the minimum eigenvalue $\lambda(P^2(t))$, of $P^2(t)$ for all large t. $\Lambda(Q^2(t))$ is easily obtained, e.g. in the case $\gamma=1$, $\eta>1$, $\Lambda(Q^2(t))=((\alpha^2\cos^2\beta t)/\beta^2t^2)+O(1/t^\nu)$, $\nu>2$ and the condition, $Q^2\leq P^2$, translates to $E[q]$ nonoscillatory when $|\alpha/\beta|<1/4$, a,b arbitrary. The other cases in which $E[q]$ is nonoscillatory are: $\gamma>1$, $\eta=1$, $|a/b|<1/4$, α,β arbitrary; $\gamma>1$, $\eta>1$, α,β,a,b arbitrary; and $\gamma=1$, $\eta=1$, $|\alpha/\beta|+|a/b|<1/4$.

Henceforth we shall concentrate exclusively on the case $\gamma=1$, $\eta>1$. In the application of Theorem 2.3 the conditions on p are all trivially satisfied and the analysis of the existence of $\bar{Q}$ and the relation $\Lambda(\bar{Q}^2)\leq\lambda(\bar{P}^2)$ is carried out by using the fact that

$$Y_q^*(s;t)Q^2(s)Y_q(s;t)\leq\Lambda(Q^2(s))Y_q^*(s;t)Y_q(s;t)$$

$$\leq\Lambda(Q^2(s))\left(\exp 2\int_t^s\Lambda(Q(\tau))\,d\tau\right)E, \qquad s>t.$$

Since $\int_t^s\Lambda(Q(\tau))\,d\tau=O(1/t)$ for $s>t$, $\bar{Q}(t)\leq((\alpha^2/2\beta^2t)+O(1/t^{\nu-1}))E$ and the condition $\bar{Q}^2\leq\bar{P}^2$ implies that $|\alpha/\beta|<1/2$, but the auxiliary condition $AB+BA\leq0$ translates into $(\alpha^2/2\beta^2)+(\alpha/\beta)\cos\beta t<3/8$ for all t large. This latter is dominant and we have $E[q]$ nonoscillatory when $|\alpha/\beta|<-1+\sqrt{7}/2\approx.323$.

Theorem 2.4 is applied using the same kind of analysis,

$$\int_t^s\Lambda(\bar{Q}+Q)\,d\tau\leq(\alpha^2/2\beta^2)\ln(s/t)+O(1/t^{\nu-2}), \qquad s>t,$$

and

$$\tilde{Q}(t)\leq(\delta^2/4(1-\delta^2)t)+O(1/t^{\nu-1}), \qquad \delta=\alpha/\beta.$$

The condition $\tilde{Q}^2\leq\tilde{P}^2$ yields $|\alpha/\beta|<\sqrt{(-1+\sqrt{17})/8}\approx.625$. The equation determined by the auxiliary condition is $\delta^4/4(1-\delta^2)+\delta^2/2+\delta-7/16<O$, a solution of which is $.3655\ldots$ So if $|\alpha/\beta|\leq.365$, $E[q]$ is nonoscillatory. Thus we have shown that the three theorems give successively better results when applied to $E[q]$ with $\gamma=1$, $\eta>1$, a,b arbitrary.

Some additional remarks regarding this example are in order. Theorem 2.2 (ii) can also be applied and yields $|\alpha/\beta|<1/2\sqrt{2}\approx.354$. Theorems 2.1 and 2.2 have a decided advantage in ease of application. The best possible result for the nonoscillation $E[q]$ is $|\alpha/\beta|<1/\sqrt{2}$ and is obtained from [5, Thm. 4.4]. The sharpness holds since, when $|\alpha/\beta|>1/\sqrt{2}$, the diagonal element, as the coefficient in a scalar problem, is oscillatory and hence $E[q]$ is oscillatory, see [14], [19]. To this point we have applied the theorems in a matrix equation versus scalar equation comparison since $p(t)$ is essentially scalar. We need to build up a set of matrix functions which satisfy the various hypothesis of the theorems, the "p" hypothesis, so as to make them comparison functions. To this end we remark that the function $q(t)$, $|\alpha/\beta|<1/\sqrt{2}$, $\gamma=1$, $\eta>1$, is one such, since in particular it satisfies the hypothesis of Theorem 2.3 (but not that of Theorem 2.1 or 2.2).

3. Taam [16] generalized the Hille comparison theorem to equations of the form

$$E[p;r] \qquad (r(t)U')'+p(t)U=0.$$

We shall assume that the coefficient of U satisfies the same conditions as earlier and that the coefficient of U' is an $n\times n$ Hermitian matrix which is positive definite for all $t\geq a$. In this section one of the differential equations will always be scalar but it will be convenient to think of a scalar differential equation of the form $E[p;r]$ as a matrix system of that same form. For (real) scalar valued functions $r(t)$, $p(t)$ we shall (abusing the notation) also denote the $n\times n$ "scalar matrices" $r(t)E$, $p(t)E$ by $r(t)$, $p(t)$ respectively, and in this case call the $n\times n$ matrix system $E[p;r]$ a *scalar system*. We remark that if a system $E[p]$ is nonoscillatory, then every scalar equation of the form $u''+\pi^*p(t)\pi u=0$, π a unit vector in C_n, is nonoscillatory (generalizations of this remark appear in [1], [9], [14], [15]). This makes it possible to obtain certain kinds of comparisons of systems directly from the scalar theorem, e.g., for p an arbitrary $n\times n$ Hermitian matrix and q a diagonal matrix (here one uses the fact that for diagonal q, $E[q]$ is nonoscillatory if every diagonal equation $u''+q_{ii}u=0$ is). It is the modification and significant generalization of these ideas which are considered in Etgen and Lewis [8]. The following results do not appear to follow from direct application of scalar theorems.

THEOREM 3.1. *Let $E[p;r]$ be a scalar system for which $P(t)=\int_t^\infty p(s)\,ds$ exists and $0<r(t)\leq kE$ on $[a,\infty)$ for some constant k. If $l(t), q(t)$ are $n\times n$ Hermitian matrices satisfying*:

(i) $r(t)\leq l(t)$,

(ii) $Q(t)l^{-1}(t)+l^{-1}(t)Q(t)\leq 2r^{-1}(t)P(t)$, *and*

(iii) $Q^2(t)\leq P^2(t)$ *on* $[a,\infty)$,

then the nonoscillation of $E[p;r]$ implies that of $E[q;l]$.

Proof. By a well-known argument the conditions on $E[p;r]$ imply the existence of a scalar matrix $V(t)$ such that $V(t)=\int_t^\infty V(s)r^{-1}(s)V(s)\,ds+P(t)$. Put $W(t)=Q(t)+\int_t^\infty Vr^{-1}V\,ds\equiv Q(t)+H(t)$; then

$$\begin{aligned}\mathfrak{R}(W)&=W'+Wl^{-1}W+q\\&=H(l^{-1}-r^{-1})H+\left((Ql^{-1}+l^{-1}Q)-2r^{-1}P\right)H+(Ql^{-1}Q-Pr^{-1}P).\end{aligned}$$

It follows from (i)–(iii) that $\mathfrak{R}(W)\leq 0$.

Condition (ii) is trivially true under the classical scalar hypothesis $|Q(t)|\leq P(t)$ and thus Theorem 3.1 is a generalization of the Taam result.

The next theorem is a generalization of a result of Hartman [10] and represents the first time a theorem has been presented which encompasses both the Sturm and Hille comparison theorems.

THEOREM 3.2. *Let $E[p]$ be a scalar system*, (2.1) *hold, and $P(t)$ be defined by* (2.2). *If $q(t)$ is an $n\times n$ Hermitian matrix and α,β are real numbers such that*

(i) $\alpha^2Q^2(t)+(1-\alpha)q(t)\leq\beta^2P^2(t)+(1-\beta)p(t)$ *and*

(ii) $\alpha Q(t)\leq\beta P(t)$,

then the nonoscillation of $E[p]$ implies that of $E[q]$.

Proof. The conditions on $E[p]$ insure the existence of a scalar matrix $v(t)$ such that $v(t)=\int_t^\infty v^2(s)\,ds+P(t)$. If $z(t)\equiv\int_t^\infty v^2(s)\,ds+(1-\beta)P(t)$, then $[z+(\beta-1)P]'=-v^2(t)=-[z+\beta P]^2$ and hence

$$z'+z^2+\beta zP+\beta Pz+(1-\beta)p+\beta^2P^2\equiv 0.$$

Now (i), (ii) imply the matrix inequality

$$q_0\equiv w'+w^2+\alpha wQ+\alpha Qw+(1-\alpha)q+\alpha^2Q^2\leq 0$$

is satisfied by the scalar matrix $w(t)=z(t)$. Let $U(t)$ be the solution of the linear system $U'=(\alpha Q+w)U$, $U(a)=E$; then U satisfies the second order linear system $U''+(q-q_0)U=0$ and hence this system is nonoscillatory, since U is a conjoined (Q and w are Hermitian) nonsingular solution. But $q-q_0\geq q$ and hence the standard Sturm comparison theorem implies the nonoscillation of $E[q]$.

Acknowledgment. The referee has called the attention of the author to the paper by Arnol'd [2], which he indicates is an important milestone in the generalization of oscillation and comparison theorems.

REFERENCES

[1] W. ALLEGRETTO AND L. ERBE, *Oscillation criteria for matrix differential inequalities*, Canad. Math. Bull. 16 (1973), pp. 5–10.

[2] V. I. ARNOL'D, *Characteristic class entering in quantization conditions*, Funct. Anal. Appl. 1 (1967), pp. 1–13.

[3] G. J. BUTLER, *Hille-Wintner comparison theorems for second order ordinary differential equations*, Proc. Amer. Math. Soc., 76 (1979), pp. 51–59.

[4] C. CONLEY, *A new statement of Wazewski's theorem and an example*, in Ordinary and Partial Differential Equations, Lecture Notes in Mathematics 564, Springer-Verlag, New York, 1976.

[5] S. B. ELIASON AND D. F. ST. MARY, *On oscillation of linear differential systems*, J. Math. Anal. Appl., 66 (1978), pp. 379–386.

[6] L. ERBE, *Hille-Wintner type comparison theorem for self-adjoint fourth order linear differential equations*, Proc. Amer. Math. Soc., 80 (1980), pp. 417–422.

[7] ———, *Comparison theorems of Hille-Wintner type for third-order linear differential equations*, Bull. Austral. Math. Soc., 21 (1980), no. 2, pp. 175–188.

[8] G. J. ETGEN AND R. T. LEWIS,.*A Hille-Wintner comparison theorem for second order differential systems*, Czechoslovak Math. J., 30(105) 1980, pp. 98–107.

[9] G. J. ETGEN AND J. F. PAWLOWSKI, *Oscillation criteria for second order self-adjoint differential systems*, Pacific J. Math., 66 (1976), pp. 99–110.

[10] P. HARTMAN, *Ordinary Differential Equations*, S. M. Hartman, Baltimore, 1973.

[11] E. HILLE, *Nonoscillation theorems*, Trans. Amer. Math. Soc., 64 (1948), pp. 234–252.

[12] R. A. JONES, *Comparison theorems for matrix Riccati equations*, SIAM J. Appl. Math., 29 (1975), pp. 77–90.

[13] Z. OPIAL, *Sur les integrales oscillantes de l'èquation differentielle $u''+f(t)u=0$*, Ann. Polon. Math., 4 (1958), pp. 308–313.

[14] D. F. ST. MARY, *On transformation and oscillation of linear differential systems*, Canad. J. Math., 14 (1977), pp. 392–399.

[15] C. A. SWANSON, *Oscillation criteria for nonlinear matrix differential inequalities*, Proc. Amer. Math. Soc., 24 (1970), pp. 824–827.

[16] C.-T. TAAM, *Non-oscillatory differential equations*, Duke Math. J., 19 (1952), pp. 493–497.

[17] D. WILLETT, *Classification of second order linear differential equations with respect to oscillation*, Advances in Math., 3 (1969), pp. 594–623.

[18] A. WINTNER, *On the comparison theorem of Kneser-Hille*, Math. Scand., 5 (1957), pp. 255–260.

[19] J. S. W. WONG, *Oscillation and nonoscillation of solutions of second order linear differential equations with integrable coefficients*, Trans. Amer. Math. Soc., 144 (1969), pp. 197–215.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics

005

GENERAL SOLUTION OF THE STOCHASTIC PRICE-DIVIDEND INTEGRAL EQUATION: A THEORY OF FINANCIAL VALUATION*

S. P. SETHI[†], N. A. DERZKO[‡] AND J. LEHOCZKY[§]

Abstract. This paper deals with the problem of the financial valuation of a firm and its shares of stock with given financing policies in a general stochastic environment. A model of the firm is described which includes the price-dividend balance integral equation whose solution yields the time path of the share price, the number of outstanding shares and the value of the firm. These are shown to be the unique conditional expectations of certain stochastic processes. A broad class of firms for which the solution formula yields finite valued solutions is characterized. This paper represents a rigorous mathematical treatment, as well as a significant stochastic extension of the Miller–Modigliani theory of financial valuation. It is also shown that the cash-flow approach and the dividend approach to valuation of a firm are not equivalent in general. A precise condition, which makes them equivalent, is also obtained.

Introduction. In this paper, we generalize the earlier work [1] of the first two authors in order to study the valuation of a firm in a general stochastic environment. The financing policies of the firm are defined by a pair of real-valued stochastic processes denoting the rates of total dividends paid out and the external equity raised at each time $t \geq 0$. The total dividend process is assumed to be nonnegative. A positive value of the external equity at time t implies that the firm is issuing new stock at that time, while a negative value means that the firm is buying back its own stock. All transaction costs are assumed to be zero in the model.

The rate of dividend per share at time t is given by the rate of total dividends divided by the number of outstanding shares at that time. The rate of change in the number of outstanding shares at time t is given by the total rate of external equity divided by the price of a share at that time. With the initial number of outstanding shares being given, the above procedure defines the stochastic process denoting the number of outstanding shares over time, provided that the price per share over time is known.

The crucial piece of information for the valuation problem is, therefore, the price per share over time. This requires some assumptions about the economy in which the firm operates. We shall assume that there exists, in the economy, a spot rate of interest at which money can be borrowed or lent. The interest rate process will be assumed to be a stochastic process. If the agents in this economy are risk-neutral, then the price of a share at time t can be defined as the expected total discounted value of future dividends per share payments given the information available through time t.

In the absence of the risk-neutrality assumption, the reformulation of the problem is accomplished by taking the expectation with respect to an appropriate probability measure that is absolutely continuous with respect to the given underlying probability measure. This idea will be discussed in detail in §5. For now, it suffices to state that the

*Received by the editors June 30, 1982, and in revised form May 10, 1983.

[†] Faculty of Management Studies, University of Toronto, Toronto, Ontario, Canada M5S 1V4. The research of this author was supported in part by the Natural Sciences and Engineering Research Council under grant A-4619.

[‡] Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 1V4.

[§] Department of Statistics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. The research of this author was supported in part by the National Science Foundation under grant ECS-8101576.

entire analysis in this paper remains valid for any probability measure that is absolutely continuous with respect to the given measure.

Within this context, the integral equations governing the price per share and the number of outstanding shares are developed and solved under fairly minimal assumptions.

Moreover, we assume that some market processes, and other information processes, have influence on the dividend, external equity and discount rate processes. These could be economic indicators, technological forecasts, the announcement of the firm's future plans, etc. These are also taken into account implicitly in valuing the firm and its shares.

This paper represents an important advance over the seminal work of Miller and Modigliani [4] (MM hereafter). It provides a rigorous mathematical foundation for the MM theory in a general stochastic environment. MM claimed that the cash flow approach and the dividend approach to valuation are equivalent. This is not true in general. In fact, we show that the cash flow approach can provide valuation for a larger class of firms than can the dividend approach. We also provide the precise additional restriction under which the two approaches are equivalent.

In the next section, we specify the notation. The model is developed in §2. The solution of the model and the main results of the paper are obtained in §3. The financial interpretations and discussion of results are provided in §4. In §5, we discuss how the model can be extended to more general economies. §6 concludes the paper.

1. Notation. Let $\{\Omega, \mathcal{F}, \pi\}$ denote the underlying probability space and let $\Upsilon = [0, \infty)$. Let $\mathcal{L}_1(\Omega, \mathcal{F}, \pi)$ denote the space of integrable random variables over $(\Omega, \mathcal{F}, \pi)$. Let the nondecreasing family $\{\mathcal{F}_t,\ t \in \Upsilon\}$ of sub-σ-algebras be given and be right-continuous. Assume that $\mathcal{F}_0$ consists of Ω and all the π-null sets, and $\mathcal{F}_\infty \equiv \sigma\{\bigcup_{t\in\Upsilon}\mathcal{F}_t\} = \mathcal{F}$.

Let $\{\bar{D}(t),\ t \in \Upsilon\}$, $\{\bar{E}(t),\ t\in\Upsilon\}$ and $\{\rho(t),\ t\in\Upsilon\}$ be *real-valued* right-continuous adapted stochastic processes defined on the probability space; these represent, respectively, the rate of total dividends issued by the firm at t, the rate of total external equity raised by the firm at t and the interest rate at t. Furthermore,

$$(1.1)\qquad 0 \le \bar{D}(t) < \infty, \quad -\infty < \bar{E}(t) < \infty, \quad 0 \le \rho(t) < \infty, \quad t \in \Upsilon, \quad \pi\text{-a.s.}$$

The notation π-a.s. or π-almost surely means that the inequalities hold with probability $\pi = 1$, i.e. $\pi(0 \le \bar{D}(t) < \infty) = 1$, etc. From now on, all equalities and inequalities relationships between random variables will be understood to be π-almost sure relationships, unless otherwise specified. The symbol $\equiv$ is used to mean equals by definition.

It should be remarked that we begin our analysis with a given probability space $\{\Omega, \mathcal{F}, \pi\}$ and a family $\{\mathcal{F}_t,\ t\in\Upsilon\}$ of sub-σ-algebras. We do not require $\mathcal{F}_t$ to be the (smallest) σ-algebra generated by the processes $\{\bar{D}(\tau), \bar{E}(\tau), \rho(\tau), 0 \le \tau \le t\}$. Thus, additional information not implied by these processes may be contained in $\mathcal{F}_t$. However, no explicit representation of this additional information is used in this paper.

Let $\mathcal{E}_t$ denote the conditional expectation operator with respect to the σ-algebra $\mathcal{F}_t$, i.e., for any random variable $X \in \mathcal{L}_1(\Omega, \mathcal{F}, \pi)$, we have the notation

$$\mathcal{E}_t X = \mathcal{E}[X|\mathcal{F}_t].$$

Note that X also belongs to $\mathcal{L}_1(\Omega, \mathcal{F}, \mu)$, where μ is a π-continuous probability measure. We also define $\mathcal{E}^\mu$ as the expectation operator and $\mathcal{E}_t^\mu$ as the conditional expectation operator with respect to μ. It should be emphasized that all the results obtained in this paper remain valid if we replace $\mathcal{E}_t$ by $\mathcal{E}_t^\mu$ for some π-continuous probability measure μ.

It is convenient to perform the entire analysis in present-value terms. For this we define

$$K_{st} \equiv \exp\left\{-\int_s^t \rho(\xi)\,d\xi\right\}, \qquad t \geq s, \quad s,t \in \Upsilon, \tag{1.2}$$

having the property

$$K_{su}(\omega) = K_{st}(\omega)K_{tu}(\omega), \qquad s \leq t \leq u, \tag{1.3}$$

on each sample path $\omega \in \Omega$.

We can now define the present-valued processes (i.e., values discounted to $t=0$) denoting total dividends and external equity as

$$D(t) \equiv K_{0t}\bar{D}(t) \quad \text{and} \quad E(t) \equiv K_{0t}\bar{E}(t). \tag{1.4}$$

Note that K_{0t}, $D(t)$, and $E(t)$ are $\mathcal{F}_t$-measurable.[1]

By convention in this paper, the processes denoting values will, henceforth, be in present-value terms. The qualification "present-valued" or "discounted" for these processes will be automatically implied, unless otherwise specified.

The main object of the paper is the determination of the share price $P(t)$, the number of outstanding shares $N(t)$ and the value of the firm $V(t)$. Let

$$N(0) \equiv N_0; \tag{1.5}$$

it is possible to set $N(0)$ arbitrarily to any positive number without any loss of generality. We also define the obvious relation between $P(t)$, $N(t)$ and $V(t)$ as

$$V(t) \equiv P(t)N(t). \tag{1.6}$$

The quantities $P(t)$, $N(t)$ and $V(t)$, must be observable in period t, once the history up to time t is specified. This implies the requirement that the processes $P(t)$, $N(t)$, $V(t)$, $t \in \Upsilon$, be adapted to the σ-field $\{\mathcal{F}_t, t \in \Upsilon\}$, i.e.,

$$P(t), \quad N(t) \quad \text{and} \quad V(t) \quad \text{be } \mathcal{F}_t\text{-measurable.} \tag{1.7}$$

In the next section, we take up the analysis of the infinite horizon firm. See [7] for the analysis of the infinite horizon as well as a π-a.s. finite horizon firm in a discrete time framework, and see [8] for its special deterministic version.

2. Valuation equations for the infinite horizon firm. An infinite horizon firm shall be denoted by the pair of stochastic processes

$$\{D(t), E(t), t \in \Upsilon\},$$

satisfying certain assumptions to be specified later. Note that (1.1) implies

$$0 \leq D(t) < \infty, \qquad -\infty < E(t) < \infty, \qquad t \in \Upsilon. \tag{2.1}$$

The equation giving the number of outstanding shares in period t is

$$N(t) = N_0 + \int_0^t \frac{E(\tau)}{P(\tau)}\,d\tau, \qquad t \in \Upsilon. \tag{2.2}$$

[1] The analysis in this paper can be generalized to allow for lump-sum dividends and external equities. In such a case, one defines processes of bounded variations representing cumulative dividends and cumulative external equity. Also, one can assume the discount process K_{0t} to be of bounded variation.

The dividend-stream approach equates the share price to the expected value of the total dividend streams paid to that share. Thus,

$$P(t)=\mathcal{E}_t\int_t^\infty \frac{D(\tau)}{N(\tau)}d\tau, \qquad t\in\Upsilon. \tag{2.3}$$

Note that (2.3) implies $\lim_{t\to\infty}P(t)=0$. In writing (2.2) and (2.3), we assume $\{E(t)/P(t),\ t\in\Upsilon\}$ to be locally integrable and $\{D(t)/N(t),\ t\in\Upsilon\}$ to be integrable processes.

By a formal manipulation of (2.3), we can derive from it the arbitrage equation

$$P(t)=\mathcal{E}_t\left[P(t+\Delta t)+\int_t^{t+\Delta t}\frac{D(\tau)}{N(\tau)}d\tau\right], \qquad \Delta t>0, \quad t\in\Upsilon.$$

This arbitrage equation states that the expected capital loss per share from selling a share at time $t+\Delta t$ instead of at time t is equal to the expected value of dividends received in the interval $[t,t+\Delta t]$.

In the deterministic case [1], there would not be any conditional expectation on the right-hand side of the arbitrage equation, and it would be possible to convert it into a differential equation by dividing both sides by Δt and letting Δt approach zero. This is not possible in our general stochastic environment. What is possible, however, is to add $\int_0^t(D(\tau)/N(\tau))\,d\tau$ to both sides of the arbitrage equation and derive the condition that

$$M(t)\equiv P(t)+\int_0^t\frac{D(\tau)}{N(\tau)}d\tau \quad \text{is a martingale.} \tag{2.4}$$

This generalizes the arbitrage equation of MM to a continuous-time stochastic environment. Here, $M(t)$ represents the sum of the price of a share at time t and all the dividends that have so far accrued to the share. This sum is a constant (i.e., $M(t)=P(0)$, $t\in\Upsilon$) in the deterministic case and is a martingale (i.e., $\mathcal{E}_0M(t)=P(0)$, $t\in\Upsilon$) in the stochastic case. Since $\mathcal{E}_tM(T)=M(t)$ for any $T\geq t$, it states that in collecting the dividends during $[t,T]$ and then selling the share at time T, the owner of the share can expect on the average to be neither wealthier nor poorer than he is at time t. In other words, there are no arbitrage opportunities in the trading of the firm's shares.

We shall now define two systems of valuation:

Dividend system ≡ Equations (2.2), (2.3), (1.6),

P-arbitrage system ≡ Equations (2.2), (2.4), (1.6) and $\lim_{t\to\infty}P(t)=0$.

The P-arbitrage system is named because of the boundary condition on $P(t)$. Later on, we shall define another arbitrage system with another boundary condition.

For a given system, a triple $\{P(t),N(t),V(t),\ t\in\Upsilon\}$ satisfying the system (1.7) and

$$0<P(t)<\infty, \quad 0<N(t)<\infty, \quad 0<V(t)<\infty, \quad t\in\Upsilon, \tag{2.5}$$

is termed a *positive solution* for which the system is well defined. For the dividend system, e.g., it means that the integral in (2.3) is well defined. Note also that in view of (1.6), the conditions in (2.5) need only be satisfied for any two of the three quantities. In the following theorem, we establish that the two valuation systems defined above are equivalent.

THEOREM 1. *$\{P(t),N(t),V(t),\ t\in\Upsilon\}$ is a positive solution of the dividend system if and only if it is a positive solution of the P-arbitrage system.*

Proof. It is obvious that every solution of the dividend system satisfies the P-arbitrage system.

Now suppose $\{P(t), N(t), V(t),\ t\in\Upsilon\}$ is a solution of the P-arbitrage system. By definition, the integral in (2.4) is well defined for this solution. Moreover, (1.7), (2.1) and (2.5) imply that $M(t)$ is a positive $\mathfrak{L}_1$-martingale. By the convergence theorem for nonnegative $\mathfrak{L}_1$-martingales [3], $X(t)$ has a limit as t approaches infinity, thus

$$\lim_{t\to\infty} M(t)=M(\infty)>0. \tag{2.6}$$

Since $\lim_{t\to\infty} P(t)=0$ for our solution, we have

$$M(\infty)=\int_0^\infty \frac{D(\tau)}{N(\tau)}d\tau \tag{2.7}$$

from (2.4) and (2.6), and therefore the integral in (2.7) is well defined. With the martingale property $M(t)=\mathcal{E}_t M(\infty)$, we can use (2.4) and (2.7) to derive

$$P(t)=\mathcal{E}_t\left[M(\infty)-\int_0^t \frac{D(\tau)}{N(\tau)}d\tau\right]=\mathcal{E}_t\int_t^\infty \frac{D(\tau)}{N(\tau)}d\tau,$$

which is the same as (2.3) and establishes that $\{P(t), N(t), V(t),\ t\in\Upsilon\}$ is a solution of the dividend system. This completes the proof.

Theorem 1 does not provide us with a solution. This, we shall do next.

3. Solution of the infinite horizon firm. In this section we specify the precise assumptions for the firm and derive the main results of our paper. Along the way, we introduce two more valuation systems, which will facilitate us in obtaining a solution.

For a positive solution, it is clear from (2.2) that $N(t)$ is of *bounded variation*. We also know from (2.4) that $M(t)$ is a martingale. It follows [9, Thm. 1.2.8, p. 26] that the process $Y(t)$ defined as

$$Y(t)\equiv M(0)N_0+\int_0^t N(s)\,dM(s)=M(t)N(t)-\int_0^t M(s)\,dN(s) \quad \text{is a martingale.} \tag{3.1}$$

It is noted in passing that $\mathcal{E}_t Y(t)=M(0)N_0=P(0)N_0$ represents the value of the firm at time zero. A more meaningful interpretation of $Y(t)$ will be provided following (3.3). Substituting for $M(t)$ from (2.4) and using the definition (1.6), we can write (3.1) as

$$\begin{aligned} Y(t)&=N(t)P(t)+N(t)\int_0^t \frac{D(s)}{N(s)}ds-\int_0^t\left[P(s)+\int_0^s \frac{D(\tau)}{N(\tau)}d\tau\right]\frac{E(s)}{P(s)}ds \\ &=V(t)+\int_0^t [D(s)-E(s)]\,ds \quad \text{is a martingale.}\end{aligned} \tag{3.2}$$

To solve (3.2) for $V(t)$, we require some assumptions only on

$$C(t)\equiv D(t)-E(t), \qquad t\in\Upsilon, \tag{3.3}$$

and a suitable boundary condition on $V(t)$. It is important to note that $C(t)$ is the net cash outflow in period t from the firm to the society. Also, note that $Y(t)$ is the sum of the firm's value at time t and all its net cash outflows in the interval $[0,t]$. Furthermore, valuation of the stream of these cash flows can be used as the basis of the valuation of the firm. For a meaningful valuation, we impose the boundary condition

$$\lim_{t\to\infty} V(t)=0. \tag{3.4}$$

It is now possible to define two more valuation systems:

Cash flow system ≡ Equations (2.2), (3.2), (3.4), (1.6),

V-arbitrage system ≡ Equations (2.2), (2.4), (3.4), (1.6).

Like the definition of the P-arbitrage system, the definition of the cash flow system is based on the martingale property (3.2). The definition of the V-arbitrage system is inspired by the boundary condition (3.4), i.e., the boundary condition (3.4) is on $V(t)$ rather than a condition being imposed on $P(t)$ in the P-arbitrage system.

THEOREM 2. *$\{P(t), N(t), V(t),\ t\in\Upsilon\}$ is a positive solution of the cash flow system if, and only if, it is a positive solution of the V-arbitrage system.*

Proof. Since (3.2) was derived from (2.2), (2.4), (1.6), it is clear that any solution of the V-arbitrage system will be a solution of the cash flow system.

To prove the converse, let $\{P(t), N(t), V(t),\ t\in\Upsilon\}$ be a solution of the cash flow system. From (3.2), $Y(t)$ is a martingale. From (3.1) and the fact that $N(t)>0$, we conclude that $M(t)$ is a martingale. This completes the proof.

We have now shown that the cash flow system and V-arbitrage system are equivalent. In Theorem 3, we derive a formula for $N(t)$ in terms of $E(\tau)$ and $V(\tau)$, $0\le\tau\le t$. In Theorem 4, we specify fairly minimal assumptions on the firm $\{D(t), E(t),\ t\in\Upsilon\}$ and obtain a unique solution for the cash flow system.

THEOREM 3. *For any positive solution,*

$$N(t)=N_0\exp\left[\int_0^t\frac{E(\tau)}{V(\tau)}d\tau\right],\qquad t\in\Upsilon. \tag{3.5}$$

Proof. From (2.2) and (1.6), we have

$$\frac{N'(t)}{N(t)}=\frac{E(t)}{V(t)},\qquad N(0)=N_0, \tag{3.6}$$

which integrates to (3.5).

THEOREM 4. *Assume*

$$D(t)\ge 0,\qquad t\in\Upsilon, \tag{A1}$$

$$\int_0^\infty\left(D(\tau)+|E(\tau)|\right)d\tau\in\mathcal{L}_1(\Omega,\mathcal{F},\pi), \tag{A2}$$

$$\mathcal{E}_t\int_t^\infty\left[D(\tau)-E(\tau)\right]d\tau>0,\qquad t\in\Upsilon. \tag{A3}$$

Then the cash flow system has a unique solution $\{P(t), N(t), V(t),\ t\in\Upsilon\}$ given by

$$V(t)=\mathcal{E}_t\int_t^\infty\left[D(\tau)-E(\tau)\right]d\tau\equiv\mathcal{E}_tU(t), \tag{3.7}$$

$$N(t)=N_0\exp\left[\int_0^t\frac{E(\tau)}{V(\tau)}d\tau\right]=N_0\exp\left[\int_0^t\frac{E(\tau)}{\mathcal{E}_\tau\int_\tau^\infty\left[D(s)-E(s)\right]ds}d\tau\right], \tag{3.8}$$

$$P(t)=\frac{V(t)}{N(t)}. \tag{3.9}$$

Solution (3.7)–(3.9) is a positive solution and shall henceforth be termed the *financial solution*.

Note that we have defined

$$U(t)\equiv\int_t^{\infty}[D(\tau)-E(\tau)]\,d\tau. \tag{3.10}$$

Proof. We need only to derive (3.7) and show that it is unique and positive. Because of (3.4) and (A2), we can take the limit of (3.2) as $t\to\infty$ and write

$$Y(\infty)=\int_0^{\infty}[D(s)-E(s)]\,ds. \tag{3.11}$$

But $Y(t)$ is a martingale. Therefore, $\mathcal{E}_t Y(\infty)=Y(t)$; thus (3.7) and uniqueness follow. That $V(t)$ of (3.7) is positive follows from (A3).

A corollary of Theorem 4 deals with the valuation of a firm without a corporate structure. From (3.3) and (3.7), we can see that such a firm can be valued under much weaker conditions, since the solution consists only of $V(t)$, $t\in\Upsilon$. From (A1)–(A3), we can derive these weaker conditions to be

$$\int_0^{\infty}|C(\tau)|\,d\tau\in\mathcal{L}_1(\Omega,\mathcal{F},\pi),\qquad \mathcal{E}_t\int_t^{\infty}C(\tau)\,d\tau>0,\qquad t\in\Upsilon.$$

Before we proceed any further, it would be instructive to solve an example explicitly.

Example 1. In this example, a firm pays out total dividends at a constant rate of \$1 per unit time until time ω, which is exponentially distributed. For $t\geq\omega$, the firm pays out total dividends at the variable rate of $1+(t-\omega)$ dollars per unit time. The discount rate is $\rho(t)\equiv 1$, $t\in\Upsilon$. The main purpose is to explicitly obtain the solution $\{P(t),N(t),V(t),\ t\in\Upsilon\}$. Another purpose is to demonstrate that the primitive price defined by the ratio $U(t)/N(t)$ need not equal the primitive price defined by the integral $\int_t^{\infty}(D(\tau)/N(\tau))\,d\tau$. However, the conditional expectations with respect to information at t of these primitive prices are precisely the prices $P(t)$. We need the following notation to state the example.

Let $\Omega=[0,\infty)$ and $\mathcal{F}_t=\sigma$-algebra generated by Borel sets in $[0,t]$. Let $\mathcal{F}=\bigcup_{t>0}\mathcal{F}_t$. Let the probability measure $\pi(d\omega)=e^{-\omega}d\omega$. Let

$$D(t,\omega)=\begin{cases}e^{-t}, & t<\omega,\\ (t-\omega+1)e^{-t}, & t\geq\omega,\end{cases}\qquad E(t,\omega)=e^{-t},\quad N_0=1.$$

Then, using (3.7)–(3.9), we obtain the following financial solution:

$$V(t,\omega)=e^{-t/2},\quad N(t,\omega)=e^{2t},\quad P(t,\omega)=e^{-3t/2},\qquad t<\omega;$$
$$V(t,\omega)=(t+1-\omega)e^{-t},\quad N(t,\omega)=(t+1-\omega)e^{2\omega},\quad P(t,\omega)=e^{-t-2\omega},\qquad t\geq\omega.$$

Moreover, we can compute the primitive

$$\frac{U(t,\omega)}{N(t,\omega)}=\begin{cases}e^{-\omega-2t}, & t<\omega,\\ e^{-t-2\omega}, & t\geq\omega,\end{cases}$$

which is not equal to the primitive

$$\int_t^{\infty}\frac{D(\tau,\omega)}{N(\tau,\omega)}=\begin{cases}\dfrac{e^{-3t}+2e^{-3\omega}}{3}, & t<\omega,\\ e^{-t-2\omega} & t\geq\omega.\end{cases}$$

Nevertheless, their conditional expectations with respect to $\mathcal{F}_t$ is the price $P(t,\omega)$.

So far, we have established an equivalence between the dividend system and the P-arbitrage system. Henceforth, these two terms will be used interchangeably. We have also established an equivalence between the cash flow system and the V-arbitrage system and have obtained a unique positive solution, termed the financial solution, for the two systems under conditions (A1)–(A3). Henceforth, the terms for these two systems will be used interchangeably.

In the next two theorems, we show that the dividend system (P-arbitrage system) and the cash flow system (V-arbitrage system) are not equivalent under (A1)–(A3). In other words, the financial solution is not a solution of the dividend system, in general. This will be shown by showing that, in general, $P(t)$ of (3.7)–(3.9) does not approach zero as t approaches infinity. Moreover, when the financial solution does not solve the dividend system, then the dividend system does not have any positive solution.

In Theorem 5, we show that any positive solution of the dividend system is a positive solution of the cash flow system, i.e., a financial solution. In Theorem 6, we obtain a precise condition under which these two systems are equivalent. These results are surprising and quite remarkable.

THEOREM 5. *Assume* (A1)–(A3). *Every positive solution* (*whenever it exists*) *of the dividend system is a positive solution of the cash flow system*, *i.e. the financial solution.*

Proof. Let $\{P(t), N(t), V(t), t\in\mathfrak{T}\}$ be a solution of the dividend system. From the derivation of (3.2), we know that the solution satisfies (3.2). From the martingale property of $Y(t)$, therefore, we have

$$Y(t)=\mathfrak{E}_tV(\theta)+\mathfrak{E}_t\int_0^\theta[D(s)-E(s)]\,ds \quad \text{for } \theta\geq t. \tag{3.12}$$

From (3.11), we can derive

$$\mathfrak{E}_tV(\theta_1)-\mathfrak{E}_tV(\theta_2)=\mathfrak{E}_t\int_{\theta_1}^{\theta_2}[D(s)-E(s)]\,ds \tag{3.13}$$

for $\theta_1,\theta_2\geq t$, implying that $\mathfrak{E}_tV(\theta)$ is a Cauchy sequence in θ. Taking the limit as $\theta\to\infty$ gives

$$Y(t)=\beta(t)+\mathfrak{E}_t\int_0^\infty[D(s)-E(s)]\,ds=\beta(t)+\mathfrak{E}_tU(0), \tag{3.14}$$

where

$$\beta(t)\equiv\lim_{\theta\to\infty}\mathfrak{E}_tV(\theta)=\mathfrak{E}_t\lim_{\theta\to\infty}V(\theta). \tag{3.15}$$

But for $s\leq t$ we have

$$\mathfrak{E}_s\beta(t)=\mathfrak{E}_s\lim_{\theta\to\infty}\mathfrak{E}_tV(\theta)=\lim_{\theta\to\infty}\mathfrak{E}_s\mathfrak{E}_tV(\theta)=\lim_{\theta\to\infty}\mathfrak{E}_sV(\theta)=\beta(s), \tag{3.16}$$

and therefore $\beta(t)$ is a martingale. Furthermore, since $V(\theta)>0$ and is in $\mathfrak{L}_1(\Omega,\mathfrak{F},\pi)$, $\beta(t)$ is a nonnegative $\mathfrak{L}_1$-martingale. Thus

$$\lim_{t\to\infty}\beta(t)=\beta\geq 0 \tag{3.17}$$

and

$$\beta(t)=\mathfrak{E}_t\beta, \tag{3.18}$$

and (3.13) can be rewritten as

$$Y(t)=\mathfrak{E}_t[\beta+U(0)]. \tag{3.19}$$

From (3.2) and (3.19), we obtain

$$V(t)=\mathcal{E}_t[\beta+U(t)]=\mathcal{E}_t\left[\beta+\int_t^{\infty}[D(s)-E(s)]\,ds\right]. \tag{3.20}$$

We shall now show that (3.20) with (3.5) and (1.6) will solve the dividend system only if $\beta=0$. The proof is by contradiction, i.e., by showing that

$$\lim_{t\to\infty}P(t)\equiv\lim_{t\to\infty}\frac{V(t)}{N(t)}\neq 0.$$

Suppose then that $\beta>0$. Then, from (3.5) and (3.20), we have

$$N(t)=N_0\exp\left[\int_0^t\frac{E(\tau)\,d\tau}{\mathcal{E}_\tau\beta+\mathcal{E}_\tau\int_\tau^\infty U(s)\,ds}\right]. \tag{3.21}$$

From (A3) and $\beta>0$, it follows that we can choose a large T so that for $\tau\geq T$

$$-\frac{|E(\tau)|}{\beta}\leq\frac{E(\tau)}{\mathcal{E}_\tau\beta+\mathcal{E}_\tau\int_\tau^\infty U(s)\,ds}\leq\frac{|E(\tau)|}{\beta}. \tag{3.22}$$

Using (A2), we can conclude that

$$-\infty<\int_0^\infty\frac{E(\tau)\,d\tau}{\mathcal{E}_\tau\beta+\mathcal{E}_\tau\int_\tau^\infty U(s)\,ds}<\infty, \tag{3.23}$$

implying that

$$0<\lim_{t\to\infty}N(t)=N_\infty<\infty. \tag{3.24}$$

Then

$$\lim_{t\to\infty}P(t)=\lim_{t\to\infty}\frac{V(t)}{N(t)}=\frac{\lim V(t)}{\lim N(t)}=\frac{\beta}{N_\infty}>0. \tag{3.25}$$

Thus $\{P(t),\ t\in\mathrm{T}\}$ does not solve (2.3) and, therefore, $\{P(t), N(t), V(t),\ t\in\mathrm{T}\}$ is not a solution of the dividend system. This contradiction implies that $\beta=0$. From (3.20), we now have

$$V(t)=\mathcal{E}_tU(t)=\mathcal{E}_t\int_t^\infty[D(\tau)-E(\tau)]\,d\tau, \tag{3.26}$$

which is the same as (3.7).

This completes the proof of Theorem 5. The converse of this theorem does not hold in general. In Theorem 6, we obtain the necessary and the sufficient conditions under which the converse holds.

THEOREM 6. *Assume* (A1)–(A3). *The financial solution is the unique positive solution of the dividend system if and only if*

$$\ln F(\infty)\equiv\int_0^\infty\frac{D(\tau)}{V(\tau)}\,d\tau=\int_0^\infty\frac{D(\tau)\,d\tau}{\mathcal{E}_\tau\int_\tau^\infty[D(s)-E(s)]\,ds}=\infty. \tag{B1}$$

Also a sufficient, but not a necessary, condition for the financial solution to be the unique solution of the dividend system is

$$\inf_{t\in\mathrm{T}}\int_0^t\frac{E(\tau)\,d\tau}{\mathcal{E}_\tau\int_\tau^\infty[D(s)-E(s)]\,ds}>-\infty. \tag{B2}$$

Proof. To prove this theorem, we need to examine the limiting behavior of $P(t)=V(t)/N(t)$ of (3.7)–(3.9). Since $\lim V(t)=0$, it is obvious from (3.8) and (3.9) that $\lim P(t)=0$ if $\inf N(t)>0$, i.e., (B2) holds.

To obtain (B1), however, the price formula (3.9) is not suitable. So we derive an alternate formula for $P(t)$. For this we need to define processes $F(t)$ and $R(t)$, which will be interpreted economically in §4. We let

$$F(t)\equiv\exp\left[\int_0^t\frac{D(s)}{V(s)}ds\right], \tag{3.27}$$

where we note that $D(s)/V(s)$ represents the *dividend yield* at time s. Since $F(t)$ is of bounded variation and $M(t)$ in (2.4) is a martingale, it follows [9, Thm. 1.2.8, p. 26] that

(3.28)

$$R(t)=M(0)F(0)+\int_0^t F(s)\,dM(s)=M(t)F(t)-\int_0^t M(s)\,dF(s)\quad\text{is a martingale.}$$

Substituting for $M(t)$ and $F(t)$ and simplifying gives

$$R(t)=P(t)F(t)=P(t)\exp\left[\int_0^t\frac{D(s)}{V(s)}ds\right]\quad\text{is a martingale.} \tag{3.29}$$

Moreover, it is obvious that $R(t)$ is a positive $\mathcal{L}_1$-martingale, and so it converges. Thus

$$R(\infty)=\lim_{t\to\infty}P(t)\exp\left[\int_0^t\frac{D(s)}{V(s)}ds\right], \tag{3.30}$$

which, it should be noted, is uniquely defined and $\mathcal{F}_\infty$-measurable. Moreover, $\mathcal{E}_t R(\infty)=R(t)$ and $\mathcal{E}_0 R(\infty)=R(0)=P(0)$. From (3.29), therefore,

$$P(t)=\mathcal{E}_t[R(\infty)]\cdot\exp\left[-\int_0^t\frac{D(s)}{V(s)}ds\right]. \tag{3.31}$$

It follows that

$$\lim_{t\to\infty}P(t)=0\Leftrightarrow\int_0^\infty\frac{D(s)}{V(s)}ds\equiv\ln F(\infty)=\infty,\quad\text{i.e., (B1) holds.} \tag{3.32}$$

This completes the proof.

4. Financial interpretations and discussion of results. In this section, first we provide the financial interpretation of formula (3.7) for the value of the firm and the price formula (3.31). Then, we discuss the significance of Theorem 6 and condition (B1).

The interpretation of (3.7) is that the value of a firm at time t is the conditional expectation of the total discounted future net cash outflow from the firm to the society, given the information available by time t. Formula (3.7) can also be interpreted as the expected present value of the total future dividends accruing to the stockholders of record at time t conditioned on $\mathcal{F}_t$. The integral of the first term in the integrand represents the expected total present value of dividends issued by the firm in the interval $[t,\infty)$ given $\mathcal{F}_t$. A portion of the total future dividends is obviously going to stock issued in the interval (t,∞). In the absence of arbitrage possibilities, the expected value of this portion, conditional on $\mathcal{F}_t$, must equal $\mathcal{E}_t\int_t^\infty E(\tau)\,d\tau$. Clearly, the residual represented by the right-hand side of (3.7), which belongs to the stockholders of record t, can now be interpreted as the present value of the firm at time t given $\mathcal{F}_t$.

Moreover, for $s<t$, we can define $\mathcal{E}_sV(t)$ to be the value of the firm at time t given the information up to time s. It should be obvious that $\mathcal{E}_tP(s)=P(s)$ for $s\leq t$.

For the interpretation of the price formula (3.31), we first note that

$$\frac{D(t)}{V(t)}=\frac{D(t)/N(t)}{P(t)} \tag{4.1}$$

is the *dividend yield* at time t. It can also be interpreted as the *share growth rate* under a *dividend reinvestment plan*. Now we rewrite (4.1) as

$$P(t)\exp\left[\int_0^t\frac{D(s)}{V(s)}ds\right]=\mathcal{E}_t[R(\infty)], \tag{4.2}$$

and draw the following diagram:

$$\begin{tikzcd}
\text{time } 0: & R(\infty) \arrow[d, "\text{More information becomes available as time progresses}"] \arrow[r, "{\mathcal{E}_0[R(\infty)]=P(0)}"] & P(0) \arrow[d, "\text{No. of shares grows under a dividend reinvestment plan}"] \\
\text{time } t: & \mathcal{E}_t[R(\infty)] \arrow[r, "\text{equal by (4.2)}"] & P(t)\exp\left[\int_0^t\frac{D(s)}{V(s)}ds\right].
\end{tikzcd}$$

At time zero, a risk-neutral investor is indifferent between the random amount $R(\infty)$ and one share of the firm valued at $P(0)$. At time t, the value of the random amount is $\mathcal{E}_t[R(\infty)]$, based on the information available by time t. On the other hand, the number of shares has grown by time t to $F(t)=\exp[\int_0^t(D(s)/V(s))\,ds]$ by having the one share at time 0 in a dividend reinvestment plan and, therefore, the value of the resulting portfolio is $R(t)=P(t)F(t)=P(t)\exp[\int_0^t(D(s)/V(s))\,ds]$. Equation (4.2) says that the risk-neutral investor is also indifferent between the choices at time t arising out of the two indifferent courses of action at time zero.

We also remark that the above interpretation remains valid in the economy with risk-averse agents provided the conditional expectation is taken with respect to an appropriate martingale measure; see §5.

We shall now discuss condition (B1). We know from Theorem 6 that when (B1) holds, the financial price $P(t)$ also solves the dividend system and, therefore, it is equal to its future *dividend content* (i.e., the present value of the future per share dividends) as in (2.3).

When (B1) does not hold, the financial price $P(t)$ does not satisfy the dividend system. In fact, in this case, the dividend system has no solution. Note, however, that since $P(t)$ solves the V-arbitrage system, we can express it as

$$P(t)=\mathcal{E}_t\left[\alpha+\int_t^\infty\frac{D(s)}{N(s)}ds\right], \tag{4.3}$$

where

$$\alpha=\lim_{t\to\infty}P(t) \tag{4.4}$$

and $\alpha>0$ when (B1) does not hold. (4.3) can be related to (2.3). It states that $P(t)$ of a share exceeds its future dividend content by an amount $\mathcal{E}_t\alpha$.

It is important to note that in the case when (B1) does not hold, the firm is unable to pay out all of its full value in the form of dividends. It does, however, pay out its full value by retiring a part of its value by repurchasing its own stock. Note that $\lim_{t\to\infty} V(t)=0$, and since $\lim_{t\to\infty} P(t)=\alpha>0$ in this case, it is obvious that $\lim_{t\to\infty} N(t)=0$. In the limit there are no remaining shares outstanding. In a way, the firm buys back all its shares "eventually". Thus, even though the price of a share does not equal its future dividend content, the case when (B1) does not hold should not be considered pathological.

Thus (B1) can be interpreted as the precise restriction under which the firm must operate to be valued as a dividend system. In words, (B1) states that the sum of the dividend yields, $\ln F(\infty)$, cannot be too small, i.e., the firm, in the long run, must issue sufficient dividends in relation to its value.

It should be noted that condition (B2) prevents $N(t)$ from approaching zero, with the consequence that $\lim V(t)=0 \Rightarrow \lim P(t)=0$. Condition (B2), therefore, is sufficient for the financial solution to be a solution of the dividend system. It is obvious that (B2) implies (B1).

We shall now present two deterministic examples in which $N(t)\to 0$, but (B1) holds only in one of them.

Example 2. $\ln F(\infty)<\infty$.

$$D(t)=\left(\tfrac{1}{2}\right)^{1+2t}, \quad E(t)=\left(\tfrac{1}{2}\right)^{1+2t}-\left(\tfrac{1}{2}\right)^{1+t}, \quad t\geq 0, \quad N(0)=1,$$

$$V(t)=\frac{\left(\frac{1}{2}\right)^{1+t}}{\ln 2}, \qquad N(t)=\left(\tfrac{1}{2}\right)^{t}\exp\left[1-\left(\tfrac{1}{2}\right)^{t}\right]\to 0,$$

$$P(t)=(1/2\ln 2)\exp\left[\left(\tfrac{1}{2}\right)^{t}-1\right], \qquad \alpha=\lim P(t)=1/2e\ln 2>0.$$

Example 3. $\ln F(\infty)=\infty$.

$$D(t)=\left(\tfrac{1}{2}\right)\left(\tfrac{1}{3}\right)^{t}, \quad E(t)=-\left(\tfrac{1}{2}\right)\left(\tfrac{1}{3}\right)^{1+t}, \quad N(0)=1,$$

$$V(t)=\left(\tfrac{1}{3}\right)^{t+1}/\ln 3, \; N(t)=\left(\tfrac{1}{3}\right)^{t/2}\to 0,$$

$$P(t)=\left(\tfrac{1}{3}\right)^{1+t/2}/\ln 3\to 0.$$

In both these examples (B2) does not hold. In Example 2, (B1) does not hold and $\alpha=1/2e\ln 2>0$. In Example 3, (B1) holds and $\alpha=0$. These examples clearly indicate the significance of (B1) and that condition (B2) is too strong.

In relating our results to those of MM [4], we proceed as follows. We have shown that the class of firms that can be valued (i.e., defined by (A1)–(A3)) can be partitioned into two subclasses. The first subclass defined by condition (B1) consists of the firms that can be valued by either the dividend approach or the cash flow approach. That is, in valuing these firms, the two approaches are equivalent. The other subclass defined by the negation of condition (B1) consists of the firms that cannot be valued by the dividend approach. The dividend approach is simply meaningless for this subclass. Thus, for this subclass, the question of whether the two approaches are equivalent cannot even arise.

To deal with the question whether the dividend policy is irrelevant, we must understand that the question applies only to the valuation $V(t)$ of the firm and not to the share price $P(t)$ and the number of outstanding shares $N(t)$. The formula for $V(t)$ is obtained in (3.7). Observe that $V(t)$ depends only on the cash flow trajectory $C(t)=D(t)-E(t)$, $t\in\Upsilon$. It does not depend on how $C(t)$ is divided between $D(t)$ and

$-E(t)$. Therefore, in the sense of MM, the dividend policy is *irrelevant* for the firm's valuation.[2] In other words, if $\{D_1(t), E_1(t),\ t\in\mathfrak{T}\}$ and $\{D_2(t), E_2(t),\ t\in\mathfrak{T}\}$ are two different firms such that $D_1(t)-E_1(t)=D_2(t)-E_2(t)$, $t\in\mathfrak{T}$, then (3.7) gives $V_1(t)=V_2(t)$, $t\in\mathfrak{T}$. Of course, it is possible for these firms to belong to the two different subclasses defined above. In that case, the dividend approach will be meaningful for one firm and meaningless for the other. Even so, the dividend policy will be irrelevant in the MM sense.

It is extremely important to emphasize that the above discussion concerning the irrelevancy of dividends assumes that the underlying family $\{\mathfrak{F}_t,\ t\in\mathfrak{T}\}$ of sub-σ-algebras is given a priori.

If we assume, on the other hand, that $\mathfrak{F}_t$ is the sub-σ-algebra generated by $\{\bar{D}(\tau), \bar{E}(\tau), \rho(\tau),\ 0\leq\tau\leq t\}$, then the dividend policy will no longer be irrelevant. This should be obvious, because in the new definition of $\mathfrak{F}_t$, the dividend policy contains relevant information for the purpose of valuation.

5. Extension to more general economies. So far, we have discussed the valuation problem of a firm in an economy with only risk-neutral agents. We now consider the case where agents need not be risk-neutral. In such an economy, if there are no arbitrage opportunities in a market and if the markets are complete, there must exist a unique positive linear functional that can value risky streams [5], [6]. Moreover, if the discount rate process $\{\rho(t),\ t\in\mathfrak{T}\}$ is interpreted as the process of spot interest rates in the economy, then a valuation functional need only be defined on the space $\mathfrak{L}_1(\Omega,\mathfrak{F},\pi)$. Let $\psi(\cdot)$ denote this valuation functional. It can be easily shown (e.g., see [6]) that this valuation functional has a representation in terms of an expectation operator E^{μ} defined on the space $\mathfrak{L}_1(\Omega,\mathfrak{F},\pi)$. The probability measure μ, which is absolutely continuous with respect to π, is uniquely determined by ψ. Thus,

$$V(0)=\psi[U(0)]=\mathcal{E}^{\mu}U(0),$$

where $U(0)$ is defined in (3.10). Moreover, we can define the conditional expectation $\mathcal{E}_t^{\mu}$: $\mathfrak{L}_1(\Omega,\mathfrak{F},\pi)\rightarrow\mathfrak{L}_1(\Omega,\mathfrak{F},\mu)$ in terms of which we can write

$$V(t)=\mathcal{E}_t^{\mu}U(t)=\mathcal{E}^{\mu}\left[U(t)|\mathfrak{F}_t\right].$$

Comparing this to (3.7), it is obvious that the entire analysis of the paper remains valid if $\mathcal{E}_t$ is replaced by $\mathcal{E}_t^{\mu}$. Note also that (3.2) is replaced by

$$V(t)+\int_0^t[D(s)-E(s)]\,ds \quad \text{is a } \mathfrak{L}_1(\Omega,\mathfrak{F},\mu)\text{-martingale.}$$

Thus, the probability measure μ is also known as an (equivalent) *martingale measure* [2].

It should be noted that we do not address the issue of the conditions under which a market is complete [2]. Nor do we construct the valuation functional ψ. Our scope here has been limited.

In the same limited way, we can discuss an economy in which agents have different expectations regarding the future. This requires the definition of the appropriate family $\{\mathfrak{F}_t,\ t\geq0\}$ of sub-σ-algebras. We proceed as follows [6]. Let J denote the set of agents. Let the nondecreasing family $\{\mathfrak{F}_t^j,\ t\geq0\}$ of sub-σ-algebras represent the information set of agent $j\in J$. Assume that processes $\{\bar{D}(t),\ t\geq0\}$, $\{\bar{E}(t),\ t\geq0\}$ and $\{\rho(t),\ t\geq0\}$ are

[2] It should be noted that $C(t)=X(t)-I(t)$, $t\in\mathfrak{T}$, where $\{X(t),\ t\in\mathfrak{T}\}$ denotes the total earnings process and $\{I(t),\ t\in\mathfrak{T}\}$ denotes the total investment process [1], [4]. Once again, it is immaterial for the firm's valuation how net cash outflow $C(t)$ is divided between $X(t)$ and $-I(t)$.

adapted to $\{\mathcal{F}_t^j, t \geq 0\}$ for every $j \in J$. We can now define the appropriate family $\{\mathcal{F}_t, t \geq 0\}$ by

$$\mathcal{F}_t = \bigcap_{j \in J} \mathcal{F}_t^j, \qquad t \geq 0.$$

With this definition for $\mathcal{F}_t$, the measurability requirement for the solution variables in (1.7) is now defined. A martingale measure μ corresponding to a given valuation functional ψ can be obtained as before.

Needless to say, the issue of the market completeness and the construction of the valuation functional for a fairly general economy is a very difficult problem. Even more difficult is the general equilibrium problem where the interest rate process $\{\rho(t), t \geq 0\}$ is not exogenous, but is to be determined within the model.

6. Concluding remarks. In this paper, we have developed conditions under which a firm may be valued. We have also obtained precise conditions under which the different valuation systems specified by MM[4] are equivalent in a fairly general stochastic environment. Our paper, therefore, represents both mathematical and stochastic extension of the Miller–Modigliani theory.

We have also developed a mathematical framework within which several other extensions of our model may be easily addressed. These would include incorporation of debt-financing and the valuation of an almost surely finite horizon firm. The latter extension has already been dealt with in a discrete-time framework [7].

Finally, in solving the valuation problem of a firm in a general stochastic environment, we have communicated an important economic problem to mathematicians and, we hope, at the same time enriched the economists' understanding of martingale methods.

Acknowledgment. Thanks are due to S. R. S. Varadhan for his helpful suggestions.

REFERENCES

[1] N. Derzko and S. P. Sethi, *General solution of price-dividend integral equation*, this Journal, 13 (1982), pp. 106–111.

[2] J. M. Harrison and S. R. Pliska, *Martingales and stochastic integrals in the theory of continuous trading*, Stochastic Processes Appl., 11 (1981), pp. 215–260.

[3] R. S. Liptser and A. N. Shiryaev, *Statistics on Random Processes* I: *General Theory*, Springer-Verlag, New York, 1977.

[4] M. H. Miller and F. Modigliani, *Dividend policy, growth, and the valuation of shares*, J. Business, 34 (1961), pp. 411–433.

[5] S. A. Ross, *A simple approach to valuation of risky streams*, J. Business, 5 (1978), pp. 453–474.

[6] S. P. Sethi, *A further simplified approach to valuation of risky streams*, Working Paper, Univ. of Toronto, Toronto, Ontario, Canada, 1982.

[7] S. P. Sethi, N. Derzko and J. Lehoczky, *Mathematical analysis and stochastic extensions of the Miller–Modigliani theory*, Working Paper, Univ. of Toronto, Toronto, Ontario, Canada, March 1982.

[8] ———, *Mathematical analysis of the Miller–Modigliani theory*, Oper. Res. Lett., 1 (1982), pp. 148–152.

[9] D. W. S. Stroock and S. R. S. Varadhan, *Multi-Dimensional Diffusion Processes*, Springer-Verlag, New York, 1979.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics
006

QUENCHING IN TIME-DELAY SYSTEMS: A SUMMARY AND A COUNTEREXAMPLE*

RAY REDHEFFER† AND REINHARD REDLINGER‡

Abstract. This paper combines the salient features of two separate investigations. The first pertains to existence theorems for parabolic-functional systems and is based upon the Karlsruhe dissertation of the second author. The relevance of these results is that they permit a time delay and they do not require the functions to be Lipschitzian. Thus the equation for the unknown function $u(t,x)$ can contain such terms as $u(t-\mu(t),x)$, $|u(t,x)|^{\lambda}$ and $|\text{grad}\, u(t,x)|^{\lambda}$, all of which play a role in our analysis.

The second line of thought pertains to the theory of quenching in time-delay systems as developed since 1976. (Here the word "quenching" has its usual English significance and means that the solution vanishes after a finite time t^*.) As a rule one must have a singularity such as $|u|^{\lambda}$ to induce quenching and, in the absence of time-delay, the proof of quenching then involves little more than a comparison argument of standard type. But if the equation contains such terms as $u(t-\mu(t),x)$, the question whether the solution does or does not quench hinges on a delicate and nonobvious relationship between the *memory-function* μ and the *control parameter* λ. Study of this relationship forms the essence of the memory-quenching problem.

Here we give a simplified exposition of the two topics above, without proofs, and we use the existence theorems to show that the quenching theory applies (nonvacuously) to equations as well as inequalities. With this as background, we construct a counterexample to show that the quenching theorems remain sharp when applied to equations with Dirichlet boundary conditions. The example settles a question that has been open in the theory of quenching since its inception.

1. Introduction. For broad classes of parabolic operators P the solutions of the inequality

$$(\operatorname{sgn} u)Pu+|u|^{\lambda}\leq 0$$

quench; that is, they vanish after a time t^* which depends on the initial condition and on the constant λ, $0\leq\lambda<1$. The proof of this well-known fact involves little more than construction of a solution $u(t,x)=\rho(t)$ of the opposite inequality which itself quenches, together with an argument of Nagumo–Westphal type to conclude that $|u(t,x)|\leq\rho(t)$. If $\rho(t)=0$ for $t\geq t^*$, the same is true of $u(t,x)$.

Let us apply this familiar line of thought when the parabolic operator P incorporates a functional as part of its structure. The strength of this functional is measured by a *memory-function* $\mu(t)\geq 0$; roughly speaking, the functional has memory μ if it can be assessed by

$$\sup_{\xi}\ \sup_{t-\mu(t)\leq\tau\leq t}|u(\tau,\xi)|.$$

The question is: If the initial conditions and λ are such that the solutions quench when P is purely parabolic, what conditions on the memory-function μ ensure that the solutions quench also in the parabolic-functional case? Although this bears a superficial resemblance to the familiar phenomenon outlined above, it is in fact an entirely different question which, so far as we know, was not asked before the first author started work on it in 1976. To underline the difference between the *quenching problem*

*Received by the editors June 11, 1982, and in revised form July 14, 1983.

† Department of Mathematics, University of California, Los Angeles, California 90024. The research of this author was supported in part by the University of Karlsruhe under auspices of the Deutsche Forschungsgemeinschaft and in part by the National Science Foundation under grant MCS79-03544.

‡ Mathematisches Institut I, Universität Karlsruhe, West Germany.

and the *memory-quenching problem* a brief bibliography of the former is given in [3], [4], [5], [6], [7], [8], [11], [12], [17]. These references embrace ordinary differential equations as well as partial differential equations of both elliptic and parabolic type, but they do not address the question of time-lag, which is the feature of the principal interest in [18].

One aspect which these investigations have in common is the presence of a suitable singularity, in virtually all cases a function of the form $C(u)=|u|^{\lambda}\operatorname{sgn} u$, $0\leq\lambda<1$. Such a term is refered to in [18] as the *control*. At first glance this terminology is perhaps unusual, but it is suggested by the fact that $C(u)$ always tends to push u toward the equilibrium position $u=0$. The chief difference between $C(u)$ and a true *control* in the sense of control theory is that the switching locus of the latter (the locus where $C(u)$ changes sign) is subject to outside manipulation, whereas here it is constrained to be the same as the locus in which u itself changes sign. The special case $C(u)=\operatorname{sgn} u$ obtained when $\lambda=0$ is effectively a bang-bang control, subject to the above proviso regarding the switching locus. As shown in [18] the theory of quenching for this case is vastly simpler than in the general case and involves little more than the obvious condition $t-\mu(t)\rightarrow\infty$ as $t\rightarrow\infty$.

After these preliminary remarks we can describe the purpose of this paper, which is threefold. The theory in [18] is developed in a degree of generality which makes it difficult to read and the problem is compounded by the fact that [18] is written in an unattractive style which is highly condensed. Our first objective is to present a simplified version, without proofs, in which the main features are more readily discerned.

Since [18] is developed within the context of differential inequalities, questions of existence do not play a significant role. This approach has the advantage of allowing a good deal of generality in the parabolic-functional operator P, but at the same time it leaves open the question whether the theorems might hold under weaker hypotheses on μ for the corresponding equations. (Recall, for instance, that the Harnack inequalities hold for equations, or for two-sided inequalities, but not for the one-sided inequalities that form the basis for the theory of subharmonic functions.) A comprehensive existence theory for parabolic-functional equations is given in [23]. As our second objective, we describe a class of operators P that satisfy the hypotheses of both theories, quenching and existence, and within this class we show that the use of differential equations rather than inequalities has no significant effect on the structure of the quenching problem. Reference [23] is of broad applicability but is somewhat long, and we present a simplified version, free of proofs, which is more readily accessible. Of course the literature on existence theory for parabolic problems is extremely comprehensive, and references [1], [2], [4] [9], [10], [13], [14], [15], [16], [25], [26], [27], [28] are only a small sample. However, no prior work known to us admits both a time-delay and a class of nonlinearities which are non-Lipschitzian in u and $\operatorname{grad} u$, as do the results [23]. The full force of this generality is used in meeting the third objective of this paper, discussed next.

Our third, and principal, objective is to answer a question that has been open ever since the research leading to [18] was initiated about seven years ago. The question pertains to the construction of counterexamples with a view to showing that the *sufficient* conditions for quenching are also *necessary*; in other words, that the quenching theorems are sharp. Only if this is so can it be asserted that the memory-function has been correctly characterized. The theorems in [18] are developed in a context of general boundary conditions, in which the Neumann condition $u_{\nu}=0$ is allowed as a special case. When this condition is imposed the construction of examples is trivial, because one can choose u to be a function of t alone. Thus it is seen that the hypotheses in [18] are in fact sharp when the conclusions are asserted in the degree of generality there given. But this observation sheds no light on the most important case of all—the case of the Dirichlet boundary condition, $u=0$. It is with a view to filling this gap that we have undertaken the present investigation. The relevant result, which has no overlap with [18] or [23], is presented in §7. It is a source of satisfaction to us that the equation in the counterexample, although highly nonlinear in $\operatorname{grad} u$ as well as u, nevertheless satisfies the hypotheses of the theorems in [23]. Thus the example shows that the quenching theorems are sharp not only when Dirichlet conditions are imposed, but also when the inequality is replaced by an equation.

2. Notation. Throughout this paper points of R^{n+1} are written in the form (t,x) with $t\in R$ and $x\in R^n$, Ω is a bounded domain in R^n with $C^{1+\alpha}$ boundary $\partial\Omega$, and

$$G=(0,\infty)\times\Omega,\quad \Gamma_0=[0]\times\overline{\Omega},\quad \Gamma_1=(0,\infty)\times\partial\Omega.$$

We refer to G as the parabolic interior and to $\Gamma=\Gamma_0\cup\Gamma_1$ as the parabolic boundary. Derivatives are written u_t, u_x, u_{xx}, where u_t is a left-hand derivative, u_x is the gradient and u_{xx} is the Hessian. These expressions denote the value of the function at (t,x), while $u(\cdot)$ is the function itself. Thus,

$$u\in R,\quad u_t\in R,\quad u_x\in R^n,\quad u_{xx}\in S^n,\quad u(\cdot)\in X,$$

where S^n is the class of real symmetric n by n matrices and X is the class of continuous functions $\overline{G}\to R$. We denote by W the subclass of functions $u\in X$ for which u_t and u_x admit a continuous extension to $\overline{G}$.

For fixed $T>0$ let $G(T)$ denote the part of G in which $t\le T$, thus $G(T)=(0,T]\times\Omega$. We write Z for the subclass of functions $\phi\in X$ which satisfy a Hölder condition

$$|\phi(s,x)-\phi(t,y)|\le K\left(|s-t|^{\alpha/2}+|x-y|^{\alpha}\right),\qquad \alpha>0$$

in $G(T)$ for each T, where K and α can depend on (ϕ,T). The subclass of functions $\phi\in X$ for which the above condition holds with $s=t$ is denoted by Y. Thus, $X\supset Y\supset Z$.

A similar definition is used for functions with domain G instead of $\overline{G}$ and for functions with range in R^n or S^n; in the latter case $|\cdot|$ is the Euclidean norm. Superscripts on X, Y, Z mean that corresponding conditions are imposed on the x derivatives. For example, $\Psi\in Z^2$ means $\Psi\in Z$, $\Psi_x\in Z$, $\Psi_{xx}\in Z$ where the functions are of the forms $\overline{G}\to R$, $G\to R^n$ and $G\to S^n$, respectively. The side condition $t\le T$ remains in force, T being fixed but arbitrarily large.

3. An existence theorem. A parabolic-functional operator is defined by

$$Pu=u_t-\eta\Delta u-f(t,x,u(\cdot)), \tag{1}$$

where η is a positive constant and f is a real-valued functional at each $(t,x)\in G$, whose properties will be described later. The problem to be considered is

$$Pu+J|u|^{\lambda}\operatorname{sgn} u=0\quad\text{in } G,\qquad u=\Psi\in Z^1\quad\text{in }\Gamma \tag{2}$$

where J and λ are constant, $J>0$, $0<\lambda<1$. The term with $J|u|^{\lambda}$ is a control which makes quenching a possibility; the strength of the control is measured by J and λ. We want to give conditions under which the following statements are both true:

(i) The problem has at least one solution.

(ii) Every solution has compact support.

Both questions involve a measure of the extent to which f depends on the past history of the function u. Let μ: $[0,\infty)\to R$ be a given function such that

$$0\leq\mu(t)\leq t,\quad 0\leq t<\infty,$$

and for $u\in X$ let

$$|u|_{\mu,t}=\sup\{|u(\tau,\xi)|: t-\mu(t)\leq\tau\leq t, \xi\in\Omega\}.$$

Important special cases are $|u|_{t,t}$ and $|u|_{0,t}$. The first involves the entire past history while the second does not involve the past at all. We refer to μ as the *memory-function*, and a functional which can be assessed by $|u|_{\mu,t}$, is said to be of memory μ; more correctly, of memory $\leq\mu$. The two measures above pertain to functionals of memory t and of memory 0, respectively.

The following result gives an affirmative answer to the first question posed above:

THEOREM 1. *With P as in* (1) *let f be a continuous function* $G\times W\to X$ *which satisfies a closure condition of the form*

$$u\in X\cap Y^1 \;\Rightarrow\; f(t,x,u(\cdot))\in Y \tag{3}$$

and is of linear growth in the sense

$$|f(t,x,u(\cdot))|\leq(\text{const})\left(1+|u|_{t,t}+|u_x|_{t,t}\right). \tag{4}$$

Then (2) *has a solution* $u\in C(\bar{G})\cap C^2(G)$.

This follows from the results in [23]. The term in J is not included there, but since the functions need not be Lipschitzian, Theorem 1 is brought within the scope of [23] by redefining f. Condition (4) is required only in $G(T)$ for each T and the constant can depend on T.

4. A condition for quenching. We denote by ρ a continuous nonnegative function of t and we write ρ_μ as an abbreviation for $|\rho|_{\mu,t}$. Thus,

$$\rho_\mu=\sup\{\rho(\tau): t-\mu(t)\leq\tau\leq t\}.$$

The simplest example of the type of problem introduced in the foregoing discussion is

$$\rho(0)=\alpha,\quad \rho_t=\rho_\mu-\rho^\lambda,\quad t>0.$$

Here ρ_t is a left derivative and α is a positive constant. The problem is: What conditions on (λ,α,μ) ensure $\rho(t)=0$ for large t? Since $\rho_t\geq\rho-\rho^\lambda$ it is clear that $\alpha<1$ and $\lambda<1$ are necessary. Suppose next that $\lambda<1$ and μ are given. If the condition $\alpha<1$ is *sufficient* to ensure $\rho(t)=0$ for large t, we say $\mu\in Q(\lambda)$. Thus, $\mu\in Q(\lambda)$ means

$$\alpha<1\Rightarrow\rho(t)=0\quad\text{for } t\geq t^*(\alpha).$$

In [18], where this definition is introduced, it is seen that the hypothesis $\mu\in Q(\lambda)$ ensures quenching in a comprehensive class of parabolic-functional inequalities. Since $\rho_t\leq\rho_\mu-\rho^\lambda$ is itself a member of this class, the condition $\mu\in Q(\lambda)$ is sharp (cf. §6 below).

Use of the hypothesis $\mu\in Q(\lambda)$ in connection with (1), (2) requires a finer measure of growth than the condition (4). This is described by

$$u\in W,\, u_x=0 \;\Rightarrow\; (\operatorname{sgn} u)f(t,x,u(\cdot))\leq A|u|+B|u|_{0,t}+C|u|_{\mu,t},$$

where A, B, C are continuous functions $G \to R$ with $C \geq 0$. We assume also that $A+B$ and C are bounded above in G. In that case the constant

$$K = \sup_G (A+B+C)$$

is finite and is called the *growth constant* of f. (If a unique K is desired one can take the inf over the possible choices of A, B, C, but this is not essential.) The following holds:

THEOREM 2. *Under the hypothesis of Theorem* 1 *let f admit the growth constant K and let $|\Psi| \leq I$, where the constant I is unrestricted if $K \leq 0$ but*

$$I < \left(\frac{J}{K}\right)^{\gamma}, \qquad \gamma = \frac{1}{1-\lambda}$$

if $K > 0$. Suppose further that $\mu \in Q(\lambda)$. Then u has compact support if, and only if, Ψ has compact support.

The result follows from those in [18] or can also be proved by a simplified version of the comparison argument given there. (Comparison theorems for parabolic-functional inequalities can be found in [19], [20], [21], [22], [24].) In the course of the proof it is seen that $|u| \leq I$, a fact which gives the following corollary.

COROLLARY 1. *Under the hypothesis of Theorem* 2, *the growth conditions of Theorems* 1 *and* 2 *are needed only for those functions $u \in W$ that satisfy $|u| \leq I$ in G.*

The corollary allows such terms as e^u or $u^5|u_x|$ and greatly increases the scope of the results.

5. An outline of certain generalizations. It is convenient to denote the class of real symmetric n by n matrices by S^n and to write

$$ab = \sum a_{ij} b_{ij}, \qquad \xi a \xi = \sum \xi_i a_{ij} \xi_j$$

for a, $b \in S^n$ and $\xi \in R^n$. In this notation the term $\eta \Delta u$ in (1) can be replaced by $a(t,x)u_{xx}$, where a: $G \to S^n$ satisfies $a_{ij} \in Y$ together with an ellipticity condition of the form

$$\xi a(t,x) \xi \geq \eta(T)|\xi|^2, \qquad \eta(T) > 0, \quad (t,x) \in G(T).$$

Under these conditions, Theorems 1 and 2 hold with

$$Pu = u_t - a(t,x)u_{xx} - f(t,x,u(\cdot)). \tag{5}$$

The main results also extend to systems. For functions u: $\bar{G} \to R^m$ we use $|u|$ to denote the sup norm $|u| = \max_j |u^j|$ and we interpret regularity conditions such as $u \in X$ componentwise; that is, in the sense $u^j \in X$, $j = 1, 2, \cdots, m$, where the latter condition has the meaning previously assigned. For twice-differentiable functions $u \in X$ an operator $P = (P^1, P^2, \cdots, P^m)$ is defined by

$$P^k u = u_t^k - a^k(t,x)u_{xx}^k - f^k(t,x,u(\cdot)) \qquad k = 1, 2, \cdots, m,$$

where each a^k: $G \to S^n$ has the properties imposed on $a(t,x)$ in (5) and each f^k is a functional analogous to f in (1). The system

$$P^k u + J^k |u^k|^{\lambda} \operatorname{sgn} u^k = 0 \quad \text{in } G, \qquad u^k = \Psi^k \quad \text{in } \Gamma$$

can be written in the condensed form

$$Pu + J|u|^{\lambda} \operatorname{sgn} u = 0 \quad \text{in } G, \qquad u = \Psi \quad \text{in } \Gamma,$$

where $\operatorname{sgn} u$ denotes the matrix $\operatorname{diag}(\operatorname{sgn} u^1, \operatorname{sgn} u^2, \cdots, \operatorname{sgn} u^m)$ and where $J|u|^\lambda$ is the vector

$$J|u|^\lambda = \left(J^1|u^1|^\lambda, J^2|u^2|^\lambda, \cdots, J^m|u^m|^\lambda\right).$$

Theorems 1 and 2 apply to this problem with virtually no change. It is remarkable that the control term involves only $|u^k|^\lambda$ in the kth equation, and not $|u|^\lambda$ as one might expect. The underlying reason for this is the fact that the kth equation is used only at points where $|u^k| = |u|$.

So far we have assumed $\mu(t) \le t$, so that the history of u prior to $t=0$ is not involved. If, instead, $\mu(t) \le t+r$, where r is fixed, we redefine Γ_0 to be $[-r,0] \times \bar{\Omega}$ and proceed as before. The initial condition associated with the class $Q(\lambda)$ is now $\rho(t) = \alpha$ for $-r \le t \le 0$, rather than $\rho(0) = \alpha$, and some of the classes X, Y, Z must be referred to $[-r, T] \times \Omega$ rather than to $[0, T] \times \Omega$. Otherwise there is no significant change.

Finally, it is possible to extend the theory to more general boundary conditions, including those of Neumann type, and to unbounded regions including those for which no boundary conditions are needed, e.g., the Cauchy problem. These extensions increase the technical complexity of the proofs but again the results are virtually the same as before. As explained next, such extensions facilitate the construction of counterexamples.

6. Simple counterexamples. By choosing all coordinates of u to be the same, one can obtain a counterexample for the vector case $m > 1$ from a corresponding example in the scalar case, $m = 1$. In view of this fact [18] we consider the case $m = 1$ here. We will show that the hypothesis $I < (J/K)^\gamma$ and the hypothesis $\mu \in Q(\lambda)$ are both necessary if Theorem 2 is to apply to the whole class of operators P considered there.

Let us begin by considering the equation

$$u_t - \eta\Delta u - \left(au + bu + cu_\mu\right) + J|u|^\lambda = 0, \tag{6}$$

where a, b, c are functions $G \to R$ and u_μ is an abbreviation for $u(t - \mu(t), x)$. For the moment we operate within the context of the last-described generalization in §5, so that the Neumann boundary condition $u_\nu = 0$ is permissible or, in the case of the Cauchy problem, no boundary condition is needed. Thus u can be a function of t alone.

In order to satisfy the prescribed continuity and monotonicity conditions we assume

$$a \le A, \quad |b| \le B, \quad 0 \le c \le C. \tag{7}$$

The factor u in bu in (6) could be replaçed by a zero-memory functional such as F: $u(t,x) \to u(t, \alpha(t,x))$, but this extension is not needed for the counterexample. Thus, the theorem allows operators P of considerable generality, but the counterexample applies even when P is severely restricted.

Let us inquire under what conditions (6) admits a constant solution $u = I > 0$, which of course does not quench. By inspection the sole condition needed is $kI = JI^\lambda$, where $k = a + b + c$. Clearly this can be satisfied, in the presence of (7), whenever $I \ge (J/K)^\gamma$, $K = A + B + C > 0$. Therefore the initial-value inequality $I < (J/K)^\gamma$ of Theorem 2 is sharp. We need not consider the case $K \le 0$ because Theorem 2 asserts that there is no counterexample in that case.

Since the choice $u = I$ satisfies $u_x = u_{xx} = 0$, we could add to the left side of (6) any function

$$f(t, x, u_x, u_{xx}, u(\cdot)), \qquad f(t, x, 0, 0, u(\cdot)) = 0.$$

Thus the seemingly special example (6) leads to a variety of additional examples. The same applies when $u(t,x)=\rho(t)$, as assumed next.

Under the harmless assumption that μ is continuous, we now show that the hypothesis $\mu\in Q(\lambda)$ is also sharp. In the contrary case the problem

$$\rho_t=\rho_\mu-\rho^\lambda, \qquad t>0, \quad \rho(0)=\alpha, \quad 0<\alpha<1$$

has a solution which does not quench. We define

$$u(t,x)=J^\gamma\rho(t)$$

and consider (6) with $a+b=A+B$, $c=C$, where $A+B=0$, $C=1$. Thus $K=1$ and the initial value satisfies

$$u(0,x)=J^\gamma\alpha<\left(\frac{J}{K}\right)^\gamma$$

since $\alpha<1$. (It is here that we use the fact that μ is not in $Q(\lambda)$.) Equation (6) is satisfied, as seen by the fact that the left side is

$$u_t-u_\mu+Ju^\lambda=J^\gamma\left(\rho_t-\rho_u+\rho^\lambda\right)=0.$$

Thus the condition $\mu\in Q(\lambda)$ is sharp even when $I<(J/K)^\gamma$.

7. Dirichlet boundary conditions. No function of the form $u(t,x)=\rho(t)$ can provide a counterexample in the presence of Dirichlet boundary conditions such as those in Theorem 2. Namely, if $\rho=\Psi$ on the boundary, where Ψ has compact support, then ρ vanishes for large t and hence ρ quenches. As explained in the introduction, construction of a counterexample satisfying Dirichlet boundary conditions is a major objective of this paper.

For simplicity let Ω be the sphere $|x|<1$ where, here and below, $|\cdot|$ denotes the Euclidean norm. We consider solutions of the form

$$u(t,x)=J^\gamma\rho(t)\sigma(x), \qquad \sigma(x)=1-|x|^2,$$

where $\rho_t=\rho_\mu-\rho^\lambda$, and we note that $u(t,x)$ vanishes on the boundary. The analysis of §6 serves to motivate the analysis given here.

With a and h chosen so that

$$a\left(\sigma+|\operatorname{grad}\sigma|\right)=-\eta\Delta\sigma, \qquad h|\operatorname{grad}\sigma|^\lambda=J(\sigma^\lambda-\sigma),$$

it is seen that

$$u_t-\eta\Delta u-a\left(u+|\operatorname{grad}u|\right)-h|\operatorname{grad}u|^\lambda-u_\mu+J|u|^\lambda=0.$$

Indeed, because of the definition of a, the terms involving η and a cancel. There remains

$$J^\gamma\sigma\left(\rho_\mu-\rho^\lambda\right)-J^{\gamma\lambda}\rho^\lambda h|\operatorname{grad}\sigma|^\lambda-J^\gamma\rho_\mu\sigma+JJ^{\gamma\lambda}\rho^\lambda\sigma^\lambda.$$

The two terms in ρ_μ cancel, and the three terms in ρ^λ cancel by the definition of h when we recall that $\gamma=1+\gamma\lambda$. Thus, the final result is 0.

Since the continuity inequality involving A,B,C is required only at points where $\operatorname{grad}u=0$, it holds if $a\leq A+B$ and $C=1$. Hence $K=1+a$ is an admissible choice and the initial-value inequality of Theorem 2 holds if

$$J^\gamma \alpha < \left(\frac{J}{1+a}\right)^\gamma, \qquad \rho(0)=\alpha.$$

Since $a\to 0$ uniformly as $\eta\to 0$ this holds for small η if $\alpha<1$. The latter condition in turn is possible for a ρ which does not quench, provided μ is continuous and does not belong to $Q(\lambda)$. Thus we get a counterexample in that case.

It remains to show that the coefficients a and h satisfy the continuity conditions imposed in Theorem 1, so that the example falls within the scope of the existence as well as the quenching part of the theory. Since

$$1+2|x|>|x|^2, \qquad 0\le |x|\le 1,$$

the coefficient $a=2n\eta/(1+2|x|-|x|^2)$ has the same continuity properties as $|x|$ and is admissible. The only difficulty with h occurs in the neighborhood of the point $x=0$ where grad σ vanishes. Expanding $\sigma^\lambda-\sigma$ in powers of $|x|$ we see that, near 0,

$$h(x)|2x|^\lambda = J(1-\lambda)|x|^2+\cdots$$

so that $h(x)=2^{-\lambda}J(1-\lambda)|x|^{2-\lambda}+\cdots$. This shows that h is also admissible and completes the construction of the counterexample when the initial condition holds but μ is not in $Q(\lambda)$.

Suppose next that $I>(J/K)^\gamma$. In this case we carry out the above calculation with $\rho=1$. Here $\rho_t=\rho_\mu-\rho^\lambda$ for every choice of μ, including $\mu=0$, and a counterexample is obtained for small η as before. If $\eta=0$ we can take $a=0$ and also rule out the case $I=(J/K)^\gamma$. The condition $\eta=0$ is allowed in the quenching part of the theory [18], though of course not in the existence theorems of [23].

The discussion of counterexamples in this and the preceding section sheds light on the structure of the quenching problem. The problem involves two classes: a class $Q(\lambda)$ of memory functions and a class $P(I,\lambda)$ of parabolic-functional operators together with initial conditions. The class $Q(\lambda)$ is described by a specific delay differential equation, while $P(I,\lambda)$ is distinguished by its extreme generality; within the context of differential inequalities, which is the natural setting for the theory, the class is vastly more general than any of the classes discussed in §5. The final result takes the following form: If $\mu\in Q(\lambda)$ then quenching occurs for every operator $P\in P(I,\lambda)$ with initial values dominated by I. But if μ is not in $Q(\lambda)$ then there exist operators $P\in P(I,\lambda)$, with initial values dominated by I, whose solutions in

$$(\operatorname{sgn} u)Pu+|u|^\lambda\le 0 \tag{8}$$

do not quench. The present paper shows that the same behavior is found for solutions of the corresponding equation, a matter which is not addressed in [18].

8. Characterization of the class $Q(\lambda)$. One of the most interesting features of the memory-quenching problem is that the same basic hypothesis, $\mu\in Q(\lambda)$, is encountered in all the extensions outlined above. Furthermore, as we have seen, the hypothesis is sharp. Because of this it was though worthwhile to characterize $Q(\lambda)$ by explicit structural properties in [18], and the main features are summarized here without proof.

In the first place, if $\mu\in Q(\lambda)$, there must be an increasing sequence $\{t_n\}$ such that $t_n\to\infty$ and

$$\mu(t)\le t-t_n, \qquad t\ge t_n. \tag{9}$$

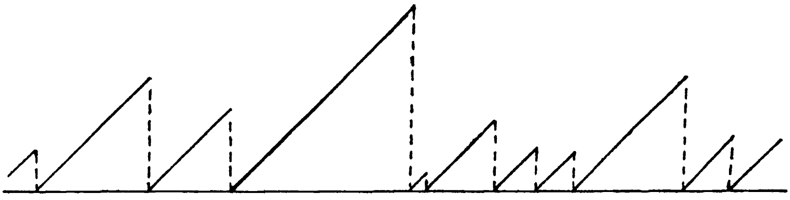


FIG. 1. *The graph of μ must lie below the slanting lines.*

This means that the graph of μ lies below an infinite set of lines of slope 1 as shown in Fig. 1.

Another necessary condition restricts the growth of μ at the left of the points t_n, and, if the restriction is sufficiently stringent, it is also sufficient. For example [18], it suffices to have

$$\mu(t) \le |t-t_1| \wedge |t-t_2| \wedge |t-t_3| \wedge |t-t_4| \cdots, \qquad 0 \le t < \infty$$

where $\{t_n\}$ is any unbounded sequence whatever; that is, if this holds for such a sequence then $\mu \in Q(\lambda)$. (Here $a \wedge b = \min(a,b)$, as usual.) The above condition agrees with (9) near t_n+, but is unnecessarily stringent near t_n-.

If we confine our attention to growth which is measured by a power of $t_n - t$, the appropriate restriction is of the form

$$\mu(t) \le r_n(t_n - t)^\lambda, \qquad t_n - s_n \le t \le t_n, \tag{10}$$

where r_n and s_n are positive numbers. Within this framework a sufficient condition for $\mu \in Q(\lambda)$ is

$$\limsup_{n\to\infty} \lambda^n(\log s_n - \gamma \log^+ r_n) \ge 0, \tag{11}$$

where $\gamma = 1/(1-\lambda)$ and where it is assumed that $\{t_n\}$ is separated in the sense that $\inf(t_n - t_{n-1}) > 0$. Conditions (10) and (11) are sharp in that they do not imply $\mu \in Q(\lambda)$ if weakened in any one of the following ways:

(a) The exponent λ in (10) is replaced by some $\tilde{\lambda} < \lambda$.

(b) The lim sup in (11) is negative.

(c) The factor λ^n in (11) is replaced by $\lambda^{n+\phi(n)}$ with $\phi(n) \to \infty$.

(d) The constant γ in (11) is less than $1/(1-\lambda)$.

(e) The sequence $\{t_n\}$ is subjected to no separation condition.

With regard to (e) the condition $\inf(t_n - t_{n-1}) > 0$ imposed above can be substantially weakened, and it is here that the principal difficulty in characterizing the class $Q(\lambda)$ is found. The weakened separation condition is discussed at length in [18].

9. Double-exponential decay. An interesting feature of the memory-quenching problem is that the simple hypothesis (9) alone leads to an astonishingly fast rate of decay for $u(t,x)$, in general, even if the more delicate criteria involving λ are not fulfilled. Here the phrase "in general" means that the initial-value inequality of Theorem 2 is imposed, and that $\{t_n\}$ satisfies the mild separation condition to which allusion was made at the end of the preceding section. In particular, $\inf(t_n - t_{n-1}) > 0$ suffices.

Under these conditions u satisfies an inequality of the form

$$|u(t,x)| \leq \theta^{1/\lambda^n}, \qquad t \geq t_n \gg 1,$$

where $\theta < 1$ is a constant depending on the initial value I and on the structure of the operator P. One could have for example $\theta = 0.1$, $\lambda = 0.1$ so that

$$|u(t,x)| \leq 10^{-10^n}, \qquad t \geq t_n.$$

The surprising nature of this inequality is appreciated when we note that for $n = 100$, say, there are not just 100 zeros before the decimal point, but 10^{100} zeros before the decimal point. As seen in [18], iterated exponentials pervade the theory of quenching in the presence of memory.

REFERENCES

[1] H. AMANN, *Invariant sets and existence theorems for semilinear parabolic and elliptic systems*, J. Math. Anal. Appl. 65 (1978), pp. 432–467.

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

[3] A. BENSOUSSAN AND J. L. LIONS, *On the support of the solution of some variational inequalities of evolution*, J. Math. Soc. Japan, 28 (1976), pp. 1–17.

[4] P. BENILAN, H. BRÉZIS AND M. G. CRANDALL, *A semilinear equation in $L^1(R^n)$*, Ann. Scuola Norm. Pisa, Cl. 3 Sci., (4) 2 (1975), pp. 523–555.

[5, 6] L. D. BERKOVITZ AND HARRY POLLARD, *A nonclassical variational problem arising from an optimal filter problem*, Arch. Rat. Mech. Anal., 26 (1967), pp. 281–304; II, 38 (1970), pp. 161–172.

[7] H. BRÉZIS, *Solutions with compact support of variational inequalities*, Uspekhi Mat. Nauk., 29 (1974), pp. 103–108; Russian Math. Surveys, 29 (1974), pp. 103–108.

[8] J. ILDEFONSO DÍAZ DÍAZ, *Resultados y metodos sobre la propriedád de extincion en tiempo finito para ecuaciones de evolucion*, preprint.

[9] S. D. EIDEL'MAN, *Parabolic Systems*, North-Holland, Amsterdam, 1969.

[10] W. E. FITZGIBBON, *Semilinear functional differential equations in Banach space*, J. Differential Equations, 29 (1978), pp. 1–14.

[11, 12] MAGNUS HESTENES AND R. M. REDHEFFER, *On the minimization of certain quadratic functionals* I, Arch. Rat. Mach. and Anal., 56 (1974), pp. 1–14; II, 56, (1974), pp. 15–33.

[13] A. M. IL'IN, A. S. KALASHNIKOV AND O. A. OLEINIK, *Linear equations of the second order of parabolic type*, Russian Math. Survey, 17 (1962), pp. 1–143.

[14] H. J. KUIPER, *Existence and comparison theorems for nonlinear diffusion systems*, J. Math. Anal. Appl., 60 (1977), pp. 1075–1103.

[15] O. A. LADYZHENSKAYA, V. A. SOLONNIKOV AND N. N. URAL'TSEVA, *Linear and Quasilinear Equations of Parabolic Type*, AMS Transl. Math. Monographs 23, Providence, RI, 1968.

[16] G. LAMOTT, *Ein Existenz- und Konvergenzsatz für schwach gekoppelte Systeme parabolischer Differentialgleichungen mit Hilfe der Linienmethode*, Dissertation, Univ. Karlsruhe, 1976.

[17] RAY REDHEFFER, *On a nonlinear functional of Berkovitz and Pollard*, Arch. Rat. Mech. Anal., 50 (1973), pp. 1-9.

[18] ______, *The dependence of quenching on memory*, ms. available at UCLA.

[19] RAY REDHEFFER AND W. WALTER, *Uniqueness, stability and error estimation for parabolic functional-differential equations*, Ber. 5, Univ. Karlsruhe 1976; also in Jubilee Volume for the 70th birthday of I. N. Vekua, Nauka, Moscow 1978.

[20] ______, *Das Maximumprinzip in unbeschränkten Gebieten für parabolische Ungleichungen mit Funktionalen*, Math. Ann., 226 (1977), pp. 155–170.

[21] ______, *Comparison theorems for parabolic functional inequalities*, Pacific J. Math., 85 (1979), pp. 447-470.

[22] ______, *Stability of the null solution of parabolic functional inequalities*, Trans. Amer. Math. Soc., (1980), pp. 285-302.

[23] REINHARD REDLINGER, *Existenzsätze für semilineare parabolische Systeme mit Funktionalen*, Dissertation, Univ. Karlsruhe, 1982.

[24] J. SZARSKI, *Strong maximum principle for non-linear parabolic differential-functional equations in arbitrary domains*, Ann. Polon. Math., 31 (1975), pp. 197–203.

[25] C. C. TRAVIS AND G. F. WEBB, *Existence and stability for partial functional differential equations*, Trans. AMS, 200 (1974), pp. 395–418.

[26] H. UGOWSKI, *On a certain non-linear initial-boundary value problem for integro-differential equations of parabolic type*, Ann. Polon. Math., 28 (1973), pp. 249–259.

[27] W. WALTER, *Differential and Integral Inequalities*, Vol. 55, Springer-Verlag, New York, 1970.

[28] D. WENDLAND, *Existenz- und Konvergenzsätze für das Cauchy-Problem bei schwach gekoppelten parabolischen Systemen mit der Linienmethode*, Dissertation, Univ. Karlsruhe, 1981.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics
007

EXISTENCE OF SOLUTIONS TO SINGULAR CONSERVATION LAWS*

MARIA ELENA SCHONBEK[†]

Abstract. We consider the existence of solutions to singular scalar conservation laws of the form $u_t+f(u)_x+\phi(u)/x=0$. We prove existence by regularizing the equation and taking a singular limit using the recently developed theory of compensated compactness. This theory allows us to pass to the limit without gradient estimates.

1. Introduction. The existence theory of inhomogeneous systems of conservation laws

$$u_t+f(u)_x=g(x,u)$$

has been developed primarily with regular forcing terms g [5]. There is presently no existence theory for singular inhomogeneous systems. In particular there is no existence theory for singular equations of the form

$$u_t+f(u)_x+\frac{\phi(u)}{x}=0, \tag{1.1}$$

where f: $R^n\to R^n$, ϕ: $R^n\to R^n$, even in the scalar case $n=1$. Algebraic singularities of the type (1.1) arise, for example, in the equations of fluid dynamics with spherical or cylindrical symmetry. The main difficulty in estimating solutions stems from the well-known fact that bounded initial data does not give rise to bounded solutions due to the focusing of waves at the origin. This paper is concerned with the existence of solutions to the Cauchy problem for scalar conservation laws

$$u_t+f(u)_x+\frac{\phi(u)}{x}=0. \tag{1.2}$$

Here f and ϕ are smooth maps from R to R. We note that the model equation (1.2) retains the essential feature of waves focusing at the origin. We establish global existence of weak solutions with initial data vanishing at infinity. The class of flux functions f and forcing functions ϕ considered will satisfy certain sign conditions at infinity (cf. §2). This class includes a model situation studied by Whitham [15], $\phi=c_0u$ and $f=c_1u^2/2+c_2u$. Existence of solutions is established by regularizing (1.2) and passing to the limit. The regularization employed takes on one of two forms according to the sign of ϕ at infinity. If $u\phi(u)>0$ for u large, (1.2) is regularized by adding a dissipative terms and removing the singularity at $x=0$

$$u_t+f(u)_x+\frac{\phi(u)}{x+\delta}=\varepsilon u_{xx}, \qquad \varepsilon,\delta>0. \tag{1.3}$$

If $u\phi(u)<0$ for large values of u, only the singularity at $x=0$ is removed

$$u_t+f(u)_x+\frac{\phi(u)}{x+\delta}=0, \tag{1.4}$$

* Received by the editors December 3, 1982, and in revised form May 15, 1983. This research was supported by the N.tional Science Foundation under grant MCS-81021405-33781 and grant RII 84087 53.

[†] Department of Mathematics, Duke University, Durham, North Carolina 27706.

where the solutions of (1.4) are distributional solutions in L^∞ obtained by the viscosity method. For details on the existence and uniqueness of such solutions we refer the reader to [10] and [14].

With regard to the general problem of taking a singular limit we recall that the classical approach is first to obtain uniform bounds on the amplitude and the derivatives of the solution, then appeal to standard compactness arguments to extract a subsequence which converges in the strong topology. We note that in nonlinear problems it is necessary in general to establish strong convergence since nonlinear maps are typically not continuous with respect to the weak topology, i.e. if u^ε converges weakly to $\bar{u}, f(u^\varepsilon)$ does not need to converge weakly to $f(\bar{u})$. We also recall that uniform bounds on the amplitude alone yield by classical methods only weakly convergent subsequences. In this paper we use the recently developed theory of compensated compactness [12] to pass to the limit without uniform control on the derivatives. The theory of compensated compactness provides a description of the weak limits and in certain cases the conditions under which weak limits become strong. Several results in measure theory and compensated compactness will be used in this paper. The results in measure theory describe the weak limit of continuous functions as the expected value of a family of probability measures. To be more specific the first result establishes for any L^∞-bounded sequence u^ε: $R^n \to R^m$, the existence of a subsequence u^{ε_k} and an associated family of probability measure $\{\nu_x(\lambda) : x \in R^n, \lambda \in R^m\}$ such that for any $f \in C$

$$\Big(\lim_{\varepsilon_k \to 0} f(u^{\varepsilon_k})\Big)(x) = \langle \nu_x, f(\lambda)\rangle = \int f(\lambda)\, d\nu_x(\lambda), \quad \text{a.e. in } \mathbb{R}^n,$$

where the limit is taken in L^∞ weak $*$. From this theorem it follows that strong convergence is equivalent to having as associated measures $\{\nu_x\}$ point masses, i.e. $\nu_x = \delta_{\bar{u}(x)}$ if and only if u^{ε_k} converges strongly to $\bar{u}$. The theory of compensated compactness is used to show that the associated measures to the sequences of solutions of (1.3) and (1.4) reduce to point masses.

The program in this paper will be first to establish a priori bounds on the amplitude of the approximate solutions and then show that the associated measures reduce to a point mass. In order to do this the notion of entropy for hyperbolic conservation laws is used together with the information that certain nonlinear functions are continuous in the weak topology. The problem of showing that the associated measures to solutions of approximate hyperbolic equations (equations of the type (1.3) and (1.4) for example) are point masses is reduced, by compensated compactness, to obtaining control on the rate of entropy production (cf. §3).

The paper is divided into two sections. In the first local L^∞ a priori bounds are obtained either by maximum principles for parabolic equations (1.3), appropriate entropy inequalities or by estimating the solutions along generalized backward characteristics introduced by C. Dafermos [2], for solutions of (1.4). In the last section we use the a priori bounds in conjunction with results of the theory of compensated compactness in order to obtain the existence of a solution of (1.2) for the Cauchy problem, that is we obtain a subsequence of solutions of (1.3) (or (1.4)) which converges pointwise a.e. to a solution of (1.2). For general background on compensated compactness the reader is referred to Dacorogna [1], Murat [6], [7], [8], and Tartar [12], [13].

2. A priori bounds. In this section we obtain a priori bounds on the amplitude of the solutions of the equations

$$u_t + f(u)_x + \frac{\phi(u)}{x+\delta} = \varepsilon u_{xx}, \tag{2.1}$$

$$u_t+f(u)_x+\frac{\phi(u)}{x+\delta}=0. \tag{2.2}$$

Here f and ϕ are smooth functions satisfying certain sign conditions at infinity which will be specified below. In addition it will be required in general that f is strictly convex or strictly concave. The a priori bounds on the amplitude will be uniform in ε and δ and they will be used in conjunction with results of the theory of compensated compactness in order to pass to the limit in equations (2.1) and (2.2) obtain a weak solution of the conservation law

$$u_t+f(u)_x+\frac{\phi(u)}{x}=0.$$

The a priori bounds will be derived from entropy inequalities or from estimates on generalized backward characteristics [2]. For completeness we recall the definition of entropy associated to a hyperbolic conservation law.

DEFINITION 2.1. A pair of real valued functions (η, q) is called an entropy pair for a conservation law

$$u_t+f(u)_x=0 \tag{2.3}$$

if all smooth solutions of (2.3) satisfy an additional equation of the form

$$\eta(u)_t+q(u)_x=0.$$

We note that (η, q) is an entropy pair if and only if the compatibility condition

$$\eta'(u)f'(u)=q'(u) \tag{2.4}$$

is satisfied.

First we shall establish an L^∞ estimate for solutions of (2.1) if $\phi(u)$ satisfies the following sign condition at infinity: There exists $M>0$ such that

$$\text{if } |u|\geq M, \qquad u\phi(u)\geq 0. \tag{2.5}$$

We note that this condition holds in the case discussed by Whitham [16] where $\phi(u)=c_0u$, $c_0>0$. In what follows we suppose that the functions ϕ and f are smooth.

THEOREM 2.1. *Let $u_0(x)$ be a smooth function vanishing at zero and infinity. Let u_δ^ε be a sequence of solutions of* (2.1) *with initial and boundary data $u_\delta^\varepsilon(x,0)=u_0(x)$ and $u_\delta^\varepsilon(0,t)=0$. If ϕ satisfies* (2.5), *then*

$$|u_\delta^\varepsilon(\cdot,t)|_\infty\leq C$$

where C is independent of ε and δ.

Proof. We recall that in [10] and [11] Oleinik has established the existence of bounded solutions with continuous derivatives for the Cauchy problem

$$u_t+f(u,x,t)_x+\psi(x,t)=\varepsilon u_{xx},$$
$$u(x,0)=u_0(x),$$

where f and ψ are smooth and $u_0(x)$ is a bounded measurable function. Here the assumptions on $u_0(x)$ insure that $\lim_{x\to\infty} u_\delta^\varepsilon(x,t)=0$ pointwise a.e. for each fixed t. To establish an L^∞ bound we construct the entropy

$$\eta(u)=\frac{(|u|-M)^2}{2} \quad \text{for } |u|\geq M, \qquad \eta(u)=0 \text{ otherwise,}$$

where M is such that $u\phi(u)\geq 0$ for all $|u|\geq M$ and $|u_0|_\infty\leq M$. Multiplying equation (2.1) by $\eta'(u)$ we obtain

$$\eta(u)_t+q(u)_x=-\eta'(u)\frac{\phi(u)}{x+\delta}+\varepsilon\eta'(u)u_{xx}, \tag{2.6}$$

where q is the corresponding entropy flux, cf. Definition 2.1. Equation (2.6) can be rewritten as

$$\eta(u)_t+q(u)_x=-\eta'(u)\frac{\phi(u)}{x+\delta}+\varepsilon(\eta'(u)u_x)_x-\varepsilon\eta''(u)u_x^2.$$

One integration in space yields

$$\frac{d}{dt}\int_0^\infty\eta(u)\,dx=-q(u)|_0^\infty-\int_0^\infty\eta'(u)\frac{\phi(u)}{x+\delta}dx \tag{2.7}$$
$$+\varepsilon\eta'(u)u_x|_0^\infty-\varepsilon\int_0^\infty\eta''(u)u_x^2dx.$$

The hypotheses on the initial data $u_0(u)$ imply that $\lim_{x\to\infty}u_\delta^\varepsilon(x,t)=\lim_{x\to\infty}u_\delta^\varepsilon(x,t)=0$, and the definition of q and η_p' imply that $q(0)=\eta'(0)=0$, thus the boundary terms in (2.8) vanish. It follows from (2.5) and the definition of η that $\eta'\phi\geq 0$ for $|u|\geq M$. Combining this with the fact that $\eta''\geq 0$, we obtain

$$\frac{d}{dt}\int_0^\infty\eta(u)\,dx\leq 0.$$

Therefore

$$\int_0^\infty\eta(u)\,dx\leq\int_0^\infty\eta(u_0)\,dx,$$

and since $|u_0|_\infty\leq M$, $\eta(u_0)=0$. Hence $\eta(u)=0$ which means that $|u|_\infty\leq M$.

We recall the following two ordering principles which are corollaries of the maximum principle for parabolic equations.

LEMMA 2.1. *Let u_δ^ε be a solution of* (2.1) *with smooth initial data $u_0(x)$ and boundary data $u(0,t)=0$. If $\phi(0)=0$ and if $u_0(x)\geq 0$ for all x or if $u_0(x)\leq 0$ for all x then $u_\delta^\varepsilon(x,t)\geq 0$ or $u_\delta^\varepsilon(x,t)\leq 0$ for all (x,t), respectively.*

For a proof see [4, Lemma 5, p. 43].

LEMMA 2.2. *Let $u(x,t)$ be a solution for the Cauchy problem for* (2.2), *with smooth initial data $u_0(x)$ with compact support. If $\phi(0)=0$ and if $u_0(x)$ is nonnegative for all $x\geq 0$ or nonpositive for all $x\geq 0$ then $u(x,t)\geq 0$ or $u(x,t)\leq$ for all (x,t) respectively.*

Proof. A straightforward modification of Volpert's argument [14, p. 264], shows that for fixed δ there exists a sequence of solutions u_δ^ε of (2.1) with smooth initial data $u_0(x)$ and zero boundary data which converges to a solution u_δ of (2.2) in the following sense:

$$\lim_{\varepsilon>0}\int_{0\leq x<r}|u_\delta(x,t)-u_\delta^\varepsilon(x,t)|\,dx=0$$

for any $r>0$, $t>0$. Hence integrating in time implies by Lebesgue's dominated convergence theorem that

$$\lim_{\varepsilon\to 0}\int_0^T\int_{0\leq x<r}|u_\delta(x,t)-u_\delta^\varepsilon(x,t)|\,dx\,dt=0$$

for all $T>0$. It follows that u_δ^ε converges to u_δ pointwise a.e. and u_δ is a distributional solution of (2.1) in L^∞. Moreover the initial data $u_0(x)$ is taken on in the weak topology, i.e. $\lim_{t\to\infty}\int(u_\delta(x,t)-u_0(x))\psi(x)\,dx=0$ for all $\psi(x)\in C_0^\infty(R)$. Hence the conclusion of Lemma 2.2 is an immediate consequence of Lemma 2.1.

Lemma 2.1 and 2.2 ensure the existence of nonpositive and nonnegative solutions of equation (2.1) and (2.2). For such solutions the following corollaries of Theorem 2.1 hold.

COROLLARY 2.1. *Let u_δ^ε be a sequence of nonnegative solutions of* (2.1) *with smooth initial data vanishing at zero and infinity and with $u_\delta^\varepsilon(0,t)=0$. If $\phi(u)\geq 0$ for $u\geq M$, then*

$$|u_\delta^\varepsilon(\cdot,t)|_\infty\leq C,$$

where C is independent of ε and δ.

Proof. We use the same line of argument as in Theorem 2.1 with entropies of the form

$$\eta(u)=\frac{(u-M)^2}{2}\quad\text{for } u\geq M,\qquad \eta(u)=0 \text{ otherwise,}$$

where $u\phi(u)>0$ for all $u\geq M$.

COROLLARY 2.2. *lThe conclusion of Corollary* 2.1 *holds if the u_δ^ε are nonpositive and ϕ satisfies $\phi(u)\leq 0$ if $u\leq -M$.*

Proof. Use the entropies

$$\eta(u)=\frac{(u+M)^2}{2}\quad\text{for } u\leq -M,\qquad \eta(u)=0 \text{ otherwise.}$$

COROLLARY 2.3. *Let $u_0(x)$ be a smooth function vanishing at zero and infinity. Fix δ and suppose that there exist M such that for $|u|\geq M$, $u\phi(u)\geq 0$, then the solution u_δ of* (2.2) *with initial data $u_0(x)$ satisfies*

$$|u_\delta(\cdot,t)|_\infty\leq C$$

where C depends only on f,ϕ and the L^∞ norm of the data.

Proof. By Theorem 2.1 there exists a uniformly bounded sequence of solutions u_δ^ε of (2.1) with initial data $u_0(x)$ and zero boundary data. Standard methods will give a bound on the spatial total variation of u_δ^ε independent of ε. For background material, we refer the reader to [14, §§17 and 18]. Thus Helly's theorem [9] can be applied to obtain strong convergence of the family u_δ^ε to a solution u_δ of (2.2) as ε vanishes. The L^∞ bound on u_δ^ε which is independent of ε and δ yield the desired result.

COROLLARY 2.4. *Let $u_0(x)$ be a smooth function vanishing at zero and infinity. If $u_0(x)\geq 0$ and $u\phi(u)\geq 0$ for $u\geq M$ or if $u_0(x)\leq 0$ and $u\phi(u)\geq 0$ for $u\leq -M<0$ then the conclusion of Corollary* 2.3 *holds.*

The next cases that will be considered have forcing term $\phi(u)$ satisfying

$$\begin{aligned} &u\phi(u)<0 \quad \text{for } u\geq M \quad \text{and/or}\\ &u\phi(u)<0 \quad \text{for } u\leq -M \end{aligned}\tag{2.8}$$

for some $M\geq 0$.

We now estimate the solutions of (2.2) on the generalized backward characteristics introduced by Dafermos [2]. Consider a solution $u(x,t)$ to the equation

$$u_t+f(u)_x+g(x,t,u)=0$$

where f and g are smooth and f is either strictly convex or strictly concave in u. Through each point (x,t) there exists backward generalized characteristics which consist either of a single classical characteristic or an infinite number of curves spanning a funnel confined between two classical characteristics. The backward classical characteristics are globally defined on the common domain of definition of f and g. In our case the backward characteristics either runs into the x-axis or runs into the t-axis. For equations which admit only outgoing waves, i.e. $f'>0$, we obtain the following result.

THEOREM 2.2. *Suppose that $f'>0$ and $f''\neq 0$. Let $u_0(x)$ be a smooth function with compact support contained in $(0,\infty)$. If there are constants M and α such that for $|u|\geq M$ $u\phi(u)<0$ and $-(f'(u)/\phi(u))(u/|u|)\geq\alpha>0$ then there exists a sequence of solutions u_δ of* (2.2) *with initial data $u_0(x)$ satisfying*

i) $\lim_{x>0} u_\delta(x,t)=0$ *for all* $t\geq 0$,

ii) $|u_\delta(x,t)|\leq \mathrm{const}(1+|\ln x|)$ *for all* $(x,t)\in R^+\times R^+$, *where the constant is independent of* δ.

Proof. Since $u_0(x)$ has compact support and $f'>0$, the t-axis is noncharacteristic for (2.2). Arguments similar to the ones used by Oleinik [10] will yield the existence of a sequence u_δ^ε of solutions of (2.2) which are uniformly small for $x\ll 1$. Hence i) follows. We shall denote by $C_{x,t}$ the set of all classical backward characteristics originating at (x,t). We note that the definite sign of f'' insures the existence of generalized backward and forward characteristics [2]. In order to estimate the solutions on classical backward characteristics we recall that if $(x(t),t)$ belongs to a classical backward characteristic

$$\lim_{y\to x^+} u_\delta(u,t)=\lim_{y\to x^-} u_\delta(y,t)=u_\delta(t)$$

where $u_\delta(t)$ is absolutely continuous in t and satisfies the following system of differential equations.

$$\frac{du}{dt}=-\frac{\phi(u)}{x+\delta},\qquad \frac{dx}{dt}=f'(u), \tag{2.9}$$

or equivalently since $f'>0$ implies $x'>0$

$$\frac{du}{dx}=-\frac{\phi(u)}{f'(u)}\frac{1}{x+\delta},\qquad \frac{dt}{dx}=\frac{1}{f'(u)}. \tag{2.10}$$

Let $(\bar{x},\bar{t})$ be an arbitrary point in $[0,\infty)\times[0,\infty)$. The following two cases need to be considered.

i) The intersection of $C_{\bar{x},\bar{t}}$ with the support of $u_0(x)$ is nonempty.

ii) The intersection of $C_{\bar{x},\bar{t}}$ with the support of $u_0(x)$ is empty.

Case i. Let $(x_0,0)\in C_{\bar{x},\bar{t}}\cap\{\operatorname{supp}u_0(x)\times[0,\infty)\}$. Let M be such that $u\phi(u)<0$, $-(u/|u|)f'(u)/\phi(u)\geq\alpha>0$ for $|u|\geq M$ and $|u_0(x)|<M$ for all x. Then either $|u(x,t)|<M$ for all $(x,t)\in C_{\bar{x},\bar{t}}$ and we are done or there exists $(x_1,t_1)\in C_{\bar{x},\bar{t}}$ such that $|u(x_1,t_1)|=M$. We suppose first that $u(x_1,t_1)=M$. In what follows for any point $(x,t)\in C_{\bar{x},\bar{t}}$ we shall use the notation $S(x,t)$ to indicate the classical backward characteristic in $C_{\bar{x},\bar{t}}$ containing (x,t) and we shall write $s(x,t)$ for the section of $S(x,t)$ joining (x,t) and $(\bar{x},\bar{t})$. After separating variables in the first equation of (2.10), we integrate over $s(x_1,t_1)$ to obtain

$$-\int_M^{\bar{u}}\frac{f'}{\phi}ds=\int_{x_1}^{\bar{x}}\frac{dx}{x+\delta}\leq \ln\bar{x}/x_1,$$

where $\bar{u}=u(\bar{x},\bar{t})$. We note that $f'>0$ implies that $0<k=\min \operatorname{supp} u \leq x(0)=x_0 \leq x_1$ and since $\phi<0$ for $u \geq M$ it follows that $\bar{u} \geq M$, hence

$$\alpha(\bar{u}-M) \leq -\int_M^{\bar{u}} f'/\phi \, ds \leq \ln \bar{x}/k,$$

and

$$\bar{u} \leq M+\frac{1}{\alpha} \ln \bar{x}/k.$$

If $u(x_1,t_1)=-M$ the proof is analogous and will be omitted.

Case ii. By uniqueness of solutions for ordinary differential equations it follows that $u(x,t)$ is zero in $C_{\bar{x},\bar{t}}$. Since $(\bar{x},\bar{t})$ was an ordinary point in $[0,\infty)\times[0,\infty)$ the proof of the theorem is complete.

For incoming waves, i.e. $f'<0$, the following result holds.

THEOREM 2.3. *Suppose $f'<0$ and $f''\neq 0$. Let $u_0(x)$ be a smooth function with compact support. If there exist constants M and α such that for $|u|\geq M$, $u\phi(u)<0$ and $(f'(u)/\phi(u))u/|u|\geq\alpha>0$ then there exists a sequence u_δ of solutions of* (2.2) *with initial data $u_0(x)$ such that for any $(x,t)\in R^+\times R^+$.*

$$|u_\delta(x,t)| \leq \mathrm{const}(1+|\ln x|)$$

where the constant is independent of δ.

Proof. Two cases need to be considered.

Case 1. The intersection of $C_{x,t}$ with the support of $u_0(x)$ is empty. Here as in the former theorem the local uniform bound on the u_δ is a consequence of the uniqueness of ordinary differential equations.

Case 2. The intersection of $C_{x,t}$ with the support of $u(x)$ is nonempty. The bound is obtained integrating over classical backward characteristics. The analysis is the same as in Case i of Theorem 2.2 and will be omitted.

Remark. By Lemmas 2.1 and 2.2, if $\phi(0)=0$ and $u_0(x)$ is either nonnegative or nonpositive and if the hypotheses on ϕ and f' hold for either $u\geq M$ or $u\leq -M$ respectively, then the conclusions of Theorem (2.2) and (2.3) hold.

Example 2.1. Let

$$\phi(u)=c_0 u, \qquad f'(u)=c_1 u+c_2.$$

This is a model situation considered by Whitham [16].

Case (i). $c_0>0$. Theorem 2.1 and Corollary 2.3 apply with arbitrary c_1 and c_2 yielding uniform bounds on the solutions u_δ^ε of (2.1) and u_δ of (2.2).

Case (ii). $c_0<0$, $c_1>0$, $c_2>0$. For positive data the characteristics corresponding to

$$u_t+(c_1 u+c_2)u_x+\frac{c_0 u}{x+\delta}=0$$

are all outgoing since $c_1 u+c_2>0$. Since $\lim_{u\to\infty} f'(u)/\phi(u)=c_1/c_0<0$, Theorem 2.2 applies.

Case (iii). $c_0<0$, $c_1>0$, $c_2<0$. Here $\lim_{u\to-\infty} f'(u)/\phi(u)=c_1/c_0<0$. For negative data the characteristics are all incoming, i.e. $c_1 u+c_2<0$, and Theorem 2.3 applies.

We now consider solutions of (2.2) where the speed of propagation f' does not have a definite sign for all values of u. In this case there are waves intersecting the t-axis and, no boundary data can be imposed at $x=0$. In order to bound solutions on backward characteristics intersecting the t-axis we suppose that the speed of propagation satisfies an appropriate sign condition at infinity.

THEOREM 2.4. *Let $u_\delta \in L^\infty(R^+ \times R^+)$ be a sequence of weak solutions of* (2.2) *with smooth initial data $u_0(x)$ which has compact support contained in* $(0,\infty)$. *If there exist nonnegative constants M, c_0, c_1 and r such that for $|u| \geq M$, $|\phi(u)| \leq c_0 u$ and $f'(u) \leq -c_1 |u|^r$, then for all $T>0$ and $N>0$*

$$\operatorname*{ess\,sup}_{(x,t)\in[0,N]\times[0,T]} \left| x^{c_0/c_1} u_\delta(x,t) \right| \leq \text{const},$$

where the constant is independent of δ.

Proof. It will be supposed that $u\phi(u)<0$ for $|u|\geq M$, otherwise the result follows from Theorem 2.1. First energy methods and Gronwall's inequality will be used to show that for all $s \geq 1$

$$\int_0^\infty \frac{\psi_p(x)|u_\delta+1|^{2s}}{2s} ds \leq c(s,\delta),$$

where $c(s,\delta)$ are constants such that $\lim_{s\to\infty} c(s,\delta))^{1/2s}$ is finite and independent of δ. Let $\psi_p(x)=x^p$ if $x\leq N$, $\psi_p(x)=N^{p+3}/x^2$ if $x \geq N$, p will be specified below. For notational convenience $\psi=\psi_p$ and $u=u_\delta$ will be used. Multiplying equation (2.2) by $\psi(x)(u+1)^{2s-1}$ and integrating in space and time yields

$$\int_0^\infty \psi(x)\frac{|u+1|^{2s}}{2s} dx = -\int_0^T\int_0^\infty \psi(x)(u+1)^{2s-1} f'(u)u_x \, dx\, dt$$
$$-\int_0^T\int_0^\infty \frac{\psi(x)\phi(u)(u+1)^{2s-1}}{x+\delta} dx\, dt + \int_0^\infty \frac{\psi(x)|u_0+1|^{2s}}{2s} dx.$$

Letting $q(u)=\int_0^u f'(y)(y+1)^{2s-1}dy$, we can rewrite the last equation as

(2.11)

$$\int_0^\infty \frac{\psi(x)|u+1|^{2s}}{2s} dx = -\int_0^T\int_0^\infty (\psi(x)q(u))_x dx\, dt + \int_0^T\int_0^\infty \psi'(x)q(u)\, dx\, dt$$
$$-\int_0^T\int_0^\infty \frac{\phi(u)\psi(x)(u+1)^{2s-1}}{x+\delta} dx\, dt + \int_0^\infty \frac{\psi(x)|u_0+1|^{2s}}{2s} dx$$
$$= I_1 + I_2 + I_3 + \int_0^\infty \frac{\psi(x)|u_0+1|^{2s}}{2s} dx.$$

Since $\psi(0)=0$ and $\lim_{x\to\infty}\psi(x)=0$ and since for each fixed δ $\lim q(u_\delta(x,t))=0$, we have that $\lim_{x\to\infty}\psi(x)q(u_\delta(x,t))=\lim_{x\to 0}\psi(x)q(u_\delta(x,t))=0$, hence I_1 vanishes. The plan now is to obtain the appropriate bounds on the integrals I_2 and I_3 and use Gronwall's inequality. The first estimates that will be obtained are for the restriction of the integrals I_2 and I_3 to $[N_\delta,\infty]\times[0,T]$ where N_δ is chosen so that $N_\delta \geq N$ and $|u_\delta(x,t)| \leq M$ for $x \geq N_\delta$ and $t\in[0,T]$. Without loss of generality it will be supposed that

$N\geq 3$, hence for $x\geq N$ $|\psi'(x)|=|3N^{p+1}/x^3|\leq\psi(x)$ and

$$\begin{aligned}(2.12)\qquad \int_0^T\int_{N_\delta}^\infty \psi(x)'q(u)\,dx\,dt&\leq k_1\int_0^T\int_{N_\delta}^\infty\psi(x)\int_0^u(y+1)^{2s-1}dy\,dx\,dt\\ &\leq k_1\int_0^T\int_0^\infty\psi(x)\frac{|u+1|^{2s}}{2s}\,dx\,dt\end{aligned}$$

where $k_1=\sup_{|u|\leq M}|f'(u)|$. The estimate of I_3 over $[N_\delta,\infty]\times[0,T]$ is obtained as follows:

$$(2.13)\qquad \int_0^T\int_{N_\delta}^\infty\psi(x)\frac{\phi(x)(u+1)^{2s-1}}{x+\delta}\,dx\,dt\leq k_2\int_0^T\int_0^\infty\psi(x)|u+1|^{2s}dx\,dt,$$

where $k_2=\sup_{u\leq M}|\phi(u)|$. In order to estimate I_2 and I_3 over $[0,N_\delta]\times[0,T]$ some further subdivision of the domain of integration will be needed. In what follows k will denote any constant which is independent of δ,p,q,T and N. Let

$$\begin{aligned}&A_M=\{(x,t)\colon\ |u(x,t)|\leq M\},\\ &A_M^+=A_m\cap[0,N]\times[0,T]\quad\text{and}\quad A_M^-=A_M^c\cap[0,N]\times[0,T].\end{aligned}$$

Then

$$\begin{aligned}(2.14)\qquad \iint_{A_M^+}\psi'(x)q(u)-\frac{\phi(u)\psi(x)(u+1)^{2s-1}}{x+\delta}\,dx\,dt&\\ \leq k(M+1)^{2s}\int_0^T\int_0^N p\frac{x^{p-1}}{2s}+\frac{x^p}{x+\delta}\,dx\,dt&\\ \leq kM^{2s}N^p(1/2s+1/p)T.&\end{aligned}$$

Noting that

$$\begin{aligned}\iint_{A_M^-}\psi'(x)q(u)\,dx\,dt&=\iint_{A_M^-}px^{p-1}\int_0^u f'(y)y^{2s-1}dy\,dx\,dt\\ &\leq K\frac{N^p}{2s}(M+1)^{2s+r}T-c_1\iint_{A_M^-}px^{p-1}\frac{u^{2s+r}}{2s+r}\,dx\,dt,\end{aligned}$$

it follows that

$$\begin{aligned}&\iint_{A_M^-}\left(\psi'(x)q(u)-\frac{\phi(u)(u+1)^{2s-1}}{x+\delta}\psi(x)\right)dx\,dt\\ &\leq\frac{KN^p(M+1)^{2s+r}}{2s}T+\iint_{A_M^-}x^{p-1}\left(-pc_1\frac{(u+1)^{2s+r}}{2s+r}+c_0|u+1|^{2s}\right)dx\,dt\\ &\qquad+\iint x^{p-1}\frac{\delta}{x+\delta}\phi(u)u^{2s-1}\,dx\,dt.\end{aligned}$$

Since $u\phi(u)<0$ the last integral is negative, hence it will be sufficient to bound the first integral. For this let $p=(2s+r)c_0/c_1$; then

$$(2.15)\qquad \iint_{A_M^-}\left[\psi'(x)q(u)-\frac{\phi(u)(u+1)^{2s-1}}{x+\delta}\right]dx\,dt\leq kN^p\frac{(M+1)^{2s+r}}{2s}.$$

To estimate I_2 and I_3 over $[N, N_\delta]\times[0,T]$ let $B_M^+ = A_M \cap [N, N_\delta]\times[0,T]$ and $B_M^- = A_M^c \cap [N, N_\delta]\times[0,T]$; then

$$(2.16)\quad \iint_{B_M^+}\left[\psi'(x)q(u) - \frac{\phi(u)(u+1)^{2s-1}}{x+\delta}\psi(x)\right]dx\,dt \leq K\int_0^T\int_0^\infty \psi(x)|u+1|^{2s}dx\,dt$$

and

$$\iint_{B_M^-}\psi'(x)q(u)\,dx\,dt \leq KN^{p+2}(N_\delta - N)\,\frac{(M+1)^{2s}}{2s}T + \iint_{B_M^-}\psi(x)\frac{(u+1)^{2s+r}}{2s+r}.$$

Hence

(2.17)

$$\iint_{B_M^-}\psi'(x)q(u)\,dx\,dt \leq KN^{p+2}\,\frac{(M+1)^{2s}}{2s}N^\delta T + |1+u_\delta|_\infty^r \int_0^T\int_0^\infty \psi(x)\,\frac{|u+1|^{2s}}{2s}dx\,dt.$$

And finally

$$(2.18)\qquad \iint_{B_M^-}\frac{\phi(u)\psi(x)}{x+\delta}(u+1)^{2s-1}dx\,dt \leq c_0\int_0^T\int_0^\infty \psi(x)|u+1|^{2s}dx\,dt.$$

It is in this last step where the linearity of $\phi(u)$ for large u is needed. Combining inequalities (2.11) through (2.18) yields

$$\int_0^\infty \psi(x)\frac{|u+1|^{2s}}{2s}dx \leq \int_0^\infty \psi(x)\frac{|u_0+1|^{2s}}{2s}dx + \frac{KN_\delta T}{2s}[(1+M)N]^{2s}$$
$$+|1+u_\delta|_\infty^r\int_0^T\int_0^\infty \frac{\psi(x)|u+1|^{2s}}{2s}(1+c_0 2s)\,dx\,dt.$$

By Gronwall's inequality it follows that

$$\int_0^\infty \frac{\psi(x)(u+1)^{2s}}{2s}dx \leq \frac{KN_\delta T}{2s}[(1+M)N]^{2s}\int_0^\infty \frac{\psi(x)|u_0+1|^{2s}}{2s}dx$$
$$\cdot\exp T\Big[|1+u_\delta|_\delta^r(1+c_0 2s)\Big].$$

Taking the $2s$ root on both sides of the last expression and letting $s\to\infty$ yields

$$\lim_{s\to\infty}\left(\int_0^\infty \psi(x)|u+1|^{2s}\right)^{1/2s} \leq [(1+M)N\exp c_0 T]\lim_{s\to\infty}\left[\int_0^\infty \psi(x)|u_0+1|^{2s}dx\right]^{1/2s}.$$

Since for all $x\leq N$, $\psi(x) = x^p = x^{c_0/c_1(2s+r)}$ the last inequality implies that

$$\operatorname*{ess\,sup}_{(x,t)\in[0,N]\times[0,T]} |x^{c_0/c_1}u| \leq 2|u_0|_\infty N^{1+c_0/c_1}(1+M)\exp c_0 T.$$

Theorem 2.4 applies in the following examples.

Example 2.2. Let f and ϕ be smooth functions which satisfy for $|u|\geq N$, $f'(u) = -u^{2s}$ some $s\geq 1$ and $\phi(u) = c_0 u$, $c_0 < 0$.

Example 2.3. Let f and ϕ be smooth functions such that for $|u|\geq N, f'(u)=c_1|u|+c$ and $\phi(u)=c_0u$ where $c_0<0$, $c_1<0$ and c_2 is arbitrary.

It is worth noting that this last example includes the case $\phi(u)=c_0u, f'(u)=c_1u+c_2$, $c_0<0$, $c_1<0$ for u positive. Whitham [15] considers this case as a model for spherical gases.

3. Existence of solutions. In this section we establish the existence of solutions to the initial value problem

$$u_t+f(u)_x+\frac{\phi(u)}{x}=0, \qquad u(x,0)=u_0(x), \tag{3.1}$$

where the behavior of f, ϕ and u_0 is described in the theorems of §2. A weak solution of (3.1) will be obtained as the limit of solutions u_δ^ε of (2.1) as ε and $\delta\to0$ or as the limit of solutions u_δ of (2.2) as $\delta\to0$. We note that in general it is necessary to establish strong convergence since nonlinear maps are generally not continuous in the weak topology. That is if u_δ converges weakly to $\bar{u}, f(u_\delta)$ need not to converge to $f(\bar{u})$. The classical approach to problems of this type is to obtain uniform estimates on the amplitude and on the derivatives of the solutions in an appropriate norm, then appeal to some standard compactness argument in order to pass to the limit, i.e. extraci a subsequence of solutions which converge strongly to the solution of the limiting equation. The approach here will be to use a priori bounds on the amplitude of the solutions together with results of the theory of compensated compactness [12]. No uniform gradient estimates will be required. We first state some results in measure theory and in the theory of compensated compactness. The first result characterizes composite weak limits in terms of expected probability measures. For the proof we refer the reader to [12, p. 147] and [1].

THEOREM 3.1. *Let $u^\varepsilon(x)\colon R^n\to R^m$ be a family of functions such that $|u^\varepsilon|_\infty\leq M$. There exists a subsequence u^{ε_k} and as associated family of probability measures $\{\nu_x(\lambda), x\in R^n, \lambda\in R^m\}$ with compact support, depending measurably on x such that for all continuous functions $g\colon R^n\to R$*

$$\lim g(u^{\varepsilon_k})(x)=\langle \nu_x,g\rangle=\int_{R^n} g(\lambda)\,d\nu_x(\lambda) \quad a.e.\ in\ R^n,$$

where the limit is taken in the weak-star topology of L^∞.

From Theorem 3.1 it follows that strong convergence corresponds to the statement that ν_y is a point mass a.e. [12, p. 154]:

COROLLARY 3.1. *Let u^ε converge to u in the weak-star topology of L^∞. Then u^ε converges strongly to u in L^p, $1\leq p<\infty$ if and only if $\nu_x=\delta_{u(x)}$ for almost all x.*

Proof. If u^ε converges strongly to u by the last theorem, we have

$$F(u)(x)=\lim_{\varepsilon\to0}F(u^\varepsilon)(x)=\langle \nu_x, F(\lambda)\rangle$$

for all continuous F, which shows that $\nu_x=\delta_{u(x)}$. For the converse we note that if $\nu_x=\delta_{u(x)}$ then

$$u^\varepsilon\to u \quad \text{and} \quad (u^\varepsilon)^2\to u^2 \quad \text{in } L^\infty \text{ weak star.}$$

From Theorem 3.1 it follows that the deviation between weak and strong convergence is measured by the spreading of the support of ν_x [3]. That is, let g be Lipschitz continuous; then

$$|g(\lim u^{\varepsilon}(x)-\lim g(u^{\varepsilon}(x)))|=\left|\int[g(u)-g(\lambda)]\,d\nu_x(\lambda)\right|$$

$$\leq \sup_{\lambda\in \operatorname{supp}\nu_x}|g(u)-g(\lambda)|\leq K\sup_{\lambda}|u(x)-\lambda|$$

$$\leq K \operatorname{diam}\overline{\operatorname{conv hull supp}}\,\nu_x,$$

where K is the Lipschitz constant of g. These remarks imply that it will be sufficient to show the reduction property

$$\nu_x(\lambda)=\delta_{u(x)}, \tag{3.2}$$

where ν_x are the probability measures associated to the solutions u_δ of (2.2). To pass to the limit for solutions of (2.1) we reduce the problem to finding convergent subsequences of (2.2) by letting first ε go to zero for fixed δ. The next lemma by Murat and Tartar describes a certain nonlinear function which is continuous under weak limits, it says that given two vector sequences bounded in L^2, if the rotation of one sequence and the expansion of the other are controlled then the inner product is continuous. More precisely

LEMMA 3.1. *Let p_n and q_n be two sequences uniformly bounded in $(L^2)^N$. If* $\operatorname{div} p_n$ *lies in a compact set of H^{-1}_{loc} and* $\operatorname{curl} q_n$ *lies in a compact set of H^{-1}_{loc}, then there exist subsequences p_{n_k} and q_{n_k} such that*

$p_{n_k}\cdot q_{n_k}\to p\cdot q$ in the sense of distributions where $p_{n_k}\to p$, $q_{n_k}\to q$ in $(L^2)^N$ weak.

Proof. (For details see [6], [12, Example 3, p. 167 and p. 179]). Roughly speaking, by Plancherel's formula it suffices to show that

$$\hat{p}_{n_k}\cdot\hat{q}_{n_k}-\hat{p}\cdot\hat{q}\to 0. \tag{3.3}$$

The weak convergence of the p_{n_k} and the q_{n_k} insures the convergence of (3.3) on compact sets and the differential constraints given in the hypothesis guarantee that the Fourier transform is small at infinity.

The program now is to use the results of Theorem 3.1, Corollary 3.1 and Lemma 3.1 together with the notion of entropy, cf. Definition 2.1, in order to show the reduction property (3.2) for the probability measures associated with the sequences of solutions of (2.2). In the future for notational convenience we shall write $\phi_\varepsilon\in H_c^{-1}$ and $\phi_\varepsilon\in L_b^1$ to indicate that ϕ_ε is a sequence of distributions which lie in a compact set of H^{-1}_{loc} or which lie in a bounded set of L^1_{loc}, respectively, and $\mathfrak{B}\mathfrak{M}$ will denote a set of bounded measures. The notation $\phi_\varepsilon\in H_c^{-1}+\mathfrak{B}\mathfrak{M}$ will be used to indicate that $\phi_\varepsilon=\phi_{\varepsilon_1}+\phi_{\varepsilon_2}$ where $\phi_{\varepsilon_1}\in H_c^{-1}$ and $\phi_{\varepsilon_2}\in\mathfrak{B}\mathfrak{M}$. The next theorem due to Tartar makes the connection between results on compensated compactness and the weak limits of solutions to conservation laws. It says that if a sequence of approximate solutions of

$$u_t+f(u)_x=0 \tag{3.4}$$

has the correct entropy production it will converge to an exact solution of (3.4). More precisely

THEOREM 3.2. ([12, p. 200]). *Let $f\in C^1$. If $u^\varepsilon=u^\varepsilon(x,t)$ is a sequence of approximate solutions of* (3.4) *uniformly bounded in L^∞ which satisfy the entropy condition*

$$\frac{\partial}{\partial t}\eta(u^\varepsilon)+\frac{\partial}{\partial x}q(u^\varepsilon)\in H_c^{-1}+\mathfrak{B}\mathfrak{M} \tag{3.5}$$

for all entropy pairs (η, q) *with* η *convex, then there exists a subsequence* u^{ε_k} *such that* u^{ε_k} *converges in the weak star topology of* L^∞ *to an exact solution* $\bar{u} \in C^0(0,T; H^{-1}_{\text{loc}}(R))$ *of* (3.4) *and*

i) $f(u^{\varepsilon_k}) \to f(\bar{u})$ *in the weak topology of* L^∞,

ii) $\bar{u}$ *satisfies Lax's entropy condition,*

iii) *if* $f''>0$ (*or* $f''<0$) *then* $u^{\varepsilon_k} \to \bar{u}$ *in* L^p, $1 \le p < \infty$.

Proof. Only the main ideas will be presented. For details we refer the reader to [12, p. 200]. Let (η_1, q_1) and (η_2, q_2) be two entropy pairs with η_1 and η_2 convex; by condition (3.5) we have

$$\operatorname{div}(\eta_1, q_1) = \eta_{1_t} + q_{1_x} \in H_c^{-1} + \mathcal{B}\mathcal{M},$$
$$\operatorname{curl}(-q_2, \eta_2) = \eta_{2_t} + q_{2_x} \in H_c^{-1} + \mathcal{B}\mathcal{M}.$$

In order to apply Lemma 3.1 the following result by Murat is needed. The proof can be found in [8] or [12].

LEMMA 3.2. *If* $g^\varepsilon \in H_c^{-1} + \mathcal{B}\mathcal{M}$ *and if* g^ε *lies in a bounded set of* $W^{-1,\infty}$ *then* $g^\varepsilon \in H_c^{-1}$.

By hypothesis $\eta_{i_t} + q_{i_x}$, $i=1,2$ lie in a bounded set of $W^{-1,\infty}$ hence Lemma 3.2 and condition (3.5) yield $\operatorname{div}(\eta_1, q_1) \in H_c^{-1}$ and $\operatorname{curl}(-q_2, \eta_2) \in H_c^{-1}$. By Lemma 3.1 it follows that

$$\langle \nu_y, \eta_1 q_2 - \eta_2 q_1 \rangle = \langle \nu_y, \eta_1 \rangle \langle \nu_y, q_2 \rangle - \langle \nu_y, \eta_2 \rangle \langle \nu_y, q_2 \rangle$$

where ν_y is the family of probability measures associated to u^ε. The last equation states that there exists a bilinear form B which commutes with ν_y, i.e.

$$B \circ \nu_y = \nu_y \circ B.$$

Under this condition Tartar shows that the reduction property (3.2) holds, i.e. ν_y reduces to a point mass ([12, p. 204–207]) if f is not affine on any interval.

THEOREM 3.3. *Let* ϕ, f *and* u_0 *be smooth functions which satisfy the conditions of Theorem* 2.1, *Corollary* 2.1 *or Corollary* 2.2. *If* u_δ^ε *is a sequence of solutions of* (2.1) *with initial and boundary data* $u_\delta^\varepsilon(x,0) = u_0(x)$, $u_\delta^\varepsilon(0,t) = 0$ *then there exists a subsequence* u_{μ_k} *which converges to* $\bar{u}$ *in the* L^∞ *weak star topology and*

i) $f(u_{\mu_k}) \to f(\bar{u})$ *in the* L^∞ *weak star topology,*

ii) $\bar{u}$ *is a weak solution of the Cauchy problem for the singular conservation law* (3.1) *with initial data* $u_0(x)$ *and* $\bar{u}$ *satisfies Lax's entropy condition,*

iii) *if* $f'' \ne 0$ *then* $u_{\mu_k} \to \bar{u}$ *in* $L^p([0,T] \times (\alpha, \beta))$, $0 < \alpha < \beta < \infty$, $p < \infty$.

Proof. The conclusions of Theorem 2.1 and Corollaries 2.1 and 2.3 imply that the solutions u_δ^ε have uniform a priori bounds in L^∞. In order to apply Theorem 3.2 we need to establish entropy condition (3.5). Multiplying both sides of equation (2.1) by η', we obtain

$$\eta(u_\delta^\varepsilon)_t + q(u_\delta^\varepsilon)_x + \eta'(u_\delta^\varepsilon) \frac{\phi(u_\delta^\varepsilon)}{x+\delta} = \varepsilon \eta'(u_\delta^\varepsilon)(u_\delta^\varepsilon)_{xx}.$$

Since $\varepsilon \eta'(u_\delta^\varepsilon) u^\varepsilon_{\delta xx} = \varepsilon \eta(u_\delta^\varepsilon)_{xx} - \varepsilon \eta''(u_\delta^\varepsilon)(u_\delta^\varepsilon)_x^2$ and $\eta'' \ge 0$ we have that

$$\eta(u_\delta^\varepsilon)_t + q(u_\delta^\varepsilon)_x + \frac{\eta'(u_\delta^\varepsilon)\phi(u_\delta^\varepsilon)}{x+\delta} \le \varepsilon \eta(u_\delta^\varepsilon)_{xx}.$$

Letting ε go to zero yields

$$\eta(u_\delta)_t+q(u_\delta)_x+\frac{\eta'(u_\delta)\phi(u_\delta)}{x+\delta}\le 0 \tag{3.6}$$

where u_δ is a distributional solution in L^∞ of (2.2). Let $\theta=\eta_t+q_x$; then for any set E contained in $[\alpha,\beta]\times(0,T)$ $0<\alpha<\beta<\infty$

$$|\theta(E)|\le K\int_E|\chi_t|+|\chi_x|\,d\theta\le K_1 \tag{3.7}$$

where χ is a smooth positive function which is 1 in E and zero outside a neighborhood of E and has bounded gradient. We note that by (3.6), $\eta_t+q_x+\eta'\phi/(x+\delta)$ is a nonpositive measure, hence by (3.7) it is sufficient to show that $\eta'\phi/(x+\delta)$ is a bounded measure to insure that the entropy condition (3.5) holds. Since u_δ was bounded in L^∞_{loc} (cf. §2) $\eta'(u_\delta)\phi(u_\delta)/(x+\delta)$ is contained boundedly in L^1_{loc} and hence is a bounded measure. Theorem 3.2 now insures the existence of a subsequence u_{μ_k} of solutions of (2.1) such that $u_{\mu_k}\to\bar u$ and $f(u_{\mu_k})\to f(\bar u)$ in L^∞_{loc} weak star and $\bar u$ is a solution of (3.1). Moreover, if $f''\neq 0$ the sequence u_{μ_k} converges strongly in $L^p_{\text{loc}}(R^+\times R^+)p<\infty$. Since the initial data $u_0(x)$ is taken on weakly by the solutions u_δ of (2.2) we have that for any test function ψ

$$\int\psi(\bar u(x,t)-u_0(x))\,dx\le\int\psi(\bar u(x,t)-u_{\mu_k}(x,t))\,dx$$
$$+\int\phi(u_{\mu_k}(x,t)-u(x,0))\,dx\le\varepsilon.$$

It also follows that the initial data u_0 is taken on weakly by the solution $\bar u$.

Corollary 3.3. *Let f,ϕ and u_0 be smooth functions which satisfy the conditions described in Theorems* 2.2, 2.3 *or* 2.4. *If u_δ is a sequence of solutions of* (2.2) *with initial data $u_0(x)$ then there exists a subsequence u_{δ_k} which converges in the weak star topology of L^∞ to $\bar u$ and the conclusions* i), ii), iii) *of Theorem* 3.3 *hold if we replace the sequence u_{μ_k} by u_{δ_k}.*

Proof. Since the u_δ are obtained as the limit of a subsequence of solutions u^ε_δ of (2.1) the proof follows the same steps of Theorem 3.3 and will be omitted.

REFERENCES

[1] B. Dacorogna, *Weak continuity and weak lowe, semicontinuity of nonlinear functions*, Lecture Notes in Mathematics 922, Springer-Verlag, New York,

[2] C. Dafermos, *Characteristics in hyperbolic conservation laws. A study of the structure and the asymptotic behavior of solutions*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium Vol. 1, Research Notes in Mathematics 17, R. J. Knops, ed., Pitman, London, 1979.

[3] R. DiPerna, *Convergence of approximate solutions to conservation laws*, Arch., Rat. Mech. Anal., to appear.

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

[5] T. P. Liu, *Quasilinear hyperbolic systems*, Comm. Math. Phys. 68 (1979), pp. 141–172.

[6] F. Murat, *Compacité par compensation*, Ann. Scuola Norm. Sup. Pisa Sci. Fis. Math., 5 (1978), pp. 489-507.

[7] ———, *Compacité par compensation: Condition necessaire et suffisante de continuité faible sous une hypothèse de rang constant*, Ann. Scuola Norm. Sup. Pisa 8 (1981), pp. 69–102.

[8] ———, *L'injection du cone positif de H^{-1} dans $W^{-1,q}$ est compacte pour tout $q<2$*, preprint.

[9] I. P. Natanson, *Theory of Functions of a Real Variable*, Ungar, New York, 1955.

[10] O. A. Oleinik, *Discontinuous solutions of nonlinear differential equations*, AMS Translations, Series 2, Vol. 26, pp.95–170.

[11] O. A. Oleinik and T. D. Ventcel, *The first boundary value problem and the Cauchy problem for quasilinear equations of parabolic type*, Mat. Sb. (N.S.), 41 (83) 1957, pp. 105–128.

[12] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Vol. IV, Research Notes in Mathematics 39, R. J. Knops, ed., Pitman, London, 1979.

[13] ______, *Une nouvelle méthode de résolution d'équations aux dérivées partielles nonlinéares*, Lecture Notes in Mathematics 665, Springer-Verlag, New York, 1977, pp. 228–241.

[14] A. I. Vol'pert, *The spaces BV and quasilinear equations*, Mat. Sb. 75 (115) (1967), No. 2.

[15] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley Interscience, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics

008

SOME PERTURBATION RESULTS AND THEIR APPLICATIONS TO STABILIZATION OF HYPERBOLIC SYSTEMS*

P. J. McKENNA[†] AND I. LASIECKA[†]

Abstract. Real selfadjoint perturbations B of a real symmetric operator T on $L_2(\Omega)$ are considered. Necessary and sufficient conditions for the spectrum of $T+B$ to remain real are given. This perturbation result is used to solve certain stabilization problems for hyperbolic systems.

1. Introduction. When solving differential equations for vibrating systems of the form

$$u_{tt}+Au=0, \tag{1}$$

where A is an $n\times n$ matrix, a necessary condition for stability of the solution is that all eigenvalues of A be real. Indeed, if λ_j is the eigenvalue corresponding to the eigenvector v_j, then $e^{\pm i\lambda_j t}v_j$ is a solution of (1). Thus if λ_j has nonzero imaginary part, the solution will grow exponentially.

In this paper, we consider infinite-dimensional versions of (1), where the operator $A=T+B$ with T real symmetric on $L^2(\Omega)$ and B a small but *nonselfadjoint* real perturbation. General perturbation theory predicts that if B is small, then the spectrum of B will be close to that of T. However, it does not tell us whether or not the spectrum remains real.

If T has an eigenvalue of multiplicity greater than or equal to two, then one can construct arbitrarily small perturbations B such that $T+B$ has complex spectrum. One simply chooses $B=\left[\begin{smallmatrix}0&-b\\ b&0\end{smallmatrix}\right]$ on a two-dimensional subspace of the relevant eigenspace. This works whether the space is finite- or infinite-dimensional. Clearly, this also works if a subsequence of eigenvalues becomes arbitrarily close, i.e., if $\lambda_{n_k}-\lambda_{n_{k-1}}\rightarrow 0$ for some subsequence $\{n_k\}$. Thus, in order that the spectrum of $T+B$, $\sigma(T+B)$, should remain real under arbitrarily small perturbations B, it is necessary that (a) the spectrum of T must consist *of eigenvalues λ_i of multiplicity one*, and (b) there exists $d>0$ such that $\lambda_{n+1}-\lambda_n\geq d>0$. In this paper, we show that these conditions are sufficient. This is done in §2.

In order to guarantee the stability of (1), one must show that the eigenvalues are real and in addition, that the corresponding eigenvectors form a Riesz basis. By modifying arguments given in [1], we show that this is true under the same hypotheses as before, namely eigenvalues of multiplicity one and the gap condition. This is done in §3. In §4, we apply these theorems to some problems in stabilization and in §5, we give examples of obvious physical importance, where the gap condition is satisfied.

Throughout the paper, Ω will be a smooth bounded region in R^n with smooth boundary. The span of a finite number of vectors $h_1,\cdots,h_n$ is denoted $\{h_1,\cdots,h_n\}$.

2. The main theorems. The main tool used in the sequel is the following technical lemma. Here H is a *real* Hilbert space.

LEMMA 1. *Let T: $H\rightarrow H$ be a bounded linear operator having eigenvalue* 1 *with eigenvector $h_1\in H$. Let $H=\{h_j\}\oplus H_2$, where $h_1\perp H_2$, and assume that $\|T|_{H_2}\|<1$. Let $T_2=T|_{H_2}$ and let $0<\varepsilon<(1-\|T_2\|)/4$. Then if $\|B\|<\varepsilon$, the operator $T+B$ has a real*

*Received by the editors February 14, 1983.

[†] Department of Mathematics, University of Florida, Gainesville, Florida 32611.

eigenvalue λ_1 *satisfying* $|1-\lambda|<2\varepsilon$. *Furthermore, if* λ_2 *is any other eigenvalue of* T *then* $|\lambda_2-\|T_2\||<2\varepsilon$.

Proof. We write

$$B=\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

where

$$B_{11}:\{h_1\}\to\{h_1\},\quad B_{12}:H_2\to\{h_1\},$$
$$B_{21}:\{h_1\}\to H_2,\quad B_{22}:H_2\to H_2.$$

We define a (nonlinear) map on the space

$$B=\{u\in\mathfrak{L}(\{h_1\},H):\|u\|\le 1\}\quad\text{by}$$
$$F(u)(y)=(T_2u+B_{21}+B_{22}u)(I+B_{11}+B_{12}u)^{-1}(y).$$

Clearly Fu is a linear map with domain $\{h_1\}$ and range H_2. We verify that $\|\mathrm{F}u\|\le 1$. This is the case since $\|(Fu)y\|\le(\|T_2\|+2\varepsilon)(1-2\varepsilon)^{-1}\|y\|$ and since $\varepsilon<(1-\|T_2\|)/4$, this guarantees that $(\|T_1\|+2\varepsilon)(1-2\varepsilon)^{-1}\le 1$. Next we verify that F is a contraction on B. Let $y\in\{h\}$, $\|y\|=1$,

$$(Fu_1)y=(T_2u_1+B_{21}+B_{22}u_1)(I+B_{11}+B_{12}u_1)^{-1}y,$$
$$(Fu_2)y=(T_2u_2+B_{21}+B_{22}u_2)(I+B_{11}+B_{12}u_2)^{-1}y.$$

Then, an elementary calculation shows that

$$\|Fu_1-Fu_2\|\le\|T_2\|(1-\varepsilon)^{-2}\|u_2-u_1\|.$$

Thus F: $B\to B$ is a contraction and has a fixed point. Therefore, there exists u: $\{h\}\to H_2$ with $\|u\|\le 1$ and

$$y=(I+B_{11}+B_{12}u)(\tilde{y}),\qquad u(y)=(Au+B_{21}+B_{22}u)(\tilde{y}).$$

Without loss of generality, let $\|\tilde{y}\|=1$. Since u: $\{h\}\to H_2$ has rank one, it follows that $u(y)=u(\lambda\tilde{y})=\lambda h_2$ for some $h_2\in H_2$ with $\|h_2\|<1$. Thus we have

$$\begin{bmatrix}\lambda\tilde{y}\\ \lambda h_2\end{bmatrix}=\begin{bmatrix} I+B_{11} & B_{12} \\ B_{21} & T_2+B_{22}\end{bmatrix}\begin{bmatrix}\tilde{y}\\ h_2\end{bmatrix},$$

and we may conclude that the vector $\tilde{y}+h_2$ is an eigenvector of $T+B$ and λ is the corresponding eigenvalue. Since all the spaces in question are real, λ is real. Furthermore

$$|(1-\lambda)|=\|y-\tilde{y}\|=\|(B_{11}+B_{12}u)\tilde{y}\|<2\varepsilon.$$

This proves the existence of a real eigenvalue within 2ε of 1. Now suppose there exists $\tilde{\lambda}$ and h_2 such that

$$\tilde{\lambda}\begin{bmatrix}\tilde{y}\\ \tilde{h}_2\end{bmatrix}=\begin{bmatrix} I+B_{11} & B_{12} \\ B_{21} & T_2+B_{22}\end{bmatrix}\begin{bmatrix}\tilde{y}\\ \tilde{h}_2\end{bmatrix}.$$

Then we can check that

$$\|h_2\|\le\varepsilon|\tilde{\lambda}-\|T_2\|-\varepsilon|^{-1}.$$

In particular $|\lambda-\|T_2\||\geq 2\varepsilon\Rightarrow\|\tilde{h}\|\leq\frac{1}{2}$. This would allow us to define another linear map $\tilde{u}$ by $\tilde{u}(e\tilde{y})=c\tilde{h}_2$, with $\|\tilde{u}\|\leq\frac{1}{2}$. This would violate the uniqueness of the fixed point of F. Thus if $\tilde{\lambda}$ is another eigenvalue of $T+B$, we have

$$\lambda-\|T_2\|\leq 2\varepsilon.$$

This concluses the proof of the lemma.

We are now in a position to prove our first theorem.

THEOREM 1. *Let T be selfadjoint T: $H\to H$, where H is a real Hilbert space. Assume that $\sigma(T)=\{\lambda_i\}_{i=1}^{\infty}$ where $\lambda_i<\lambda_{i+1}\to\infty$, $\inf|\lambda_i-\lambda_j|\geq d>0$ and each eigenvalue is of multiplicity one. Then if $\|B\|<\frac{1}{24}d$, the spectrum of $T+B$ is real and consists of eigenvalues λ_i^* of multiplicity one satisfying $|\lambda-\tilde{\lambda}_i|\leq d/3$.*

Proof. We shall show that for each i, there exists precisely one real eigenvalue $\tilde{\lambda}_i$ such that $|\lambda_i-\tilde{\lambda}_i|\leq d/3$.

Let $k=\lambda_i-d/3$ and let $S=d/3(T-kI)^{-1}$. Note that $\|S\|=1$, 1 is an eigenvalue of S with associated eigenvector ϕ_i, the ith eigenvector of T. If $H_2=\{\phi_i\}^{\perp}$, we have $\|S|_{H_2}\|=(d/3)\sup|\lambda_i-\lambda_j-d/3|^{-1}\leq\frac{1}{2}$. Thus the hypotheses of the lemma apply to S. Note that λ^* is an eigenvalue of $T+B$ if and only if $(d/3)(\lambda^*-k)^{-1}$ is an eigenvector of $(d/3)(T+B-kI)^{-1}$. However,

$$\text{(2)}\quad \left\|\frac{d}{3}\left[(T+B-kI)^{-1}-(T-kI)^{-1}\right]\right\|=\left\|\frac{d}{3}\left[(T+B-kI)^{-1}B(T-kI)^{-1}\right]\right\|$$
$$\leq\frac{d}{3}\left(\frac{d}{3}-\|B\|\right)^{-1}\|B\|\frac{3}{d}=\left(\frac{d}{3}-\|B\|\right)^{-1}\|B\|.$$

Since $\|B\|((d/3)-\|B\|)^{-1}\leq\frac{1}{9}$ the hypotheses of the lemma apply with $\varepsilon'=\frac{1}{9}$. Then $(d/3)(T+B-kI)^{-1}$ has precisely one real eigenvalue $\tilde{\lambda}_i$ satisfying $|\tilde{\lambda}-1|<2\varepsilon'=\frac{2}{9}$.

Thus λ_i^*, defined by

$$\frac{d}{3}(\lambda_i^*-k)^{-1}=\tilde{\lambda}_i,$$

is an eigenvalue of $T+B$, and recalling that $k=\lambda_i-d/3$, an elementary calculation yields that

$$|\lambda_i^*-\lambda_i|<\frac{d}{3}\left(\frac{2\varepsilon'}{1-2\varepsilon'}\right)=\frac{2d}{21}.$$

THEOREM 2. *Let T be as in Theorem 1, and assume in addition that $\lambda_{i+1}-\lambda_i=d_i\to+\infty$ as $i\to\infty$. Then for any $B\in\mathfrak{L}(H)$, there exists an integer N such that if λ_i^* is an eigenvalue of $T+B$, and $i\geq N$, then λ_i^* has multiplicity one and is real.*

Proof. The proof is practically identical to that of Theorem 1 so we include only a sketch. The main point to observe is that in (2), the expression $(d_i/3-\|B\|)^{-1}\|B\|$ can be made arbitrarily small by choosing i sufficiently large. The rest of the calculations then go through as before, showing that close to each eigenvalue of T there is an eigenvalue of $T+B$ of multiplicity one.

THEOREM 3. *Let T be as in Theorem 1. Let B be a compact operator. Then if λ_i is an eigenvalue of T_1 there exists an integer N_B such that $i\geq N$ implies that near λ_i there is exactly one eigenvalue λ_i^* of $T+B$. Moreover, λ_i^* is real.*

Proof. The proof in this case is again similar to that of Theorem 1. The only change is to observe in (2) that if B is compact then the norm of $B(T-(\lambda_i-d/3))^{-1}$ may be made arbitrarily small by choosing i large enough. The rest of the theorem goes through as before.

The next theorem shows that the requirement that T be selfadjoint can be relaxed.

THEOREM 1′. *Let H be a real Hilbert space and let T: $H \to H$ be such that*:

(i) *The resolvent of T is compact.*

(ii) *$\sigma(T)=\{\lambda_i\}_{i=1}^{\infty}$, where $|\lambda_i-\lambda_j|\geq d>0$ and each λ_i is real of multiplicity one.*

(iii) *The corresponding eigenfunctions ϕ_i are similar to an orthonormal basis, i.e., the eigenprojections P_i corresponding to eigenfunctions ϕ_i and eigenprojections Q_i corresponding to some orthonormal basis Ψ_i are related to each other by the similarity transformations $P_i=W^{-1}Q_iW$.*

Then there exists $\varepsilon>0$ such that if $\|B\|<\varepsilon$, then $\sigma(T+B)$ is real.

Proof. Let $k=\lambda_i-d/\alpha$ for α a constant to be determined later. Let $S=(d/\alpha)(T-kI)^{-1}$. Observe that 1 is an eigenvalue of S, and that if $H_2=\{\phi_i\}$, then for an appropriate choice of α, we have $\|S|_{H_2}\|<1$. Indeed, since $\Sigma_1^{\infty}|(x,\phi_i)|^2\leq C\|x\|^2$, we have

$$\left\|S|_{H_2}(x)\right\|\leq\frac{d}{\alpha}\,\frac{\|W^{-1}\|\|W\|C\|x\|^2}{d|1-1/\alpha|}=\frac{\|W^{-1}\|\|W\|C\|x\|^2}{\alpha-1}.$$

Now choosing α so that $\|W^{-1}\|\|W\|C(\alpha-1)^{-1}<1$, we conclude that $\|S|_{H_2}\|<1$. Thus the hypotheses of Lemma 1 are satisfied.

Notice that λ^* is an eigenvalue of $T+B$ if and only if $(\lambda^*-k)^{-1}(d/\alpha)$ is an eigenvalue of $(d/\alpha)R(k,T+B)$. Since k is real, it is enough to show that $(d/\alpha)(\lambda^*-k)^{-1}$ is real. However,

$$\frac{d}{\alpha}R(k,T+B)=\frac{d}{\alpha}(T+B-kI)^{-1}=P+S,$$

where (as before)

$$\begin{aligned}\|P\|&=\left|\frac{d}{\alpha}\right|\left\|(T+B-kI)^{-1}B(T-kI)^{-1}\right\|\\&=\left|\frac{d}{\alpha}\right|\left\|(T-kI)^{-1}\left[I+B(T-kI)^{-1}\right]^{-1}B(T-kI)^{-1}\right\|\\&\leq\frac{d}{\alpha}\|B\|\|(T-kI)\|^{-1}\left(1-\|B\|\|(T-kI)^{-1}\|\right)^{-1}.\end{aligned}$$

Since,

$$\|(T-kI^{-1}\|\leq\frac{C\|W^{-1}\|\|W\|}{d(1-1/\alpha)}\leq\frac{C\|W^{-1}\|\|W\|}{(d/\alpha)(\alpha-1)}<\frac{\alpha}{d},$$

we have

$$\|P\|<\frac{d}{\alpha}\|B\|\frac{\alpha}{d}\,\frac{1}{1-\|B\|\alpha/d}=\frac{\|B\|}{1-\|B\|\alpha/d}.$$

In order to satisfy the assumptions of Lemma 1, we must require that

$$4\|B\|\left(1-\|B\|\frac{\alpha}{d}\right)^{-1}<\|W^{-1}\|\|W\|\frac{C}{\alpha-1}, \tag{3}$$

which is accomplished by choosing $\|B\|$ sufficiently small. By Lemma 1, we conclude that $(d/\alpha)(\lambda^*-k)^{-1}$ is real and therefore λ^* is real. This procedure, applied to each i, yields the result of Theorem 1′, namely that all eigenvalues are real.

3. A perturbation result in complex Hilbert space. In the applications of §2, the requirement that the spectrum of $T+B$ be real was merely *necessary* for stability. For sufficiency, we shall need another theorem which is similar to one appearing in the literature. Rather than repeat an entire proof on account of one small technical variation, we shall outline the main steps and refer to Kato [1] for the details.

THEOREM 4. *Let T be normal, with compact resolvent and simple eigenvalue λ_i, $|\lambda_i-\lambda_j|\geq d>0$. Let $B\in\mathfrak{L}(H)$ with $\|B\|\leq\varepsilon<d/2$. Then we have*

a) $\sigma(T+B)=\{\mu_k\}_1^\infty$, *satisfying* $|\lambda_k-\mu_k|<d/3$, $k=1,2,3,\cdots$.

b) *If Q_k are the eigenprojections of $T+B$ and P_k are the eigenprojections (corresponding to λ_k) of T, then there exists an invertible $W\in\mathfrak{L}(H)$ such that*

$$Q_k=W^{-1}P_kW.$$

Proof. The proof is similar to that of [1, Thm. 4.15a, p. 293]. We shall use the following facts (referenced if not elementary).

(i) T normal $\Rightarrow\|T\|=\operatorname{spr}(T)$ where $\operatorname{spr}(T)$ is the spectral radius of T.

(ii) $\operatorname{spr}(T-\xi I)^{-1}=\operatorname{dist}(\xi,\sigma(T))^{-1}$.

(iii) $\|(T-\lambda I)^{-1}\|=\operatorname{dist}(\lambda,\sigma(A))^{-1}$.

(iv) T normal and B bounded imply that if $|\lambda-\lambda_k|=d_k/2$. Then $\|(A+B-\lambda I)^{-1}\|\leq[(d_k/2)-\|B\|]^{-1}$.

(v) If $B(\lambda_i,d/3)=\{\lambda C\colon |\lambda-\lambda_i|\leq d/3\}$, and if $\|B\|<3/d$, then $B(\lambda_i,d/3)$ contains exactly one eigenvalue of $T+B$ of multiplicity 1 and no other points of $\sigma(T+B)$ see [1, Thms. 317, 318, p. 214]. We will use the following lemma due to Kato.

LEMMA 2. *Let $\{P_j\}_{j=0,1,\dots}$ be a complete family of orthogonal projections, and let $\{Q_j\}_{j=0,1,\dots}$ be a family of (not necessarily orthogonal) projections such that $Q_jQ_k=\delta_{jk}$, Q_j. Assume that*:

(i) $\dim p_0=\dim Q_0=m<\infty$,

(ii) $\sum_{j=1}^\infty\|P_j(Q_j-P_j)u\|^2\leq c\|u\|^2$, $c<1$.

Then there exists an invertible W such that $Q_j=W^{-1}P_jW_j$, $j=0,1,2,\cdots$.

Therefore in order to complete the proof, it is enough to establish the validity of (ii) (we take $P_0=0$) with P_j and Q_j as follows:

$$Q_j=\frac{1}{2\pi i}\int_{\partial B_j}R(\lambda,T+B)\,d\lambda,\qquad P_j=\frac{1}{2\pi i}\int_{\partial B_j}R(\lambda,T)\,d\lambda.$$

After setting:

$$Z_j=\frac{1}{2\pi i}\int_{\partial B_j}\frac{R(\lambda,T)}{\lambda-\lambda_j}d\lambda,\qquad Z_j'=\frac{1}{2\pi i}\int_{\partial B_j}\frac{R(\lambda,T+B)}{\lambda-\mu_j}d\lambda,$$

we have

$$Q_j-P_j=-Q_jBZ_j-Z_j'BP_j;\tag{4}$$

see [1, (4.38), p. 296]. We observe that since

$$Q_k=\frac{1}{2\pi i}\int_{\partial B_k}R(\lambda,T+B)\,d\lambda,$$

$$\|Q_k\|\leq\frac{1}{2\pi}\int_{\partial B_k}(d-\|B\|)\,d\lambda\leq C_1\tag{5}$$

for some C_1 independent of k. Furthermore,

$$\|Z_k'\| \le \frac{1}{2\pi} \int_{\partial B_k} \frac{\|R(\lambda, T+B)\|}{|\lambda - \mu_k|} d\lambda \le C_2 \tag{6}$$

for some C_2 independent of k. Since, by [1, p.40],

$$\sum_{k=1}^{\infty} \|Z_k u\|^2 = \sum_{k=1}^{\infty} \sum_{j \ne k} |\lambda_j - \lambda_k|^{-2} \|P_j u\|^2$$

and since

$$\sup_j \sum_{\substack{k=1\\ k \ne j}}^{\infty} |\lambda_j - \lambda_k|^{-2} \le \frac{1}{d^2} \sup \sum_{k=1}^{\infty} \left(\frac{1}{j-k} \right)^2 \le C_3$$

for some $C_3 > 0$, we have

$$\sum \|A_k u\|^2 \le C_3 \sum \|P_j u\|^2 = C_3 \|u\|^2. \tag{7}$$

Equations (3), (4) and (5) immediately give

$$\sum_{k=1}^{\infty} \|Q_k B Z_k u\|^2 \le C_4 \|B\|^2 \|u\|^2 \tag{8}$$

and

$$\sum_{k=1}^{\infty} \|Z_k' B P_k u\|^2 \le C_5 \|B\|^2 \|u\|^2 \tag{9}$$

for suitable positive constants C_4, C_5. This together with (4) completes the proof for suitably small $\|B\|$. Theorem 4 yields the following:

COROLLARY 1. *Let*

$$\hat{T} \triangleq \begin{bmatrix} T & 0 \\ 0 & M \end{bmatrix}.$$

Assume T: $H \to H$ satisfies the hypotheses of Theorem 4 with eigenvectors ϕ_1, ϕ_2, $\cdots$, and M stands for a bounded linear operator from $H_n \to H_n$, where H_n is n-dimensional real Hilbert space. Assume that all eigenvalues γ_i of M are distinct and different from the eigenvalues of T.

Then the assertions of Theorem 4 hold on $H \times H_n$. More precisely with $P_0(Q_0)$ standing for the eigenprojection of $T(T+B)$ corresponding to γ_i (perturbed γ_i), we have: $\exists W$ invertible such that

$$Q_k = W^{-1} P_k W, \qquad k = 0, 1, 2, \cdots.$$

To prove the corollary it is enough to apply Lemma 2 with Q_0 being defined as a total projection on perturbed eigenvalues γ_i. Notice that due to the completeness of the system P_i, $i = 0, 1, 2, \cdots$, on $H \times H_n$, by Corollary 1, the system $Q_i = i = 0, 1, \cdots$, is also complete (it is similar to P_i). Consequently the eigenvectors of $T+B$ constitute a Riesz basis in $H \times H_n$.

Remark 1. The conclusions of Theorem 1 may also be obtained by modifying and expanding Kato's perturbation theory [1, p. 293] on the complex plane. Indeed, by modifying Kato's argument along the lines of our proof of Theorem 4, we obtain the eigenvalues in small circles $B(\alpha/3,\lambda_r)$. Since A and B are real operators, one can immediately deduce that if $\tilde{\lambda}_r$ is an eigenvalue, so is the complex conjugate. By using the fact that in each circle there is one eigenvalue of multiplicity one we conclude that $\tilde{\lambda}_r$ must be real. This method relies heavily on the machinery of [1]. We feel our proof is more self-contained, and gives rise to an easier estimation of the constants, using as it does, only the constructive contraction fixed point theorem.

4. Applications.

4.1. Statement of results. In order to illustrate some of the applications of the previous perturbation results we present the following stabilization problem for a vibrating string.

Let Ω be a bounded open domain in R^n with a boundary Γ. Let A stand for a *selfadjoint* generator of a strongly continuous semigroup. Assume also that the resolvent of A is *compact*. Consider the following model

$$(10)\qquad \frac{d^2x(t)}{dt^2}=Ax(t)+g\langle x(t),w\rangle_{L_2(\Omega)},\qquad x(0)=x_0,\quad x_t(0)=x,$$

where the vectors g,w belong to $L_2(\Omega)$.

Remark 2. As a canonical example of A one can take:

$$Ax=A(\xi,\partial)x,\qquad x\in D(A)$$

with $A(\xi,\partial)$ formally a selfadjoint strongly elliptic operator and

$$D(A)=\left\{x\in L_2(\Omega),A(\xi,\partial)x\in L_2(\Omega),x|_\Gamma \text{ or } \left(\frac{\partial x}{\partial\eta}\Big|_\Gamma\right)=0\right\}.$$

It is well known that due to the compactness of the resolvent of A, the spectrum of A consists of a sequence of isolated eigenvalues $\{\lambda_i\}$, with $-\lambda_i\to\infty$. Since A is assumed selfadjoint, all λ_i are real, and the corresponding eigenvectors ϕ_i form an orthonormal basis in $L^2(\Omega)$.

We assume that the first N eigenvalues of A are strictly positive (for example $A=\Delta+C^2I$). Hence the "free system" (when $g=w=0$) blows up exponentially as

$$\|x(t)\|+\|\dot{x}(t)\|\to\infty\quad \text{as } t\to\infty.$$

By introducing the appropriate vectors w and $g\in L^2(\Omega)$, we wish to "stabilize" system (10). By "stabilization", we mean the restoration of the oscillatory character of the system. It is worth noticing that unless feedback on velocity is introduced, this is the most we can expect. In fact the solution of (10) will not decay to zero unless a feedback term involving $x_t(t)$ is introduced. Stabilization is achieved by shifting the N positive eigenvalues of A to the left half of the real line.

THEOREM 5. *Assume*:

(H1) *The eigenvalues* $\{\lambda_i\}$ *of* A *are simple and satisfy* $|\lambda_{i+1}-\lambda_i|\geq d>0$.

(H2) *The coordinates* $w_i=\langle w_i,\phi_i\rangle$ *of* w *satisfy* $w_i\neq 0$, $i=1,2,\cdots,N$.

Then there exists $\delta>0$ *such that for every* $x_0\in D(A^{1/2})$, $x_1\in L^2(\Omega)$ *and for every* $g\in L^2(\Omega)$ *satisfying* $\|g_s\|_{L^2(\Omega)}\leq\delta$, *the solution of the system* (10) *can be written*

$$x(t,x_0,x_1)=\sum_i\sin(\eta_it)\alpha_i(x_0,x_1)\psi_i+\sum_i\cos(\eta_it)\tilde{\alpha}_i(x_0,x_1)\psi_i,$$

where the above series are convergent in $D(A^{1/2})$, η_i *are real*, ψ_i *are a Riesz basis in* $D(A^{1/2})$, *and* $\alpha(x_0,x_1)$ *are bounded linear functionals on* $D(A^{1/2})\times L_2(\Omega)$. *Analogously*

$$x_t(t,x_0,x_1)=\sum_i \sin(\eta_i t)\alpha_i(x_0,x_1)\eta_i\psi_i+\sum\cos(\eta_i t)\tilde{\alpha}_i(x_0,x_1)\eta_i\psi_i$$

where convergence takes place in $L_2(\Omega)$.

As an immediate corollary of Theorem 5 we obtain:

COROLLARY 2. *Under the assumptions of Theorem* 5, *we have*

$$\|x(t)\|_{D(A^{1/2})}+\|x_t(t)\|_{L_2(\Omega)}\le C\big[\|x_0\|_{D(A^{1/2})}+\|x_1\|_{L_2(\Omega)}\big]$$

where the constant C does not depend on t, x_0, x_1.

4.2. Preliminaries to the proof of Theorem 5. Set

$$E^{(1)}=D(A^{1/2})\times L_2(\Omega).$$

It is immediately verified that the original system (10) can be rewritten equivalently as a first order system on E, i.e.,

$$\frac{d}{dt}y(t)=\mathcal{A}y(t)+\mathcal{B}y(t),\qquad y(0)=y_0\quad \text{with}\, y=(x_1x_t)^T, \tag{11}$$

$$\mathcal{A}\doteq\begin{bmatrix}0 & I\\ A & 0\end{bmatrix},\qquad \mathcal{B}\doteq\begin{bmatrix}0 & 0\\ B & 0\end{bmatrix}, \tag{12}$$

where $By\doteq\langle y,w\rangle g$ and $D(\mathcal{A})=D(A)\times D(A^{1/2})$. It is well known that $\mathcal{A}$ is the generator of a continuous group on E. Since $\mathcal{B}$ is a bounded operator on E, by standard perturbation theorem $\mathcal{A}+\mathcal{B}$ also generates a continuous semigroup $e^{(\mathcal{A}+\mathcal{B})t}$ and

$$\|e^{(\mathcal{A}+\mathcal{B})t}\|_{E\to E}\le Ce^{wt}.$$

By elementary calculations one can easily check that the spectrum of $\mathcal{A}$ consists of the points:

$$\tilde{\lambda}_i=\pm\sqrt{\lambda_i},\quad i=1,\cdots,N,\qquad \tilde{\lambda}_i=\pm\sqrt{-\lambda_i},\quad i=N+1,\cdots$$

with the corresponding eigenvectors:

$$\tilde{\phi}_i=\big[\tilde{\lambda}_i^{-1}\phi_i,\phi_i\big],$$

where the last vectors constitute an orthonormal basis on $D(A^{1/2})\times L_2(\Omega)\doteq E$. According to a standard decomposition theorem (see [1, p.178]) one can decompose the space $X\doteq L_2(\Omega)$ into two orthogonal subspaces $P_u(X)$ and $P_s(X)\doteq X-P_u(X)$ with P_u standing for an orthogonal projection onto span $\{\phi_i\}i=1,\cdots,N$. Accordingly, one decomposes E into $\mathcal{P}_u(E)$ and $\mathcal{P}_s(E)$, where $\mathcal{P}_u$ is an orthogonal projection onto span $\{\tilde{\phi}_i i=1,2,\cdots,N\}$. Let $\mathcal{P}_u(A_u)$ and $\mathcal{P}_s(A_s)$ be the restriction of $\mathcal{A}(A)$ to $\mathcal{P}_u(E)$ $(P_u(X))$ and $\mathcal{P}_s(E)(P_s(X))$, respectively. After setting

$$y_s\doteq\mathcal{P}_s y,\qquad y_u\doteq\mathcal{P}_u y$$

we rewrite (11) as follows:

$$\frac{d}{dt}\begin{bmatrix}y_s\\ y_u\end{bmatrix}=\begin{bmatrix}\mathcal{A}_s & 0\\ 0 & \mathcal{A}_u\end{bmatrix}\begin{bmatrix}y_s\\ y_u\end{bmatrix}+\begin{bmatrix}\mathcal{B}_s & \mathcal{B}_{su}\\ \mathcal{B}_{us} & \mathcal{B}_u\end{bmatrix}\begin{bmatrix}y_s\\ y_u\end{bmatrix},$$

where

$$\mathcal{A}_s \doteq \begin{bmatrix} 0 & I_s \\ A_s & 0 \end{bmatrix}, \qquad \mathcal{A}_u \doteq \begin{bmatrix} 0 & I_u \\ A_u & 0 \end{bmatrix},$$

$$\mathcal{B}_s \doteq \begin{bmatrix} 0 & 0 \\ B_s & 0 \end{bmatrix}, \qquad \mathcal{B}_{su} \doteq \begin{bmatrix} 0 & 0 \\ B_{su} & 0 \end{bmatrix},$$

$$\mathcal{B}_{us} \doteq \begin{bmatrix} 0 & 0 \\ B_{us} & 0 \end{bmatrix}, \qquad \mathcal{B}_u \doteq \begin{bmatrix} 0 & 0 \\ B_u & 0 \end{bmatrix},$$

$$B_s \doteq P_s g \langle P_s w, \cdot \rangle \doteq g_s \langle w_s, \cdot \rangle, \qquad B_{us} \doteq g_u \langle w_s, \cdot \rangle,$$

$$B_u \doteq P_u g \langle P_u w, \cdot \rangle = g_u \langle w_u, \cdot \rangle, \qquad B_{su} \doteq g_s \langle w_u, \cdot \rangle.$$

Notice that the eigenvalues $\tilde{\lambda}_i i=1,\cdots,N$ of a $2N \times 2N$ matrix $\mathcal{A}_u$ are real ($2N$ positive and $2N$ negative), while the eigenvalues of $\mathcal{A}_s$ are purely imaginary. Therefore the purpose of introducing a perturbation is precisely to *shift* N eigenvalues of A_u to *the left* (hence eigenvalues of $\mathcal{A}_u$ to imaginary axis) without perturbing "too much" the spectrum of $\mathcal{A}_s$ (i.e., leaving it on the imaginary axis).

For sake of clarity of exposition we start by outlining a brief plan of the proof of Theorem 5.

(i) By selecting appropriate vectors $g_u = P_u g$ we force the eigenvalues of $\mathcal{A}_u + \mathcal{B}_u$ to lie on the imaginary axis. This can be done, due to Hypothesis (H2) by using rather standard arguments in finite-dimensional control theory.

(ii) By requiring that $\|g_s\|$ be "small" enough, and using Hypothesis (H1), we make sure that the spectrum of $\mathcal{A} + \mathcal{B}$ is purely imaginary.

(iii) Having established the location of the spectrum of $\mathcal{A} + \mathcal{B}$ we must finally show that the corresponding eigenvectors constitute a Riesz basis in E. To accomplish this we will make use of Corollary 1. Notice that in the case of an analytic semigroup, after having established the location of the spectrum, one can automatically conclude stability of the system. This is however not the case in our present situation ($e^{(\mathcal{A}+\mathcal{B})t}$ is certainly not analytic). Correct location of the spectrum is only a necessary condition for stability but not by any means sufficient. That's why one must show that corresponding eigenvectors form a basis.

Let us proceed with the details.

4.3. Proof of Theorem 5. Let us consider the following system of N equations:

$$\frac{d}{dt} y_u = \mathcal{A}_u y_u + \mathcal{B}_u y_u. \tag{13}$$

By (12)

$$\mathcal{A}_u + \mathcal{B}_u = \begin{bmatrix} 0 & I_u \\ A_u + B_u & 0 \end{bmatrix}.$$

Without *any* loss of generality, we can assume that the $N \times N$ matrix A_u is *diagonal* with distinct real eigenvalues. Then $A_u + B_u = \Lambda + \overline{G}W$, where

$$\Lambda = \operatorname{diag}[\lambda_i]_{i=1,\cdots,N}, \quad G = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_N \end{bmatrix}, \quad W = [w_1, w_2, \cdots, w_N]$$

(recall $g_i=(g,\phi_i)$, $w_i=(w,\phi_i)$). Provided that the $N\times N$ matrix

$$[W^T,\Lambda W^T,\cdots,\Lambda^{N-1}W^T] \tag{14}$$

is of full rank, a well-known result in system theory (see [5]) guarantees the existence of a matrix (vector) G such that A_u+B_u has an arbitrarily preassigned set of eigenvalues.

It is then readily seen (see [5]) that hypothesis (H2) is *necessary* and *sufficient* for (14) to be satisfied. Therefore we are in a position to *select* a vector $g_u=[g_1,g_2,\cdots,g_N]$ such that the eigenvalues of A_u+B_u say $-\mu_k^2$ are all *distinct*, *negative*, and *different* from λ_i, $i=N+1, N+2,\cdots$. Thus

$$\sigma(\mathcal{A}_u+\mathcal{B}_u)=\Big\{\pm i\mu_k;k=1,\cdots,N,\mu_k \text{ are real, distinct and different from } \sqrt{-\lambda_k}\,,k=N+1,\cdots\Big\}. \tag{15}$$

To proceed with part (ii), let us rewrite the original system (12) as:

$$\frac{d}{dt}\begin{bmatrix} ys \\ y_u\end{bmatrix}=\begin{bmatrix}\mathcal{A}_s & 0\\ 0 & \mathcal{A}_u+\mathcal{B}_u\end{bmatrix}\begin{bmatrix}y_s\\ y_u\end{bmatrix}+\begin{bmatrix}\mathcal{B}_s & \mathcal{B}_{su}\\ \mathcal{B}_{us} & 0\end{bmatrix}\begin{bmatrix}y_s\\ y_u\end{bmatrix}. \tag{16}$$

Notice that the spectrum of the first operator is purely imaginary. Our aim is to guarantee that the perturbations

$$\begin{bmatrix}\mathcal{B}_s & \mathcal{B}_{su}\\ \mathcal{B}_{us} & 0\end{bmatrix}$$

will perserve this characteristic of the spectrum. Therefore it is natural to view the last operator as a perturbation of

$$\begin{bmatrix}\mathcal{A}_s & 0\\ 0 & \mathcal{A}_u+\mathcal{B}_u\end{bmatrix}$$

and then apply our Theorem 1. By exploiting explicit forms for $\mathcal{A}_s$, $\mathcal{B}_s$, $\mathcal{A}_u$ and $\mathcal{B}_u$ we rewrite the eigenvalue problem for (16) as follows:

$$\begin{bmatrix}0 & I & 0 & 0\\ A_s+B_s & 0 & 0 & 0\\ 0 & 0 & 0 & I\\ 0 & 0 & A_u+B_u & 0\end{bmatrix}\begin{bmatrix}\psi_{1s}\\ \psi_{2s}\\ \psi_{1u}\\ \psi_{2u}\end{bmatrix}+\begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 0 & B_{su} & 0\\ 0 & 0 & 0 & 0\\ B_{us} & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\psi_{1s}\\ \psi_{2s}\\ \psi_{1u}\\ \psi_{2u}\end{bmatrix}=iu\begin{bmatrix}\psi_{1s}\\ \psi_{2s}\\ \psi_{1u}\\ \psi_{2u}\end{bmatrix}.$$

The above yields:

$$\psi_{2s}=i\mu\psi_{1s}, \tag{17a}$$

$$\psi_{2u}=i\mu\psi_{1u}, \tag{17b}$$

$$\begin{bmatrix}A_s & 0\\ 0 & A_u+B_u\end{bmatrix}\begin{bmatrix}\psi_{1s}\\ \psi_{1u}\end{bmatrix}+\begin{bmatrix}B_s & B_{ssu}\\ B_{us} & 0\end{bmatrix}\begin{bmatrix}\psi_{1s}\\ \psi_{1u}\end{bmatrix}=-\mu^2\begin{bmatrix}\psi_{1s}\\ \psi_{1u}\end{bmatrix}. \tag{17c}$$

Equation (17c) can be viewed as a perturbation of

$$\begin{bmatrix}A_s & 0\\ 0 & A_u+B_u\end{bmatrix}$$

on the *real* Hilbert space $L_2(\Omega)$. Notice that all the requirements of Theorem 1′ are fulfilled. In fact all the eigenvalues of A_s and A_u+B_u are real, negative and distinct (see (15)) and the "gap condition" is satisfied with

$$d=\min\left\{|\lambda_i-\lambda_j|, ij=N+1,\cdots; |\lambda_i-\mu_k^2|, i=N+1,\cdots, k=1,2,\cdots,N\right\}.$$

The eigenvectors of

$$\begin{bmatrix} A_s & 0 \\ 0 & A_u+B_u \end{bmatrix}$$

generate a Riesz basis. This can be readily seen by recalling that A_s has a system of orthonormal eigenvectors which are a basis in $P_s(L_2(\Omega))$ and A_u+B_u has distinct eigenvectors, which generate N linearly independent eigenvectors in $P_u[L_2(\Omega)]$. Therefore we are in a position to refer to Theorem 1′ in order to claim that for $\|B_s\|$, $\|B_{su}\|$, $\|B_{us}\|$ sufficiently small, $-\mu^2$ are *real and negative* and consequently *the eigenvalues* of $\mathcal{A}+\mathcal{B}$ are *purely imaginary and equal to* $\pm i\mu$. Next we observe that the condition that $\|B_s\|$, $\|B_{su}\|$, $\|B_{us}\|$ be small, can be slightly relaxed so as to demand only that $\|g_s\|$ is small (i.e., $\|B_{su}\|$). In fact after rewriting (17c) we arrive at

$$(A_s+B_s)\Psi_{1s}+B_{us}\Psi_{1u}=-\mu^2\Psi_{1s}, \qquad (A_u+B_u)\Psi_{1u}+B_{su}\Psi_{1s}=-\mu^2\Psi_{1u}.$$

Hence after noticing that $[A_u+B_u+\mu^2 I]$ is invertible for all $-\mu^2\notin\sigma(A_u+B_u)$ (it is enough to consider only such μ since otherwise our assertion is proved), we have

$$\text{(18)} \qquad \left[(A_s+B_s)-B_{us}\left(A_u+B_u+\mu^2\right)^{-1}B_{su}\right]\Psi_{1s}=-\mu^2\Psi_{1s}.$$

It is readily seen that (18) depends on the product of B_{us} and B_{su}, therefore for all $-\mu^2\notin\sigma(A_u+B_u)$ it is enough to demand that only one term, namely B_{su} have a small norm. This way we arrive at the following lemma which completes the proof of (ii).

LEMMA 3. *With $g_u=P_u g$ selected so as to guarantee* (15) *and $g_s=P_s g$ such that $\|g_s\|$ is sufficiently small, we have*

$$\sigma(\mathcal{A}+\mathcal{B})=\{i\mu_k; \mu_k \text{ real}\}.$$

To prove (iii) we simply refer to Corollary 1 applied with

$$T=\mathcal{A}+\mathcal{B}=\begin{bmatrix} \mathcal{A}_s & 0 \\ 0 & \mathcal{A}_u+\mathcal{B}_u \end{bmatrix}+\begin{bmatrix} \mathcal{B}_s & \mathcal{B}_{us} \\ \mathcal{B}_{su} & 0 \end{bmatrix},$$

where $T=\mathcal{A}_s$, $M=\mathcal{A}_u+\mathcal{B}_u$, $H=\mathcal{P}_s E$, $H_n=\mathcal{P}_u E$ and $\dim H_n=2N$. It is left to the reader to verify, by using, arguments similar to those above, that all the hypotheses of Corollary 1 are satisfied. (Recall that the eigenvectors of a selfadjoint operator generate a basis in $\mathcal{P}_s E$.) Therefore Corollary 1 yields

LEMMA 4. *With $g_s=P_s g$ such that $\|g_s\|$ has sufficiently small norm, the eigenvectors ψ_k of $\mathcal{A}+\mathcal{B}$; $\psi_k=[i\mu_k\Psi_k]$ constitute a Reisz basis in E.*

Finally, using Lemmas 3 and 4 we are in a position to complete the proof of Theorem 5. In fact, by virtue of Lemma 4 we write

$$y(t)=e^{(\mathcal{A}+\mathcal{B})t}y_0=e^{(\mathcal{A}+\mathcal{B})t}\sum_{k=1}^{\infty}\alpha_k(y_0)\Psi_k,$$

where $\alpha_k(y_0)$ are linear bounded functionals defined on $E=D(A^{1/2})\times L_2(\Omega)$ which converges to a limit also in E. Lemma 4 yields

$$y(t)=\sum_{k=1}^{\infty} e^{i\mu_k t}\alpha_k(y_0)\Psi_k, \qquad \mu_k \text{ real.}$$

which completes the proof of Theorem 5.

5. Examples of operators satisfying a gap condition. In this section we present some examples of dynamic systems satisfying the condition

$$|\lambda_i-\lambda_j|\geq d>0 \quad \text{if } i\neq j.$$

The first and most obvious example is when $\Omega\subseteq\Re^1$, that is when we consider a Sturm–Liouville operator on an interval. For example, take

$$Ay=y'' \quad \text{with } y(0)=y(\pi)=0.$$

Then the eigenvalues are $\lambda_n=-n^2$, and are of multiplicity one. This operator satisfies $(\lambda_{n+1}-\lambda_n)\to\infty$. We now give some examples where the gap does not go to plus infinity. Consider the operator

$$A\begin{bmatrix} x \\ y \end{bmatrix}=\begin{bmatrix} x'' \\ y'' \end{bmatrix}, \qquad x(0)=x(\pi)=0, \quad y'(0)=0, \quad y'(\pi)=y(\pi).$$

As before, the $\lambda_n^{(1)}=n^2$ are eigenvalues with corresponding eigenvectors $(x,y)=(\sin nt,0)$. However, in addition, the numbers $\lambda_n^{(2)}=\gamma_n^2$ defined by $x_n=\operatorname{cotan}\pi x_n$, $\lambda_n>0$ are also eigenvalues, with corresponding eigenvectors $(x,y)=\cos x_n t$. It is easy to check that for large n, the solution of $x_n=\operatorname{cotan}\pi x_n$ is approximately $x=1/(n-x)$ and we can verify that $\lambda_n\sim(n+\sqrt{n^2-4})/2$. Thus the gap between any two eigenvalues is

$$n^2-\lambda_n^2=\left(\frac{3n+\sqrt{n^2-4}}{2}\right)\left(\frac{n-\sqrt{n^2-4}}{2}\right)=\left(\frac{3n+\sqrt{n^2-4}}{2}\right)\left(\frac{2}{n+\sqrt{n^2+4}}\right)\to 2$$

as $n\to\infty$.

As another example, consider the biharmonic operator $Ay=\Delta^2 y$ on the region $\Omega=\{(x,y),0\leq x\leq\pi,0\leq y\leq 2^{1/4}\pi\}$ with $y=0$ on $\partial\Omega$.

In this case, the spectrum consists of eigenvalues $\lambda_{nm}=(n^2+\sqrt{2}\,m^2)^2$ of multiplicity one and we can check that

$$\begin{aligned}|\lambda_{mn}-\lambda_{kl}|&=\left|\left(n^2+\sqrt{2}\,m^2\right)^2-\left(k^2+\sqrt{2}\,l^2\right)^2\right|\\&=\left|(n^2-k^2)-\sqrt{2}\,(m^2-l^2)\left(n^2+k^2+\sqrt{2}\,(m^2+l^2)\right)\right|\\&=\frac{(n^2-k^2)^2-2(m^2-l^2)^2}{(n^2-k^2)+\sqrt{2}\,(m^2-l^2)}\cdot(n^2+k^2)+\sqrt{2}\,(m^2+l^2).\end{aligned}$$

However $|(n^2-k^2)^2-2(m^2-l^2)^2|\geq 1$ since $\sqrt{2}$ is irrational, and therefore we may conclude that $|\lambda_{mn}-\lambda_{kl}|\geq 1$. If for all integers q, we have that $|\alpha-(p/q)|\geq\varepsilon/q^2$, we can verify that when $\Omega=\{(x,y)\in R^2,0\leq x\leq\pi,0\leq y\leq\alpha^{1/2}\pi\}$ the biharmonic with zero boundary data satisfies a gap condition. The set of α satisfying this condition is of the same cardinality as the continuum [3]. Thus, the vibrating plate equation fits into the framework of (10).

As a fourth example, we consider a problem in heat conduction which occurs in [2]. We consider a bar, of negligible depth on the interval $0 \leq x \leq \pi$. The conductivity k of this bar is small compared to its specific heat R. The region $\Omega = \{0 \leq x \leq 1, z < 0\}$ is filled with a material whose specific heat r is small compared to its conductivity. The heat flow equations are

$$(19) \qquad R\frac{\partial T}{\partial \tau} = k\frac{\partial^2 T}{\partial x^2} - K\frac{\partial \tau}{\partial z}, \qquad r\frac{\partial \tau}{\partial \tau} = K\left(\frac{\partial^2 \tau}{\partial x^2} + \frac{\partial^2 \tau}{\partial z^2}\right),$$

where $T(x,t)$ is the temperature on the bar and $\tau(x,z,t)$ is the temperature in Ω. Along with (19) we consider the boundary conditions

$$(20) \qquad \begin{aligned} &\frac{\partial T}{\partial x}(0,t)=0, \quad \frac{\partial T}{\partial x}(1,t)=0, \quad \frac{\partial \tau}{\partial x}(0,z,t)=0, \quad \frac{\partial \tau}{\partial x}(1,z,t)=0, \\ &\lim_{z\to+\infty}\frac{\partial \tau}{\partial z}(x,z,t) = \lim_{z\to-\infty}\tau(x,z,t)=0, \quad \tau(x,0,t)=T(x,t), \quad 0 \leq x \leq 1. \end{aligned}$$

If we assume k, r very small compared to K, R, we may use a simplified model for (19)

$$\frac{\partial T}{\partial \tau} = -\frac{K}{R}\frac{\partial \tau}{\partial z}, \qquad \frac{\partial^2 \tau}{\partial x^2} + \frac{\partial^2 \tau}{\partial z^2} = 0$$

together with boundary conditions (20). Russell [2] showed that (21) can be written as

$$\frac{dT}{dt} = A(t) \quad \text{where } AT = -\frac{K}{R}\frac{\partial \tau}{\partial z} \quad \text{with } (A) = H^1(0,1).$$

In addition, it was shown that A is the square root of the Sturm-Liouville operator

$$ST = -\frac{K^2}{R^2}\frac{d^2T}{dx^2}, \qquad D(S) = \left\{T \in H^2(0,1), \frac{dT}{dx}(0) = \frac{dT}{dx}(1) = 0\right\},$$

and thus A has spectrum

$$\lambda_n = -\left(\frac{n^2 R^2 \pi^2}{K^2}\right)^{1/2},$$

and the gap condition is satisfied.

Acknowledgments. We wish to acknowledge helpful conversations with Robert Long and D. Russell.

REFERENCES

[1] T. Kato, *Perturbation Theory of Linear Operators*, Springer-Verlag, New York, 1976.

[2] L. F. Ho and D. L. Russell, *Admissible input elements for systems in Hilbert space and a Carleson measure criterion*, SIAM J. Control Optim.,21 (1983), pp. 614–640.

[3] K. B. Stolarsky, *Algebraic Numbers and Diophantine Equations*, Dekkar, New York, 1974.

[4] R. Triggiani, *Boundary feedback stabilization of parabolic equations*, Appl. Math. Optim., 6 (1980), pp. 201–220.

[5] M. Wonham, *On pole assignment in multi-input controllable linear systems*, IEEE Trans-AC, 12(b) (1967), pp. 660–665.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics
009

BOUNDARY PROBLEMS FOR THE BIHARMONIC OPERATOR IN A SQUARE WITH L^p-DATA*

LORENZA DIOMEDA[†] AND BENEDETTA LISENA[†][‡]

Abstract. Applied boundary value problems for the bi-Laplacian often actually involve domains with nonsmooth boundaries. The second fundamental problem of plane elasticity, the viscous flow and Stokes problem are representative examples of problems which can be reduced to the interior or exterior Dirichlet problem for the bi-Laplacian. In this paper we study the Dirichlet problem for Δ^2 in a plane square with L^p-data, looking for a solution in potential form. By the method of pseudodifferential operators we show an existence theorem provided the data are in L^p, $1<p<3$, and satisfy a proper compatibility condition. We also solve two further boundary value problems which are related to the Dirichlet problem.

Introduction. In this paper we study the Dirichlet boundary problem for the operator Δ^2 in a plane square, with L^p-data.

In polygonal regions analogous problems were studied by many authors, for example P. Grisvard [2], [3], J. E. Lewis–C. Parenti [4], [5], M. Merigot [7], J. Necas [8], S. M. Nikol'skii [9], K. Rektorys–V. Zahradnik [10].

For the Dirichlet problem for the Laplace operator in a polygon Ω Grisvard [2] obtained some results which concern the injectivity of the operator $\mathcal{P}_\Omega u=(\Delta u, u_{|\Gamma_1},\cdots,u_{|\Gamma_n})$ where $\bigcup_{i=1}^n\Gamma_i=\partial\Omega$, and the index of this operator in L^2.

Merigot studies the regularity of the variational solutions of boundary problems for $2m$-order operators and he obtains results similar to Grisvard in L^p spaces. Grisvard [3] considered a boundary value problem for the Navier–Stokes equations with data given in H^m studying solution behaviour in a neighbourhood of the corners by using the behaviour of the biharmonic equation's solutions.

For boundary problems on the operators Δ and Δ^2, a numerical approach can be found, for example, in Nikol'skii and Rektorys–Zahradnik's works [9], [10].

Necas studies a boundary problem for biharmonic operator with data given in the L^p subspace whose elements are transforms of functions in $W_2^2(\Omega)$ by a linear map and gives an appropriate definition of problem's solution.

The Dirichlet problem for the biharmonic equation in a C^1-domain of the plane, with boundary conditions in L^p-sense, has been recently solved by J. Cohen–J. Gosselin in [11].

In our last paper [1], we solved the Dirichlet problem for the operator Δ^2 in a plane sector with L^p-data by a potential representation of the solution and by using the theory of pseudodifferential operators on $L^p(\mathbb{R}_+)$ defined by Lewis–Parenti in [4].

One of the motivations of [5] is the study of single and double potentials to solve the Dirichlet problems in bounded polygonal domains with L^p-data for the Laplace operator. In §1 we state some definitions and theorems developed in [5] to which we refer for more details.

In this paper, we study exactly the following problem.

Let $\Omega=\{(x,y)\mathbb{R}^2|0<x<1, 0<y<1\}$,

$$\text{(I)}\qquad \Delta^2u=0 \quad \text{in}\,\Omega,\quad \left.\frac{\partial u}{\partial\tau}\right|_{\partial\Omega}=h\in L^p(\partial\Omega),\quad \left.\frac{\partial u}{\partial n^+}\right|_{\partial\Omega}=l\in L^p(\partial\Omega),$$

* Received by the editors October 26, 1982, and in revised form May 30, 1983.

[†] Istituto di Analisi Matematica, University of Bari, 70121 Bari, Italy.

[‡] The research of this author was partially supported by C.N.R. Gruppo G.N.A.F.A.

where τ is the unit tangent to $\partial\Omega$ and n^+ is the unit interior normal to $\partial\Omega$ and the equalities are to consider as boundary limits on L^p.

Writing the solution in potential form, we obtain a boundary operator A, which is in the new class of pseudodifferential operators introduced in [5]. The operator A is elliptic and it has finite index for every $p \neq 3$.

We also calculate the dimension of $\operatorname{Ker}(A)$ for $1<p<3$, determining the "adjoint" problem of (I). Therefore, by Fredholm's theory, we show that (I) has a solution if the data h verifies an appropriate compatibility condition. These results are proved in §2 of this work.

By the potential technique, in §3, we study two further problems for Δ^2 with second order boundary conditions. The first problem has a unique solution for all boundary data, whereas the second is solvable provided boundary data verify a proper compatibility condition. The extension of these results to any polygonal domain in $\mathbb{R}^2$, presents only technical difficulties.

1. In this section we define a class of pseudodifferential operators (pdo's) acting on $L^p([0,1])$. We find this type of operators studying previous boundary problems. The algebra of these pdo's has been introduced by Lewis and Parenti in [5] so we refer to this work for details.

We shall denote the open half line $(0,+\infty)$ as $\mathbb{R}_+$ and $[0,+\infty)$ as $\overline{\mathbb{R}}_+$. If a is a real number the integer $(a]$ will be defined as the greatest integer smaller than a.

DEFINITION 1.1. Let $-\infty<a<b<+\infty$. By $\mathfrak{F}_{a,b}$ we denote the class of functions $f \in C^\infty(\mathbb{R}_+)$ such that:

1. For $i=0,1,\cdots,(-a]$, there are scalars f_{j0} such that for every k and every $\delta>0$,

$$\left(-t\frac{d}{dt}\right)^k\left(f(t)-\sum_{j=0}^{(-a]}\frac{1}{j!}f_{j0}t^j\right)=O(t^{-a-\delta}), \qquad t\to 0^+.$$

2. For $j=0,1,\cdots,(b]$, there are scalars $f_{j\infty}$ such that for every k and every $\delta>0$,

$$\left(-t\frac{d}{dt}\right)^k\left(f(t)-\sum_{j=0}^{(b]}\frac{1}{j!}f_{j\infty}t^{-j}\right)=O(t^{-b+\delta}), \qquad t\to +\infty.$$

DEFINITION 1.2. For $-\infty\leq c<d\leq+\infty$ and $m\in\mathbb{R}$ the class $\theta^m_{c,d}$ consists of those functions $a(z)$ holomorphic in the strip

$$S_{c,d}=\{z\in\mathbb{C}\,|\,c<\operatorname{Re}z<d\}$$

such that for every $(c',d')\subset(c,d)$, for every k

$$\frac{d^k}{dz^k}a(z)=O\left(|\operatorname{Im}z|^{m-k}\right) \quad \text{as } |\operatorname{Im}z|\to+\infty,\ z\in S_{c',d'}.$$

DEFINITION 1.3. For $-\infty\leq c<d\leq\infty$ and $m\in\mathbb{R}$, the symbol class $\Sigma^m_{c,d}$ is defined by

$$\Sigma^m_{c,d}=\bigcap_{(c',d')\subset(c,d)}\mathfrak{F}_{c-c',d-d'}\left(\theta^m_{c',d'}\right).$$

DEFINITION 1.4. A symbol $a(t,z)$ is in the class $\Sigma_{1/p}$. iff for some c,d with $0\leq c<1/p<d\leq 1$

1. $a(t,z)\Sigma^0_{c,d}$.

2. There are functions $a_+(t),\ a_-(t)\in\mathcal{F}_{c-d,d-c}$ such that

$$a(t,z)-a_+(t)\theta(z)-a_-(t)(1-\theta(z))\in\Sigma_{c,d}^{-1}$$

where $\theta(z)=(1-e^{2\pi iz})^{-1}$.

If $a(t,z)\in\Sigma_{1/p}$, the functions $a_+(t)$, $a_-(t)$ are uniquely determined by the relations

$$a_+(t)=a\left(t,\frac{1}{p}+i\infty\right),\qquad a_-(t)=a\left(t,\frac{1}{p}-i\infty\right).$$

DEFINITION 1.5. If $a(t,z)\in\Sigma_{1/p}$ and

$$Af(t)=\frac{1}{2\pi i}\int_{\operatorname{Re}z=1/p}t^{-z}a(t,z)\tilde{f}(z)\,dz$$

where $\tilde{f}(z)$ is the Mellin transform of $f\in C_0^\infty(\mathbb{R})$, then the principal symbol of $A,\sigma(A)$, is the function $a(t,z)$ restricted to the boundary of the compact rectangle

$$R_{1/p}=\left\{(t,z)\middle|0\le t\le+\infty,z=\frac{1}{p}+i\xi,\xi\in\overline{\mathbb{R}}\right\}.$$

Now we are able to define the class of operators acting on $L^p([0,1])$ which interest us directly. For $f\in L^p([0,1])$, define

$$Tf(t)=f(1-t). \tag{1.1}$$

DEFINITION 1.6. A bounded operator A on $L^p([0,1])$ is a pdo of class $\mathrm{OP}\Sigma_{1/p}([0,1])$ iff

1. If $\phi,\psi\in C_0^\infty(\mathbb{R})$ have disjoint supports then the map

$$f\to(\phi A\psi)f,$$

is a compact operator on $L^p(\mathbb{R}_+)$.

2. If $\phi,\psi\in C_0^\infty([0,1])$, there is an operator $A_{\phi\psi}\in\mathrm{OP}\Sigma_{1/p}$ and a compact operator $K_{\phi\psi}$ on $L^p([0,1])$ such that

$$\phi A\psi=A_{\phi\psi}+K_{\phi\psi}.$$

3. The operator TAT satisfies conditions 1 and 2.

An example of an operator of class $\mathrm{OP}\Sigma_{1/p}([0,1])$ is the finite Hilbert transform

$$HF(t)=\frac{1}{\pi}\overset{*}{\int_0^1}\frac{f(s)}{t-s}ds$$

and Hardy operators[1] on $L^p([0,1])$.

The principal symbol of an operator $A\in\mathrm{OP}\Sigma_{1/p}([0,1])$ is a continuous function defined on the boundary of the compact rectangle $R_{1/p,[0,1]}$ in Fig. 1.1.

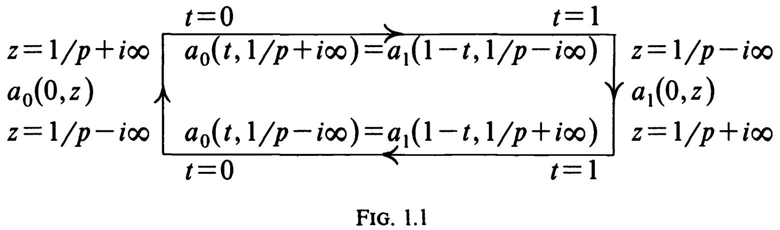


FIG. 1.1

[1] For the definition of Hardy operator see [4].

The functions $a_0(t,z)$ and $a_1(t,z)$ are obtained in the following way: Let $\phi,\psi\in C_0^\infty([0,1])$ and $A^0_{\phi\psi}$ and $A^1_{\phi\psi}$ be the two operators of class $\mathrm{OP}\Sigma_{1/p}(\mathbb{R}_+)$ provided by Definition 1.6,

$$\phi A\psi = A^0_{\phi\psi}+K^0_{\phi\psi},$$
$$\phi(TAT)\psi = A^1_{\phi\psi}+K^1_{\phi\psi}.$$

Then

$$\sigma_p\left(A^0_{\phi\psi}\right)(t,z)=a_0(t,z)\phi(t)\psi(t),$$
$$\sigma_p\left(A^1_{\phi\psi}\right)(t,z)=a_1(t,z)\phi(t)\psi(t).$$

The principal symbol of an $N\times N$ system of pdo's of class $\mathrm{OP}\Sigma_{1/p}([0,1])$ is defined as the matrix of principal symbols.

DEFINITION 1.7. An $N\times N$ system of pdo's of class $\mathrm{OP}\Sigma_{1/p}([0,1])$ is elliptic on $(L^p([0,1]))^N$ iff the determinant of the matrix of principal symbols does not vanish on $\partial R_{1/p,[0,1]}$.

If A is a system of pdo's of class $\mathrm{OP}\Sigma_{1/p}([0,1])$ which is elliptic on L^p, we denote

$$\operatorname{ind}_p(A)=\dim\operatorname{Ker}A-\dim\operatorname{Ker}A^*, \tag{1.2}$$

$$n_{1/p}(\sigma(A))=\frac{1}{2\pi}\Delta_{\partial R_{1/p,[0,1]}}\operatorname{Arg}\det\sigma(A). \tag{1.3}$$

In [5] it has been proved that

$$\operatorname{ind}_p(A)=n_{1/p}(\sigma(A)), \tag{1.4}$$

the change in argument of $\det\sigma(A)$ being taken as $\partial R_{1/p,[0,1]}$ in Fig. 1.1 is traversed in the clockwise direction.

Among applications studied by Lewis and Parenti in [5] there is the resolution of the Dirichlet problem for Laplace's equation in a polygon. We consider this problem in the square,

$$\Omega=\{(x,y)\mathbb{R}^2|0<x<1,0<y<1\}$$

with boundary $\partial\Omega=\Gamma_1\cup\Gamma_2\cup\Gamma_3\cup\Gamma_4$ and vertices labeled $P_0,P_1,P_2,P_3,P_4=P_0$ as $\partial\Omega$ is traversed in the clockwise direction.

For $t\in[0,1]$,

$$P_t=P_{t,1}=tP_0+(1-t)P_1\in\Gamma_1,$$
$$P_t=P_{t,2}=tP_2+(1-t)P_1\in\Gamma_2,$$
$$P_t=P_{t,3}=tP_2+(1-t)P_3\in\Gamma_3,$$
$$P_t=P_{t,4}=tP_4+(1-t)P_3\in\Gamma_4.$$

This parametrization is shown in Fig. 1.2.

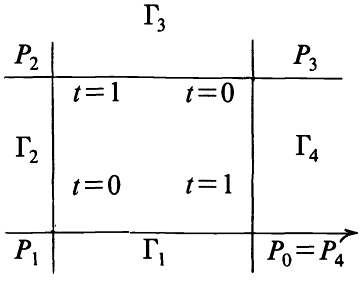


FIG. 1.2

The arclength $d\sigma$ on Γ_i is given by

$$d\sigma=(-1)^i dt. \tag{1.5}$$

For $\phi\in L^p(\partial\Omega)$ define the double layer potential

$$u(X)=\frac{1}{\pi}\int_{\partial\Omega}\frac{\langle X-Q,n_Q\rangle}{|X-Q|^2}\phi(Q)\,d\sigma_Q, \tag{1.6}$$

where n_Q is the interior unit normal to a point $Q\in\partial\Omega$ and we seek ϕ such that u is a solution of the problem:

$$\Delta v=0 \quad \text{in}\,\Omega, \quad v|_{\partial\Omega}=\psi, \quad \psi\in L^p(\partial\Omega). \tag{1.7}$$

The boundary condition of (1.7) turns into the following system:

$$\begin{pmatrix} I & K & Z & K_1 \\ K & I & K_1 & Z \\ Z & K_1 & I & K \\ K_1 & Z & K & I \end{pmatrix}\begin{pmatrix}\phi_1\\ \phi_2\\ \phi_3\\ \phi_4\end{pmatrix}=B\phi=\psi, \tag{1.8}$$

where $\phi_i(t)=\phi(P_t)$, $P_t\in\Gamma_i$,

$$\begin{aligned} K\phi_j(x)&=\frac{1}{\pi}\int_0^1\frac{x}{x^2+s^2}\phi_j(s)\,ds,\\ Z\phi_j(x)&=\frac{1}{\pi}\int_0^1\frac{1}{1+(1-x-t)^2}\phi_j(t)\,dt, \end{aligned} \tag{1.9}$$

and $K_1=TKT$ with T defined by (1.1).

K is a Hardy operator, Z is a compact operator so that B is a matrix of operators in $\mathrm{OP}\Sigma_{1/p}([0,1])$.

It has been proved that B is elliptic for $p\neq\frac{3}{2}$ and in this case we also have

$$\operatorname{ind}_p(B)=\begin{cases}4 & \text{if } 1<p<\frac{3}{2},\\ 0 & \text{if } p>\frac{3}{2}.\end{cases}$$

For $p>\frac{3}{2}$ $\operatorname{Ker}(B)=\{0\}$ and B is invertible on $(L^p([0,1]))^4$; hence, for $p>\frac{3}{2}$, the problem (1.7) has a unique solution in the form of double layer potential.

2. In this section we solve the problem (I) in Ω. We use the notation of the previous section relative to the Dirichlet problem in a square. In particular, $\partial\Omega$ is parametrized as in Fig. 1.2.

We seek a solution of the problem as the following potential with densities $f,g\in L^p(\partial\Omega)$:

$$u(X)=\frac{1}{2}\int_{\partial\Omega}\Delta F(X-Q)f(Q)\,d\sigma_Q+\int_{\partial\Omega}\frac{\partial^2F}{\partial\tau_Q\partial n_Q^+}(X-Q)g(Q)\,d\sigma_Q, \tag{2.1}$$

where $F(X)$ is the fundamental solution of Δ^2 defined as

$$F(X)=\frac{1}{4\pi}|X|^2\log|X|^2. \tag{2.2}^2$$

[2] See [6] for the definition of F.

The unit tangent τ_Q to a point $Q\in\partial\Omega$ has these values:

$$\tag{2.3}\begin{aligned}&\tau_Q=\big((-1)^i,0\big), && Q\in\Gamma_i, && i=1,3,\\ &\tau_Q=\big((0,(-1)^i)\big), && Q\in\Gamma_i, && i=2,4,\end{aligned}$$

and n_Q^+ is the unit interior normal to a point $Q\in\partial\Omega$.

Then, for $P\in\partial\Omega$,

$$\tag{2.4}\begin{aligned}&\lim_{\substack{X\to P\\ X\in\Omega}}\frac{\partial u}{\partial\tau_P}(X)=G\mathbf{f}(P)+M\mathbf{g}(P),\\ &\lim_{\substack{X\to P\\ X\in\Omega}}\frac{\partial u}{\partial n_P^+}(X)=(I-C)\mathbf{f}(P)+N\mathbf{g}(P),\end{aligned}$$

where the limits in (2.4) are taken in $L^p(\partial\Omega)$ as $X\to P$ and the operators G,M,C,N, are so defined:

$$G=\begin{bmatrix}H & K & -R & -K_1\\ K & H & -K_1 & -R\\ -R & -K_1 & H & K\\ -K_1 & -R & K & H\end{bmatrix},$$

$$M=\begin{bmatrix}0 & -K'+S' & -V & -R_1'+S_1'\\ -K'+S' & 0 & -K_1'+S_1' & -V\\ -V & -K_1'+S_1' & 0 & -K'+S'\\ -K_1'+S_1' & -V & -K'+S' & 0\end{bmatrix},$$

$$\tag{2.5} C=\begin{bmatrix}0 & K' & Z & K_1'\\ K' & 0 & K_1' & Z\\ Z & K_1' & 0 & K'\\ K_1' & Z & K' & 0\end{bmatrix},$$

$$N=\begin{bmatrix}H & K-S & -R+L & -K_1+S_1\\ K-S & H & -K_1+S_1 & -R+L\\ -R+L & -K_1+S_1 & H & K-S\\ -K_1+S_1 & -R+L & K-S & H\end{bmatrix}.$$

Here H is the Hilbert transform on $L^p([0,1])$, K,Z,K_1 are defined by (1.9), K',S,S' are the following Hardy operators:

$$\tag{2.6}\begin{aligned}&K'\phi(t)=\frac{1}{\pi}\int_0^1\frac{s}{s^2+t^2}\phi(s)\,ds, \qquad S\phi(t)=\frac{1}{\pi}\int_0^1\frac{2ts^2}{(s^2+t^2)^2}\phi(s)\,ds,\\ &S'\phi(t)=\frac{1}{\pi}\int_0^1\frac{2t^2s}{(s^2+t^2)^2}\phi(s)\,ds.\end{aligned}$$

K_1',S_1,S_1' are obtained from K',S,S' in the following way:

$$\tag{2.7} K_1'=TK'T,\quad S_1=TST,\quad S_1'=TS'T,$$

where T is the intertwining operator (1.1).

Finally R, V, L are the following compact operators:

$$
\begin{aligned}
R\phi(t)&=\frac{1}{\pi}\int_0^1\frac{1-t-s}{1+(1-t-s)^2}\phi(s)\,ds,\\
V\phi(t)&=\frac{1}{\pi}\int_0^1\frac{1-(1-t-s)^2}{\left(1+(1-t-s)^2\right)^2}\phi(s)\,ds,\\
L\phi(t)&=\frac{1}{\pi}\int_0^1\frac{2(1-t-s)}{\left(1+(1-t-s)^2\right)^2}\phi(s)\,ds.
\end{aligned}
\tag{2.8}
$$

Therefore the boundary conditions of problem (I) turn into the following system of integral equations:

$$
A\begin{pmatrix}\mathbf{f}\\ \mathbf{g}\end{pmatrix}=\begin{pmatrix}G & M\\ I-C & N\end{pmatrix}\begin{pmatrix}\mathbf{f}\\ \mathbf{g}\end{pmatrix}=\begin{pmatrix}\mathbf{h}\\ \mathbf{l}\end{pmatrix}.
\tag{2.9}
$$

Now we have to study the properties of the boundary operator A.

PROPOSITION 2.1. *A is a matrix of operators in* $\mathrm{OP}\Sigma_{1/p}([0,1])$. *Moreover A is elliptic in* $(L^p([0,1]))^8$ *for every* $p\in]1,+\infty[$, $p\neq 3$, *and its index has the following values*:

$$
\operatorname{ind}_p(A)=\begin{cases}4, & 1<p<3,\\ -4, & p>3.\end{cases}
$$

Proof. First we remark, that in the matrix A there are the compact operators R, V, L, Z, the Hilbert transform H, the identity operator I, the Hardy operators K, K', S, S' and the operators K_1, K_1', S_1, S_1', obtained from K, K', S, S' by translation. All these operators are in $\mathrm{OP}\Sigma_{1/p}([0,1])$ by Definition 1.6.

The principal symbols of the compact operators R, V, L, Z are zero. Since $THT=-H$, the principal symbol of H is given, on $\partial R_{1/p'[0,1]}$, as shown in Fig. 2.1.

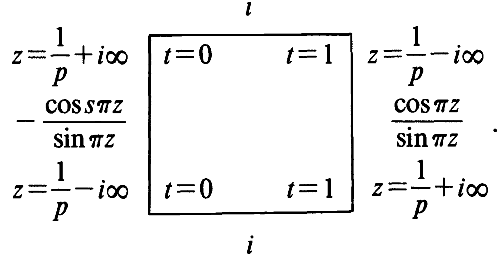


FIG. 2.1

The principal symbol of the Hardy operator K is given in Fig. 2.2,

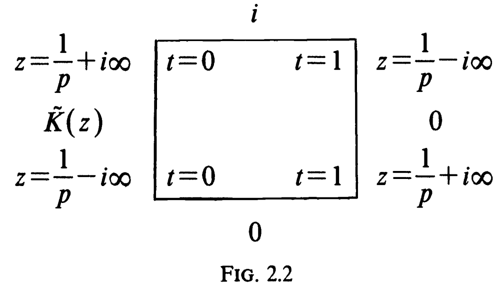


FIG. 2.2

where $\tilde{K}(z)$ is the Mellin transform of $K(t)=t/(1+t^2)$, the kernel of operator K, i.e.

$$K\phi(t)=\frac{1}{\pi}\int_0^1\frac{t}{t^2+s^2}\phi(s)\,ds=\frac{1}{\pi}\int_0^1 K\left(\frac{t}{s}\right)\phi(s)\,\frac{ds}{s}.$$

The principal symbols of K', S, S' have an analogous representation. Since $K_1=TKT$ the principal symbol of K_1 may be represented as in Fig. 2.3, and likewise for K_1', S_1, S_1'.

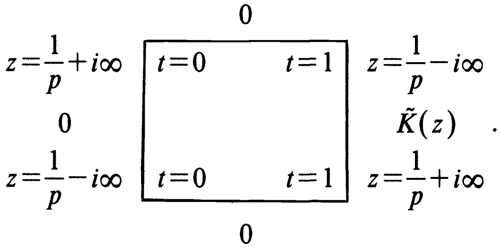


FIG. 2.3

Hence we can show $\sigma_p(A)(t,z)$ as in Fig. 2.4

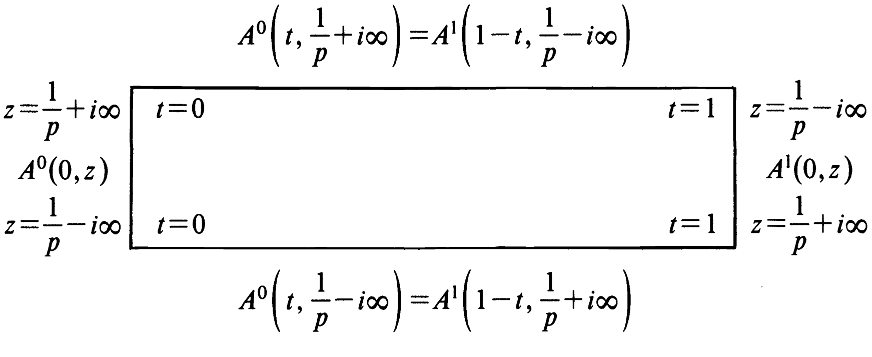

$A^0\left(t,\frac{1}{p}+i\infty\right)=A^1\left(1-t,\frac{1}{p}-i\infty\right)$
$z=\frac{1}{p}+i\infty$
$A^0(0,z)$
$z=\frac{1}{p}-i\infty$
$t=0$
$t=1$
$t=0$
$t=1$
$z=\frac{1}{p}-i\infty$
$A^1(0,z)$
$z=\frac{1}{p}+i\infty$
$A^0\left(t,\frac{1}{p}-i\infty\right)=A^1\left(1-t,\frac{1}{p}+i\infty\right)$

FIG. 2.4

where the matrices $A^0(t,1/p\pm i\infty)$ are independent of t and their determinant is equal to 1; the matrix $A^0(0,z)$ becomes a 2×2 block diagonal matrix by an even number of row and column transpositions and the two blocks on the diagonal are both equal to

$$\begin{pmatrix} \tilde{h}(z) & \tilde{k}(z) & 0 & (-\tilde{k}'(z)+\tilde{s}'(z)) \\ \tilde{k}(z) & \tilde{h}(z) & (-\tilde{k}'(z)+\tilde{s}'(z)) & 0 \\ 1 & -\tilde{k}'(z) & \tilde{h}(z) & (\tilde{k}(z)-\tilde{s}(z)) \\ -\tilde{k}'(z) & 1 & (\tilde{k}(z)-\tilde{s}(z)) & \tilde{h}(z) \end{pmatrix},$$

where

$$\tilde{h}(z)=-\frac{\cos\pi z}{\sin\pi z},\quad \tilde{k}(z)=\frac{\sin(\pi z/2)}{\sin\pi z},\quad \tilde{k}'(z)=\frac{\cos(\pi z/2)}{\sin\pi z},$$

$$\tilde{s}(z)=\frac{(1-z)\sin(\pi z/2)}{\sin\pi z},\quad \tilde{s}'(z)=\frac{z\cos(\pi z/2)}{\sin\pi z}.$$

The matrix of symbols $A^1(0,z)$ is of the same type of $A^0(0,z)$. Furthermore

$$\det A^0(0,z)=\det A^1(0,z)=\left(\frac{\sin^2\pi z/2-z^2}{\cos^2\pi z/2}\right)^2\left(\frac{4\sin^2\pi z/2-1}{4\sin^2\pi z/2}\right)^4$$

and this function is zero only at $z=\frac{1}{3}$, $0<\mathrm{Re}\,z<1$. By Definition 1.7 the system (2.9) is elliptic on $(L^p([0,1]))^8$ for $p\neq 3$.

Finally, by (1.3) and (1.4), we can readily compute the index of A.

When $1<p<3$, we will determine the dimension of Ker(A) in $(L^p([0,1]))^8$. To this end we compute the dimension of the kernel of A^*, the adjoint operator of A, in $(L^q([0,1]))^8$, $q>\frac{3}{2}$.

The operator A^* is obtained from the following exterior problem:

$$(2.10)\quad \Delta^2u=0 \quad \text{in } \mathbb{R}^2-\bar{\Omega},\quad \left(-\frac{1}{2}\Delta u\right)\Big|_{\partial\Omega}=h\in L^q(\partial\Omega),\ \frac{\partial^2 u}{\partial u\partial n^-}\Big|_{\partial\Omega}=l\in L^q(\partial\Omega),$$

where n_Q^- is the exterior unit normal to a point $Q\in\partial\Omega$ and τ_Q is defined as (2.3).

To solve the problem (2.10), we introduce the following potential with densities f, $g\in L^q(\partial\Omega)$:

$$(2.11)\qquad u(X)=\int_{\partial\Omega}\frac{\partial F}{\partial\tau_Q}(X-Q)f(Q)\,d\sigma_Q+\int_{\partial\Omega}\frac{\partial F}{\partial n_Q^+}(X-Q)g(Q)\,d\sigma_Q$$

where F is the fundamental solution of Δ^2 defined in (2.2).

Then for $P\in\partial\Omega$:

$$(2.12)\qquad \begin{aligned}&\lim_{\substack{X\to P\\ X\in\mathbb{R}^2-\bar{\Omega}}}-\frac{1}{2}\Delta u(X)=G^*\mathbf{f}(P)+(I-C^*)\mathbf{g}(P),\\ &\lim_{\substack{X\to P\\ X\in\mathbb{R}^2-\bar{\Omega}}}\frac{\partial^2 u}{\partial\tau_P\partial n_P^-}(X)=M^*\mathbf{f}(P)+N^*\mathbf{g}(P),\end{aligned}$$

where G^*, C^*, M^*, N^* are the adjoint operators of G, C, M, N defined by (2.5)[3].

The boundary conditions of the problem (2.10) turn into the following system of integral equations:

$$(2.13)\qquad A^*\begin{pmatrix}\mathbf{f}\\ \mathbf{g}\end{pmatrix}=\begin{pmatrix}G^* & I-C^*\\ M^* & N^*\end{pmatrix}\begin{pmatrix}\mathbf{f}\\ \mathbf{g}\end{pmatrix}=\begin{pmatrix}\mathbf{h}\\ \mathbf{l}\end{pmatrix}.$$

If $q>\frac{3}{2}$ we will show $\dim_q(\mathrm{Ker}(A^*))=1$.

It is easy to verify that $(\mathbf{a},\mathbf{0})\in\mathrm{Ker}A^*$ if $\mathbf{a}$ is the vector $(1,-1,1,-1)$. To show that, modulo multiplicative constants, $(\mathbf{a},\mathbf{0})$ is the unique element of Ker(A^*), we state a lemma concerning the kernel of the operator G^*.

[3] Note that K', S', $-H$ are the adjoint operators of K, S, H respectively, and the compact operators R, V, L, Z are selfadjoint.

LEMMA 2.1. *For $q>\frac{3}{2}$, the kernel of G^* is spanned by* $(1,-1,1,-1)$, *in* $(L^q([0,1]))^4$.

Proof. Let $\mathbf{f}=(f_1,f_2,f_3,f_4)\in \operatorname{Ker} G^*$ and consider the following potential w with density $\mathbf{f}$:

$$w(x,y)=\frac{1}{\pi}\left\{\int_0^1 \frac{x-t}{(x-t)^2+y^2}f_1(t)\,dt+\int_0^1\frac{y-s}{(y-s)^2+x^2}f_2(s)\,ds\right.$$
$$\left.+\int_0^1\frac{1-x-t}{(1-x-t)^2+(1-y)^2}f_3(t)\,dt+\int_0^1\frac{1-y-s}{(1-y-s)^2+(1-x)^2}f_4(s)\,ds\right\}.$$

One can easily verify that w satisfies the following exterior Dirichlet problem:

$$\Delta w=0 \quad \text{in } \mathbb{R}^2-\bar{\Omega}, \quad w|_{\partial\Omega}=0, \quad \lim_{|(x,y)|\to\infty} w(x,y)=0.$$

By the maximum principle w is equal to zero in $\mathbb{R}^2-\Omega$ and, therefore, it is also a solution of the following exterior problem:

$$\begin{array}{ll} \Delta w=0 \quad \text{in } \mathbb{R}^2-\bar{\Omega}, & \\ \displaystyle\lim_{y\to 0^-}\int_{-\infty}^{x}\frac{\partial w}{\partial y}(v,y)\,dv=0, & \displaystyle\lim_{x\to 0^-}\int_{-\infty}^{y}\frac{\partial w}{\partial x}(x,v)\,dv=0, \\ \displaystyle\lim_{y\to 1^+}\int_{1-x}^{+\infty}-\frac{\partial w}{\partial y}(v,y)\,dv=0, & \displaystyle\lim_{x\to 1^+}\int_{1-y}^{+\infty}-\frac{\partial w}{\partial x}(x,v)\,dv=0. \end{array} \tag*{(2.14)[4]}$$

The boundary conditions (2.14) lead to the following system of integral equations:

$$\begin{array}{l} -If_1(x)-Kf_2(x)+Zf_3(x)-K_1f_4(x)=0, \\ -Kf_1(y)-If_2(y)-K_1f_3(y)+Zf_4(y)=0, \\ Zf_1(x)-K_1f_2(x)-If_3(x)-Kf_4(x)=0, \\ -K_1f_1(y)+Zf_2(y)-Kf_3(y)-If_4(y)=0. \end{array} \tag{2.15}$$

Put $\phi=(f_1,-f_2,f_3,-f_4)$, so (2.15) can be written synthetically as

$$(-I+C^*)\Phi=0.$$

$(-I+C^*)$ is the boundary operator which one obtains solving the exterior Dirichlet problem for Δ by a solution like (1.6). In [5] it has been shown that, for $q>\frac{3}{2}$,

$$\operatorname{Ker}(-I+C^*)\cap L^q(\partial\Omega)=\{\text{const.}\}.$$

Therefore, modulo multiplicative constants, we have:

$$f_1=f_3=1, \qquad f_2=f_4=-1.$$

LEMMA 2.2. *If u is the solution of the homogeneous boundary problem* (2.10) *in form* (2.11), *then u has the following expression*:

$$u(x,y)=c(x^2-y^2)+p(x,y), \qquad (x,y)\in\mathbb{R}^2-\bar{\Omega},$$

where $c\in\mathbb{R}$ and $p(x,y)$ is a first degree polynomial.

[4] These boundary conditions can be also expressed by the inverse operator of $\Lambda: L^p_1(\mathbb{R})\to L^p(\mathbb{R})$ s.t. $\mathfrak{F}(\Lambda f)(\xi)=|\xi|\mathfrak{F}f(\xi)$, where $\mathfrak{F}$ is the Fourier transform.

Proof. Since u is a solution of (2.10) in form (2.11), we have that $v=\Delta u$ is a solution of the following problem:

$$\Delta v=0 \quad \text{in } \mathbb{R}^2-\overline{\Omega}, \quad v|_{\partial\Omega}=0, \qquad \lim_{|(x,y)|\to+\infty} v(x,y)=0. \tag{2.16}$$

By the maximum principle, v is equal to zero in $\mathbb{R}^2-\overline{\Omega}$, i.e., u is harmonic in $\mathbb{R}^2-\overline{\Omega}$. On the other hand $w=\partial^2u/\partial x\partial y$ also satisfies the problem (2.16) and therefore w is equal to zero in $\mathbb{R}^2-\overline{\Omega}$.

From $\partial^2u/\partial x\partial y\equiv 0$ in $\mathbb{R}^2-\Omega^2$ it follows that

$$u(x,y)=a(x)+b(y) \quad \forall (x,y)\in\mathbb{R}^2-\overline{\Omega},$$

with a and b twice differential functions.

Moreover, u being harmonic in $\mathbb{R}^2-\overline{\Omega}$, the statement of Lemma 2.2 is true.

PROPOSITION 2.2. *The kernel of A^*, in $(L^q([0,1]))^8$, with $q>\frac{3}{2}$, is spanned by* $(\mathbf{a},\mathbf{0})$, $(\mathbf{a}=(1,-1,1,-1))$.

Proof. Let $(\mathbf{f},\mathbf{g})\in\operatorname{Ker}(A^*)$ and consider the function u defined by (2.10) with densities $\mathbf{f}$ and $\mathbf{g}$. By Lemma 2.2

$$u(x,y)=c(x^2-y^2)+p(x,y) \quad \forall (x,y)\in\mathbb{R}^2-\overline{\Omega}.$$

Consequently, putting

$$w(x,y)=u(x,y)-c(x^2-y^2)-p(x,y) \quad \forall (x,y)\in\mathbb{R}^2,$$

we have

$$w\equiv 0 \quad \text{in } \mathbb{R}^2-\overline{\Omega}$$

and

$$\frac{\partial w}{\partial x}\equiv\frac{\partial w}{\partial y}\equiv 0 \quad \text{in } \mathbb{R}^2-\overline{\Omega}.$$

Since w, $\partial w/\partial x$, $\partial w/\partial y$ are continuous on $\partial\Omega$, we have

$$\lim_{\substack{(x,y)\to P\\(x,y)\in\Omega}} w(x,y)=\lim_{\substack{(x,y)\to P\\(x,y)\in\Omega}}\frac{\partial w}{\partial x}(x,y)=\lim_{\substack{(x,y)\to P\\(x,y)\in\Omega}}\frac{\partial w}{\partial y}(x,y)=0, \qquad P\in\partial\Omega.$$

Therefore w is a solution of the following problem

$$\Delta^2 w=0, \quad w|_{\partial\Omega}=0, \quad \left.\frac{dw}{dn^+}\right|_{\partial\Omega}=0. \tag{2.17}$$

Apply Green's formula to the pair of functions w and Δw:

$$\iint_\Omega(\Delta w\Delta w-w\Delta(\Delta w))\,dx\,dy=\int_{\partial\Omega}\left(\Delta w\frac{dw}{dn^+}-w\frac{d}{dn^+}(\Delta w)\right)d\sigma. \tag{2.18}$$

The integral on the r.h.s. of (2.18), is well defined because the function $(\Delta w(dw/dn^+)-w(d/dn^+)(\Delta w))$ is summable.

The behaviour of this function, in a neighbourhood of vertices of Ω, may be obtained in the same way as in [1, §3].

By (2.17), the boundary integral in (2.18) is zero and therefore

$$\int\int_{\Omega}(\Delta w)^2 dx\,dy=0,$$

from which $\Delta w=0$ in Ω.

Since w is harmonic in Δ, zero on $\partial\Omega$, w is zero in Ω, i.e.

$$u(x,y)=c(x^2-y^2)+p(x,y)\quad \forall(x,y)\in\Omega.$$

By this equality we have that u is harmonic in Ω and $\partial^2u/\partial x\partial y\equiv 0$ in Ω.

Therefore, for $P\in\partial\Omega$

$$(2.19)\qquad \begin{aligned}&\lim_{\substack{X\to P\\ X\in\Omega}}\left(-\frac{1}{2}\Delta u\right)(X)=G^*\mathbf{f}(P)+(-I-C^*)\mathbf{g}(P)=0,\\ &\lim_{\substack{X\to P\\ X\in\Omega}}\frac{\partial^2 u}{\partial\tau_P\partial n_P^+}(X)=M^*\mathbf{f}(P)+N^*\mathbf{g}(P)=0.\end{aligned}$$

Subtracting (2.19) from (2.4) we obtain $\mathbf{g}\equiv\mathbf{0}$. Since $(\mathbf{f},\mathbf{g})\in\operatorname{Ker}A^*$, from (2.13) we deduce, in particular $G^*\mathbf{f}=0$, i.e. $\mathbf{f}\in\operatorname{Ker}G^*$. Applying Lemma 2.1 we can assert that $\mathbf{f}=c\mathbf{a}$, $c\in\mathbb{R}$. Since $\mathbf{a}\in\operatorname{Ker}M^*$ too, we can conclude that $(\mathbf{a},\mathbf{0})$ spans $\operatorname{Ker}A^*$ in $(L^q([0,1]))^8$, $q>\frac{3}{2}$.

PROPOSITION 2.3. *The kernel of A has dimension* 5 *in* $(L^p([0,1]))^8$, $1<p<3$.

Proof. By Proposition 2.1, $\operatorname{ind}_p(A)=4$, $1<p<3$, and by the previous proposition $\dim_q(\operatorname{Ker}(A^*))=1$, $q>\frac{3}{2}$.

Then

$$\dim_p(\operatorname{Ker}(A))=\operatorname{ind}_p(A)+\dim_q\operatorname{Ker}(A^*)=5.$$

PROPOSITION 2.4. *For* $1<p<3$, *the problem* (I) *has a solution as* (2.1), *provided that the data* $h\in L^p(\partial\Omega)$ *satisfies the following compatibility condition*:

$$\int_{\partial\Omega}h\,d\eta\equiv 0.$$

Proof. Let $1<p<3$. Since A is elliptic, A has closed range and so we can apply the Fredholm theory to A.

In particular the equation

$$A\begin{pmatrix}\mathbf{f}\\ \mathbf{g}\end{pmatrix}=\begin{pmatrix}\mathbf{h}\\ \mathbf{l}\end{pmatrix}$$

is solvable in $(L^p([0,1]))^8$ provided $(\mathbf{h},\mathbf{l})$ is orthogonal to the unique element of $\operatorname{Ker}A^*$. Then the following condition must be satisfied:

$$(2.20)\qquad \int_0^1\left(\sum_{i=1}^4 h_i(t)a_i(t)\right)dt=0$$

where a_i are the coordinates of vector $\mathbf{a}$, $(\mathbf{a}=(1,-1,1,-1))$.

Keeping in mind that $\partial\Omega$ is traversed in the clockwise direction and the parametrization of the sides of $\partial\Omega$, the condition (2.20) can be written as follows:

$$\begin{aligned}0&=\int_0^1 h_1(t)\,dt-\int_0^1 h_2(t)\,dt+\int_0^1 h_3(t)\,dt-\int_0^1 h_4(t)\,dt\\&=\int_1^0 h_1(t)\,dt-\int_0^1 h_2(t)\,dt-\int_1^0 h_3(1-t)\,dt-\int_0^1 h_4(1-t)\,dt\\&=-\int_{\partial\Omega} h\,d\eta.\end{aligned}$$

3. In this section we study two further boundary problems for Δ^2 with second order boundary conditions. About the first problem which is the corresponding interior problem of (2.9), we show that there is an unique solution as a proper potential for every boundary data. By the same potential we solve the second problem, provided boundary data verify a proper compatibility condition.

Consider the first problem:

$$(3.1)\qquad \Delta^2u=0\quad\text{in }\Omega,\quad \left.\frac{\partial^2 u}{\partial\tau\partial n^+}\right|_{\partial\Omega}=h\in L^p(\partial\Omega),\quad \left.\frac{1}{2}\Delta u\right|_{\partial\Omega}=l\in L^p(\partial\Omega),$$

where τ_Q is the unit tangent to a point $Q\in\partial\Omega$ defined as (2.3) and n_Q^+ is the interior unit normal to a point $Q\in\partial\Omega$.

For $\mathbf{f},\mathbf{g}\in(L^p([0,1]))^4$ define the following potential[5]

$$\begin{aligned}u(x,y)=\frac{1}{2\pi}\Bigg\{&\int_0^1\left(\int_t^x\log\left((v-t)^2+y^2\right)dv\right)f_1(t)\,dt\\&+\int_0^1\left(\int_s^y\log\left(x^2+(v-s)^2\right)dv\right)f_2(s)\,ds\\&+\int_0^1\left(\int_t^{1-x}\log\left((v-t)^2+(1-y)^2\right)dv\right)f_3(t)\,dt\\&+\int_0^1\left(\int_s^{1-y}\log\left((1-x)^2+(v-s)^2\right)dv\right)f_4(s)\,ds\\&+\int_0^1 y\log\left((x-t)^2+y^2\right)g_1(t)\,dt\\&+\int_0^1 x\log\left(x^2+(y-s)^2\right)g_2(s)\,ds\\&+\int_0^1(1-y)\log\left((1-x-t)^2+(1-y)^2\right)g_3(t)\,dt\\&+\int_0^1(1-x)\log\left((1-x)^2+(1-y-s)^2\right)g_4(s)\,ds\Bigg\}.\end{aligned}\qquad(3.2)$$

[5] The first four terms of u can be regarded as

$$\int_{\partial\Omega}(\Lambda^{-1}\Delta F)(X-Q)f(Q)\,d\sigma_Q$$

(see footnote 4); the last four terms of u represent

$$\int_{\partial\Omega}\frac{\partial F}{\partial n_Q^+}(X-Q)g(Q)\,d\sigma_Q.$$

For $P \in \partial\Omega$ we obtain

$$
\begin{aligned}
&\lim_{\substack{X \to P \\ X \in \Omega}} \frac{\partial^2 u}{\partial \tau_P \partial n_P^+}(X) = B\mathbf{f}(P) - N^*\mathbf{g}(P), \\
&\lim_{\substack{X \to P \\ X \in \Omega}} \frac{1}{2}(\Delta u)(X) = O\mathbf{f}(P) + B\mathbf{g}(P)
\end{aligned}
\tag{3.3}
$$

where B is defined by (1.8) and N^* denotes the adjoint matrix operator of N defined by (2.5).

Then the boundary conditions (3.1) turn into the following system of integral equations:

$$
\begin{pmatrix} B & -N^* \\ O & B \end{pmatrix} \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} = \begin{pmatrix} \mathbf{h} \\ \mathbf{l} \end{pmatrix}.
\tag{3.4}
$$

For this problem we have the following proposition.

PROPOSITION 3.1. *The matrix operator defined in* (3.4) *is elliptic in* $(L^p([0,1]))^8$ *for every* $p \neq \frac{3}{2}$. *Its index is equal to zero for* $p > \frac{3}{2}$ *and is equal to* 8 *for* $1 < p < \frac{3}{2}$. *Furthermore for* $p > \frac{3}{2}$, *its kernel has dimension zero so the matrix* (3.4) *is invertible for* $p > \frac{3}{2}$.

Proof. Since the determinant of the matrix of principal symbol of the operator (3.4), is equal to the square of $\det \sigma_p(B)(t,z)$, the proposition is true by the results of §1 about operator B.

PROPOSITION 3.2. *The problem* (3.1) *has a unique solution in the form* (3.2) *for all data h and l belonging to* $L^p(\partial\Omega)$, *with* $p > \frac{3}{2}$.

Now we study the following problem

$$
\Delta^2 u = 0 \quad \text{in } \Omega, \quad \left.\frac{\partial^2 u}{\partial \tau^2}\right|_{\partial\Omega} = h \in L^p(\partial\Omega), \quad \left.\frac{1}{2}\Delta u\right|_{\partial\Omega} = l \in L^p(\partial\Omega),
\tag{3.5}
$$

where τ_Q is the unit tangent to a point $Q \in \partial\Omega$ defined as (2.3).

We seek a solution u of problem (3.5) as the potential (3.2). When we impose the boundary conditions of problem (3.5) to this u we have for $P \in \partial\Omega$

$$
\begin{aligned}
&\lim_{\substack{X \to P \\ X \in \Omega}} \frac{\partial^2 u}{\partial \tau_P^2}(X) = D\mathbf{f} + E\mathbf{g}, \\
&\lim_{\substack{X \to P \\ X \in \Omega}} \frac{1}{2}\Delta u(X) = O\mathbf{f} + B\mathbf{g},
\end{aligned}
\tag{3.6}
$$

where D and E are the following matrices of operators in $\mathrm{OP}\Sigma_{1/p}([0,1])$
(3.7)

$$
D = \begin{bmatrix} H & K' & R & -K_1' \\ K' & H & -K_1'2 & R \\ R & -K_1' & H & K' \\ -K_1' & R & K' & H \end{bmatrix}, \quad E = \begin{bmatrix} 0 & K+S & V & K_1+S_1 \\ K+S & 0 & K_1+S_1 & V \\ V & K_1+S_1 & 0 & K+S \\ K_1+S_1 & V & K+S & 0 \end{bmatrix}
$$

and B is defined by (1.8).

Then the boundary conditions of (3.5) turn into the following system of integral equations:

$$J\begin{pmatrix}\mathbf{f}\\ \mathbf{g}\end{pmatrix}=\begin{pmatrix}D & E\\ O & B\end{pmatrix}\begin{pmatrix}\mathbf{f}\\ \mathbf{g}\end{pmatrix}=\begin{pmatrix}\mathbf{h}\\ \mathbf{l}\end{pmatrix}. \tag{3.8}$$

Proposition 3.3. *The matrix of operators* (3.8) *is elliptic in* $(L^p([0,1]))^8$ *for every* $p\neq\frac{3}{2}$ *and its index is equal to zero when* $p>\frac{3}{2}$, *and to eight when* $1<p<\frac{3}{2}$.

Proof. By Definition 1.7 we must compute the determinant of the matrix of principal symbols of J on $\partial R_{1/p,[0,1]}$.

This determinant is equal to 16 on the top and the bottom of $\partial R_{1/p,[0,1]}$ and is equal to $((4\sin 2\pi/2z-3)/\cos 2\pi z/2)^4$ on the right-hand side and the left-hand side of $\partial R_{1/p,[0,1]}$.

The zeros of $4\sin^2\pi z/2-3$ occur only at $z=\frac{2}{3}$ and therefore J is elliptic for $p\neq\frac{3}{2}$. From (1.3) and (1.4) it follows

$$\operatorname{ind}_p(J)=\begin{cases}0, & p>\frac{3}{2},\\ 8, & 1<p<\frac{3}{2}.\end{cases}$$

Now we will study the kernel of the boundary operator J when $p>\frac{3}{2}$.

It is easy to see that $(\mathbf{1},\mathbf{0})$ $(L^p([0,1]))^8$ is in the kernel of J. We will show that $\dim_p\operatorname{Ker}(J)=1$.

Lemma 3.1. *Let* $(\mathbf{f},\mathbf{g})\in(L^p([0,1]))^8$ *be in* $\operatorname{Ker}(J)$, *then*

$$\mathbf{g}=\mathbf{0}.$$

Proof. By (3.8) if $(\mathbf{f},\mathbf{g})\in\operatorname{Ker}J$, g satisfies in particular the homogeneous system $B\mathbf{g}=0$. Then $g\in\operatorname{Ker}B\cap(L^p([0,1]))^4$, $p>3/2$, and by results of §1, $\mathbf{g}=\mathbf{0}$.

Proposition 3.4. *Let be* $p>\frac{3}{2}$, *then*

$$\operatorname{Ker}J\cap(L^p([0,1]))^8=\{(\mathbf{c},\mathbf{0})|c\in\mathbb{R}\}. \tag{3.9}$$

Proof. By Lemma 3.1, if $(\mathbf{f},\mathbf{g})\in\operatorname{Ker}J\cap(L^p([0,1]))^8$, $p>3/2$, then $\mathbf{g}\equiv\mathbf{0}$. Then $\mathbf{f}\in\operatorname{Ker}D\cap(L^p([0,1]))^4$. Comparing the matrix D with the matrix G^* and then using Lemma 2.1 we can state that $\mathbf{f}=c\mathbf{1}$.

In order to apply Fredholm theory to the operator J we consider the following exterior problem for Δ^2:

$$\begin{aligned}&\Delta^2u=0 \quad \text{in } \mathbb{R}^2-\bar{\Omega},\\ &-\frac{1}{2}\Lambda^{-1}(\Delta u)\Big|_{\partial\Omega}=h\in L^q(\partial\Omega),\\ &\frac{\partial u}{\partial n^-}\Big|_{\partial\Omega}=l\in L^q(\partial\Omega).\end{aligned} \tag{3.10^6}$$

Writing the solution as

$$u(X)=\frac{1}{2}\int_{\partial\Omega}\frac{\partial^2F}{\partial\tau_Q^2}(X-Q)f(Q)\,d\sigma_Q+\int_{\partial\Omega}\Delta F(X-Q)g(Q)\,d\sigma_Q,$$

[6] See footnote 4 for the definition of Λ^{-1}.

we obtain the adjoint operator of J as a boundary operator. By Proposition 3.4 we have

$$\dim(\operatorname{Ker} J^*) \cap \left((L^q([0,1]))^8\right) = 1 \quad \text{when } 1<q<3.$$

Let us denote by (ω, ω') an element of $\operatorname{Ker} J^*$.

Applying Fredholm's theory we can state the following proposition.

PROPOSITION 3.5. *If the data of problem* (3.5), *h and l, are in $L^p(\partial\Omega)$ with $p>\frac{3}{2}$ and satisfy the compatibility conditions*

$$\int_{\partial\Omega} h\omega d\sigma = 0 = \int_{\partial\Omega} l\omega' d\sigma,$$

then the problem (3.5) *has a solution in the form* (3.2).

Acknowledgment. We wish to thank Prof. J. Lewis for useful conversations about this work.

REFERENCES

[1] L. DIOMEDA AND B. LISENA, *Problemi di Dirichlet e di derivata obliqua per l'operatore Δ^2 in un settore piano*, Rendiconti di Matematica, 1 2, Ser. VII (1982), pp. 189–217.

[2] P. GRISVARD, *Alternative de Fredholm relative au problème de Dirichlet dans un polygone ou un polyedre*, Boll. U.M.I., 5 (1972), pp. 132–164.

[3] ______, *Singularitè des solutions du problème de Stokes dans un polygone*, to appear.

[4] J. E. LEWIS AND C. PARENTI, *Pseudodifferential operators and Hardy kernels on $L^p(\mathbb{R}_+)$*, Ann. Scuola Normale di Pisa, VII (3), (1980), pp. 481–503.

[5] ______, *Pseudodifferential operators of Mellin type*, Comm. PDE, to appear.

[6] F. JOHN, *Plane waves and spherical means applied to partial differential equations*, Interscience Tracts in Pure and Applied Mathematics, 2, Wiley Interscience, New York, 1955.

[7] M. MERIGOT, *Solutions en normes L^p des problèmes elliptiques dans des polygones plans*, These pour docteur d'ètat en Math, Universitè Nice, 1974.

[8] J. NECAS, *L'extension de l'espace des conditions aux limites du problème biharmonique pour les domains à points anguleux*, Czechslovak Math. J., 9 (1959), pp. 339–371.

[9] S. M. NIKOL'SKII, *Boundary properties of functions defined on a region with angular points* II–*Harmonic functions on rectangular regions.* AMS (2), 83, (1969), pp. 121–141.

[10] K. REKTORYS AND V. ZAHRADNIK, *Solutions of the first biharmonic problem by the method of least squares on the boundary*, Aplikace Matematiki, 19 (1974), pp. 101–131.

[11] J. COHEN AND J. GOSSELIN, *The Dirichlet problem for the biharmonic equation in a C^1-domain in the plane*, to appear.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics

010

ON A POLYNOMIAL BASIS IN SOME SPACES OF ANALYTIC FUNCTIONS*

YU. A. KAZMIN†

Abstract. In the article we consider for polynomial systems the problem of stability of the property of forming a basis relative to some perturbations in a number of spaces of analytic functions. Some sufficient conditions are obtained in this direction. It is shown that the conditions mentioned above are sharp in some particular cases.

Introduction. In this article we consider three (in some senses similar) Banach spaces of functions defined as follows:

$$\text{I.}\qquad l_1^+(r) \overset{\text{def}}{=} \left\{ x^+(t)=\sum_{k=0}^{\infty} x_k^+ t^k : \|x^+\|_r=\sum_{k=0}^{\infty} |x_k^+| r^k < +\infty \right\};$$

$$\text{II.}\qquad l_1^-(r) \overset{\text{def}}{=} \left\{ x^-(t)=\sum_{k=0}^{\infty} \frac{x_k^-}{t^{k+1}} : \|x^-\|_r=\sum_{k=0}^{a} \frac{|x_k^-|}{r^{k+1}} < +\infty \right\};$$

$$\text{III.}\qquad l_1(r) \overset{\text{def}}{=} \left\{ x(t)=\sum_{-\infty}^{\infty} x_k t^k : \|x\|_r=\sum_{-\infty}^{\infty} |x_k| r^k < +\infty \right\}.$$

These three spaces of absolutely convergent power series are defined for every positive parameter r, $0<r<\infty$ (but r is fixed for each space). It is obvious that, for given r, $0<r<+\infty$, we have $l_1(r)=l_1^+(r)\oplus l_1^-(r)$. In a similar way we can introduce the spaces $l_p^{\pm}(r)$ and $l^p(r)$ for all r, $0<r<\infty$ and all p, $1\le p\le\infty$. All these spaces are Banach spaces. For example,

$$l_\infty^-(r) \overset{\text{def}}{=} \left\{ y^-(t)=\sum_{k=0}^{\infty} \frac{y_k^-}{t^{k+1}} : \|y^-\|_{r,\infty}=\sup_k \frac{|y_k^-|}{r^{k+1}} < +\infty \right\}.$$

Of course, for example, the series

$$\sum_{k=0}^{\infty} \frac{y_k^-}{r^{k+1}} \in l_\infty^-(1)$$

are, in the general case, formal power series on the unit circle $|z|=1$, but in each case they define functions

$$y^-(t)=\sum_{k=0}^{\infty} \frac{y_k^-}{z^{k+1}}$$

which are analytic for all z, $|z|>1$.

Let us note that the space $l_\infty^-(r)$ is the dual space of $l_1^+(r)$. This will be very important later on.

Let B, $B\neq\varnothing$, be a set of functions in $l_1^+(r)$ with the following properties:

(i) $\forall\varphi(t)=\sum_{k=0}^{\infty}\varphi_k t^k\in B$ is equal to 1 at the origin, i.e. $\varphi_0=\varphi(0)=1$;

* Received by the editors July 14, 1983.

† Department of Mathematics, Moscow State University, Moscow 117 234 USSR.

(ii) for any sequence $\Phi=\{\varphi_k(t)\}_{k=0}^{\infty}$, $\varphi_k(t)\in B$, $\forall k$, the system

$$(1)\qquad \{t^k\varphi_k(t)\}_{k=0}^{\infty}$$

forms a quasi-power basis in $l_1^+(r)$.

The last statement means that if we define the operator T_Φ corresponding to the sequence $\Phi=\{\varphi_k(t)\}$, $\varphi_k\in B$ in the following way:

$$\text{For } \forall x^+(t)=\sum_{k=0}^{\infty} x_k^+t^k\in l_1^+(r) \text{ let } T_\Phi\circ x^+=\sum_{k=0}^{\infty} x_k^+t^k\varphi_k(t),$$

then T_Φ is a linear continuous operator that effects a one-to-one mapping of $l_1^+(r)$ into itself. In other words, T_Φ is an automorphism of $l_1^+(r)$ for any $\Phi=\{\varphi_k\}_{k=0}^{\infty}$, $\varphi_k\in B$, $\forall k$.

We can now formulate the main result of this paper.

THEOREM 1. *Every sequence $\Phi=\{\varphi_k(t)\}_{k=0}^{\infty}$, $\varphi_k\in B$, $\forall k$, where B is any set in $l_1^+(r)$ with properties* (i) *and* (ii), *generates a polynomial system*

$$(2)\qquad P_k^\Phi\left(\frac{1}{z}\right)=\frac{1}{2\pi i}\int_{|t|=r}\frac{\varphi_k(t)\,dt}{t^{k+1}(t-z)},\qquad |z|>r\quad (k=0,1,2,\cdots),$$

which is a quasi-power basis in $l_1^-(r)$.

Section 1 contains the proof of Theorem 1. In §2 we consider a number of applications of that theorem. For example, we have the following corollary.

COROLLARY 1. *Every polynomial system $\{(z+\alpha_n)^n\}$, $\alpha_n\in R$, $-1\le\alpha_n\le 1$ forms a quasi-power basis in the space of entire functions of exponential type* $[1;\pi/4)$. *Here the constant $\pi/4$ is sharp.*

Here (as usual) $[1;\pi/4)$ is the space of entire functions of order at most 1 and of type less than $\pi/4$ if of order 1. For references concerning polynomial bases, see, for example, [1], [2], [3] and [4].

1. Proof of Theorem 1. We precede the proof of the theorem with some auxiliary propositions.

LEMMA 1. *Under the hypotheses of Theorem* 1, *the linear continuous operators*

$$T_\Phi\left[\sum_{k=0}^{\infty} x_kt^k\right]=\sum_{k=0}^{\infty} x_kt^k\varphi_k(t)$$

are uniformly bounded for $\forall\Phi=\{\varphi_k\}$, $\varphi_k\in B$, $\forall k$.

The proof is obvious, because B is a bounded set in the space $l_1^+(r)$ and also satisfies condition (i). Hence we have

$$1\le\sup_{\varphi\in B}\|\varphi\|_r=M<+\infty.$$

From this it is easy to obtain $\|T_\Phi\circ x^+(t)\|_r\le M\|x^+(t)\|_r$ for $\forall x^+\in l_1^+(r)$ and $\forall\Phi=\{\varphi_k\}$, $\varphi_k\in B$, $\forall k$.

LEMMA 2. *Under the hypotheses of Theorem* 1, *the linear continuous operators T_Φ^{-1}, $\forall\Phi$ are also uniformly bounded.*

Proof. It is sufficient to show that $\exists m$, $0<1/m<+\infty$, such that $\|T_\Phi^{-1}\|<1/m$ for $\forall\Phi=\{\varphi_k\}_{k=0}^{\infty}$, $\varphi_k\in B$, $\forall k$. Let us suppose for simplicity that $r=1$. Under this assumption the dual space of $l_1^+(1)$ is the space $l_\infty^-(1)$ of all formal power series $y^-(t)=\sum_{k=0}^{\infty}y_k^-/t^{k+1}$, $\|y^-\|=\sup_k|y_k^-|<+\infty$ (as described in the Introduction). Let us notice that the system biorthogonal to a system (1), namely $\{t^k\varphi_k(t)\}$, $\varphi_k(0)=1$, is always some polynomial sequence of the form

$$Q_n^\Phi\left(\frac{1}{t}\right)=\begin{vmatrix} 1 & \varphi_0'(0) & \dfrac{\varphi_0''(0)}{2!} & \cdots & & \dfrac{\varphi_0^{(n)}(0)}{n!} \\ 0 & 1 & \varphi_1'(0) & \cdots & & \dfrac{\varphi_1^{(n-1)}(0)}{(n-1)!} \\ 0 & 0 & 1 & \cdots & & \dfrac{\varphi_2^{(n-2)}(0)}{(n-2)!} \\ & \cdots & & & & \cdots \\ 0\cdot & 0 & 0 & \cdots & 1 & \varphi_{n-1}(0) \\ \dfrac{1}{t} & \dfrac{1}{t^2} & \dfrac{1}{t^3} & \cdots & \dfrac{1}{t^n} & \dfrac{1}{t^{n+1}} \end{vmatrix},$$

and $Q_n^\Phi(1/t)=(T_\Phi^{-1})^*\circ t^{-n-1}$, where $(T_\Phi^{-1})^*$ is the operator conjugate to T_Φ^{-1}.

If the family $\{T_\Phi^{-1}\}$, $\forall\Phi$ were not bounded, then the family of conjugate operators $\{(T_\Phi^{-1})^*\}$, $\forall\Phi$ would also be unbounded. By the Banach–Steinhaus theorem there would exist a fixed element $y_0^-(t)$, $y_0^-(t)\in l_\infty^-(1)$ (this space, as mentioned previously, is the dual space of $l_1^+(1)$) and a sequence of operators $(T_n^{-1})^*$ (the elements of the sequence belong to the family $\{(T_\Phi^{-1})^*\}$, $\forall\Phi$) such that we would have

$$\text{(3)}\qquad \|(T_n^{-1})^*\circ y_0^-\|_{l_\infty^-(1)}\to\infty \quad\text{when } n\to\infty.$$

But (3) implies that there exists a subsequence $\{1/t^{k_n+1}\}$ such that also

$$\text{(4)}\qquad \left\|(T_n^{-1})^*\circ\frac{1}{t^{k_n+1}}\right\|_{l_\infty^-(1)}\to\infty \quad\text{when } n\to\infty.$$

For if (4) does not hold then (3) does not hold, and we have a contradiction.

It is not difficult to deduce from (4) that there are numbers j_{k_n}, $0\le j_{k_n}<k_n$, such that

$$\text{(5)}\qquad \left|\left\langle\left[(T_n^{-1})^*\circ\frac{1}{t^{k_n+1}}\right]\cdot t^{j_{k_n}}\right\rangle\right|=\left|\frac{1}{2\pi i}\int_{|t|=1}\left[(T_n^{-1})^*\circ\frac{1}{t^{k_n+1}}\right]\cdot t^{j_{k_n}}dt\right|\to\infty.$$

Now let $j_{k_n}^*=\max\{j_k\}$ for which (5) is true (maximum over k for each fixed n). If we consider the structure of the polynomials $Q_n^\Phi(1/t)=(T_n^{-1})^*(1/t^{n+1})$, we see that $j_{k_n}^*$ is always finite and moreover $0\le j_{k_n}^*<k_n$. From the relation

$$\left|\left\langle\left[(T_n^{-1})^*\circ\frac{1}{t^{k_n+1}}\right]\cdot t^{j_{k_n}^*}\right\rangle\right|\to\infty$$

we easily obtain

$$\left|\left\langle\left[(T_n^{-1})^*\circ\frac{1}{t^{k_n+1}}\right]\cdot t^{j_{k_n}^*}\varphi_{j_{k_n}^*}(t)\right\rangle\right|\to\infty.$$

But this means that

$$\delta_{k_n j^*_{k_n}} = \left\langle \left[(T_n^{-1})^* \frac{1}{t^{k_n+1}} \right] \cdot T_n t^{j^*_{k_n}} \right\rangle \to \infty,$$

where δ_{kj} is the Kronecker symbol. We recall that $j^*_{k_n}$ is always less than k_n, and consequently we have $0 \to \infty$. This contradiction completes the proof of Lemma 2.

LEMMA 3. *Under the hypotheses of Theorem* 1, *for* $\forall \Phi = \{\varphi_k\}_0^\infty$, $\varphi_k \in B$, $\forall k$ $(B \subset l_1^+(r))$ *and* $\forall x^+(t) \in l_1^+(r)$, *we have*

$$m\|x^*\|_r \leq \|T_\Phi \circ x^+\|_r \leq M\|x^+\|_r,$$

where $0 < m \leq M < +\infty$.

We emphasize that the constants m and M are the same for all T_Φ, $\forall \Phi$.

This lemma is an immediate consequence of Lemmas 1 and 2.

Now we can prove Theorem 1. For the sake of simplicity we assume that $r = 1$.

We introduce an operator S_Φ, defined initially only on the basis vectors t^{-k-1} by the formula

$$S_\Phi \circ \frac{1}{t^{k+1}} = P_k^\Phi\left(\frac{1}{t}\right), \qquad k = 0, 1, 2, \cdots,$$

where the $P_k(1/t)$, $k = 0, 1, \cdots$, are the polynomials (2) generated by a sequence $\Phi = \{\varphi_k\}_{k=0}^\infty$, $\varphi_k \in B$, $\forall k$. It is easy to see that the operator S_Φ can be extended to a linear continuous operator on the entire space $l_1^-(1)$, mapping $l_1^-(1)$ into itself. We do this as follows. We have

$$S_\Phi \circ \left(\sum_{k=0}^\infty \frac{x_k^-}{t^{k+1}} \right) = \sum_{k=0}^\infty x_k^- P_k^\Phi\left(\frac{1}{t}\right)$$

for any element

$$x^-(t) = \sum_{k=0}^\infty \frac{x_k^-}{t^{k+1}} \in l_1^-(1).$$

It is evident that the family of operators $\{S_\Phi\}$, $\forall \Phi$, $\Phi = \{\varphi_k\}$, $\varphi_k \in B$, $\forall k$ is uniformly bounded. We also notice that the range $S_\Phi \circ (l_1^-(1))$ of each operator S_Φ is everywhere dense in $l_1^-(1)$. Hence we conclude that if there exists a sequence $\Phi = \{\varphi_k\}_{k=0}^\infty$, $\varphi_k \in B$, $\forall k$ for which the system (2) is not a basis, then there exists

$$x_0^-(t) = \sum_{k=0}^\infty \frac{x_k}{t^{k+1}}, \qquad x_0^-(t) \in l_1^-(1), \qquad \|x_0^-\| = \sum |x_k^-| > 0$$

such that $S_\Phi \circ x_0^-(t) = 0$, i.e.

$$\sum_{k=0}^\infty x_k^- P_k\left(\frac{1}{t}\right) = 0.$$

However, from this relation we have

$$\sum_{k=0}^\infty x_k^- \frac{\varphi_k(t)}{t^{k+1}} = -x^+(t), \tag{6}$$

where $x^+(t)=\Sigma_{k=0}^{\infty}x_k^+t^k\psi_k(t)$ is some element of $l_1^+(r)$. Here $\psi(t)=\{\psi_k\}$, $\forall\psi_k\in B$, $\forall k$ is a sequence of elements of B; and in (6)–(7) we are to understand that all equations are to be interpreted in the sense of the metric of the space $l_1(1)$.

From (6) we obtain

$$\sum_{k=0}^{\infty}\frac{x_k^-\varphi_k(t)}{t^{k+1}}+\sum_{k=0}^{\infty}x_k^+t^k\psi_k(t)\equiv 0,$$

and also, $\forall n$, $n\in N$,

$$\sum_{k=0}^{\infty}\frac{x_k^-\varphi_k(t)}{t^{k-n+1}}+t^n\sum_{k=0}^{\infty}x_k^+t^k\psi_k(t)\equiv 0. \tag{7}$$

This means that

$$\left\|\sum_{k=0}^{n+1}x_k^-\varphi_k(t)t^{n-k-1}+t^n\sum_{k=0}^{\infty}x_k^+t^k\psi_k(t)\right\|_r\to 0$$

when $n\to\infty$.

However, by Lemma 3 we have

$$\left\|\sum_{k=0}^{n+1}x_k^-t^{n-k+1}\varphi_k(t)+t^n\sum_{k=0}^{\infty}x_k^+t^k\psi_k(t)\right\|\geq m\left[\sum_{k=0}^{n+1}|x_k^-|+\sum_{k=0}^{\infty}|x_k^+|\right]$$

$\forall n$, $n\in N$. But $\Sigma_{k=0}^{\infty}|x_k^-|>0$ and $m>0$. Hence we find that the last relation contradicts (4). This completes the proof of Theorem 1.

2. Some applications. We now turn to some applications of Theorem 1. The space $A(|z|<R)$ is defined as the space of functions that are analytic in the disk $|z|<R$, with the usual topology of uniform convergence on compact subsets K of the disk. It is known that if the system $\{z^n\varphi_n(z)\}_{n=0}^{\infty}$ is a quasi-power basis in $A(|z|<R)$ then it is also a quasi-power basis in every $A(|z|<r)$ $\forall r$, $0<r\leq R$. Moreover, it is also a quasi-power basis in every $l_1^+(r)$ $\forall r$, $0<r<R$. It is also obvious that if $\{z^n\varphi_n(z)\}$ is a quasi-power basis in any $l_1^+(r)$, $\forall r$, $0<r<R$, then it is simultaneously a quasi-power basis in every $A(|z|<r)$, $\forall r$, $0<r\leq R$.

From results concerning the Abel-Goncharov problem, it follows that all systems of the form $\{z^ne^{\alpha_nz}\}_{n=0}^{\infty}$, $\alpha_n\in[-1,1]$ are quasi-power bases in the space $A(|z|<\pi/4)$; and the constant $\pi/4$ is sharp. It is also well known that the same systems, but with complex α_n, $|\alpha_n|\leq 1$, are always quasi-power bases in $A(|z|<W)$, where $W=0.7377\ldots$ is the so-called Whittaker constant.

Let $[1;\sigma)$ be the space of entire functions of order at most 1 and of type less than σ if of order 1. In other words,

$$[1;\sigma)\overset{\text{def}}{=}\left\{f(z)=\sum\frac{a_k}{k!}z^k:\ \limsup_{k\to\infty}|a_k|^{1/k}<\sigma\right\}.$$

We now consider the problem of when a polynomial system of the form

$$\left\{(z+\alpha_n)^n\right\}_{n=0}^{\infty},\qquad -1\leq\alpha_n\leq 1 \tag{8}$$

is a quasi-power basis; these systems were mentioned in the Introduction (see Corollary 1). We have the representation

$$(9) \qquad \frac{(z+\alpha_n)^n}{n!}=\frac{1}{2\pi i}\int_{|t|=r}\frac{e^{t(z+\alpha_n)}}{t^{n+1}}dt.$$

If the system (8) is not a basis in the space $[1,\pi/4)$, then there exists a sequence of complex numbers a_k^* such that $\limsup_{k\to\infty}|a_k^*|^{1/k}=\rho<\pi/4$ and $\{a_k^*\}\not\equiv\{0\}$, for which we have identically

$$(10) \qquad \sum_{k=0}^{\infty}\frac{a_k^*}{k!}(z+\alpha_k)^k\equiv 0$$

for some sequence $\{\alpha_k\}$, $-1\leq\alpha_k\leq 1$. By using the representation (9), we can rewrite (10) in the form

$$(11) \qquad \frac{1}{2\pi i}\int_{|t|=\sigma}e^{zt}\left[a_k^*\frac{e^{\alpha_k t}}{t^{k+1}}\right]dt=0.$$

But from (11) we obtain, identically in the $l^1(r)$ sense,

$$(12) \qquad \sum_{k=0}^{\infty}a_k^*\frac{e^{\alpha_k t}}{t^{k+1}}=x^+(t),$$

where $x^+(t)\in l_1^+(r)$, $\forall r$, $\rho<r<\pi/4$. Then, by Theorem 1, if (12) holds, there are nonbasic sequences among the systems $\{z^n e^{\alpha_n z}\}$, $-1\leq\alpha_n\leq 1$. But this contradicts the results quoted above. Hence the positive part of Corollary 1 is established. Also, the example

$$\sin\frac{\pi}{4}(z-1)+\cos\frac{\pi}{4}(z+1)\equiv 0$$

shows that the constant $\pi/4$ in Corollary 1 is sharp. Similar considerations for the family of polynomial systems

$$(13) \qquad \{(z+\alpha_n)^n\}_{n=0}^{\infty}, \qquad |\alpha_n|\leq 1,$$

where α_n are complex numbers, let us obtain the following corollary.

COROLLARY 2. *The system* (13) *is a quasi-power basis in the space* $[1, W)$, $W=0.7377\ldots$.

In [3] and [4] we previously considered the problem of when systems

$$(14) \qquad \{(z+\omega^n)^n\}_{n=0}^{\infty}, \qquad |\omega|=1$$

are quasi-power bases. Our main results were as follows. The system (14) is a quasi-power basis in the space $[1;|\sigma_\omega|)$, where σ_ω is the root closest to the origin of the equation

$$\sum_{n=0}^{\infty}\frac{\omega^{-n(n+1)/2}}{n!}t^n=0.$$

The system (14) is not a basis in the space $[1,|\sigma_\omega|]$, and thus is not a basis in any space $[1,\sigma)$, $\sigma>|\sigma_\omega|$. We define

$$\Omega=\inf_{|\omega|=1}|\sigma_\omega|.$$

It is known that $\Omega=0.7377\ldots$, and it is conjectured that $W=\Omega$ (but this is still unproved). Now we can easily obtain a proposition which is in some sense a negative addition to Corollary 2.

COROLLARY 3. *Among the polynomial systems* (14) *there exists a system that is not a basis in the space* $[1,\Omega]$ *and thus also not in any* $[1,\Omega+\varepsilon)$ $\forall\varepsilon$, $\varepsilon>0$.

If we return to the beginning of the present article and consider polynomial systems $\{P_n(1/z)\}$ from a different point of view, we can interpret Theorem 1 as a proposition concerning the stability of the property of being a basis for generalized Appell polynomials. Let us recall the definition of these polynomials. One version of the definition is as follows. Let $\varphi(z)$ be an element of the space $A(|z|<R)$, $\varphi(z)\not\equiv 0$. Then

$$P_n^\varphi\left(\frac{1}{z}\right)=\frac{1}{2\pi i}\int_{|t|=R-\varepsilon}\frac{\varphi(t)\,dt}{t^{n+1}(t-z)},\qquad |z|>R \tag{15}$$

is a system of polynomials. If we now form the Borel transform of (15), i.e.

$$A_n^\varphi(z)=\frac{1}{2\pi i}\int_{|t|=R}e^{zt}P_n^\varphi\left(\frac{1}{t}\right)dt,\qquad n=0,1,2,\cdots, \tag{16}$$

we obtain the classical Appell polynomials $\{A_n(z)\}_{n=0}^{\infty}$ generated by $\varphi(t)$. The system (16) is a quasi-power basis in the space $[1;R)$ if and only if $\varphi(z)$ has no zeros in the disk $|z|<R$. Now the system (15) is a special case of the systems (2) investigated in the present paper, and the corresponding sequences (similar to (16)) generated by (2) can be considered as generalized systems of Appell polynomials.

Now we observe that when $-1\le\alpha_n\le 1$, there is an identity, in the sense of the space $l_1(r)$,

$$\sum_{k=0}^{\infty}a_k\frac{e^{\alpha_k t}}{t^{k+1}}=x^+(t), \tag{17}$$

where $\{a_k\}$ is a sequence of complex numbers with $\limsup_{k\to\infty}|a_k|^{1/k}=\pi/4$, and $\{\alpha_k\}$ is a sequence of real numbers belonging to the interval $[-1,1]$, and $x^+(t)\in l_1^+(r)$. If we now define

$$e_n(t)=\sum_{k=0}^{n}\frac{t^k}{k!},\qquad n=0,1,\cdots,$$

and notice that

$$\frac{1}{t^{k+1}}=\frac{(-1)^k}{k!}\frac{d^k}{dt^k}\left(\frac{1}{t}\right),$$

then the following results are easily obtained from (17).

I. There exists a linear homogeneous differential equation with polynomial coefficients

$$\sum_{k=0}^{\infty}\frac{a_k e_k(\alpha_k t)}{k!}y^{(k)}(t)=0, \tag{18}$$

with $\limsup_{k-\infty}|a_k|^{1/k}=r\ge\pi/4$ and $-1\le\alpha_k\le 1$, and $\deg e_k(t)=k$, that has the solution $y(t)=1/t$.

II. There are no linear homogeneous differential equations of type (18) with $\limsup_{k\to\infty}|a_k|^{1/k}<\pi/4$ that have the solution $y(t)=1/t$.

REFERENCES

[1] J. M. Whittaker, *Sur les séries de base de polynômes quelconques*, Gauthier-Villars, Paris, 1949.

[2] R. P. Boas and R. C. Buck, *Polynomial Expansions of Analytic Functions*, Springer-Verlag, Berlin, Göttingen, Heidelberg, 1964, 1967.

[3] Yu. A. Kazmin, *On expansions in series of powers* $(z-q^n)^n$, Mat. Zametki, 1, 6 (1967), pp. 683–688.

[4] ______, *On the values of M. T. Eweida's constants*, Vestnik Moskov. Univ. Ser. Mat. Meh., 2 (1978), pp. 88–90.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics

011

ASYMPTOTICS FOR ORTHOGONAL POLYNOMIALS ASSOCIATED WITH exp($-x^4$)*

PAUL NEVAI†

Abstract. A Plancherel–Rotach type asymptotic expansion is given for the orthogonal polynomials associated with exp($-x^4$).

AMS-MOS subject classification (1980). Primary 42C05

Key words. orthogonal polynomials, asymptotics

Let $w(x)=\exp(-x^4)$, $x\in\mathbb{R}$, and let $p_n(x)=\gamma_n x^n+\cdots$, $\gamma_n>0$, denote the orthonormal polynomials corresponding to w. The properties of these polynomials p_n have been reviewed in [4]. In particular, it was pointed out that they satisfy the recurrence formula

$$xp_n=a_{n+1}p_{n+1}+a_np_{n-1}, \tag{1}$$

$n=0,1,2,\cdots$ where the recursion coefficients a_n can successively be determined from the equation

$$n=4a_n^2(a_{n+1}^2+a_n^2+a_{n-1}^2), \tag{2}$$

$n=1,2,\cdots$, $a_0^2=0$ and $a_1^2=\Gamma(\frac{3}{4})/\Gamma(\frac{1}{4})$. Lew and Quarles [3] have found an asymptotic series for a_n in (1). We will only use a simplified version of this asymptotic series which states that

$$a_n^2=\left(\frac{n}{12}\right)^{1/2}[1+24^{-1}n^{-2}+O(n^{-4})] \tag{3}$$

uniformly for $n=1,2,\cdots$ It was also shown in [4] that if

$$\phi_n(x)=a_{n+1}^2+a_n^2+x^2 \tag{4}$$

and

$$z(x)=p_n(x)[\phi_n(x)]^{-1/2}\exp\left(-\frac{x^4}{2}\right), \tag{5}$$

then z satisfies the differential equation

$$z''+f_nz=0 \tag{6}$$

where f_n is defined by

$$\begin{aligned} f_n(x)=4a_n^2\big[4\phi_n(x)\phi_{n-1}(x)+1-4a_n^2x^2-4x^4-2x^2\phi_n(x)^{-1}\big]\\ -4x^6-4x^4\phi_n(x)^{-1}-3x^2\phi_n(x)^{-2}+6x^2+\phi_n(x)^{-1}. \end{aligned} \tag{7}$$

* Received by the editors April 5, 1983. This material is based upon research supported by the National Science Foundation under grant MCS 81-01720.

† Department of Mathematics, Ohio State University, Columbus, Ohio 43210.

Richard Askey has recently informed me that both (2) and (6) were known to J. Shohat who published them in [5, p. 407]. My proof of (6) is identical to the one found by Shohat. Applying an improvement of Liouville-Stekloff's method in [4], I used the differential equation (6) to obtain asymptotics for p_n. Taking (3) into consideration this asymptotic expansion takes the form

$$(8)\qquad \exp\left(-\frac{x^4}{2}\right)p_n(x)$$
$$=An^{-1/8}\sin\left\{\left(\frac{64}{27}\right)^{1/4}n^{3/4}x+(12)^{-1/4}n^{1/4}x^3-(n-1)\frac{\pi}{2}\right\}+o(n^{-1/8}),$$

$n=1,2,\cdots$, with some positive constant A, and (8) holds uniformly when x belongs to a fixed interval.

In this paper we will find Plancherel–Rotach type asymptotics for the polynomials p_n. This means that if X_n denotes the largest zero of p_n and $0<c<1$, then the asymptotic expression wil be valid uniformly for $|x|\leq cX_n$. Since G. Freud [2] proved that $\lim X_n(4n/3)^{-1/4}=1$ (see also [4]), we might as well consider asymptotics for $|x|\leq c(4n/3)^{1/4}$. One of the advantages of our Plancherel–Rotach type asymptotics is that it enables us to find the value of the constant A in (8) which turns out to be equal to $12^{1/8}\pi^{-1/2}$.

THEOREM 1. *Let $0<\varepsilon<\pi/2$ be fixed and let $x=(4n/3)^{1/4}\cos\theta$. Then the asymptotic formula*

$$(9)$$
$$p_n(x)\exp\left(-\frac{x^4}{2}\right)=12^{1/8}\pi^{-1/2}n^{-1/8}(\sin\theta)^{-1/2}$$
$$\cdot\cos\left[n12^{-1}(12\theta-4\sin 2\theta-\sin 4\theta)+\theta 2^{-1}-\pi 4^{-1}\right]+O(n^{-9/8})$$

holds uniformly for $n=1,2,\cdots$ and $\varepsilon\leq\theta\leq\pi-\varepsilon$.

Note that by putting $x=0$ in (8) and (9), we obtain that $A=12^{1/8}\pi^{-1/2}$ in (8).

THEOREM 2. *If $0<\varepsilon<\pi/2$ is fixed and $x=(4n/3)^{1/4}\cos\theta$, then*

$$n^{-3/4}\sum_{k=0}^{n-1}p_k^2(x)\exp(-x^4)=2(12)^{1/4}(3\pi)^{-1}\sin\theta(1+2\cos^2\theta)+O(n^{-1})$$

holds uniformly for $n=1,2,\cdots$ and $\varepsilon\leq\theta\leq\pi-\varepsilon$.

For fixed values of x, Theorem 2 was proved in [4, Thm. 13]. First we have to prove three auxiliary propositions.

LEMMA 1. *There exists a constant $A>0$ such that*

$$p_n(0)=A\cos\left(\frac{n\pi}{2}\right)n^{-1/8}\left[1+O(n^{-1})\right].$$

Proof. Since w is even, $p_n(0)=0$ when n is odd. If n is even, then we can apply the recurrence formula (1) repeatedly with $x=0$ to obtain

$$p_n(0)=(-1)^{n/2}\gamma_0\prod_{k=1}^{n/2}a_{2k-1}\prod_{k=1}^{n/2}a_{2k}^{-1}$$

which we rewrite as

$$p_n(0)=(-1)^{n/2}\gamma_0 \prod_{k=1}^{n/2} \left[a_{2k-1}(2k-1)^{-1/4}\right] \prod_{k=1}^{n/2} \left[a_{2k}^{-1}(2k)^{1/4}\right]$$

$$\cdot (n!)^{1/4} 2^{-n/4}[(n/2)!]^{-1/2}$$

and now the lemma follows from (3) and Stirling's formula.

Lemma 2. *There exists a constant* $c>0$ *such that for every* $0<\varepsilon<\pi/2$,

$$\limsup_{n\to\infty} \int_{\cos\varepsilon \le |x|(4n/3)^{-1/4} \le 1} \left[1-x^2(4n/3)^{-1/2}\right] p_n^2(x) \exp(-x^4)\,dx \le c(1-\cos\varepsilon).$$

Proof. While proving [4, Thm. 8], we established the inequality

(10)

$$n^{-1}\sum_{k=0}^{n-1} p_k^2(x)$$

$$\ge 16 a_n^4 n^{-1}\left\{\left[\left(|x|2^{-1}a_n^{-1}\right)^2+\frac{1}{2}\right]\left[1-|x|2^{-1}a_n^{-1}\right]+\frac{1}{4}\left[a_{n-1}^2 a_n^{-2}-1\right]\right\} p_n^2(x)$$

$$+16 a_n^4 n^{-1}\left\{\left[\left(|x|2^{-1}a_n^{-1}\right)^2+\frac{1}{2}\right]\left[1-|x|2^{-1}a_n^{-1}\right]+\frac{1}{4}\left[a_{n+1}^2 a_n^{-2}-1\right]\right\} p_{n-1}^2(x).$$

Since by (3) $a_n^2 \ge (n/12)^{1/2}$,

$$a_{n+1}^2 a_n^{-2}-1=(2n)^{-1}+O(n^{-2})>0,$$
$$a_{n-1}^2 a_n^{-2}-1=-(2n)^{-1}+O(n^{-2})>-n^{-1},$$

and

$$1-|x|2^{-1}a_n^{-1} \ge 1-\left(\frac{n}{12}\right)^{1/4} a_n^{-1}=(48n^2)^{-1}+O(n^{-4})>0$$

for $|x|\le(4n/3)^{1/4}$ and $n\ge n_0$, we obtain

$$n^{-1}\sum_{k=0}^{n-1} p_k^2(x) \ge \frac{2}{3}\left(1-|x|2^{-1}a_n^{-1}\right)p_n^2(x)-(3n)^{-1}p_n^2(x)$$

for $|x|\le(4n/3)^{1/4}$ and $n\ge n_0$. Moreover, from $a_n^2\ge(n/12)^{1/2}$ follows

$$1-|x|2^{-1}a_n^{-1} \ge 1-|x|(4n/3)^{-1/4} \ge \left(1-x^2(4n/3)^{-1/2}\right)/2$$

so that

$$\left[1-x^2\left(\frac{4n}{3}\right)^{-1/2}\right] p_n^2(x) \le 3n^{-1}\sum_{k=0}^{n} p_k^2(x)$$

for $n\ge n_0$. Now the lemma follows from the inequality of G. Freud [1]

$$n^{-1}\sum_{k=0}^{n} p_k^2(x) \le B n^{-1/4}\exp(x^4), \qquad x\in\mathbb{R}$$

which holds with an absolute constant B.

LEMMA 3. *Let* $0<\varepsilon<\pi/2$ *be fixed and let* g *be defined by*

$$g(\theta)=1-\frac{2}{3}\cos 2\theta-\frac{1}{3}\cos 4\theta. \tag{11}$$

Then for the function f_n *defined by* (7) *we have*

$$\left(\frac{4n}{3}\right)^{1/2}\sin^2\theta f_n(x)=\left[ng(\theta)+\frac{1}{2}\right]^2+O(1)$$

uniformly for $\varepsilon\leq\theta\leq\pi-\varepsilon$ *where* $x=(4n/3)^{1/4}\cos\theta$.

Proof. It follows from (3) and (4) that

$$\begin{aligned}\phi_n(x)^{-1}&=\left[\left(\frac{n}{12}\right)^{1/2}[1+O(n^{-2})]+\left(\frac{n+1}{12}\right)^{1/2}[1+O(n^{-2})]+\left(\frac{4n}{3}\right)^{1/2}\cos^2\theta\right]^{-1}\\&=\left(\frac{n}{3}\right)^{-1/2}[1+2\cos^2\theta]^{-1}+O(n^{-3/2}),\end{aligned}$$

and therefore by (3)

$$\begin{aligned}&4a_n^2(1-2x^2\phi_n(x)^{-1})-4x^4\phi_n(x)^{-1}+6x^2\\&=2\left(\frac{n}{3}\right)^{1/2}\left[1-4\cos^2\theta(1+2\cos^2\theta)^{-1}-8\cos^4\theta(1+2\cos^2\theta)^{-1}+6\cos^2\theta\right]+O(n^{-1/2})\\&=2\left(\frac{n}{3}\right)^{1/2}[1+2\cos^2\theta]+O(n^{-1/2})\end{aligned}$$

so that

$$\begin{aligned}&\left(\frac{4n}{3}\right)^{1/2}\sin^2\theta\left[4a_n^2(1-2x^2\phi_n(x)^{-1})-4x^4\phi_n(x)^{-1}+6x^2\right]\\&\qquad=\left(\frac{4n}{3}\right)\sin^2\theta[1+2\cos^2\theta]+O(1).\end{aligned}\tag{12}$$

Now we simplify

$$S=\phi_n(x)\phi_{n-1}(x)-a_n^2x^2-x^4. \tag{13}$$

Taking (4) into consideration we get

$$S=x^2(a_{n+1}^2+a_n^2+a_{n-1}^2)+(a_{n+1}^2+a_n^2)(a_n^2+a_{n-1}^2),$$

and by (2) and (3)

$$\begin{aligned}S&=n(4a_n^2)^{-1}x^2+\left[\left(\frac{n+1}{12}\right)^{1/2}(1+O(n^{-2}))+\left(\frac{n}{12}\right)^{1/2}(1+O(n^{-2}))\right]\\&\qquad\cdot\left[\left(\frac{n}{12}\right)^{1/2}(1+O(n^{-2}))+\left(\frac{n-1}{12}\right)^{1/2}(1+O(n^{-2}))\right]\\&=n(4a_n^2)^{-1}x^2+\frac{n}{3}+O(n^{-1}).\end{aligned}\tag{14}$$

Combining (3), (13) and (14) we obtain

$$(15)\quad \begin{aligned} &4a_n^2[4\phi_n(x)\phi_{n-1}(x)-4a_n^2x^2-4x^4]-4x^6\\ &=4nx^2+8\left(\frac{n}{3}\right)^{3/2}-4x^6+O(n^{-1/2})\\ &=\left(\frac{4n}{3}\right)^{3/2}(3\cos^2\theta+1-4\cos^6\theta)+O(n^{-1/2})\\ &=\left(\frac{4n}{3}\right)^{3/2}\sin^2\theta(1+2\cos^2\theta)^2+O(n^{-1/2}).\end{aligned}$$

Applying (3) we see that

$$(16)\qquad \left(\frac{4n}{3}\right)^{1/2}\sin^2\theta\left[\phi_n(x)^{-1}-3x^2\phi_n(x)^{-2}\right]=O(1).$$

Therefore by (7), (12), (15) and (16)

$$\begin{aligned}\left(\frac{4n}{3}\right)^{1/2}\sin^2\theta f_n(x)&=\left(\frac{4n}{3}\right)^2\sin^4\theta(1+2\cos^2\theta)^2+\left(\frac{4n}{3}\right)\sin^2\theta(1+2\cos^2\theta)+O(1)\\ &=\left[n\left(\frac{4}{3}\right)\sin^2\theta(1+2\cos^2\theta)+\frac{1}{2}\right]^2+O(1)\end{aligned}$$

from which the lemma follows immediately.

Proof of Theorem 1. First we let $x=(4n/3)^{1/4}\cos\theta$ in (5) and (6) and define u by

$$(17)\qquad u(\theta)=z\left(\left(\frac{4n}{3}\right)^{1/4}\cos\theta\right).$$

With this substitution (6) takes the form

$$(18)\qquad u_{\theta\theta}-\cot\theta\, u_\theta+\left(\frac{4n}{3}\right)^{1/2}\sin^2\theta f_n(x)u=0.$$

Now we replace u and θ in (18) by

$$(19)\qquad v(\theta)=u(\theta)\left[g(\theta)+(2n)^{-1}\right]^{1/2}[\sin\theta]^{-1/2}$$

and

$$(20)\qquad \tau=\int_{\pi/2}^{\theta}\left[g(t)+(2n)^{-1}\right]dt$$

respectively where g is defined by (11). Then

$$u_\theta=v_\tau\left[g(\theta)+(2n)^{-1}\right]^{1/2}[\sin\theta]^{1/2}+v\left\{[\sin\theta]^{1/2}\left[g(\theta)+(2n)^{-1}\right]^{-1/2}\right\}_\theta$$

and

$$\begin{aligned}u_{\theta\theta}=v_{\tau\tau}&\left[g(\theta)+(2n)^{-1}\right]^{3/2}[\sin\theta]^{1/2}+v_\tau g_\theta[\sin\theta]^{1/2}\left[g(\theta)+(2n)^{-1}\right]^{-1/2}\\ &+2v_\tau\left[g(\theta)+(2n)^{-1}\right]\left\{[\sin\theta]^{1/2}\left[g(\theta)+(2n)^{-1}\right]^{-1/2}\right\}_\theta\\ &+v\left\{[\sin\theta]^{1/2}\left[g(\theta)+(2n)^{-1}\right]^{-1/2}\right\}_{\theta\theta}.\end{aligned}$$

Since

$$\cot\theta\big[g(\theta)+(2n)^{-1}\big]^{1/2}[\sin\theta]^{1/2}$$
$$=g_\theta[\sin\theta]^{1/2}\big[g(\theta)+(2n)^{-1}\big]^{-1/2}$$
$$+2\big[g(\theta)+(2n)^{-1}\big]\Big\{[\sin\theta]^{1/2}\big[g(\theta)+(2n)^{-1}\big]^{-1/2}\Big\}_\theta,$$

we can rewrite (18) in terms of v and τ as

$$v_{\tau\tau}\big[g(\theta)+(2n)^{-1}\big]^{3/2}[\sin\theta]^{1/2}+v\Big\{[\sin\theta]^{1/2}\big[g(\theta)+(2n)^{-1}\big]^{-1/2}\Big\}_{\theta\theta}$$
$$-v\cot\theta\Big\{[\sin\theta]^{1/2}\big[g(\theta)+(2n)^{-1}\big]\Big\}_\theta$$
$$+v\left(\frac{4n}{3}\right)^{1/2}\sin^2\theta f_n(x)\big[g(\theta)+(2n)^{-1}\big]^{-1/2}[\sin\theta]^{1/2}=0,$$

and, applying Lemma 3, we obtain

$$v_{\tau\tau}+n^2v=O(1)v \tag{21}$$

which holds uniformly for $\varepsilon\le\theta\le\pi-\varepsilon$. It was proved in [4, Thm. 8] that

$$\exp(-x^4/2)|p_n(x)|=O(n^{-1/8})$$

uniformly for $\varepsilon\le\theta\le\pi-\varepsilon$. Thus by (3), (4), (5), (11), (17) and (19)

$$|v(\tau)|=O(n^{-3/8})$$

uniformly for $\varepsilon\le\theta\le\pi-\varepsilon$. Therefore we obtain from (21) the equation

$$v_{\tau\tau}+n^2v=O(n^{-3/8}) \tag{22}$$

uniformly for $\varepsilon\le\theta\le\pi-\varepsilon$. Considering (22) as a nonhomogeneous second order equation we can solve it and we get

$$v(\tau)=v(0)\cos n\tau+n^{-1}v_\tau(0)\sin n\tau+n^{-1}\int_0^\tau O(n^{-3/8})\sin[n(t-\tau)]\,dt$$

so that

$$v(\tau)=v(0)\cos n\tau+n^{-1}v_\tau(0)\sin n\tau+O(n^{-11/8}) \tag{23}$$

uniformly for $\varepsilon\le\theta\le\pi-\varepsilon$. By (4), (5), (11), (17) and (19)

$$v(0)=p_n(0)\phi_n(0)^{-1/2}\left[\frac{4}{3}+(2n)^{-1}\right]^{1/2} \tag{24}$$

and

$$v_\tau(0)=-\left(\frac{4n}{3}\right)^{1/4}p_n'(0)\phi_n(0)^{-1/2}\left[\frac{4}{3}+(2n)^{-1}\right]^{-1/2} \tag{25}$$

Since

$$p_n'(0)=4a_n\phi_n(0)p_{n-1}(0)$$

[4, formula (12)], we can simplify (25) to

$$v_\tau(0) = -4\left(\frac{4n}{3}\right)^{1/4} a_n p_{n-1}(0)\phi_n(0)^{1/2}\left[\frac{4}{3}+(2n)^{-1}\right]^{-1/2}. \tag{26}$$

It follows from (3), (4), (24), (26) and Lemma 1 that

$$v(0) = A\left(\frac{4}{3}\right)^{1/2}\phi_n(0)^{-1/2}n^{-1/8}\cos(n\pi/2)+O(n^{-11/8}) \tag{27}$$

and

$$n^{-1}v_\tau(0) = -A\left(\frac{4}{3}\right)^{1/2}\phi_n(0)^{-1/2}n^{-1/8}\sin(n\pi/2)+O(n^{-11/8}). \tag{28}$$

Substituting (27) and (28) into (23), we obtain

$$v(\tau) = A\left(\frac{4}{3}\right)^{1/2}\phi_n(0)^{-1/2}n^{-1/8}\cos\left[n\tau+\frac{n\pi}{2}\right]+O(n^{-11/8}) \tag{29}$$

uniformly for $\varepsilon \le \theta \le \pi-\varepsilon$. We can write (29) in terms of p_n as

$$\begin{aligned} p_n(x)\exp\left(-\frac{x^4}{2}\right) &= A[\phi_n(x)/\phi_n(0)]^{1/2}\sin\theta[g(\theta)+(2n)^{-1}]^{-1/2}\left(\frac{4}{3}\right)^{1/2} \\ &\quad\cdot n^{-1/8}(\sin\theta)^{-1/2}\cos\left[n\tau+\frac{n\pi}{2}\right]+O(n^{-9/8}) \end{aligned} \tag{30}$$

uniformly for $\varepsilon \le \theta \le \pi-\varepsilon$, where we used the fact that $\phi_n(x)^{1/2}=O(n^{1/4})$ which follows from (3) and (4). Moreover by (3), (4) and (11)

$$\left[\frac{\phi_n(x)}{\phi_n(0)}\right]^{1/2}\sin\theta[g(\theta)+(2n)^{-1}]^{-1/2}\left(\frac{4}{3}\right)^{1/2} = 1+O(n^{-1}),$$

so that (30) becomes

$$p_n(x)\exp\left(-\frac{x^4}{2}\right) = An^{-1/8}(\sin\theta)^{-1/2}\cos\left[n\tau+\frac{n\pi}{2}\right]+O(n^{-9/8}), \tag{31}$$

and by (20) the theorem will be proved if we show that $A=12^{1/8}\pi^{-1/2}$. First let us remark that if h is a continuous function, then

$$\int_{\varepsilon}^{\pi-\varepsilon} h(\theta)\cos(2n\tau+n\pi)\,d\theta = \int_{\varepsilon_1}^{\varepsilon_2} h(\theta)\left[g(\theta)+(2n)^{-1}\right]^{-1}\cos(2n\tau+n\pi)\,d\tau$$

where

$$\varepsilon_1 = \int_{\pi/2}^{\varepsilon}\left[g(t)+(2n)^{-1}\right]dt$$

and

$$\varepsilon_2 = \int_{\pi/2}^{\pi-\varepsilon}\left[g(t)+(2n)^{-1}\right]dt.$$

Since h and g^{-1} are continuous functions of τ as well for $\varepsilon\leq\theta\leq\pi-\varepsilon$, we can apply Riemann–Lebesgue's lemma to conclude that

$$\lim_{n\to\infty}\int_{\varepsilon}^{\pi-\varepsilon}h(\theta)\cos(2n\tau+n\pi)\,d\theta=0. \tag{32}$$

Now we will prove

$$A^2\leq 12^{1/4}\pi^{-1}. \tag{33}$$

From

$$\int_{|x|\leq\cos\varepsilon(4n/3)^{1/4}}p_n^2(x)\exp(-x^4)\,dx\leq 1$$

and (31) we obtain

$$A^2(\pi-2\varepsilon)12^{-1/4}+A^2 12^{-1/4}\int_{\varepsilon}^{\pi-\varepsilon}\cos(2n\tau+n\pi)\,d\theta\leq 1+O(n^{-1})$$

and applying (32) with $h(\theta)\equiv 1$, we get

$$A^2(\pi-2\varepsilon)12^{-1/4}\leq 1.$$

Since $0<\varepsilon<\pi/2$ is arbitrary, inequality (33) follows. The next step is to show

$$A^2\geq 12^{1/4}\pi^{-1}. \tag{34}$$

Applying the recurrence formula (1), we obtain

$$1-(a_{n+1}^2+a_n^2)\left(\frac{4n}{3}\right)^{-1/2}=\int_{-\infty}^{\infty}\left[1-x^2\left(\frac{4n}{3}\right)^{-1/2}\right]p_n^2(x)\exp(-x^4)\,dx$$

so that

$$1-(a_{n+1}^2+a_n^2)\left(\frac{4n}{3}\right)^{-1/2}\leq\int_{|x|\leq(4n/3)^{1/4}}\left[1-x^2\left(\frac{4n}{3}\right)^{-1/2}\right]p_n^2(x)\exp(-x^4)\,dx. \tag{35}$$

It follows from the asymptotics (31) that

$$\int_{|x|\leq\cos\varepsilon(4n/3)^{1/4}}\left[1-x^2\left(\frac{4n}{3}\right)^{-1/2}\right]p_n^2(x)\exp(-x^4)\,dx$$
$$=A^2 12^{-1/4}\int_{\varepsilon}^{\pi-\varepsilon}\sin^2\theta\,d\theta+A^2 12^{-1/4}\int_{\varepsilon}^{\pi-\varepsilon}\sin^2\theta\cos(2n\tau+n\pi)\,d\theta+O(n^{-1})$$

and letting here $n\to\infty$ and using (32) with $h(\theta)\equiv\sin^2\theta$, we obtain

$$\limsup_{n\to\infty}\int_{|x|\leq\cos\varepsilon(4n/3)^{1/4}}\left[1-x^2\left(\frac{4n}{3}\right)^{-1/2}\right]p_n^2(x)\exp(-x^4)\,dx\leq A^2 12^{-1/4}\pi/2. \tag{36}$$

Combining (35) and (36) with (3) and Lemma 2, we see that

$$\frac{1}{2}\leq A^2 12^{-1/4}\frac{\pi}{2}+c(1-\cos\varepsilon)$$

and by letting $\varepsilon\to 0$, inequality (34) follows. By (33) and (34) $A^2=12^{1/4}\pi^{-1}$ and since $A>0$ by Lemma 1, we have completed the proof of the theorem.

Proof of Theorem 2. If $x=(4n/3)^{1/4}\cos\theta_1$ where $0<\varepsilon_1<\pi/2$, then there exist n_1 and ε_2 depending on ε_1 such that $0<\varepsilon_2<\pi/2$ and if $x=(4(n-1)/3)^{1/4}\cos\theta_2$ and $n>n_1$, then $\varepsilon_2\leq\theta_2\leq\pi-\varepsilon_2$. Simple computation shows that

$$\theta_1=\theta_2+(4n)^{-1}\cot(\theta_2)+O(n^{-2}). \tag{37}$$

Let

$$\tau_1=\int_{\pi/2}^{\theta_1}\left[g(t)+(2n)^{-1}\right]dt \tag{38}$$

and

$$\tau_2=\int_{\pi/2}^{\theta_2}\left[g(t)+(2(n-1))^{-1}\right]dt \tag{39}$$

where g is defined by (11). First we will prove that

$$n\tau_1+\frac{n\pi}{2}=(n-1)\tau_2+\frac{(n-1)\pi}{2}+\theta_1+O(n^{-1}). \tag{40}$$

We have

$$\tau_1-\tau_2=\int_{\theta_2}^{\theta_1}g(t)\,dt+\left(\theta_1-\frac{\pi}{2}\right)(2n)^{-1}-\left(\theta_2-\frac{\pi}{2}\right)(2(n-1))^{-1}$$

so that by (37)

$$\tau_1-\tau_2=\int_{\theta_2}^{\theta_1}g(t)\,dt+O(n^{-2})$$

and by applying (37) again, we obtain

$$\tau_1-\tau_2=(\theta_1-\theta_2)g(\theta_2)+O(n^{-2})=(4n)^{-1}\cot(\theta_2)g(\theta_2)+O(n^{-2}).$$

Hence

$$n\tau_1-(n-1)\tau_2+\frac{\pi}{2}=\frac{1}{4}\cot(\theta_2)g(\theta_2)+\int_{\pi/2}^{\theta_2}g(t)\,dt+\frac{\pi}{2}+O(n^{-1}).$$

A somewhat tedious but elementary computation yields

$$\frac{1}{4}\cot(\theta_2)g(\theta_2)+\int_{\pi/2}^{\theta_2}g(t)\,dt=\theta_2-\frac{\pi}{2}=\theta_1-\frac{\pi}{2}+O(n^{-1})$$

where we also used (37). Hence (40) follows. Formula (37) also implies

$$(\sin\theta_1)^{-1}=(\sin\theta_2)^{-1}+O(n^{-1}). \tag{41}$$

Using the notation of (38) and (39), we can rewrite the asymptotic formula (9) as

$$p_n^2(x)\exp(-x^4)\sin\theta_1=12^{1/4}\pi^{-1}n^{-1/4}\cos^2\left[n\tau_1+\frac{n\pi}{2}\right]+O(n^{-5/4}) \tag{42}$$

and

(43)

$$p_{n-1}^2(x)\exp(-x^4)\sin\theta_2=12^{1/4}\pi^{-1}n^{-1/4}\cos^2[(n-1)\tau_2+(n-1)\pi/2]+O(n^{-5/4}).$$

Applying (40) and (41), we obtain from (43) the asymptotics

$$(44)\quad p_{n-1}^2(x)\exp(-x^4)\sin\theta_1=12^{1/4}\pi^{-1}n^{-1/4}\cos^2[n\tau_1+n\pi/2-\theta_1]+O(n^{-5/4}).$$

Using some trigonometric identities, we can conclude from (42) and (44) that

$$(45)\qquad \begin{aligned}&[p_n^2(x)-2\cos\theta_1 p_n(x)p_{n-1}(x)+p_{n-1}^2(x)]\exp(-x^4)\\ &\qquad =12^{1/4}\pi^{-1}n^{-1/4}\sin\theta_1+O(n^{-5/4}).\end{aligned}$$

During the proof in [4, Thm. 8] we obtained

$$(46)\qquad \begin{aligned}\sum_{k=0}^{n-1}p_k^2(x)&=4a_n^2\phi_{n-1}(x)p_n^2(x)-4a_nx(2a_n^2+x^2)p_n(x)p_{n-1}(x)\\ &\quad+4a_n^2\phi_n(x)p_{n-1}^2(x).\end{aligned}$$

By (3) and (4)

$$\left.\begin{matrix}\phi_n(x)\\ \phi_{n-1}(x)\\ 2a_n^2+x^2\end{matrix}\right\}=\left(\frac{n}{3}\right)^{1/2}(1+2\cos^2\theta_1)+O(n^{-1/2})$$

and, by using (3) and (9), we get from (46) the estimate

$$(47)\quad \begin{aligned}\sum_{k=0}^{n-1}p_k^2(x)&=\frac{2n}{3}(1+2\cos^2\theta_1)\\ &\quad\cdot[p_n^2(x)-2\cos\theta_1 p_n(x)p_{n-1}(x)+p_{n-1}^2(x)]+O(n^{-1/4})\exp(x^4).\end{aligned}$$

Now the theorem follows from (45) and (47).

Note added in proof. If one applies Theorem 1 for fixed values of x, then the asymptotic formula may be strengthened in a considerable way. Namely, for a given interval Δ

$$\begin{aligned}&p_n(x)\exp\left(-\frac{x^4}{2}\right)\\ &\quad=12^{1/8}\pi^{-1/2}n^{-1/8}\\ &\qquad\cdot\Big[\big(1+B_1(x)n^{-1/2}\big)\cos\big((64/27)^{1/4}xn^{3/4}+12^{-1/4}x^3n^{1/4}-n\pi/2\big)\\ &\qquad\quad+\big(B_2(x)n^{-1/4}+B_3(x)n^{-3/4}\big)\\ &\qquad\qquad\cdot\sin\big((64/27)^{1/4}xn^{3/4}+12^{-1/4}x^3n^{1/4}-n\pi/2\big)\Big]+O(n^{-9/8})\end{aligned}$$

uniformly for $n=1,2,\cdots$ and $x\in\Delta$, where

$$\begin{aligned}B_1(x)&=\frac{1}{8}\left(\frac{3}{4}\right)^{1/2}\left(x^2+\frac{9}{10}x^6-\frac{81}{400}x^{10}\right),\\ B_2(x)&=\frac{1}{2}\left(\frac{3}{4}\right)^{1/4}\left(-x+\frac{9}{20}x^5\right),\\ B_3(x)&=\frac{1}{4}\left(\frac{3}{4}\right)^{3/4}\left(-\frac{3}{4}x^3+\frac{163}{560}x^7+\frac{81}{1600}x^{11}-\frac{243}{32000}x^{15}\right).\end{aligned}$$

Similar asymptotic expansions for orthogonal polynomials associated with $\exp(-x^6)$ are given by R. Sheen (see [6]).

REFERENCES

[1] G. Freud, *On polynomial approximation with the weight* $\exp(-x^{2k}/2)$, Acta Math. Acad. Sci. Hungar., 24 (1973), pp. 363–371.

[2] G. Freud, *On the greatest zero of an orthogonal polynomial*, manuscript, 1978.

[3] J. S. Lew and D. A. Quarles, Jr., *Nonnegative solutions of a nonlinear recurrence*, J. Approximation Th., 38 (1983), 357–379.

[4] P. Nevai, *Orthogonal polynomials associated with* $\exp(-x^4)$, in Second Edmonton Conference on Approximation Theory, Can. Math. Soc. Conf. Proc., 3 (1983), pp. 263–285.

[5] J. Shohat, *A differential equation for orthogonal polynomials*, Duke Math. J., 5 (1939), pp. 401–417.

[6] P. Nevai, *Two of my favorite ways of obtaining asymptotics for orthogonal polynomials*, in Functional Analysis and Approximation by P. L. Butzer and B. Sz.-Nagy, eds., ISNM, Birkhauser Verlag, Basel, 1984.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics

012

ON A SUFFICIENT CONDITION FOR BEST L^1-APPROXIMATION*

FRANZ PEHERSTORFER[†]

Abstract. Let $\{g_1,\cdots,g_n\}$, $n\in\mathbb{N}$, be a Chebyshev-system of continuous functions on $I:=[a,b]$, put $G_n:=\{\sum_{i=1}^n a_ig_i\,|\,a_i\in\mathbb{R}\}$ and let $Z\subset I$ be a set of positive measure.It is known that the condition $\int_{I\setminus Z}|g|\le\int_Z|g|$ for all $g\in G_n$ implies that 0 is a best L^1-approximation from G_n to every function $f\in L^1(I)$ which vanishes on a set including Z. In this paper we characterize those sets Z of the form $\sum_{j=1}^l[\alpha_j,\beta_j]\subset I$ which satisfy the above condition. The results extend and generalize those of the author [J. Math. Anal. Appl., 84 (1981), pp. 170–177], where the special case $Z=[\alpha,\beta]\subset I$ is discussed.

1. Introduction and notation. Let $\{g_1,\cdots,g_n\}$, $n\in\mathbb{N}$, be a Chebyshev system of continuous functions on $I:=[a,b]$ and put $G_n:=\{\sum_{i=1}^n a_ig_i\,|\,a_i\in\mathbb{R}\}$. For $J\subset I$ and $h\in L^1(I)$ let $S(h,J)$ denote the number of strong sign changes of h on J (see [14, Definition 13.1]). $S^-(h,J)$ denotes the number of points of int(J) at which h changes sign (compare, e.g., [13]). Note that $S^-(h,J)\le S(h,J)$.

In §2 of this paper we characterize those sets Z of the form $\sum_{j=1}^l[\alpha_j,\beta_j]$, where $a\le\alpha_1<\beta_1<\alpha_2<\cdots<\alpha_l<\beta_l\le b$, which satisfy the condition

$$(1)\qquad \int_{I\setminus Z}|g|\le\int_Z|g| \quad \text{for all } g\in G_n.$$

The results extend and generalize those of the author [11], where the special case $Z=[\alpha,\beta]\subset I$ is discussed. As is well known (see, e.g., [14]) condition (1) implies that 0 is a best L^1-approximation to every function $f\in L^1(I)$ which vanishes on a set including Z.

In §3 we study the following problem, first considered by Motzkin and Walsh [6]: What are necessary conditions on Z for the existence of a function $f\in L^1(I)$ with the following properties: f vanishes exactly on Z, $S(f,I)\le n-k$, $k\in\{1,\cdots,n\}$, and 0 is a best L^1-approximation to f from G_n. For the special case $Z=[\alpha,\beta]\subset I$ this problem was solved (using different methods) by the author in [10].

Finally, in §4, we state a sufficient condition for $f\in L^1(I)$ to have 0 as best approximation from $R_1^n[-1,+1]$, where $R_1^n[-1,+1]$ is the family of quotients p/q of polynomials p of degree$\le n$ by polynomials q of degree≤ 1 which are positive on $[-1,+1]$.

Henceforth a function φ is called a sign function on I, if $\varphi\in L_\infty(I)$ and $\varphi^2=1$ a.e. on I. If $h\in L^1(I)$ is such that $\int_I gh=0$ for all $g\in G_n$, then we write $h\perp G_n$.

2. Characterizations. In this section let Z be of the form

$$Z=\sum_{j=1}^{l}\left[\alpha_j,\beta_j\right],\quad \text{where } a\le\alpha_1<\beta_1<\alpha_2<\cdots<\alpha_l<\beta_l\le b.$$

LEMMA 1. *Let $\gamma\in\mathbb{R}^+$. If $\int_{I\setminus Z}|g|\le\gamma\int_Z|g|$ for all $g\in G_n$ and $\int_{I\setminus Z}|g^*|=\gamma\int_Z|g^*|$, then $\int_{I\setminus Z}g\operatorname{sgn}g^*-\gamma\int_Z g\operatorname{sgn}g^*=0$ for all $g\in G_n$, and $S^-(g^*,I\setminus Z)=0$.*

*Received by the editors August 17, 1981, and in revised form January 27, 1982.

[†] Institut für Mathematik, Universität Linz, A-4040 LINZ-Auhof, Austria.

Proof. Since the proof is a repetition of the arguments used in proving [4, Lemma 4(a) and 4(b)], we omit it.

THEOREM 1. *The following three properties are equivalent*:

(1) $\int_{I\setminus Z}|g|\leq\int_Z|g|$ *for all* $g\in G_n$.

(2) *For every sign function* ψ *on* I *with* $\psi\perp G_n$ *and* $S(\psi,Z)\leq n-1$ *the following condition holds*: *If* $g\in G_n$ *is such that* $\operatorname{sgn} g=\psi$ *a.e. on* Z, *then* $\operatorname{sgn} g=-\psi$ *a.e. on* $I\setminus Z$.

(3) *There exist no* $g^*\in G_n$ *resp.* $\gamma\in(1,\infty)$, *such that*

$$\int_{I\setminus Z} g\operatorname{sgn} g^*-\gamma\int_Z g\operatorname{sgn} g^*=0 \quad \textit{for all } g\in G_n.$$

Proof. For the implication (1)⇒(2), assume that there exist a sign function ψ on I with $\psi\perp G_n$ and $S(\psi,Z)\leq n-1$ and an element $\bar g\in G_n$, such that $\operatorname{sgn}\bar g=\psi$ a.e. on Z and $\operatorname{sgn}\bar g\neq-\psi$ on a subset of positive measure of $I\setminus Z$. Using the fact that $\psi\perp G_n$, we obtain

$$\int_Z|\bar g|=\int_Z\bar g\psi=-\int_{I\setminus Z}\bar g\psi<\int_{I\setminus Z}|\bar g|$$

in denial of (1).

Concerning the implication (2)⇒(3), let us assume that (2) holds and that there exist a $g^*\in G_n$ and a $\gamma\in(1,\infty)$, such that

$$\frac{1}{\gamma}\int_{I\setminus Z} g\operatorname{sgn} g^*-\int_Z g\operatorname{sgn} g^*=0 \quad \text{for all } g\in G_n.$$

By [12, Lemma 2], see also [2], there exists a sign function ψ on I, such that $\psi=-\operatorname{sgn} g^*$ a.e. on Z and $\psi\perp G_n$. In view of (2) we obtain that $\psi=\operatorname{sgn} g^*$ a.e. on $I\setminus Z$. Thus

$$\frac{1}{\gamma}\int_{I\setminus Z} g^*\operatorname{sgn} g^*=-\int_Z g^*\psi=\int_{I\setminus Z} g^*\psi=\int_{I\setminus Z} g^*\operatorname{sgn} g^*,$$

which is a contradiction to $\gamma\in(1,\infty)$.

For the implication (3)⇒(1), assume that (1) is false. Then there exist a $g^*\in G_n$ and a $\gamma\in(1,\infty)$, such that

$$\int_{I\setminus Z}|g|\leq\gamma\int_Z|g| \quad \text{for all } g\in G_n \quad \text{and} \quad \int_{I\setminus Z}|g^*|=\gamma\int_Z|g^*|.$$

In view of Lemma 1 the implication is proved.

COROLLARY 1. *Suppose that* $\int_{I\setminus Z}|g|\leq\int_Z|g|$ *for all* $g\in G_n$. *Let* ψ *be a sign function such that* $\psi\perp G_n$ *and* $S(\psi,Z)\leq n-1$. *Then*

(a) $S^-(\psi,I\setminus Z)=0$.

(b) ψ *changes sign at each boundary point of* $I\setminus Z$, *which is different from the endpoints* a,b *of the interval* I.

(c) $S^-(\psi,Z)\geq n-2+\delta_{I\setminus Z}$, *where* $\delta_{I\setminus Z}=0$ *if* $a,b\notin I\setminus Z$ *and* $\delta_{I\setminus Z}=1$ *otherwise*.

Proof. (a) Assume that ψ changes sign on $I\setminus Z$. Then, since $S(\psi,Z)\leq n-1$, there exists a $\bar g\in G_n$, such that $\operatorname{sgn}\bar g=\psi$ a.e. on Z and $\bar g$ has no simple zero on $I\setminus Z$. Hence $\operatorname{sgn}\bar g\neq-\psi$ on $I\setminus Z$ which is a contradiction to (2) of Theorem 1.

Simil one demonstrates (b) and (c).

In order to show that condition (1) does not hold for a given set Z the following theorem seems to be very useful.

THEOREM 2. *$\int_{I\setminus Z}|g|<\int_Z|g|$ for all $g\in G_n$ if and only if every sign function ψ with $\psi\perp G_n$ satisfies $S(\psi,Z)\geq n$.*

Proof. Necessity. Let $\int_{I\setminus Z}|g|<\int_Z|g|$ for all $g\in G_n$ and assume that there exists a sign function ψ, such that $\psi\perp G_n$ and $S(\psi,Z)\leq n-1$. Then it follows with the help of Corollary 1(a) and 1(b) that there exists a $g^*\in G_n$, such that

$$\psi=\begin{cases}-\operatorname{sgn} g^* \text{ a.e.} & \text{on } Z,\\ \operatorname{sgn} g^* \text{ a.e.} & \text{on } I\setminus Z.\end{cases}$$

Thus we obtain that

$$\int_{I\setminus Z}|g^*|=\int_{I\setminus Z}g^*\psi=-\int_Z g^*\psi=\int_Z|g^*|,$$

which is a contradiction.

Sufficiency. By (2) of Theorem 1 it follows immediately $\int_{I\setminus Z}|g|\leq\int_Z|g|$ for all $g\in G_n$. Now let us assume that there exists a $g^*\in G_n$, such that $\int_{I\setminus Z}|g^*|=\int_Z|g^*|$. Setting

$$\varphi=\begin{cases}-\operatorname{sgn} g^* & \text{on } Z,\\ \operatorname{sgn} g^* & \text{on } I\setminus Z,\end{cases}$$

we obtain from Lemma 1 that $\varphi\perp G_n$ and $S(\varphi,Z)\leq n-1$, which is a contradiction.

COROLLARY 2. *Let $I\setminus Z=\sum_{j=1}^{l}(\alpha_j,\beta_j)$, $a<\alpha_1<\beta_l<b$ and suppose that $\int_{I\setminus Z}|g|\leq\int_Z|g|$ for all $g\in G_n$. Furthermore let ψ be a sign function on I such that $\psi\perp G_n$ and $S(\psi,I)=n-1+k$, where $k\in\{1,\cdots,2l\}$. Then there do not exist $[(k+1)/2]$ intervals $(\alpha_{j_\nu},\beta_{j_\nu})$, $\nu\in\{1,\cdots,[(k+1)/2]\}, j_\nu\in\{1,\cdots,l\}$, which contain two consecutive changes of sign of ψ.*

Proof. Let us assume that there exist $[(k+1)/2]$ intervals $(\alpha_{j_\nu},\beta_{j_\nu})$ which contain two consecutive changes of sign of ψ. Then there exist $[(k+1)/2]$ intervals $(\alpha^*_{j_\nu},\beta^*_{j_\nu})\subset(\alpha_{j_\nu},\beta_{j_\nu})$, which contain exactly two consecutive changes of sign of ψ. Setting

$$I\setminus W=\sum_{\nu=1}^{[(k+1)/2]}(\alpha^*_{j_\nu},\beta^*_{j_\nu})$$

we obtain that

$$S(\psi,W)=S^-(\psi,W)=n-1+k-2[(k+1)/2]\leq n-1.$$

Applying Corollary 1 to W it follows that there exists a $\bar{g}\in G_n$, such that

$$\int_{I\setminus Z}|\bar{g}|>\int_{I\setminus W}|\bar{g}|>\int_W|\bar{g}|>\int_Z|\bar{g}|,$$

which is a contradiction.

Notation. Let P_n, $n\in\mathbb{N}_0$, denote the set of algebraic polynomials of degree equal or less than n. As usual U_n denotes the Chebyshev polynomial of second kind of degree n.

Example. Let $I=[-1,+1]$, $G_n=P_{n-1}$, $I\setminus Z=(\alpha_1,\beta_1)\cup(\alpha_2,\beta_2)$, where $-1<\alpha_1<\beta_2<1$. Corollary 2 implies the following two facts:

If one of the intervals (α_1,β_1), (α_2,β_2) contains two consecutive zeros of U_n or U_{n+1}, then inequality (1) does not hold.

If both intervals (α_1,β_1) and (α_2,β_2) contain two consecutive zeros of U_{n+2} or U_{n+3}, then inequality (1) does not hold.

Lemma 3. *Let $c_1,\cdots,c_n\in\mathbb{R}$ be given. There exists a function $\phi\in L_\infty(Z)$, such that $|\phi|=\eta$ a.e. on Z, $\eta\in(0,1]$, $S(\phi,Z)\le n-1$ and $\int_Z g_i\phi=c_i$ for $i=1,\cdots,n$, if and only if there exists a sign function ψ on Z, such that $S(\psi,Z)\le n$ and $\int_Z g_i\psi=c_i$ for $i=1,\cdots,n$.*

Proof. Necessity. Using the facts that $\{g_1,\cdots,g_n\}$ is a T-system on I and thus a T-system on Z and that $S(\phi,Z)\le n-1$, there exists a $g^*\in G_n$, such that

$$\eta\int_Z g_i\operatorname{sgn} g^*=c_i \quad \text{for } i=1,\cdots,n,$$

from which it follows (see [3]) that

$$\eta=\max_{a_i\in\mathbb{R}}\left|\sum_{i=1}^n a_ic_i\right|\Big/\left\|\sum_{i=1}^n a_ig_i\right\|_Z,$$

where $\|h\|_Z:=\int_Z|h|$ for $h\in L^1(Z)$.

Now in view of [5] there exists an element $g_{n+1}\in C(I)$, such that $\{g_1,\cdots,g_n,g_{n+1}\}$ is a T-system on I. Defining the function

$$\bar\eta\colon\mathbb{R}^{n+1}\to\mathbb{R} \text{ by } \bar\eta(d_1,\cdots,d_{n+1})=\max_{a_i\in\mathbb{R}}\left|\sum_{i=1}^{n+1} a_id_i\right|\Big/\left\|\sum_{i=1}^{n+1} a_ig_i\right\|_Z,$$

we obtain from [3, p. 177], that $\bar\eta$ is continuous. Since the function $\bar\eta(c_1,\cdots,c_n,\cdot)$ is not bounded on $\mathbb{R}$ and since $\bar\eta(c_1,\cdots,c_n,c_{n+1})=\eta<1$, where $c_{n+1}:=\eta\int_Z g_{n+1}\operatorname{sgn} g^*$, we deduce from the continuity of $\bar\eta(c_1,\cdots,c_n,\cdot)$ that there exists a $d^*_{n+1}\in\mathbb{R}$, such that $\bar\eta(c_1,\cdots,c_n,d^*_{n+1})=1$. From [3] it follows again, that there exists an element $\sum_{i=1}^{n+1}b_i^*g_i$, such that

$$\int_Z g_i\operatorname{sgn}\left(\sum_{i=1}^{n+1}b_i^*g_i\right)=c_i \quad \text{for } i=1,\cdots,n$$

and

$$\int_Z g_{n+1}\operatorname{sgn}\left(\sum_{i=1}^{n+1}b_i^*g_i\right)=d^*_{n+1}.$$

Taking into account the fact that $S(\sum_{i=1}^{n+1}b_i^*g_i,Z)\le n$, the assertion follows by putting $\psi=\operatorname{sgn}(\sum_{i=1}^{n+1}b_i^*g_i)$.

Sufficiency. We have only to consider the case $S(\psi,Z)=n$. Choose g_{n+1} as above and let $\sum_{i=1}^{n+1}b_i^*g_i$ be such that $\psi=\operatorname{sgn}(\sum_{i=1}^{n+1}b_i^*g_i)$. Setting $d^*_{n+1}=\int_Z g_{n+1}\psi$, it follows that $\bar\eta(c_1,\cdots,c_n,d^*_{n+1})=1$, which implies the existence of a number $\eta\in(0,1]$, such that

$$\max_{a_i\in\mathbb{R}}\left|\sum_{i=1}^n a_ic_i\right|\Big/\left\|\sum_{i=1}^n a_ig_i\right\|_Z=\eta.$$

In view of [3] the sufficiency is proved.

Theorem 3. *The following three properties are equivalent*:

(1) $\int_{I\setminus Z}|g|\le\int_Z|g|$ *for all* $g\in G_n$.

(4) *For every sign function ψ on $I\setminus Z$ with $S^-(\psi,I\setminus Z)=0$ there exists a sign function φ on I, such that $\varphi=\psi$ on $I\setminus Z$, $\varphi\perp G_n$ and $n-2+\delta_{I\setminus Z}\le S(\varphi,Z)\le n$.*

(5) *For every sign function ψ on $I\setminus Z$ with $S^-(\psi,I\setminus Z)=0$ there exists a sign function φ on I, such that $\varphi=\psi$ on $I\setminus Z$ and $\varphi\perp G_n$.*

Proof. (1)⇒(4). Let ψ be a sign function on $I\setminus Z$. Then (1) implies that 0 is a best approximation to ψ^*, where

$$\psi^*(x)=\begin{cases}\psi(x) & \text{on } I\setminus Z,\\ 0 & \text{on } Z.\end{cases}$$

By [4, Thm. 1] there exist a $\gamma\in(0,1]$ and an element $\bar{g}\in G_n$, such that

$$\int_{I\setminus Z} g\psi-\gamma\int_Z g\operatorname{sgn}\bar{g}=0 \quad \text{for all } g\in G_n.$$

Hence, in view of Lemma 3, there exist a sign function φ with $\varphi=\psi$ on $I\setminus Z$, such that $\varphi\perp G_n$ and $S(\varphi,Z)\le n$. Furthermore it follows from Corollary 1, that $S(\varphi,Z)\ge n-2+\delta_{I\setminus Z}$.

The implication (4)⇒(5) is trivial.

For the implication (5)⇒(1), assume that (1) is false and that (5) holds. Since (1) is false, there exist a $\gamma\in(1,\infty)$ and a $g^*\in G_n$, such that

$$\int_{I\setminus Z} g\operatorname{sgn} g^*-\gamma\int_Z g\operatorname{sgn} g^*=0 \quad \text{for all } g\in G_n,$$

and g^* has no change of sign on $I\setminus Z$. Thus, in view of (5), there exists a sign function φ on I, such that $\varphi=\operatorname{sgn} g^*$ on $I\setminus Z$ and $\varphi\perp G_n$, which implies that

$$\gamma\int_Z|g^*|=\int_{I\setminus Z}|g^*|=\int_{I\setminus Z}g^*\varphi=-\int_Z g^*\varphi\le\int_Z|g^*|$$

which is a contradiction to $\gamma\in(1,\infty)$.

Example. Let $I=[-1,+1]$ and let

$$I_{1,n}=[0,\cos(2n-1)\pi/4n] \quad\text{and}\quad I_{2,n}=[\cos\pi/4,\cos(n-1)\pi/4n] \quad \text{for } n\in\mathbb{N}.$$

Observing that $\operatorname{sgn} U_{4n-1}=\operatorname{sgn} U_{4n-5}$ on $I_{1,n}$ and $\operatorname{sgn} U_{4n-1}=-\operatorname{sgn} U_{4n-5}$ on $I_{2,n}$, it follows from Theorem 3 that

$$\int_{I_{1,n}\cup I_{2,n}}|p|\le\int_{[-1,+1]\setminus(I_{1,n}\cup I_{2,n})}|p| \quad \text{for all } p\in P_{4n-6}.$$

Sign functions which are orthogonal to (trigonometric) polynomials resp. rational (trigonometric) functions have been characterized by us in [8] and [9].

COROLLARY 3. *The following two properties are equivalent:*

(1*) $\int_{I\setminus Z}|g|\le\int_Z|g|$ *for all* $g\in G_n$ *and* $\int_{I\setminus Z}|g^*|=\int_Z|g^*|$.

(5*) (5) *holds and there exists a sign function* φ *on* I *with the following properties*: $\varphi\perp G_n$, $S^-(\varphi,I\setminus Z)=0$, $n-2+\delta_{I\setminus Z}\le S^-(\varphi,Z)\le n-1$ *and* φ *changes sign at each boundary point of* $I\setminus Z$, *which is different from the endpoints* a,b *of the interval* I.

Proof. (1*)⇒(5*). That (5) holds follows immediately from Theorem 3. Setting

$$\varphi=\begin{cases}\operatorname{sgn} g^* & \text{on } I\setminus Z,\\ -\operatorname{sgn} g^* & \text{on } Z,\end{cases}$$

we conclude with the help of Lemma 1 and Corollary 1, that φ has the above cited properties.

(5*)⇒(1*). Let $g^* \in G_n$ be such that

$$\operatorname{sgn} g^* = \begin{cases} -\varphi & \text{a.e. on } Z, \\ \varphi & \text{a.e. on } I \setminus Z. \end{cases}$$

Then, since $\varphi \perp G_n$, it follows that $\int_{I\setminus Z}|g^*| = \int_Z |g^*|$. By Theorem 3 the implication is proved.

In [11] the special cases $Z=[\alpha,\beta]\subset I$ resp. $Z=I\setminus[\alpha,\beta]$ were considered. From Corollary 3 of this paper and [10, Lemma 2] one obtains [11, Thms. 2 and 4]. For the important case $G_n = P_{n-1}$ we obtain

COROLLARY 4. *Let* $I=[-1,+1]$ *and let* $-1<\alpha<\beta<1$.

(a) $\int_{I\setminus[\alpha,\beta]}|p| \leq \int_{[\alpha,\beta]}|p|$ *for all* $p\in P_{n-1}$ *if and only if there exists a* $t\in(-1,+1)$, *such that* $\alpha\leq x_1(t)$ *and* $\beta\geq x_{n+1}(t)$, *where* $x_1(t)$ $(x_{n+1}(t))$ *is the smallest (greatest) zero of the polynomial* $U_{n+1}-2tU_n+t^2U_{n-1}$.

(b) $\int_{[\alpha,\beta]}|p| \leq \int_{I\setminus[\alpha,\beta]}|p|$ *for all* $p\in P_{n-1}$ *if and only if there exists a* $t\in[-1,+1]$, *such that* $[\alpha,\beta]\subset[x_j(t),x_{j+1}(t)]$, *where* $x_j(t)$, $x_{j+1}(t)$, $j\in\{1,\cdots,n\}$, *are two consecutive zeros of the polynomial* $U_{n+1}-2tU_n+t^2U_{n-1}$.

Proof. Let φ be such a sign function that $\varphi\perp P_{n-1}$ and $S(\varphi,I)=n+1$. Then it follows from [8, Thm. 4] that there exists a $t\in(-1,+1)$, such that $\varphi=\pm\operatorname{sgn}(U_{n+1}-2tU_n+t^2U_{n-1})$ a.e. on I.

(a) Follows now from Corollary 3.

(b) Additionally using the fact that $\varphi=\pm\operatorname{sgn} U_n=\pm\operatorname{sgn}(U_{n+1}-2U_n+U_{n-1})$ a.e. on I, if $\varphi\perp P_{n-1}$ and $S(\varphi,I)=n$, the assertion follows from Corollary 3.

3. On a problem of Motzkin and Walsh.

Notation. For $f\in L^1(I)$ let $Z(f)=\{x\in[a,b]\,|\,f(x)=0\}$. Furthermore let μ denote the Lebesgue measure.

THEOREM 4. *Let* $Z\subset I$ *be such that* $\mu(Z)>0$.

(a) *There exists a function* $f\in L^1(I)$, *such that* $Z(f)=Z$, $S(f,I)\leq n-1$ *and* 0 *is a best approximation to f from* G_n *on I if and only if* $\min_{g\in G_n}\int_{I\setminus Z}|g|/\int_Z|g|\leq 1$.

(b) *If there exists a function* $f\in L^1(I)$ *such that* $Z(f)=Z$, $S(f,I)\leq n-1-k$, $k\in\{0,\cdots,n-1\}$, *and* 0 *is a best approximation to f from* G_n *on I, then* $\min_{g\in G_{n-k}}\int_{I\setminus Z}|g|/\int_Z|g|\leq 1$.

Proof. First let us note that $\int_{I\setminus Z}|g|/\int_Z|g|$ attains its minimum (see [8, p. 1230]).

(a) *Necessity.* Let $g^*\in G_n$ be such that $\operatorname{sgn} g^*=\operatorname{sgn} f$ a.e. on $I\setminus Z$. Then it follows with the aid of [14, Thm. 13-4], that $\int_{I\setminus Z}|g^*|\leq\int_Z|g^*|$, which implies the statement.

Sufficiency. Since there exists a $\gamma\in(0,1]$ and a $g^*\in G_n$, such that $\gamma\int_Z|g|\leq\int_{I\setminus Z}|g|$ for all $g\in G_n$ and $\gamma\int_Z|g^*|=\int_{I\setminus Z}|g^*|$, we obtain that

$$\gamma\int_Z g\operatorname{sgn} g^* - \int_{I\setminus Z} g\operatorname{sgn} g^* = 0 \quad \text{for all } g\in G_n.$$

Putting

$$f=\begin{cases} g^* & \text{on } I\setminus Z, \\ 0 & \text{on } Z, \end{cases}$$

the assertion follows from [4, Thm. 1].

(b) The assertion can be demonstrated as in (a).

As a consequence of Theorem 4 we obtain the following result of Motzkin and Walsh [6].

Corollary 5. *Suppose that* 0 *is a best approximation to* $f \in L^1(I)$ *from* G_n *on* I, *where* $\mu(Z(f))>0$ *and* $S(f,I)=n-1-k$, $k\in\{0,\cdots,n-1\}$. *Then*

$$\mu(Z(f))\geq\frac{1}{2}\min_{g\in G_{n-k}}\int_I|g|/\max_{x\in I}|g(x)|.$$

Proof. Observing that

$$\int_I|g|/\mu(Z(f))\max_{x\in I}|g(x)|\leq\int_I|g|/\int_{Z(f)}|g|\quad\text{for all } g\in G_{n-k},$$

the assertion follows from Theorem 4(b).

The next corollary shows the connection with the problem considered in the second section.

Corollary 6. *Let* $k\in\{0,\cdots,n-1\}$ *and suppose that* $\int_Z|g|\leq\int_{I\setminus Z}|g|$ *for all* $g\in G_n$ *and* $\int_Z|g^*|=\int_{I\setminus Z}|g^*|$ *for an element* $g^*\in G_{n-k}$. *Let* W *be a proper subset of* Z *of positive measure. Then there exists no function* $f\in L^1(I)$, *such that* $Z(f)=W$, $S(f,I\setminus W)\leq n-1-k$ *and* 0 *is a best approximation to* f *from* G_n *on* I.

Proof. Since $W\subsetneqq Z$ we get

$$\int_{I\setminus W}|g|>\int_{I\setminus Z}|g|\geq\int_Z|g|>\int_W|g|\quad\text{for all } g\in G_{n-k}.$$

By Theorem 4(b) the assertion is proved.

4. An application to rational L^1-approximation. Dunham has shown in [1] that a function $f\in L^1$ which has 0 as a best L^1-approximation from $R_1^n[-1,+1]$ must vanish on a set of positive measure. In this section we construct a set $Z\subset I$, such that every function $f\in L^1$ with $Z(f)\supset Z$ has 0 as a best L^1-approximation from $R_1^n[-1,+1]$.

Lemma 4. *Suppose* $n\in\mathbb{N}$, $n\geq 2$ *and let* $q_{n+2,d}=U_{n+2}-2d^2U_n+d^4U_{n-2}$ *for* $d\in(-1,+1)$.

(a) *Let* n *be even. Then* $q_{n+2,d}$, $d\in(-1,+1)$, *has no zero in the interval* $(-\cos(n+2)\pi/2(n+3),\cos(n+2)\pi/2(n+3))$.

(b) *Let* n *be odd. Then* $q_{n+2,d}$, $d\in(-1,+1)$, *has no zero in the intervals* $(0,\cos(n+1)\pi/2(n+3))$ *and* $(-\cos(n+1)\pi/2(n+3),0)$.

Proof. Simple calculation gives for $x=\cos i\pi/(n+3)$, $i=1,\cdots,n+2$, $U_{n+2}(x)-2d^2U_n(x)+d^4U_{n-2}(x)=(-1)^i2d^2(1-d^2\cos 2i\pi/(n+3))\sin 2i\pi/(n+3))$. Thus we obtain for n even (odd), that $q_{n+2,d}$, $d\in(-1,+1)\setminus\{0\}$, has at least one zero in each interval $(-\cos(i-1)\pi/(n+3))$, $-\cos i\pi/(n+3))$, $i=1,\cdots,n+2$, $i\neq n/2+2$ $(i\neq(n+1)/2,(n+1)/2+2)$ from which the assertion follows.

Theorem 5. *Let* $n\in\mathbb{N}$ *and assume that* $f\in L^1([-1,+1])$.

(a) *Let* $n\geq 2$ *be even and let* $[\alpha,\beta]\subset[-\cos(n+2)\pi/2(n+3),\cos(n+2)\pi/2(n+3)]$. *Suppose that* $Z(f)\supset[-1,\alpha]\cup[\beta,1]$. *Then* 0 *is a best approximation to* f *from* $R_1^n[-1,+1]$.

(b) *Let* $n\geq 3$ *be odd and let* $[\alpha,\beta]\subset[0,\cos(n+1)\pi/2(n+3)]$ *or* $[\alpha,\beta]\subset[-\cos(n+1)\pi/2(n+3),0]$. *Suppose that* $Z(f)\supset[-1,\alpha]\cup[\beta,1]$. *Then* 0 *is a best approximation to* f *from* $R_1^n[-1,+1]$.

Proof. For $\tilde{d}\in(-1,+1)$ let $G_{n,\tilde{d}}:= =\{p/1-\tilde{d}x\,|\,p\in P_n\}$. By Theorem 4 of [8] it follows that for $d=(1-\sqrt{1-\tilde{d}^2})/\tilde{d}$

$$\varphi_d:=\operatorname{sgn}\left(U_{n+2}-2d^2U_n+d^4U_{n-2}\right)\perp G_{n,\tilde{d}}.$$

Furthermore we obtain from Lemma 4 that φ_d, $d \in (-1, +1)$, has no change of sign on $[\alpha, \beta]$. Thus it follows from Theorem 3 that for every $\tilde{d} \in (-1, +1)$

$$\int_{[\alpha,\beta]} |g| \leq \int_{I \setminus [\alpha,\beta]} |g| \quad \text{for all } g \in G_{n,\tilde{d}}.$$

According to [1, p. 226] the assertion is proved.

REFERENCES

[1] C. B. DUNHAM, *Degeneracy in mean rational approximation*, J. Approx. Theory, 4 (1971), pp. 225–229.

[2] S. JA. HAVINSON AND Z. S. ROMANOVA, *Approximation properties of finite-dimensional subspaces in L_1*, Mat. Sb. 89 (1972), pp. 3–15; Math. USSR Sb., 18 (1972), pp. 1–14.

[3] M. KREIN, *The L-problem in an abstract linear normed space*, in Some Questions in the Theory of Moments, N. I. Akiezer and M. Krein, eds., Translations of Mathematical Monographs, Vol. 2, Amer. Math. Soc., Providence, R. I., 1962.

[4] A. KROO, *On the continuity of best approximations in the space of integrable functions*, Acta Math. Acad. Sci. Hungar., 32 (1978), pp. 331–348.

[5] P. LAASONEN, *Einige Sätze über Tschebycheffsche Funktionensysteme*, Ann. Acad. Sci. Fenn., 52 (1949), pp. 3–24.

[6] T. S. MOTZKIN AND J. L. WALSH, *Polynomials of best approximation on an interval*, Proc. Nat. Acad. Sci., 45 (1959), pp. 1523–1528.

[7] ______, *Best approximations within a linear family on an interval*, Proc. Nat. Acad. Sci., 46 (1960), pp. 1225–1233.

[8] F. PEHERSTORFER, *On the representation of extremal functions in the L^1-norm*, J. Approx. Theory, 27 (1979), pp. 61–75.

[9] ______, *Trigonometric polynomial approximation in the L^1-norm*, Math. Z., 169 (1979), pp. 261–269.

[10] ______, *On a L^1-approximation problem*, Anal. Math., 8 (1982), pp. 181–188.

[11] ______, *On a sufficient condition in L^1-approximation*, J. Math. Anal. Appl., 84 (1981), pp. 170–177.

[12] R. R. PHELPS, *Čebysev subspaces of finite dimension in L_1*, Proc. Amer. Math. Soc., 17 (1966), pp. 646–652.

[13] A. PINKUS AND Z. ZIEGLER, *Interlacing properties of the zeros of the error function in best L_p-approximation*, J. Approx. Theory, 27 (1979), pp. 1–18.

[14] J. RICE, *The Approximation of Functions*, Vol. 2, Addison-Wesley, Reading, MA, 1969.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics

013

OPTIMAL MONOSPLINES WITH A MAXIMAL NUMBER OF ZEROS*

R. B. BARRAR[†] AND H. L. LOEB[†‡]

Abstract. In this paper we obtain general theorems about oscillating families. Both polynomial and extended totally positive monosplines are included. The results are applied to characterize the unique monospline of minimal uniform norm that oscillates in a given manner.

Introduction. This article is devoted to the proof of several new results concerning both polynomial and extended totally positive monosplines which oscillate in a prescribed manner. Among such families we characterize the unique function of minimal uniform norm. Actually our results are shown to be corollaries of more general theorems about oscillating families. One application of our results is to the problem of finding an optimal integration formula in L_1 where the formula is to be exact for all polynomials of a prescribed maximal degree and for a class of splines of given knot multiplicities [16].

Our results extend the literature in two directions. First, they generalize the corresponding properties of Chebyshev polynomials [15], and of Chebyshev systems [6]. Second, they generalize the characterization theorem for best uniform monosplines with multiple nodes and simple zeros to allow for both multiple nodes and multiple zeros (see [9], [10], [11], [12]).

For the polynomial case our analysis is based on some recent results on interpolation [1]. For totally positive kernels we rely on the results of [13], [14].

Polynomial monosplines. In this section we develop the theory which will enable us to characterize the oscillating polynomial monospline of least uniform norm. Specifically consider the set of all monosplines of the form

$$M(x)=\int_0^1\phi_p(x,\xi)\,d\xi+\sum_{i=0}^{p-1}a_i\phi_p^{(i)}(x,0)+\sum_{i=1}^{q}\sum_{j=0}^{m_i-1}a_{ij}\phi_p^{(j)}(x,\xi_i)\tag{1}$$

where each of the following quantities are fixed;

a) m_i is an odd positive integer $i=1,\cdots,q$,

b) $p\ge 2$ is an integer.

Further

$$\phi_p^{(i)}(x,\xi)=\frac{\partial^i}{\partial\xi^i}\frac{(x-\xi)_+^{p-1}}{(p-1)!}\quad\text{with } y_+^m=\begin{cases}y^m, & y\ge 0,\\ 0, & y<0\end{cases}$$

and the "free knots" $\{\xi_i\}_{i=1}^q$ obey the inequalities, $0<\xi_1<\cdots<\xi_q<1$. Setting $N=p+\sum_{i=1}^q(m_i+1)$, consider a set of fixed positive integers $\{n_i\}_{i=1}^s$ which satisfy, $N=\sum_{i=1}^s n_i$. The basic problem can be phrased in the following manner. Let the open simplex Δ_s be defined as follows:

$$\Delta_s=\{\mathbf{x}=(x_1,\cdots,x_s):\ 0<x_1<\cdots<x_s<1\}.\tag{2}$$

*Received by the editors February 3, 1983, and in revised form June 30, 1983.

[†] Mathematics Department, University of Oregon, Eugene, Oregon 97403.

[‡] The research of this author was supported in part by the Sonderforschungsbereich 72, University of Bonn.

Under certain restrictions which will be detailed later, for each $\mathbf{x}=(x_1,\cdots,x_s)\in\Delta_s$ there exists a unique monospline $M(\cdot,\mathbf{x})$ of the form (1) which satisfies

$$\frac{\partial^j M}{\partial x_i^j}(x_i,\mathbf{x})=0, \qquad j=0,1,\cdots,n_i-1, \quad i=1,\cdots,s. \tag{3}$$

We will employ the notation $M^{(j)}(x)=(\partial^j/\partial x^j)M(x,\mathbf{x})$. Then we seek to characterize the $\mathbf{x}^*\in\Delta_s$ with the optimal property,

$$\|M(\cdot,\mathbf{x}^*)\|=\min_{\mathbf{x}\in\Delta_s}\|M(\cdot,\mathbf{x})\|, \tag{4}$$

where $\|f\|=\max_{x\in[0,1]}|f(x)|$.

As previously intimated a key ingredient in our approach is the following pair of recent results.

THEOREM 1 (Barrar, Loeb [1]). *If $m=\max_{1\le i\le q}m_i$ and $n=\max_{1\le i\le s}n_i$, then under the restrictions $m,n\le p$, there is at most one polynomial monospline of the form* (1) *which satisfies* (3). *With the further restriction that $m+n\le p$, there is exactly one monospline which satisfies* (3).

THEOREM 2 (Barrar, Loeb [1]). *For $p\ge3$, $m+n\le p-1$, and each $\mathbf{x}\in\Delta_s$ the corresponding $M(\cdot,\mathbf{x})$ has the feature; If $\mathbf{A}=(a_0,\cdots,a_{p-1},a_{10},\cdots,a_{1,m_1-1},a_{21},\cdots,a_{q,m_q-1},\xi_1,\cdots,\xi_q)\in R^N$ is the parameter set for $M(x,\mathbf{x})\equiv M(x,\mathbf{A})=M(x)$, the $N\times N$ Jacobian matrix $\partial\bar{M}(\mathbf{x})/\partial\mathbf{A}$ is nonsingular with*

$$\frac{\partial\bar{M}(\mathbf{x})}{\partial\mathbf{A}}\equiv\left[\begin{array}{ccc}
\dfrac{\partial M(x_1)}{\partial A_1} & \cdots & \dfrac{\partial M(x_1)}{\partial A_N}\\
\cdot & \cdots & \cdot\\
\cdot & \cdots & \cdot\\
\cdot & \cdots & \cdot\\
\dfrac{\partial M^{(n_1-1)}(x_1)}{\partial A_1} & \cdots & \dfrac{\partial M^{(n_1-1)}(x_1)}{\partial A_N}\\
\dfrac{\partial M(x_2)}{\partial A_1} & \cdots & \dfrac{\partial M(x_2)}{\partial A_N}\\
\cdot & \cdots & \cdot\\
\cdot & \cdots & \cdot\\
\cdot & \cdots & \cdot\\
\dfrac{\partial M^{(n_s-1)}(x_s)}{\partial A_1} & \cdots & \dfrac{\partial M^{(n_s-1)}(x_s)}{\partial A_N}
\end{array}\right]. \tag{4'}$$

The following is a useful extension of a result of Micchelli [2].

THEOREM 3. *For $m+n\le p$, there is a uniform bound on the set of parameters $\mathbf{A}\in R^N$ such that for some $\mathbf{x}\in\Delta_s$,*

$$M(x;\,\mathbf{A})\equiv M(x;\,\mathbf{x}).$$

Proof. Consider $(x_1,\cdots,x_s)\in\Delta_s$ and let

$$\mathbf{y}=\Big(\underbrace{x_1,\cdots,x_1}_{m_1},x_2,\cdots,x_{s-1},\underbrace{x_s,\cdots,x_s}_{m_s}\Big)\in R^N.$$

Choose a sequence, $\{y^{(v)}:=(y_1^{(v)},\cdots,y_N^{(v)})\}_{v=1}^{\infty}$, such that $0<y_1^{(v)}<\cdots<y_N^{(v)}<1$ and $\mathbf{y}^{(v)}\to\mathbf{y}$. By the cited result of Micchelli, there is a uniform bound on the parameter sets of $M(\cdot,\mathbf{y}^{(v)})$. It follows directly then from Rolle's theorem and Theorem 1 that $M(\cdot,\mathbf{y}^{(v)})$ converges uniformly to $M(\cdot,\mathbf{x})$. Clearly Micchelli's bound will suffice. □

We will consider a more general problem which will have (4) as a special case. Given a set of fixed positive integers $\{\mu_i\}_{i=1}^r$ and a fixed set of real numbers $\{d_i\}_{i=0}^{r+1}$ satisfying:

If $e_i=d_i-d_{i-1}$, $i=1,\cdots,r+1$, then:

$$(5)\qquad (-1)^{N-1-\sum_{j=1}^{i-1}\mu_j}e_i>0,\qquad i=1,\cdots,r+1,$$

where $N-1=\sum_{i=1}^r\mu_i$.

Then we seek to answer the following questions. Does there exist a $\mathbf{x}=(x_1,\cdots,x_r)\in\Delta_r$ and a positive number E such that for some $M(x)$ of the type (1),

$$(6a)\qquad M(x_i)=Ed_i,\quad i=0,1,\cdots,r+1,$$

$$(6b)\qquad M^{(j)}(x_i)=0,\quad j=1,\cdots,\mu_i,\quad i=1,\cdots,r,$$

with $x_0\equiv 0$ and $x_{r+1}\equiv 1$? Further, is such a $M(x)$ unique? Clearly (6) is equivalent to

$$(7a)\qquad \int_{x_{k-1}}^{x_k}M'(x)\,dx=Ee_k,\qquad k=1,\cdots,r+1,$$

$$(7b)\qquad M^{(1+j)}(x_i)=0,\qquad j=0,1,\cdots,\mu_i-1,\quad i=1,\cdots,r,$$

$$(7c)\qquad M(x)=Ed_0+\int_{x_0}^{x}M'(x)\,dx.$$

According to Theorem 1, there is a unique $M'(x)$ which satisfies (7b). We designate this $M'(x)$ as $M'(x,\mathbf{x})$.

For $\mathbf{x}\in\Delta_r$ set

$$\frac{\partial M'(x,\mathbf{x})}{\partial\mathbf{x}}=\left(\frac{\partial M'(x,\mathbf{x})}{\partial x_1},\cdots,\frac{\partial M'(x,\mathbf{x})}{\partial x_r}\right),$$

where of course $M'(x,\mathbf{x})$ has the form (1) with "p" being replaced by "$p-1$". Thus $M'(x,\mathbf{x})$ can be represented by a $(N-1)$-dimensional parameter vector $\mathbf{A}$ with the corresponding $(N-1)\times(N-1)$ Jacobian matrix $\partial M'(\mathbf{x})/\partial\mathbf{A}$ (see (4')). Let $\partial M'(x,\mathbf{A})/\partial\mathbf{A}$ be the $1\times(N-1)$ vector with the ith component, $\partial M'(x,\mathbf{A})/\partial A_i$, and let

$\partial\overline{M'(\mathbf{x})}/\partial\mathbf{x}$ be the $(N-1)\times r$ matrix defined by

$$\frac{\partial\overline{M'(\mathbf{x})}}{\partial\mathbf{x}}=\begin{bmatrix} \frac{\partial M^{(1)}}{\partial x_1}(x_1,\mathbf{A}) & \cdots & \frac{\partial M^{(1)}}{\partial x_r}(x_1,\mathbf{A}) \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \frac{\partial M^{(\mu_1)}}{\partial x_1}(x_1,\mathbf{A}) & \cdots & \frac{\partial M^{(\mu_1)}}{\partial x_r}(x_1,\mathbf{A}) \\ \frac{\partial M^{(1)}}{\partial x_1}(x_2,\mathbf{A}) & \cdots & \frac{\partial M^{(1)}}{\partial x_r}(x_2,\mathbf{A}) \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \frac{\partial M^{(\mu_r)}}{\partial x_1}(x_r,\mathbf{A}) & \cdots & \frac{\partial M^{(\mu_r)}}{\partial x_r}(x_r,\mathbf{A}) \end{bmatrix} \tag{8}$$

where the only nonzero elements are

$$\frac{\partial M^{(\mu_i)}}{\partial x_i}(x_i,\mathbf{A})=M^{(\mu_i+1)}(x_i,\mathbf{A}) \qquad (i=1,\cdots,r).$$

LEMMA 1. *For* $m+n\le p-2$

$$\frac{\partial\overline{M'(x,\mathbf{x})}}{\partial\mathbf{x}}=\frac{-\partial\overline{M'(x,\mathbf{A})}}{\partial\mathbf{A}}\left[\frac{\partial\overline{M'(\mathbf{x})}}{\partial\mathbf{A}}\right]^{-1}\frac{\partial\overline{M'(\mathbf{x})}}{\partial\mathbf{x}}; \tag{9a}$$

$$\left.\frac{\partial^j}{\partial x^j}\frac{\partial}{\partial x_i}M'(x,\mathbf{x})\right|_{x=x_l}=0, \qquad j=0,1,\cdots,\mu_l-2, \quad i,l=1,\cdots,r; \tag{9b}$$

$$\left.\frac{\partial^{\mu_l-1}}{\partial x^{\mu_l-1}}\frac{\partial}{\partial x_i}M'(x,\mathbf{x})\right|_{x=x_l}=\delta_{il}\left[-M^{(\mu_i+1)}(x_i,\mathbf{x})\right], \tag{9c}$$

where

$$\operatorname{sgn}M^{(\mu_i+1)}(x_i,\mathbf{x})=(-1)^{\Sigma_{j=i+1}^{r}\mu_j}, \qquad i,l=1,\cdots,r. \tag{9d}$$

Proof. By Theorem 2, $\partial M'(\mathbf{x})/\partial\mathbf{A}$ is nonsingular. (9a) is thus a consequence of the *implicit function theorem*. Let $\boldsymbol{\delta}_i$ be the member of the natural basis for R^r whose ith coordinate is one. Then for $0\le j\le\mu_l-1$ and $l\neq i$,

$$\left.\frac{\partial^j}{\partial x^j}\frac{\partial}{\partial x_i}M'(x,\mathbf{x})\right|_{x=x_l}=\lim_{h\to 0}\frac{M^{(j+1)}(x_l,\mathbf{x}+h\boldsymbol{\delta}_i)-M^{(j+1)}(x_l,\mathbf{x})}{h}$$

$$=\lim_{n\to 0}\frac{0-0}{h}=0.$$

Next, if $0 \leq j \leq \mu_i - 2$,

$$\left. \frac{\partial^j}{\partial x^j} \frac{\partial}{\partial x_i} M'(x,\mathbf{x}) \right|_{x=x_i} = \lim_{h\to 0} \frac{M^{(j+1)}(x_i,\mathbf{x}+h\boldsymbol{\delta}_i) - M^{(j+1)}(x_i,\mathbf{x})}{h}$$

$$= \lim_{h\to 0} \frac{M^{(j+1)}(x_i,\mathbf{x}+h\boldsymbol{\delta}_i) - M^{(j+1)}(x_i+h,\mathbf{x}+h\boldsymbol{\delta}_i)}{h}$$

$$= -\lim_{\substack{h\to 0\\ \xi_h \to x_i}} M^{(j+2)}(\xi_h,\mathbf{x}+h\boldsymbol{\delta}_i) = 0.$$

Finally,

$$\left. \frac{\partial^{\mu_i-1}}{\partial x^{\mu_i-1}} \frac{\partial}{\partial x_i} M'(x,\mathbf{x}) \right|_{x=x_i} = \lim_{h\to 0} \frac{M^{(\mu_i)}(x_i,\mathbf{x}+h\boldsymbol{\delta}_i) - M^{(\mu_i)}(x_i,\mathbf{x})}{h}$$

$$= \lim_{h\to 0} \frac{M^{(\mu_i)}(x_i,\mathbf{x}+h\boldsymbol{\delta}_i) - M^{(\mu_i)}(x_i+h,\mathbf{x}+h\boldsymbol{\delta}_i)}{h}$$

$$= -M^{(\mu_i+1)}(x_i,\mathbf{x}).$$

Now a slight generalization of [4, Thm. 6] shows that since $M'(x,\mathbf{x})$ has a full set of zeros,

$$\text{(10)} \qquad \operatorname{sgn} M'(x,\mathbf{x}) = (-1)^{\sum_{j=i}^{r}\mu_j}, \qquad x_{i-1} < x < x_i \quad (i = 1,\cdots,r+1),$$

where $x_0 = 0$ and $x_{r+1} = 1$. Hence

$$\operatorname{sgn} M^{(\mu_i+1)}(x_i,\mathbf{x}) = (-1)^{\sum_{j=i+1}^{r}\mu_j} \qquad (i = 1,\cdots,r).$$

This completes the proof. □

Remark. From (9a), $\partial M'(x,\mathbf{x})/\partial x_i$ has the form,

$$\text{(11)} \quad \frac{\partial M'}{\partial x_i}(x,\mathbf{x}) = \sum_{j=0}^{p-2} a_j^{(i)} \frac{\partial}{\partial x} \phi_p^{(j)}(x,0) + \sum_{l=1}^{q} \sum_{j=0}^{m_l} a_{lj}^{(i)} \frac{\partial}{\partial x} \phi_p^{(j)}(x,\xi_l) \qquad (i = 1,\cdots,r).$$

By [5, p. 511], the $(N-1)$-function which generates $\partial M'(x,\mathbf{x})/\partial x_i$ form a totally positive system [7, p. 4].

Let $\{v_i\}_{i=1}^{r} \subset V$ where V is a $(N-1)$-dimensional extended Chebyshev subspace of order $\mu = \max\{\mu_i\} + 1$; that is, each nonzero element of V has at most $N-2$ zeros counting multiplicities up to order $\mu - 1$ and $V \subset C^{\mu-1}[0,1]$. For $\mathbf{x} = (x_1,\cdots,x_r) \subset \Delta_r$, define

$$\text{(12a)} \qquad B\begin{pmatrix} 1,\cdots,r+1 \\ 1,\cdots,r \end{pmatrix} = \left(\int_{x_{i-1}}^{x_i} v_j(x)\,dx \right)_{i=1\, j=1}^{r+1\, r},$$

where $x_0 = 0$ and $x_{r+1} = 1$ and

$$\text{(12b)} \qquad D_{\mathbf{x}}(i) = \det B\begin{pmatrix} 1,\cdots,i-1,i+1,\cdots,r \\ 1, \quad \cdots, \quad r \end{pmatrix}, \qquad i = 1,\cdots,r.$$

Then [6, Lemma 2] can be expressed in the form

LEMMA 2. *If the set of function* $\{v_i\}_{i=1}^r$ *has the additional properties,*

$$\frac{d^j}{dx^j}v_i(x_l)=0, \qquad j=0,1,\cdots,\mu_i-2, \quad i,l=1,\cdots,r,$$

$$\frac{d^{\mu_i-1}}{d^{\mu_i-1}x}v_i(x_l)=\delta_{il}(-1)^{1+\Sigma_{j=i+1}^r\mu_j}, \qquad i,l=1,\cdots,r,$$

then

$$\operatorname{sgn} D_{\mathbf{x}}(i)=\left[(-1)^{i+\Sigma_{l=i+1}^r\mu_l}\right]\left[(-1)^{\Sigma_{j=1}^r\Sigma_{l=j+1}^r\mu_l}\right] \qquad (i=1,\cdots,r).$$

LEMMA 3. *Under the hypothesis of Lemma* 1 *if we let* $v_i(x)=\partial M'(x,\mathbf{x})/\partial x_i$ $(i=1,\cdots,r)$ *in Lemma* 2, *then*

$$\operatorname{sgn} D_{\mathbf{x}}(i)=\left[(-1)^{i+\Sigma_{l=i+1}^r\mu_l}\right]\left[(-1)^{\Sigma_{j=1}^r\Sigma_{l=j+1}^r\mu_l}\right]. \tag{13}$$

Proof. Assume first that for some i $(1\le i\le r)$, $D_{\mathbf{x}}(i)=0$. This means that there is a set $\{b_1,\cdots,b_r\}$ where $\Sigma_{j=1}^r b_j^2>0$ so that if $w(x)=\Sigma_{j=1}^r b_j\partial M'(x,\mathbf{x})/\partial x_j$ then,

$$\int_{x_{j-1}}^{x_j} w(x)\,dx=0, \qquad j=1,\cdots,i-1,i+1,\cdots,r+1. \tag{14}$$

For small $\varepsilon>0$, let $\phi_{p-1}^{(\varepsilon)}(x,y)$ be the Gaussian transform of $\phi_p(x,y)$; that is,

$$\phi_{p-1}^{(\varepsilon)}(x,y)=\frac{1}{\sqrt{2\pi\varepsilon}}\int_{-\infty}^{\infty} e^{-((x-z)/2\varepsilon)^2}\phi_{p-1}(z,y)\,dz.$$

We let $M'(x;\,\varepsilon)$ denote any "monospline" of the form (1) where $\phi_p(x,z)$ is replaced by $\phi_{p-1}^{(\varepsilon)}(x,z)$ and p by $p-1$.

With $\phi_{p-1}^{(0)}(x,z)=\phi_{p-1}(x,z)$, consider now the system of $(N-1)$ nonlinear equations,

$$\frac{d^j}{dx^j}[M'(x,A;\,\varepsilon)]_{x=x_i}\equiv 0, \qquad j=0,1,\cdots,\mu_i-1, \quad i=1,\cdots,r \tag{15}$$

in the $N-1$ unknowns which are the components of A where ε is a parameter. For $\varepsilon=0$ we have a unique solution $A(0)$ by Theorem 1. Further by Theorem 2, the Jacobian at $\varepsilon=0$ is nonsingular. Thus we can invoke the implicit function theorem to secure a solution $A(\varepsilon)$ close to $A(0)$ for small $\varepsilon>0$. Set $M'(x,\mathbf{x};\,\varepsilon)=M'(x,\mathrm{A}(\varepsilon);\,\varepsilon)$. From (9) and (14) it follows that $w(x)$ changes sign in each of the r intervals

$$(x_0,x_1),\,(x_1,x_2),\cdots,(x_{i-2},x_{i-1}),\,(x_{i+1},x_{i+2}),\cdots,(x_r,x_{r+1}).$$

For small $\varepsilon>0$ by (9a), $w^{(\varepsilon)}(x)=\Sigma_{j=1}^r b_j\partial M'(x,\mathbf{x};\,\varepsilon)/\partial x_j$ also has this property. By (9b) $w^{(\varepsilon)}(x)$ has $N-1-r$ additional zeros (including multiplicities) for a total of $N-1$ zeros. This cannot be since $w^{(\varepsilon)}(x)$ is a nonzero element of a $(N-1)$-dimensional extended Chebyshev system [7, p. 15] (see also (11)). Thus $D_{\mathbf{x}}(i)\neq 0$ $(i=1,\cdots,r)$. Thus by continuity for small $\varepsilon>0$, $\operatorname{sgn} D_{\mathbf{x}}(i)=\operatorname{sgn} D_{\mathbf{x}}^{(\varepsilon)}(i)$ where $D_{\mathbf{x}}^{(\varepsilon)}(i)$ is obtained by replacing $v_j(x)$ by $\partial M'(x,\mathbf{x};\,\varepsilon)/\partial x_j$ $(j=1,\cdots,r)$ in (12a, b). Since the $N-1$ functions which generate each $\partial M'(x,\mathbf{x};\,\varepsilon)/\partial x_j$ (note (11) again) form a $(N-1)$-dimensional extended Chebyshev system of all orders, the result follows by (9b, c) and Lemma 2. □

Now consider for $\mathbf{x}\in\Delta_s$, the implicit system of $r+1$ differential equations,

$$(16)\qquad \frac{d}{ds}\left[\int_{x_{k-1}(s)}^{x_k(s)} M'(x,\mathbf{x}(s))\,dx\right]=e_k\frac{dE(s)}{ds}-f_k \qquad (k=1,\cdots,r+1)$$

in the $r+1$ unknowns, $\{\mathbf{x}(s)=(x_1(s),\cdots,x_r(s)),\ E(s)\}$, with the initial conditions $\mathbf{x}(0)=\mathbf{x}=(x_1,\cdots,x_r)$ and $E(0)=0$. Here

$$(17)\qquad \int_{x_{k-1}}^{x_k} M'(x,\mathbf{x})\,dx=f_k,\qquad k=1,\cdots,r+1.$$

From (10) we note that for $\mathbf{x}(s)\in\Delta_r$,

$$(18)\qquad e_kM'(x,\mathbf{x}(s))>0,\quad x\in(x_{k-1}(s),x_k(s))e_kf_k>0,\qquad k=1,\cdots,r+1.$$

Expanding (16) reveals that

$$(19)\qquad \sum_{k=1}^{r}\left[\int_{x_{k-1}(s)}^{x_k(s)}\frac{\partial}{\partial x_k}M'(x,\mathbf{x}(s))\,dx\right]\frac{dx_k(s)}{ds}-\frac{e}{k}\frac{dE(s)}{ds}=-f_k \qquad (k=1,\cdots,r+1).$$

Thus for $\mathbf{x}(s)\in\Delta_r$, an application of Lemma 3, (18), and Cramer's rule shows that

$$(20)\qquad \frac{dE}{ds}=\frac{\Sigma_{k=1}^{r+1}|D_k(\mathbf{x}(s))||f_k|}{\Sigma_{k=1}^{r+1}|D_k(\mathbf{x}(s))||e_k|}\le\frac{\max f_k}{\min e_k},$$

where the Jacobian of the system (19) is nonsingular.

Using the cited result of Micchelli together with the fact that $M'(x,\mathbf{x}(s))$ has a full set of zeros when $\mathbf{x}(s)\in\Delta_r$ yields the fact that the collection of all such functions have uniformly bounded coefficients. Further by (17) and (18) $\mathbf{x}(s)\not\to\partial\Delta_r$ and by integrating (16),

$$(21)\qquad \int_{x_{k-1}(s)}^{x_k(s)} M'(x,\mathbf{x}(s))\,dx=(1-s)f_k+e_kE(s) \qquad (k=1,\cdots,r+1).$$

Since $f_ke_k>0$, the boundedness of $M'(x,\mathbf{x}(s))$ together with (20) and (21) imply that for $s\in[0,1]$, $\mathbf{x}(s)\notin\partial\Delta_r$.

These facts in consort with a maximal extension analysis [8, Chap. II, Thm. 3.1, Lemma 3.1] yields the result that the solution to (16) exists over [0, 1]. Thus at $s=1$, the solution satisfies,

$$(22)\qquad \int_{x_{k-1}(1)}^{x_k(1)} M'(x,\mathbf{x}(1))=e_kE(1) \qquad (k=1,\cdots,r+1),$$

and

$$(23)\qquad M(x)=Ed_0+\int_{x_0}^{x}M'(x,\mathbf{x}(1))\,dx.$$

Thus $(\mathbf{x}(1),E(1))$ solves the problem posed in (6a) and (6b).

Now for each $\mathbf{x}\in\Delta_r$, let $G(\mathbf{x})=(\mathbf{x}(1),E(1))$; that is, $G(\mathbf{x})$ is the solution to the system of differential equations (19) with the initial conditions $\mathbf{x}(0)=\mathbf{x}$ and $E(0)=0$. Since the set Δ_r is connected and G is continuous over Δ_r, $G(\Delta_r)$ is connected. Further if $(\mathbf{x},E)$ is a solution to the nonlinear system defined by (6a, b), as previously remarked the corresponding Jacobian Matrix is nonsingular. Hence, if W is the set of all solutions, each point of W is an isolated point by the implicit function theorem. Next, if we use $(\mathbf{x},E)\in W$ as the initial conditions for the system of differential equations, it is easy to verify that $(\mathbf{x}(s),E(s))\equiv(\mathbf{x},sE)$. Thus $G(\Delta_r)=W$ and each point of W is an isolated point; that is, W consists of exactly one point and the solution of (6a), (6b) is unique.

Using a technique similar to the one employed in (9) one can establish the results as defined in (6a) and (6b) under the weaker hypothesis that $p\geq 1$. Next, Theorems 1 and 2 of [6] can be translated into our setting and then applied to (6a) and (6b) to yield.

THEOREM 4. *Under the assumption that* $m+n\leq p$ *and* $p\geq 1$, *there is exactly one monospline of the form* (1) *satisfying* (3) *which is of minimal uniform norm. The optimal monospline* $M^*(x)$ *is completely characterized by the fact that relative to its set of zeros* $\mathbf{x}^*=(x_1^*,\cdots,x_s^*)\subset\Delta_s$, *there is a set of* $s+1$ *points*,

$$0=t_0<t_1<\cdots<t_s=1,$$

so that

$$\|M^*(\cdot,\mathbf{x}^*)\|=(-1)^{N-\sum_{j=1}^{i}n_j}M^*(t_i,\mathbf{x}^*),\qquad i=0,1,\cdots,s, \quad .$$

with $0=t_0<x_1^*<t_1<x_2^*<\cdots<x_s^*<t_s=1$.

Extended totally positive monosplines. Consider now a kernel $K(x,\xi)$ which is an extended totally positive kernel of order N in both x and ξ in $(c,d)\times[a,d']$ (see 7, p. 375]). We examine all monosplines of the form (1) where $\phi_p(x,\xi)$ is replaced by $K(x,\xi)$, the restrictions on "p" are removed, and the following new restrictions are added; $a<0<1<b$, $0<1<d'$, $c<a<b<d'$, where the free knots must satisfy

$$0\leq\xi_1<\xi_2<\cdots<\xi_q\leq d'.$$

It can be shown using as key ingredients the techniques employed in developing the results for polynomial monosplines that these results are valid also in this new setting. The proofs for this new kernel are far less intricate.

REFERENCES

[1] R. B. BARRAR AND H. L. LOEB, *Fundamental theorem of algebra for monosplines and related results*, SIAM J. Numer. Anal., 6 (1980), pp. 874–882.

[2] C. A. MICCHELLI, *The fundamental theorem of algebra for monosplines with multiplicities*, in Linear Operators and Approximation, P. L. Butzer et al., eds., Birkhauser Verlag, Basel, 1972, pp. 419–430.

[3] L. L. SCHUMAKER, *Zeros of spline functions and applications*, J. Approx. Theory, 18 (1976), pp. 152–168.

[4] R. B. BARRAR AND H. L. LOEB, *Multiple zeros and applications to optimal linear functionals*, Numer. Math., 25 (1976), pp. 251–262.

[5] S. KARLIN, *Total Positivity*, Stanford Univ. Press, Stanford, 1968.

[6] R. B. BARRAR AND H. L. LOEB, *Oscillating Tchebycheff systems*, J. Approx. Theory, 31 (1981), pp. 188–197.

[7] S. KARLIN AND W. S. STUDDEN, *Tchebycheff Systems: With Applications in Analysis and Statistics*, Interscience, New York, 1966.

[8] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[9] R. B. BARRAR AND H. L. LOEB, *On monosplines with odd multiplicity of least norm*, J. d'Analyse Math., 33 (1978), pp. 12–38.

[10] R. S. JOHNSON, *On monosplines of least deviation*, Trans. Amer. Math. Soc., 96 (1960), pp. 458–477.

[11] R. B. BARRAR, H. L. LOEB AND H. WERNER, *On the uniqueness of the best uniform extended totally positive monospline*, J. Approx. Theory, 28 (1980), pp. 20–29.

[12] S. Karlin, *On a class of best nonlinear approximation problems and extended monosplines*, in Studies in Spline Functions and Approximation Theory, S. Karlin, C. Micchelli, A. Pinkus and I. J. Schoenberg, eds., Academic Press, New York, 1976, pp. 19–59.

[13] S. Karlin and A. Pinkus, *Gaussian quadrature formulae with multiple nodes*, in Studies in Spline Functions and Approximation Theory, S. Karlin, C. Micchelli, A. Pinkus and I. J. Schoenberg, eds., Academic Press, New York, 1976.

[14] D. Barrow, *On multiple node Gaussian quadrature formulae*, Math. Comp., 32 (1978), pp. 431–439.

[15] B. D. Bajanov, *A generalization of Chebyshev polynomials*, J. Approx. Theory, 26 (1979), pp. 293–300.

[16] ———, *Uniqueness of the monosplines of least deviation*, in Numerical Integration, G. Hämmerlin, ed., ISNM 45, Birkhäuser Verlag, Basel, 1979.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics
014

ON SIMPLIFIED ASYMPTOTIC FORMULAS FOR A CLASS OF MATHIEU FUNCTIONS*

D. NAYLOR[†]

Abstract. This paper considers the asymptotic form of solutions of the equation $y_{xx}=(u^2+2h^2\cosh 2x)y$ for real values of x and h and large values of u. Attention is focussed on the solution $\psi(x,u)$ that tends to zero as $x\to\infty$ and for values of u in the half plane $\mathrm{Re}(u)\geq 0$. The basic asymptotic formulas that appear require the determination of an elliptic integral but, when u is large, it is shown how this integral can be suitably approximated by elementary functions. An asymptotic formula is derived which gives the large zeros of the function $\psi(x,u)$ regarded as a function of u, the quantity x being supposed prescribed and positive.

1. Introduction. In a previous paper [3] the author considered the integral transform defined by the equation

$$f_1(u)=\int_a^\infty K_u(kr)f(r)\,\frac{dr}{r}$$

where k, a are positive constants and $K_u(kr)$ denotes the MacDonald type Bessel function. This transform is useful when the damped wave equation $\Delta w=k^2w$ is expressed in polar coordinates r, θ and the domain of interest is the infinite region $r\geq a$ bounded internally at $r=a$. If the boundary of the domain of interest is an elliptic cylinder rather than a circular one, the natural coordinates to use would be ρ, σ where $kx=2h\cosh\rho\cos\sigma$, $ky=2h\sinh\rho\sin\sigma$ and h is a constant. Separation of the variables then leads to the pair of equations

$$F''(\rho)=(u^2+2h^2\cosh 2\rho)F(\rho),$$
$$G''(\sigma)+(u^2+2h^2\cos 2\sigma)G(\sigma)=0$$

where u^2 is a separation constant.

On rewriting the first of the above equations in a more standard notation, we find that we are led to consider the basic differential equation

$$y_{xx}=(u^2+2h^2\cosh 2x)y \tag{1}$$

where $0\leq x<\infty$. The solutions of (1) are related to the modified Mathieu functions $M_\nu^{(j)}(z), j=1,2,3,4$, which satisfy the equation

$$y_{zz}=(u^2-2h^2\cosh 2z)y.$$

Upon setting $z=x+i\pi/2$, we find that the preceding equation transforms into (1) so that solutions of the latter are the functions $M_\nu^{(j)}(x+i\pi/2)$, the notation being that of [2, p. 165]. The quantity ν (the characteristic exponent) is related to the eigenvalue parameter u^2 by a complicated equation which for large values of u can be approximated by means of the equation [2, p. 125]

$$u^2=\nu^2+O\left(\frac{h^4}{\nu^2}\right). \tag{2}$$

*Received by the editors September 16, 1982, and in revised form June 1, 1983.

[†] Department of Applied Mathematics, University of Western Ontario, London, Ontario, Canada N6A 5B9.

It is shown in [2, p. 170] that, for fixed ν,

$$(3)\qquad \begin{aligned} M_\nu^{(3)}(z)&=H_\nu^{(1)}(2h\cosh z)[1+O(\operatorname{sech} z)],\\ M_\nu^{(4)}(z)&=H_\nu^{(2)}(2h\cosh z)[1+O(\operatorname{sech} z)] \end{aligned}$$

as $\operatorname{Re}(z)\to+\infty$ in any strip $|\operatorname{Im} z|\leq$constant. In the two preceding formulas $H_\nu^{(1)}$, $H_\nu^{(2)}$ denote the Hankel functions. Since $\cosh(x+i\pi/2)=i\sinh x$,

$$H_\nu^{(1)}(ix)=(-2i/\pi)e^{-i\nu\pi/2}K_\nu(x),\qquad H_\nu^{(2)}(ix)=2e^{i\nu\pi/2}I_\nu(x)+(2i/\pi)e^{-i\nu\pi/2}K_\nu(x),$$

it follows that

$$(4)\qquad M_\nu^{(3)}\left(x+\frac{i\pi}{2}\right)=\frac{-2i}{\pi}e^{-i\nu\pi/2}K_\nu(2h\sinh x)[1+O(\operatorname{cosech} x)]$$

$$(5)\qquad M_\nu^{(4)}\left(x+\frac{i\pi}{2}\right)=\left[\frac{+2i}{\pi}e^{-i\nu\pi/2}K_\nu(2h\sinh x)+2e^{i\nu\pi/2}I_\nu(2h\sinh x)\right]\cdot[1+O(\operatorname{cosech} x)]$$

as $x\to+\infty$ and ν fixed.

It follows from (4), (5) that the equation (1) possesses essentially just one solution that tends to zero as $x\to\infty$. This solution is the function $M_\nu^{(3)}(x+i\pi/2)$ which we denote for brevity as $\psi(x,u)$. To discuss an integral transform having this function as kernel, asymptotic formulas are needed that give the behaviour of $\psi(x,u)$ with the parameter $u\to\infty$ in the half plane $\operatorname{Re}(u)\geq0$ and the results are required for unbounded $x\geq0$. Formulas giving the asymptotic behaviour of the Mathieu function $M_\nu^{(3)}(z)$ have been obtained by Sharples [8], who has applied the theory developed by Olver [4]. The desired results could in principle be obtained from those of Sharples by setting $z=x+i\pi/2$. Alternatively the formulas in question can be deduced from the results obtained by Pitts [6], [7], who considered the equation

$$(6)\qquad y_{xx}+[\lambda-q(x)]y=0.$$

Pitts obtains asymptotic forms of the solution of the preceding equation that tends to zero as $x\to\infty$ in the case when the function $q(x)$ satisfies various conditions that are all fulfilled by the function $q(x)=2h^2\cosh 2x$. This choice reduces the equation (6) to the same form as that of the equation (1). For the problem at hand the formulas developed by Pitts and Sharples both required the determination of the integral

$$(7)\qquad \int_{x_0}^{x}(\lambda-2h^2\cosh 2t)^{1/2}dt$$

where $2h^2\cosh 2x_0=\lambda$ and, in our notation, $\lambda=-u^2$. The expression (7) is an elliptic integral of the second kind. In the next two sections of this paper approximate expressions for this integral are found which are asymptotic as x or $u\to\infty$, but not uniformly with respect to h. The formulas giving the behaviour of $\psi(x,u)$ are derived first as $u\to\infty$ in the sector D: $|\arg u|\leq\pi/2-\varepsilon$, and then as $u\to\infty$ in the sectors D_1: $\pi/2-\varepsilon\leq|\arg u|\leq\pi/2$. The formulas valid in D are expressed entirely in terms of elementary functions, whilst those applicable in D_1 are expressed in terms of Hankel functions of order $1/3$.

2. Asymptotic forms in the sector *D*. The formulas giving the asymptotic behaviour of $\psi(x,u)$ as $u\to\infty$ in D are expressed in terms of the variable $\xi(x,u)$ defined by the equation

$$\xi(x,u)=\int_0^x(u^2+2h^2\cosh 2t)^{1/2}dt. \tag{8}$$

The radical appearing in this integral is chosen so that its real part is positive. That this is possible follows from the fact that $\operatorname{Re}(u\pm ih\sqrt{2\cosh 2x})=\operatorname{Re}(u)\geq 0$ so that we can select $|\arg(u\pm ih\sqrt{2\cosh 2x})^{1/2}|\leq\pi/4$ and hence $|\arg(u^2+2h^2\cosh 2x)^{1/2}|\leq\pi/2$ as desired. Therefore $\operatorname{Re}\xi(x,u)\geq 0$ for all $x\geq 0$. It is shown by Pitts [7, Lemma 2.3] that for $u\in D$ any solution of (1) that is bounded as $x\to\infty$ must reduce to a constant multiple of that solution y_1 which possesses the asymptotic form

$$y_1(x,u)=(u^2+2h^2\cosh 2x)^{-1/4}e^{-\xi(x,u)}[1+o(1)], \tag{9}$$

as $x\to\infty$ for fixed u. This solution also possesses [7, Lemma 2.1] the property that

$$y_1(x,u)=(u^2+2h^2\cosh 2x)^{-1/4}e^{-\xi(x,u)}\left[1+O(u^{-1})\right] \tag{10}$$

as $u\to\infty$ in D, uniformly for all $x\geq 0$. The solution $\psi(x,u)=M_\nu^{(3)}(x+i\pi/2)$ introduced in §1 is a multiple of y_1. If we write $\psi(x,u)=c(u)y_1(x,u)$, the coefficient $c(u)$ can be determined by comparing equations (4) and (9), both of which apply when $x\to\infty$, and appealing to the formula [1, p. 139],

$$K_\nu(2h\sinh x)=\sqrt{\frac{\pi}{4h\sinh x}}\exp(-2h\sinh x)[1+O(e^{-x})]. \tag{11}$$

It follows that

$$c(u)=-i\sqrt{\frac{2}{\pi}}\,e^{-i\nu\pi/2}\lim_{x\to\infty}\exp[\xi(x,u)-he^x]. \tag{12}$$

To determine $c(u)$ and at the same time obtain the form of $\psi(x,u)$ when u is large, we appeal to the following formula for the variable $\xi(x,u)$:

$$\begin{aligned}\xi(x,u)=&(u^2+2h^2\cosh 2x)^{1/2}-u\log\left[u+(u^2+2h^2\cosh 2x)^{1/2}\right]\\&+ux+A(u)+O(u^{-1}e^{-2x}).\end{aligned} \tag{13}$$

The function $A(u)$ appearing in this equation does not appear in the final formula for $\psi(x,u)$ and so does not need to be determined or estimated. The equation (13) also holds as $u\to\infty$ in the entire half plane $\operatorname{Re}(u)\geq 0$, and it holds uniformly for all $x\geq 0$.

To construct (13), we first use the identity $\cosh 2t=\sinh 2t+e^{-2t}$ to form the equation

$$u^2+2h^2\cosh 2t=2h^2\sinh 2t+u^2\tanh 2t+(u^2+2h^2\cosh 2t)e^{-2t}\operatorname{sech}2t.$$

This equation is divided by $(u^2+2h^2\cosh 2t)^{1/2}$ and integrated for $0\leq t\leq x$. We obtain the equation

$$\begin{aligned}\xi(x,u)=&\int_0^x\frac{2h^2\sinh 2t\,dt}{(u^2+2h^2\cosh 2t)^{1/2}}+u^2\int_0^x\frac{\sinh 2t\,dt}{\cosh 2t(u^2+2h^2\cosh 2t)^{1/2}}\\&+\int_0^x e^{-2t}(u^2+2h^2\cosh 2t)^{1/2}\operatorname{sech}2t\,dt.\end{aligned} \tag{14}$$

The first two integrals occurring on the right-hand side of (14) can be evaluated explicitly. If at the same time, the third integral is expressed as the difference of the two integrals corresponding to the domains $(0,\infty)$ and (x,∞), we find the equation

$$\begin{aligned}(15)\quad \xi(x,u)&=(u^2+2h^2\cosh 2x)^{1/2}-u\log\left[u+(u^2+2h^2\cosh 2x)^{1/2}\right]\\ &\quad+\frac{u}{2}\log(\cosh 2x)+A_1(u)-\int_x^{\infty}e^{-2t}(u^2+2h^2\cosh 2t)^{1/2}\operatorname{sech}2t\,dt\end{aligned}$$

where $A_1(u)$ is a function of u only. The integral remaining on the right-hand side of (15) is expressed as the sum of two integrals I_1 and I_2 which are defined by the equations

$$(16)\qquad I_1=\int_x^{\infty}e^{-2t}\left[(u^2+2h^2\cosh 2t)^{1/2}-u\right]\operatorname{sech}2t\,dt,$$

$$(17)\qquad I_2=u\int_x^{\infty}e^{-2t}\operatorname{sech}2t\,dt=\frac{u}{2}\log(1+e^{-4x}).$$

The quantity I_1 is readily shown to be $O(u^{-1}e^{-2x})$ by appealing to the bounds

$$(u^2+2h^2\cosh 2t)^{1/2}=\begin{cases}u+O(u^{-1}e^{2t}), & 0\le t\le \operatorname{Re}(x_0),\\ O(e^t), & \operatorname{Re}(x_0)\le t\end{cases}$$

where x_0 is defined by the equation $2h^2\cosh 2x_0=-u^2$. Since $|u|\sim|he^{x_0}|$ and this is $O(e^t)=O(u^{-1}e^{2t})$ for $t\ge\operatorname{Re}(x_0)$, we see that $(u^2+2h^2\cosh 2t)^{1/2}-u=O(u^{-1}e^{2t})$ uniformly for all $t\ge 0$. Upon inserting this bound into (16), it is seen that $I_1=O(u^{-1}e^{-2x})$ as required, so that, after using (17), we find that (15) takes the form

$$\begin{aligned}(18)\qquad \xi(x,u)&=(u^2+2h^2\cosh 2x)^{1/2}-u\log\left[u+(u^2+2h^2\cosh 2x)^{1/2}\right]\\ &\quad+ux+A(u)+I_1,\end{aligned}$$

where $A(u)=A_1(u)-(u/2)\log 2$ need not be determined. This proves (13).

It is easily shown from (18) that

$$\lim_{x\to\infty}\left[\xi(x,u)-he^x\right]=A(u)-u\log h,$$

so that (12) yields the result

$$c(u)=-i\left(\frac{2}{\pi}\right)^{1/2}\exp\left[-i\nu\pi/2+A(u)-u\log h\right].$$

On substituting this expression into $\psi(x,u)=c(u)y_1$ and recalling (10) and (18), we obtain the required formula

(19)

$$\begin{aligned}\psi(x,u)&=-i\left(\frac{2}{\pi}\right)^{1/2}(u^2+2h^2\cosh 2x)^{-1/4}\\ &\quad\cdot\exp\left[\frac{-i\nu\pi}{2}-ux-u\log h\right.\\ &\qquad\qquad+u\log\left\{u+(u^2+2h^2\cosh 2x)^{1/2}\right\}\\ &\qquad\qquad\qquad\left.-(u^2+2h^2\cosh 2x)^{1/2}\right]\left[1+O(u^{-1})\right]\end{aligned}$$

as $u \to \infty$ in $|\arg u| \le \pi/2 - \varepsilon$. This equation, though somewhat complicated, gives the asymptotic behaviour of the Mathieu function $\psi(x,u) = M_\nu^{(3)}(x + i\pi/2)$ entirely in terms of elementary functions, and it applies uniformly for all $x \ge 0$.

If x is fixed, the equation (19) simplifies to yield the formula

$$(20) \qquad \psi(x,u) = -i\left(\frac{2}{\pi u}\right)^{1/2} \exp\left[\frac{-i\nu\pi}{2} - ux + u\log\left(\frac{2u}{he}\right)\right]\left[1 + O(u^{-1})\right].$$

3. Asymptotic forms in the sectors D_1. In this section we consider the asymptotic formulas that apply in the sectors $\pi/2 - \varepsilon \le |\arg u| \le \pi/2$. We consider first the sector $-\pi/2 \le \arg u \le -\pi/2 + \varepsilon$ and to facilitate the application of the formulas derived by Pitts, we introduce the new variable $v = iu$ so that $0 \le \arg v \le \varepsilon$ in this domain. The asymptotic formulas then involve the variable

$$(21) \qquad \zeta(x,v) = \int_{x_0}^{x} (v^2 - 2h^2\cosh 2t)^{1/2}\,dt$$

where $2h^2\cosh 2x_0 = v^2$, the.value of $\operatorname{Im} x_0$ being chosen in the interval $0 \le \operatorname{Im} x_0 \le \pi/2$. The radical appearing in the integral in (21) is chosen so that its imaginary part is positive.

The basic solution y_2 that is bounded as $x \to \infty$ now possesses the asymptotic forms [7, Lemma 2.3]

$$(22) \qquad y_2 = (v^2 - 2h^2\cosh 2x)^{-1/4}\left(\frac{\pi\zeta}{2}\right)^{1/2} H_{1/3}^{(1)}(\zeta)\left[1 + o(1)\right]$$

as $x \to \infty$ for fixed $v \in D_1$, and [7, Lemma 2.2],

$$(23) \qquad y_2 = (v^2 - 2h^2\cosh 2x)^{-1/4}\left(\frac{\pi\zeta}{2}\right)^{1/2} H_{1/3}^{(1)}(\zeta)\left[1 + O(v^{-1})\right]$$

as $v \to \infty$ in D_1, uniformly for $x \ge 0$.

Alternatively the formulas (22), (23) may if desired be expressed in terms of Airy functions by means of the standard relations connecting these functions with the Bessel functions of order $\frac{1}{3}$. The resulting formulas are similar to those developed by Sharples [8]. In comparing the results, however, it must be remembered that the formulas given by Sharples refer to solutions of the basic differential equation satisfied by the modified Mathieu functions and that it is necessary to set $z = x + i\pi/2$ in his formulas to reconcile the results.

As in §2, it is required to obtain an asymptotic formula for the variable $\zeta(x,v)$ valid as $v \to \infty$ in D_1 and this can be obtained almost immediately from the formula (18) for $\xi(x,u)$ already constructed. On bearing in mind the definitions (8), (21) we find, on making the appropriate changes in (18), that

$$(24) \qquad \begin{aligned} \zeta(x,v) = {} & (v^2 - 2h^2\cosh 2x)^{1/2} - v\log\left[v + (v^2 - 2h^2\cosh 2x)^{1/2}\right] \\ & + vx + B(v) + J(x,v), \end{aligned}$$

where $B(v)$ is determined by the requirement that $\zeta(x_0, v) = 0$ and the quantity $J(x,v)$, which arises from the I_1 term in (18), is defined by the equation

$$(25) \qquad J(x,v) = \int_x^\infty e^{-2t}\left[(v^2 - 2h^2\cosh 2t)^{1/2} - v\right]\operatorname{sech} 2t\,dt.$$

The bound on I_1 established in §2 shows that $J=O(v^{-1}e^{-2x})$ whenever x is real and positive. The condition $\zeta(x_0,v)=0$ leads to the following equation for $B(v)$:

$$(26) \qquad B(v)=v\log v-vx_0-\int_{x_0}^{\infty}e^{-2t}\left[(v^2-2h^2\cosh 2t)^{1/2}-v\right]\operatorname{sech}2t\,dt.$$

The path of integration in the complex t-plane is taken to consist of (i) the straight line from x_0 to $\operatorname{Re}(x_0)$ together with (ii) the part of the real axis for which $t\geq\operatorname{Re}(x_0)$. The contribution of the second such part is $(v^{-1}e^{-2x_0})$ and this is $O(v^{-3})$ since $v\sim he^{x_0}$. The integral corresponding to part (i) is taken along a line for which $\operatorname{Re}(t)=\operatorname{Re}(x_0)\sim\log|v/h|$ so that $(v^2-2h^2\cosh 2t)^{1/2}=O(v)$ and $e^{-2t}\operatorname{sech}2t=O(v^{-4})$ thereon. Hence the integrand is $O(v^{-3})$ on the path in question and since the length of the latter is $|\operatorname{Im}(x_0)|\sim|\arg v|$, which is less than ε, we see that the corresponding integral is itself $O(v^{-3})$. Thus the integral appearing in (26) is $O(v^{-3})$ so that $B(v)=v\log v-vx_0+O(v^{-3})$. The latter equation reduces further, since the equation $v^2=2h^2\cosh 2x_0$ implies that $he^{x_0}=v+O(v^{-3})$ and this leads to the equation

$$(27) \qquad B(v)=v\log h+O(v^{-3}).$$

The equation (24) now gives, for $x\geq 0$, the desired formula:

$$(28) \qquad \zeta(x,v)=(v^2-2h^2\cosh 2x)^{1/2}-v\log\left[v+(v^2-2h^2\cosh 2x)^{1/2}\right]$$
$$+vx+v\log h+B_1(v)+O(v^{-1}e^{-2x})$$

where $B_1(v)=O(v^{-3})$.

In the sector $0\leq\arg v\leq\varepsilon$ the function ψ must reduce to a constant multiple of the solution y_2 introduced at the beginning of this section. We therefore set $\psi=D(v)y_2$ where the coefficient $D(v)$ is determined by taking the limit as $x\to\infty$. On substituting the expressions (4), (22) and appealing to the formula (11) and the result, [1, p. 139],

$$(29) \qquad H^{(1)}_{1/3}(\zeta)=\left(\frac{2}{\pi\zeta}\right)^{1/2}\exp\left(i\zeta-\frac{5i\pi}{12}\right)\left[1+O(\zeta^{-1})\right],$$

which applies as $\zeta\to\infty$ in $-\pi<\arg\zeta<2\pi$, we find that

$$(30) \qquad D(v)=\left(\frac{2}{\pi}\right)^{1/2}\lim_{x\to\infty}\exp\left[-i\zeta-he^{x}-\frac{iv\pi}{2}+\frac{i\pi}{6}\right].$$

The limit present in the preceding equation can be found from the expression (28) by noting that

$$(v^2-2h^2\cosh 2x)^{1/2}=ihe^{x}\left[1+O(e^{-2x})\right],$$
$$\log\left[v+(v^2-2h^2\cosh 2x)^{1/2}\right]=x+\log h+\frac{i\pi}{2}+O(e^{-x})$$

for v fixed and x large. It follows that

$$\lim_{x\to\infty}\left[\zeta(x,v)-ihe^{x}\right]=B_1(v)-\frac{iv\pi}{2}.$$

The insertion of this result into equation (30) leads to the formula

$$D(v)=\left(\frac{2}{\pi}\right)^{1/2}\exp\left[-\frac{iv\pi}{2}-\frac{v\pi}{2}+\frac{i\pi}{6}\right]\left[1+O(v^{-3})\right].$$

Hence, by (23), the basic solution $\psi = D(v)y_2$ possesses the asymptotic form

$$(31)\quad \psi = (v^2 - 2h^2 \cosh 2x)^{-1/4} \exp\left[-\frac{iv\pi}{2} - \frac{v\pi}{2} + \frac{i\pi}{6}\right]\left[\zeta^{1/2} H_{1/3}^{(1)}(\zeta) + O(v^{-1} e^{-\operatorname{Im}\zeta})\right]$$

as $v \to \infty$ in $0 \leq \arg v \leq \varepsilon$, uniformly for $x \geq 0$. In (30) the variable $\zeta(x, v)$ is defined by (28) in which the term $B_1(v)$, being $O(v^{-3})$, is omitted, since only the leading term in the asymptotic expansion of ψ is sought, and $O(v^{-1})$ terms are already ignored in the approximation accepted.

If x is fixed, the equations (31) and (28) simplify to give the formulas

$$(32)\qquad \psi = v^{-1/2} \exp\left[-\frac{iv\pi}{2} - \frac{v\pi}{2} + \frac{i\pi}{6}\right]\left[\zeta^{1/2} H_{1/3}^{(1)}(\zeta) + O(v^{-1} e^{-\operatorname{Im}\zeta})\right]$$

$$(33)\qquad \zeta(x, v) = v(x+1) - v \log\left(\frac{2v}{h}\right) + O(v^{-1}),$$

for $v \to \infty$ in $0 \leq \arg v \leq \varepsilon$. Since the variable ζ is now large, the Hankel function present in equation (32) may be replaced by the appropriate asymptotic expression. It is shown by Pitts [6, Lemma 5.1], that as x varies between 0 and ∞, $\arg \zeta$ varies over an interval contained within $(-\pi, \pi/2)$. If v is real and $v^2 < 2h^2 \cosh 2x$, then $\arg \zeta = \pi/2$ but if $v^2 > 2h^2 \cosh 2x$, then $\arg \zeta = -\pi$. Thus, for v large and positive and x fixed, $\arg \zeta = -\pi$. This is consistent with equation (33), the dominant term on the right-hand side of that equation being the logarithmic one. For such values of v the asymptotic formula (29) is inapplicable and it is necessary to appeal to the formulas [9, p. 75]:

$$H_w^{(1)}(ze^{-i\pi}) = J_w(ze^{-i\pi}) + iY_w(ze^{-i\pi}),$$
$$J_w(ze^{-i\pi}) = e^{-iw\pi} J_w(z),$$
$$Y_w(ze^{-i\pi}) = e^{iw\pi} Y_w(z) - 2i \cos w\pi J_w(z).$$

After slight reduction it is found that

$$H_{1/3}^{(1)}(ze^{-i\pi}) = 2e^{-i\pi/6}\left[\cos\frac{\pi}{6} J_{1/3}(z) - \sin\frac{\pi}{6} Y_{1/3}(z)\right].$$

On substituting the asymptotic formulas, [1, p. 139],

$$J_{1/3}(z) = \left(\frac{2}{\pi z}\right)^{1/2} \cos\left[z - \frac{5\pi}{12}\right]\left[1 + O(z^{-1})\right],$$
$$Y_{1/3}(z) = \left(\frac{2}{\pi z}\right)^{1/2} \sin\left[z - \frac{5\pi}{12}\right]\left[1 + O(z^{-1})\right],$$

which apply as $z \to \infty$ in $|\arg z| < \pi$, we find that

$$(34)\qquad H_{1/3}^{(1)}(ze^{-i\pi}) = 2e^{-i\pi/6}\left(\frac{2}{\pi z}\right)^{1/2} \cos\left(z - \frac{\pi}{4}\right)\left[1 + O(z^{-1})\right].$$

On setting $\zeta = ze^{-i\pi}$ in (32) and inserting the formula (34), we obtain the equation

$$(35)\qquad \psi = 2\left(\frac{2}{\pi v}\right)^{1/2} \exp\left(-\frac{iv\pi}{2} - \frac{v\pi}{2} - \frac{i\pi}{2}\right)\left[\cos\left(z - \frac{\pi}{4}\right) + O(v^{-1})\right],$$

as $v \to \infty$ in $0 \le \arg v \le \varepsilon$, where

$$z = v \log\left(\frac{2v}{h}\right) - v(x+1) + O(v^{-1}). \tag{36}$$

The final formula for ψ is therefore

$$\psi = 2\left(\frac{2}{\pi v}\right)^{1/2} \exp\left[-\frac{iv\pi}{2} - \frac{v\pi}{2} - \frac{i\pi}{2}\right]\left[\cos\left\{v\log\left(\frac{2v}{h}\right) - v(x+1) - \frac{\pi}{4}\right\} + O(v^{-1})\right].$$

4. The zeros of $\psi(x,u)$. In this section we consider the asymptotic distribution of the zeros of the function $\psi(x,u)$, the quantity x being supposed prescribed and positive. First it is noted that the zeros in question are purely imaginary. This can be verified by applying a standard procedure in which the equation (1) is multiplied by the complex conjugate $\bar{y}$ and integrated for $a \le x < \infty$, the integral of the term $\bar{y}y_{xx}$ being transformed by an integration by parts. If $y(a) = 0$, we find the equation

$$\int_a^\infty |y_x|^2 dx + 2h^2 \int_a^\infty |y|^2 \cosh 2x\, dx + u^2 \int_a^\infty |y|^2 dx = 0.$$

It follows immediately from this equation that the possible values of u^2 are negative and that the corresponding values of u are purely imaginary. The formula (20) which applies in the sector $|\arg u| \le \pi/2 - \varepsilon$ confirms that no zeros can occur in this sector for large u. Those zeros that are located on the negative imaginary axis of the complex u-plane and which are of sufficiently large magnitude can be determined with the aid of (35), since $v = iu$ is positive. Alternatively the Hankel function appearing in (32) could be expressed in terms of the Airy function and appeal made to the results of Olver [5, p. 367] where the zeros of the latter function are determined. If we follow the former procedure, we find that the large zeros are given by the approximate formula

$$z = \left(n - \frac{1}{4}\right)\pi + O(n^{-1})$$

where n is a large positive integer. By (36), the values of v are then given by the equation

$$v \log\left(\frac{2v}{h}\right) - v(x+1) = \left(n - \frac{1}{4}\right)\pi + O(n^{-1}). \tag{37}$$

The corresponding zero of the function $\psi(x,u)$ is obtained from the value of v determined by (37) by applying the formula $u = -iv$.

Acknowledgment. The author wishes to thank the referee for suggestions and improvements which simplified the final formulas for $\xi(x,u)$, $\psi(x,u)$ and which extended their domains of validity.

REFERENCES

[1] W. Magnus, F. Oberhettinger and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1965.

[2] J. Meixner and F. W. Schafke, *Mathieusche Funktionen und Spharoidfunktionen*, Springer-Verlag, Berlin, 1954.

[3] D. Naylor, *On a non-selfadjoint eigenfunction expansion*, SIAM J. Math. Anal., 9 (1978), pp. 967–978.

[4] F. W. J. OLVER, *The asymptotic solution of linear differential equations of the second order for large values of a parameter*, Phil. Trans. Roy. Soc. A., 247 (1954), pp. 307–327.

[5] ______, *The asymptotic expansions of Bessel functions of large order*, Phil. Trans. Roy. Soc. A., 247 (1954), pp. 328–368.

[6] C. G. C. PITTS, *Asymptotic approximations to solutions of a second order differential equation*, Quart. J. Math. Oxford, (2) 17 (1966), pp. 307–320.

[7] ______, *Simplified asymptotic approximations to solutions of a second order differential equation*, Quart. J. Math. Oxford, (2) 21 (1970), pp. 223–242.

[8] ALAN SHARPLES, *Uniform asymptotic forms of modified Mathieu functions*, Quart. J. Mech. Appl. Math., 20 (1967), pp. 365–380.

[9] G. N. WATSON, *Theory of Bessel Functions*, Cambridge Univ. Press, London, 1958.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics
015

EXPANSIONS OF OPERATORS RELATED TO *xD* AND THE FRACTIONAL DERIVATIVE*

R. TREMBLAY[†] AND B. J. FUGÈRE[‡]

Abstract. The operator

$$\left(D_z^{(1-\delta)\beta}\prod_{j=1}^{m}\left(zD+\alpha_j\right)^{r_j}z^{\delta\beta}\right)^n,$$

where $\delta\in\{0,1\}$, $m,n,r_j\in N\cup\{0\}$ and α_j, $\beta\in C$, and where D_z^β is the fractional derivative operator, is expanded in terms of the elementary operator

$$\left(zD-(-1)^\gamma\omega-(1-\gamma)t+1\right)_t\equiv z^{(1-2\gamma)\omega+(1-\gamma)t}D^tz^{(2\gamma-1)\omega+\gamma t},$$

where $\gamma\in\{0,1\}$ and $\omega\in C$. An explicit expression for the coefficients appearing in these expansions is obtained. The expansions contain many well-known cases in the literature as special cases. Also, many new operational formulas are given, in particular with the ordinary differential operator D and the integral operator D_z^{-1}.

Some equivalent relations are also obtained by using two different sets of parameters. An illustration is given by

$$\left(D\prod_{j=1}^{m}\left(zD+\alpha_j\right)^{r_j}\right)^n=z^{-n-1}\left(\prod_{j=1}^{m}\left(zD+\alpha_j-1\right)^{r_j}(zD-1)z\right)^nz^{-n+1}.$$

1. Introduction. The derivative of a function $F(z)$ with respect to $g(z)$ of arbitrary order α, where $\alpha\in C$ in general, is denoted by $D_{g(z)}^\alpha F(z)$ and is called a fractional derivative.

An extensive "fractional calculus" for the operator $D_{g(z)}^\alpha$ exists in the literature (see for a good survey [20]), and many examples of the use of the fractional derivative in integral [11] and differential [12] equations have also been discussed. Many representations of this operator have been proposed in the past. The most important ones have recently been reviewed [16], [17].

Reference [17] contains a large number of selected formulas and theorems concerning the "fractional calculus" such as the Leibniz's rule for the law of exponents, the generalized Taylor's series, etc.

Some of these results have been generalized by R. Tremblay [23], by considering more general operators constructed with the fractional derivative, namely:

$$D_z^\alpha\left(z^\alpha D_z^\alpha\right)^r\quad\text{and}\quad\left\{D_z^{(1-\delta)\beta}\prod_{j=1}^{m}\left(zD+\alpha_j\right)^rz^{\delta\beta}\right\}^n,$$

where $\delta\in\{0,1\}$, $\beta,\alpha_j\in C$ and $m,n,r,j\in N\cup\{0\}$. Under suitable conditions the validity of the law of exponents has been shown [23, Thm. 4.3], namely,

$$D_z^\alpha\left(z^\alpha D_z^\alpha\right)^r\cdot D_z^\beta\left(z^\beta D_z^\beta\right)^r=D_z^{\alpha+\beta}\left(z^{\alpha+\beta}D_z^{\alpha+\beta}\right)^r,\tag{1.1}$$

*Received by the editors February 4, 1981, and in revised form October 5, 1983. This research was supported by the Defense Research Board under grant CRAD 3610-607 and CRSNG A-8699.

[†] Département de Mathématiques, Université du Québec à Chicoutimi, Québec, Canada G7H 2B1.

[‡] Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, Ontario, Canada K7L 2W3.

and the following operational formula obtained [23, Thm. 4.2]:

$$(1.2)\quad \left\{D_z^{(1-\delta)}\prod_{j=1}^{m}\left(zD+\alpha_j\right)^{r_j}z^{\delta\beta}\right\}^n = D_z^{(1-\delta)\theta\beta n}z^{\delta(1-\theta)\beta n}\prod_{i=0}^{n-1}\prod_{j=1}^{m}\left[zD+\alpha_j-\beta\theta n+\beta i\right]^{r_j}$$
$$\cdot z^{\delta\theta\beta n}D_z^{(1-\delta)(1-\theta)\beta n},$$

where $\delta,\theta\in\{0,1\}$. It should be noted that in [23] the proof given is for $r_j=r$, $\forall_j$, but the result remains valid for different r_j.

The study of this class of operators will be carried out in this paper. First, the following expansion formula must be proven:

$$(1.3)\quad \prod_{j=1}^{m}\left(zD+\alpha_j+1\right)_n^{r_j}=\sum_{t=0}^{nR}C_{n,t,R}^{(\gamma)}(\alpha_M,\omega)\left(zD-(-1)^\gamma\omega-(1-\gamma)t+1\right)_t,$$

where $(A)_t=A(A+1)\cdots(A+t-1)$, $t\geq 1$; $(A)_0=1$, $R=\Sigma_{j=1}^m r_j$, where $r_j\in N$, $j\in N$, and $\gamma\in\{0,1\}$. The parameters ω and the set $\alpha_M=\{\alpha_1,\alpha_2,\cdots,\alpha_m\}\in C$ are independent of z. The coefficients $C_{n,t,R}^{(\gamma)}(\alpha_M,\omega)$ are defined by the hypergeometric form (see [22] for the definition of the hypergeometric function),

$$(1.4)\quad C_{n,t,R}^{(\gamma)}(\alpha_M,\omega)=\frac{(-1)^{(1-\gamma)t+\gamma nR}}{t!}\prod_{j=1}^{m}\left(1+\omega+(-1)^\gamma\alpha_j-\gamma n\right)_n^{r_j}$$
$$\cdot{}_{R+1}F_R\left(\left.\begin{matrix}-t,\left(1+\omega+(-1)^\gamma\alpha_M+(1-\gamma)n,r_M\right)\\ \left(1+\omega+(-1)^\gamma\alpha_M-\gamma n,r_M\right)\end{matrix}\right|1\right).$$

This contracted notation (A_M,r_M) appearing in the C coefficients defined above represents the sequence $(A_1,r_1),(A_2,r_2),\cdots,(A_m,r_m)$, in which each symbol (B,p) contains p terms equal to B. For the case $r_i=1$ for $i\in\{1,2,\cdots,m\}$, $(A_m,1)=A_M$. With this definition, the contracted notation corresponds to the usual contracted notation A_M, representing the sequence $A_1,A_2,\cdots,A_m$, which has been previously defined [18, p. 41].

Consequently, by using (1.3) and the operational formula (1.2), the main result of this paper involving the fractional derivative D_z^β is obtained:

$$(1.5)\quad \left\{D_z^{(1-\delta)\beta}\prod_{j=1}^{m}\left(zD+\alpha_j\right)^{r_j}z^{\delta\beta}\right\}^n=\sum_{t=0}^{nR}C_{1,t,nR}^{(\gamma)}(\alpha_M-\beta\theta n+\beta N-1,\omega)$$
$$\cdot D^{(1-\delta)\beta\theta n}z^{\delta(1-\theta)\beta n}\left(zD-(-1)^\gamma\omega-(1-\gamma)t+1\right)_t$$
$$\cdot z^{\delta\beta\theta n}D_z^{(1-\delta)(1-\theta)\beta n},$$

where $\beta\in C$; $\gamma,\delta,\theta\in\{0,1\}$ and $\alpha_M-\beta\theta n+\beta N-1$ represent the set of parameters $\{\alpha_j-\beta\theta n+\beta i-1|1\leq i\leq n,\ 1\leq j\leq m\}$. The other parameters have been previously defined. Of course, the choice of the parameters must be such that the hypergeometric function is well defined.

In fact, (1.4) contains eight formulas depending on the values assigned to the parameters θ,γ and δ.

In §3, the cases where the hypergeometric function in (1.4) can be summed are investigated for the ordinary differential operator D and the integral operator D_z^{-1}.

To complete §3 various identities are given by specializing the parameters. It should also be emphasized that the C coefficients considered here are independent of z. The case where the coefficients are functions of z will be the subject of a subsequent paper.

2. Sketch of the proof and some remarks on the C coefficients. As mentioned in the introduction, the main formula (1.5) will be proved if the operational formula (1.3) can be proved. To establish the proof of this operational formula, we can proceed by induction. Another alternative is to operate with each side of (1.3) on x^s. If there is equality for $s=0,1,2,3,\cdots$, then Carlitz's theorem [6] guarantees the equivalence of the operators involved. In carrying out the proof by induction, the principles given in [23] must be used.

It is interesting to notice the following properties of the C coefficients:

$$(2.1)\qquad C^{(\gamma)}_{1,t,nR}(\alpha_M-\beta\theta n+\beta N-1,\omega)=C^{(\gamma)}_{1,t,nR}\big(\alpha_M-\beta\theta n+\beta N-1+(-1)^{\gamma}\omega,0\big),$$

$$(2.2)\qquad C^{(\gamma)}_{n,t,R}(\alpha_M,\omega)=C^{(\gamma)}_{n,t,R}\big(\alpha_M+(-1)^{\gamma}\omega,0\big).$$

Also, if $m=1, r_1=r$, $\omega=\gamma=\alpha_1=0$ in (1.4), we obtain

$$(2.3)\qquad \begin{aligned} C^{(0)}_{n,t,r}(0,0)&=\frac{(-1)^t}{t!}(n!)^r{}_{r+1}F_r\left(\begin{matrix}-t, & (1+n,r)\\ & (1,r)\end{matrix}\,\middle|\,1\right)\\ &=\frac{1}{t!}\sum_{j=0}^{t}(-1)^{t-j}\binom{t}{j}(j+1)^r(j+2)^r\cdots(j+n)^r\\ &=A^{(r)}(n,t),\end{aligned}$$

where

$$(2.4)\qquad \big(D(xD)^r\big)^n=\sum_{t=0}^{nr}A^{(r)}(n,t)x^tD^{t+n}.$$

In fact the last equation generalizes Lardner's formula [15], which is given by (3.5). These operators have been previously introduced by Carlitz [5]. Other formulas involving the numbers $A^{(r)}(n,t)$ can be found in [23]. It should also be noted that (2.3) has been obtained by using a result developed in [4] and that the $A^{(r)}(n,t)$ contain as a special case the Stirling numbers of the second kind.

3. Some special cases. In this section, some special cases of the main result will be given. They contain a large number of parameters, and by specializing them many new operational formulas involving the operator D are obtained.

First, we investigate the cases where the hypergeometric function appearing in the C coefficients can be summed. To the knowledge of the authors, only the Gauss summation theorem [22, p. 49],

$$(3.1)\qquad {}_2F_1\left(\begin{matrix}a,b\\ c\end{matrix}\,\middle|\,1\right)=\frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)},\qquad \operatorname{Re}(c-a-b)>0,$$

can be applied.

We begin with the case $m=r_1=1$ in (1.5). Using (3.1), the following expansions are obtained:

$$(3.2)\quad \{D^{1-\delta}(zD+\alpha)z^{\delta}\}^n=\sum_{t=0}^{n}(-1)^{\gamma(n-t)}\binom{n}{t}\frac{\left(1+\omega+(-1)^{\gamma}(\alpha-\theta n)-\gamma n\right)_n}{\left(1+\omega+(-1)^{\gamma}(\alpha-\theta n)-\gamma n\right)_t}$$
$$\cdot D^{(1-\delta)\theta n}z^{\delta(1-\theta)n}\left(zD-(-1)^{\gamma}\omega-(1-\gamma)t+1\right)_t$$
$$\cdot z^{\delta\theta n}D^{(1-\delta)(1-\theta)n},$$

(3.3)

$$\{D_z^{\delta-1}(zD+\alpha)z^{-\delta}\}^n=\sum_{t=0}^{n}(-1)^{\gamma(n-t)}\binom{n}{t}\frac{\left(1+\omega+(-1)^{\gamma}(\alpha-1)-(1-\theta-\gamma+2\gamma\theta)n\right)_n}{\left(1+\omega+(-1)^{\gamma}(\alpha-1)-(1-\theta-\gamma+2\gamma\theta)n\right)_t}$$
$$\cdot D^{(\delta-1)\theta n}z^{\delta(\theta-1)n}\left(zD-(-1)^{\gamma}\omega-(1-\gamma)t+1\right)_t z^{-\delta\theta n}D_z^{(\delta-1)(1-\theta)n},$$

where $\delta,\theta,\gamma\in\{0,1\}$ and $\omega\in C$. It is interesting to notice that if $\delta=\gamma=\theta=\alpha=\omega=0$ then using Boole's formula [9], [6],

$$(3.4)\qquad z^tD^t=(zD-t+1)_t,$$

reduces (3.2) and (3.3) to

$$(3.5)\qquad (DzD)^n=\sum_{s=0}^{n}\binom{n}{s}\frac{n!}{s!}z^sD^{s+n},$$

$$(3.6)\qquad \left(D_z^{-1}zD\right)^n=\sum_{s=0}^{n}(-1)^{n-s}\frac{n!}{s!}z^sD^{s-n},$$

where $D_z^{-1}f(z)=\int_0^z f(t)\,dt$.

The formula (3.6) has been published by Lardner [15].

Recently, Al-Salam and Ismail [2] have given an equivalent formula of (3.6) for the finite difference operators Δ and ∇. A generalization of (3.5) has also been given by Osipov [19].

Now, by putting $\omega=(-1)^{\gamma+1}(\alpha+\theta n)+\gamma n-n$ in (3.2) and $\omega=(-1)^{\gamma}(1-\alpha)+(1-\theta-\gamma+2\gamma\theta)n-n$ in (3.3), all the terms appearing in this expansion vanish except the last one. Therefore it yields

$$(3.7)\quad \left(D^{1-\delta}(zD+\alpha)z^{\delta}\right)^n=D^{(1-\delta)\theta n}z^{\delta(1-\theta)n}\left(zD+\alpha-\theta n+1\right)_n z^{\delta\theta n}D^{(1-\delta)(1-\theta)n}$$

and

$$(3.8)\quad \left(D_z^{\delta-1}(zD+\alpha)z^{-\delta}\right)^n=D_z^{(\delta-1)\theta n}z^{\delta(\theta-1)n}\left(zD+\alpha-(1-\theta)n\right)_n z^{-\delta\theta n}D_z^{(\delta-1)(1-\theta)n}.$$

The special case $\alpha=0$ in (3.7) can be found in [23], and the case $\alpha=\delta=0$, $\theta=1$ has been previously obtained by Carlitz [5].

Coming back to (1.5) and letting $m=2$, $\gamma=r_1=r_2=1$, $\beta=1$ and $\omega=\alpha_2-\theta n$, we obtain another case and the Gauss summation theorem can be used. Among the $2n+1$

terms obtained from the expansion formula, the first n terms vanish and, after a "sliding" of the index of summation, we obtain

$$(3.9)\quad \left(D^{1-\delta}(zD+\alpha_2)(zD+\alpha_1)z^{\delta}\right)^n=\sum_{t=0}^{n}(-1)^{n-t}\binom{n}{t}\frac{(1-\alpha_2-\alpha_1)_n}{(1+\alpha_2-\alpha_1)_t} \cdot D^{(1-\delta)\theta n}z^{\delta(1-\theta)n}(zD+\alpha_2-\theta n+1)_{n+t} \cdot z^{\delta\theta n}D^{(1-\delta)(1-\theta)n},$$

and similarly from (1.5) with $\beta=-1$,

$$(3.10)\quad \left(D_z^{\delta-1}(zD+\alpha_1)(zD+\alpha_2)z^{-\delta}\right)^n=\sum_{t=0}^{n}\binom{n}{t}\frac{(1-\alpha_1-\alpha_2)_n}{(1+\alpha_1-\alpha_2)_t}D^{(\delta-1)\theta n}z^{\delta(\theta-1)n} \cdot\left(zD+\alpha_2-(1-\theta)n-t\right)_{n+t}z^{-\delta\theta n}D_z^{(\delta-1)(1-\theta)n},$$

where $\delta,\theta\in\{0,1\}$.

Finally it is interesting to notice that many identities can be obtained with different choices of parameters. For instance, by considering the substitutions

(a) $\delta=\theta=\omega=0,\ \beta=\gamma=1$,

(b) $\delta=\theta=\omega=\gamma=0,\ \beta=1,\ m\to m+1,\ \alpha_j\to\alpha_j-1,\ r_{m+1}=0$ and $\alpha_{m+1}\to-1$,

we can deduce

$$(3.11)\quad \left(D\prod_{j=1}^{m}(zD+\alpha_j)^{r_j}\right)^n=z^{-n-1}\left(\prod_{j=1}^{m}(zD+\alpha_j-1)^{r_j}(zD-1)z\right)^n z^{-n+1}.$$

Also, using (1.1) and (3.11) it it easy to obtain the triple equality

$$(3.12)\quad \left(D(zD)^r\right)^n=D^n(z^nD^n)^r=z^{-n-1}\left((zD-1)^{r+1}z\right)^n z^{-n+1}.$$

These equalities are by no means trivial.

Another example is obtained by respectively putting $\alpha_j=-j$ and $\alpha_j=j$ where $j\in\{1,2,\cdots,m\}$ in (3.11):

$$(3.13)\quad (z^{m+2}D^{m+1})^n=z^{n+1}(z^mD^{m+1})^n z^{n-1},$$

$$(3.14)\quad (z^2D^{m+1}z^m)^n=z^{n+1}(D^{m+1}z^m)^n z^{n-1}.$$

Formally, if $m=-2$ in (3.14) or (3.15), a known identity previously proposed as a problem by M. S. Klamkin [13] is rediscovered. This identity is

$$z^{n-1}\int\frac{dz}{z^2}\int\frac{dz}{z^2}\cdots\int\frac{dz}{z^2}\int z^{n-1}F(z)\,dz=\int\cdots\int F(z)(dz)^n+P(z),$$

where $P(z)$ is a polynomial of degree $n-1$. It should be noticed that by taking definite integrals, the $P(z)$ may be ignored.

It should also be pointed out that the formula (1.5) still contains a large number of particular cases to be studied. Here we have limited ourselves to the cases $\beta=\pm1$. To illustrate that, we can say that the operator (1.5) contains as a particular case the operator $(x^rD^r)^n$, which was studied a long time ago by Carlitz [3]. Also, the operator

$(xDx)^n$ was investigated by Al-Salam [1] and the operator $(x^k(xD+\lambda))^n$ was studied recently by K. R. Patil and N. R. Thakare [21]. Other known cases can be obtained. If in (1.5) we put $\beta=k-1$, $m=1$, $\alpha_1=1-k$, $r_1=1$, $\delta=1$, $\theta=0$ and (3.4), we obtain the expansions (corresponding to the values $\gamma=0$ and $\gamma=1$) of the operator $(z^kD)^n$ given by Chak [7]. An interesting use of this operator is made in [24].

We can also obtain the expansion of the operator $(x^{\alpha+1}D)$ in terms of the operator D^k used recently by Charalambides [8] and Comtet [10].

REFERENCES

[1] W. Al-Salam, *Operational representations for the Laguerre and other polynomials*, Duke Math J., 31 (1964), pp. 127–142.

[2] W. A. Al-Salam and M. E. H. Ismail, *Some operational formulas*, J. Math. Anal., 51 (1975), pp. 208–218.

[3] L. Carlitz, *On a class of finite sums*, Amer. Math. Monthly, 37 (1930), pp. 472–479.

[4] ———, *On arrays of numbers*, Amer. J. Math., 54 (1932), pp. 739–752.

[5] ———, *Some operational formulas*, Math. Nachr., 45 (1970), pp. 379–389.

[6] ———, *A theorem on differential operators*, Amer. Math. Monthly, 83 (1976), pp. 351–354.

[7] A. M. Chak, *A class of polynomials and a generalization of Stirling numbers*, Duke Math. J., 23 (1956), pp. 45–56.

[8] Ch. A. Charalambides, *A new kind of numbers appearing in the n-fold convolution of truncated binomial and negative binomial distributions*, SIAM J. Appl. Math., 33 (1977), pp. 279–288.

[9] S. K. Chatterjea, *Operational formulae for certain classical polynomials*, (I), Quart. J. Math. Oxford, 14 (1963), pp. 241–246.

[10] L. Comtet, *Une formule explicite pour les puissances successives de l'opérateur de dérivation de Lie*, C. R. Acad. Sci. Paris, Sér. A, 276 (1973), pp. 165–168.

[11] A. Erdélyi, *An integral equation involving Legendre functions*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 15–30.

[12] T. P. Higgins, *The use of fractional integral operators for solving nonhomogeneous differential equations*, Document D1-82-0677, Boeing Scientific Research Laboratories, Seattle, Washington, 1967.

[13] M. S. Klamkin, *Problem no.* 5974, Amer. Math. Monthly, 81 (1974), p. 525.

[14] M. S. Klamkin and D. J. Newman, *On the reducibility of some linear differential operators*, Amer. Math. Monthly, 66 (1959), pp. 293–295.

[15] T. S. Lardner, *Relations between* ${}_0F_3$ *and Bessel functions*, SIAM Rev., 11 (1969), pp. 69–72.

[16] J. L. Lavoie, T. J. Osler and R. Tremblay, *Fundamental properties of fractional derivatives via Pochhammer integrals*. Fractional Calculus and Its Applications, Lecture Notes in Mathematics 457, Springer-Verlag, Berlin-Heidelberg-New York 1974, pp. 323–356.

[17] ———, *Fractional derivatives and special functions*, SIAM Rev., 18 (1976), pp. 240–268.

[18] Y. L. Luke, *The Special Functions and Their Approximations*, Vol. I, Academic Press, New York, 1969.

[19] S. Osipov, *On the expansion of a polynomial of the operator*, B_α, USSR Comput. Math. and Math. Phys., 3 (1963), pp. 250–256.

[20] T. J. Osler, *A further extension of the Leibniz rule to fractional derivatives and its relation to Parseval's formula*, this Journal, 3 (1972), pp. 1–16.

[21] K. R. Patil and N. R. Thakare, *Bilateral generating function for a function defined by generalized Rodrigues' formula*, Indian J. Pure Appl. Math., 8, 4 (1977), pp. 425–429.

[22] E. D. Rainville, *Special Functions*, Macmillan, New York, 1960.

[23] R. Tremblay, *Some operational formulas involving the operators xD, x*Δ *and fractional derivatives*, this Journal, 10 (1979), pp. 933–943.

[24] O. P. Vijay, *Generalization of Bell polynomials and related operational formulas*, Publ. Inst. Math. (Beograd), 19 (33) (1975), pp. 173–180.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics
016

CONVERGENCE THEOREMS FOR MATRIX CONTINUED FRACTIONS*

DAVID A. FIELD[†]

Abstract. Convergerice theorems are proven for continued fractions of the form $K(A_n/I)$ and $K(I/B_n)$ where A_n and B_n are $n\times n$ matrices and I is the identity matrix. Geometric rules of convergence are determined for the matrix exponential.

1. Introduction. Since calculations involving matrix valued functions with matrix arguments are feasible with large computers, such functions have been used in solving many difficult problems. For example, at General Motors Research Laboratories the matrix exponential has been involved with the dynamic equations of robot control and with mathematical models of catalytic converters. In [11] Varga used Padé approximants, equivalent to continued fraction approximants, of the matrix exponential to derive and analyze certain discrete approximations to solutions of self-adjoint parabolic differential equations.

The matrix continued fraction used by Varga is an example of noncommutative continued fractions about which not much is known. Few convergence theorems for noncommutative continued fractions have been published. Two theorems are stated in [15], where Wynn reviews many aspects of the theory of continued fractions whose elements do not commute under a multiplication law. In a Banach space, extensions of Worpitsky's theorem have been proven by Fair [3], Hayden [4] and Negoescu [6], [7]. Hayden's and Negoescu's papers included the Banach space version of g-fractions. Hayden also proved a twin convergence theorem.

In this paper several convergence theorems for continued fractions whose arguments are $n\times n$ matrices are proven. These theorems[1] include the complete analogue of Worpitsky's theorem, a matrix analogue of a twin convergence theorem proven by Copp [2] for ordinary continued fractions, and a convergence theorem for continued fractions of the form $K(I/B_n)$. The second section of this paper presents the convergence theorems. A third section concludes this paper with convergence rates for the matrix exponential. Throughout this paper the norm of a matrix will be denoted by $\|\cdot\|$ and is not restricted to any particular norm.

2. Convergence theorems. The matrix continued fractions under consideration in the first two theorems in this section are of the form

$$K_2 \equiv \frac{I}{I+}\,\frac{A_2}{I+}\,\frac{A_3}{I+}\cdots, \tag{2.1}$$

where the A's are $n\times n$ matrices of complex numbers and I is the nth order identity matrix. By defining

$$K_{j,n} \equiv \frac{I}{I+}\,\frac{A_j}{I+}\,\frac{A_{j+1}}{I+}\cdots+\frac{A_{n-1}}{I+}\,\frac{A_n}{I}, \qquad n\geq j, \tag{2.2}$$

*Received by the editors November 6, 1981, and in revised form May 30, 1983.

[†] Mathematics Department, General Motors Research Laboratories, Warren, Michigan 48090-9055.

[1] A discussion with the late Professor Karl Rasmussen, Mathematics Department, University of Michigan at Dearborn, of his proof of a matrix Worpitsky theorem (the version in [2]) subsequently led to my proving these theorems which would have been applied in his research on matrix Riccati equations arising from scattering functions in physics [6], [7].

the value of K_2 is the limit of the sequence of its nth approximants, $\{K_{2,n}\}$ whenever this sequence converges.

The nth approximant $K_{2,n}$ can be evaluated in several ways. The backward recurrence method of evaluation proceeds by multiplying $(I+A_n)^{-1}$ and A_{n-1}, followed by adding I to the product, whereupon the inverse of the sum and A_{n-2} are multiplied and so on. Since multiplication of matrices is noncommutative, inverses are usually either always premultiplied, e.g. $(I+A_n)^{-1}A_{n-1}$, or always postmultiplied, e.g. $A_{n-1}(I+A_n)^{-1}$, to form pre- or post-continued fractions.

The computationally preferred method of evaluating (2.2) is by using the standard three-term recurrence relations which are equivalent to the above backward recurrence even for continued fractions with noncommutative elements, see Wynn [14], [15]. However, in the proofs of the following theorems, where an efficient evaluation of $K_{2,n}$ is not an issue, the backward recurrence algorithm will be useful in deriving truncation errors for post-continued fractions.

Crucial to each step in the backward recurrence algorithm is the existence of inverses. To guarantee the existence of these inverses the following well-known lemmas will be used. Their proofs are found in [8].

LEMMA 2.1. *If* $\|A\|<1$, *then* $(I+A)^{-1}$ *exists.*

LEMMA 2.2. *If* $\|A\|<1$, *then* $\|(I+A)^{-1}\|\leq 1/(1-\|A\|)$ *and* $\|(I+A)^{-1}-I\|\leq \|A\|/(1-\|A\|)$.

THEOREM 2.3. *The matrix continued fraction in* (2.1) *converges whenever* $\|A_i\|\leq\frac{1}{4}$.

Proof. The matrix continued fraction in (2.1) will converge if its sequence of nth approximants is a Cauchy sequence. The proof will first show the existence of the inverses needed in the backward recurrence algorithm and then show that the sequence of approximants $\{K_{2,n}\}$ is a Cauchy sequence.

Let $a_n=n/2(n+1)$, $n\geq 1$ and define $K_{j,j-1}\equiv I$, the identity matrix. The following induction argument shows that $\|A_iK_{i+1,i+n-1}\|\leq a_n<1$, $n\geq 1$, so that by Lemma 2.1 the inverses required for the backward recurrence algorithm exist; that is, $K_{i,i+n-1}$ exists for $n\geq 1$. Since $\|A_i\|\leq\frac{1}{4}$ by hypothesis, $\|A_iK_{i+1,i}\|=\|A_i\|\leq a_1=\frac{1}{4}$ and $(I+A_i)^{-1}=K_{i,i}$ exists. Similarly $K_{i+1,i+1}=(I+A_{i+1})^{-1}$ exists and by Lemma 2.2, $\|A_iK_{i+1,i+1}\|\leq\frac{1}{4}(1/(1-\frac{1}{4}))=a_2$ so that $K_{i,i+1}$ also exists. Therefore $K_{i+1,i+2}$ exists. Suppose for $i\geq 2$, $1\leq j\leq n$, $\|A_iK_{i+1,i+j-1}\|\leq a_j<1$, and $K_{i,i+j-1}$ exists. This induction hypothesis implies that $\|A_{i+1}K_{i+2,i+j}\|\leq a_j<1$ so that by Lemma 2.1, $K_{i+1,i+j}=(I+A_{i+1}K_{i+2,i+j})^{-1}$ exists. Thus by Lemma 2.2 $\|A_iK_{i+1,i+j}\|\leq\frac{1}{4}(1/(1-a_j))=a_{j+1}<1$ and by Lemma 2.1, $K_{i,i+n}$ exists and the induction is proven.

The above argument not only proves the existence of the inverses used in the backward recurrence algorithm but also proves the existence of the inverses used in the remainder of the proof.

The norm of the difference between the $(n+p)$th and the nth approximants of the matrix continued fraction in (2.1) can be written as

$$\begin{aligned}\|K_{2,n}-K_{2,n+p}\| &= \left\|(I+A_2K_{3,n})^{-1}-(I+A_2K_{3,n+p})^{-1}\right\| \\ &= \left\|(I+A_2K_{3,n})^{-1}\left[I-(I+A_2K_{3,n})(I+A_2K_{3,n+p})^{-1}\right]\right\| \\ &= \left\|(I+A_2K_{3,n})^{-1}\left[I+A_2K_{3,n+p}-(I+A_2K_{3,n})\right](I+A_2K_{3,n+p})^{-1}\right\| \\ &= \left\|(I+A_2K_{3,n})^{-1}A_2(K_{3,n}-K_{3,n+p})(I+A_2K_{3,n+p})^{-1}\right\|.\end{aligned}$$

Thus,

(2.3)

$$\|K_{2,n}-K_{2,n+p}\|\le\|I-K_{n+1,n+p}\|\prod_{i=2}^{n}\left[\|A_i\|\cdot\left\|(I+A_iK_{i+1,n})^{-1}\right\|\cdot\left\|(I+A_iK_{i+1,n+p})^{-1}\right\|\right]$$

where $K_{n+1,n}\equiv I$. Note that $\|A_iK_{i+1,n}\|\le a_{n-i+1}$. Therefore,

$$\left\|(I+A_iK_{i+1,n})^{-1}\right\|\le 1/(1-a_{n-(i-1)}) \tag{2.4}$$

and

(2.5)

$$\|K_{n+1,n+p}-I\|=\left\|(I+A_{n+1},K_{n+2,n+p})^{-1}(I-(I+A_{n+1}K_{n+2,n+p}))\right\|\le a_p(1-a_p).$$

Combining (2.3) with (2.4) and (2.5) yields

$$\begin{aligned}\|K_{2,n}-K_{2,n+p}\|&\le\frac{a_p}{1-a_p}\left(\frac{1}{4}\right)^{n-1}\prod_{i=1}^{n-1}\left(\frac{1}{1-a_i}\right)\left(\frac{1}{1-a_{i+p}}\right)\\&\le\frac{p}{p+2}\left(\frac{1}{4}\right)^{n-1}\prod_{i=1}^{n-1}\frac{2(i+1)}{(i+2)}\frac{2(i+p+1)}{i+p+2}\\&=\frac{p}{p+2}\frac{2}{n+1}\frac{p+2}{p+n+1}<\frac{2}{n+1}.\end{aligned}\tag{2.6}$$

Since $2/(n+1)$ in (2.6) is independent of p, the sequence of approximants $\{K_{2,n}\}$ of the continued fraction in (2.1) is a Cauchy sequence. Not only is the theorem proved, but (2.6) estimates the error of approximating the continued fraction in (2.1) by its nth approximant.

In the following theorem, letting $\rho=\frac{1}{2}$ in (2.7) sharpens Hayden's theorem [4, p. 369] so that $\|A_{2n-1}\|$ need not be uniformly bounded away from $\frac{1}{4}$. (2.7) is in fact the analogue of Copp's result [2] (also see [12]) for ordinary continued fractions $K(a_n/1)$. Copp's theorem guarantees convergence of $K(a_n/1)$ if

$$|a_{2n-1}|\le\rho^2,\qquad |a_{2n}|\ge(1+\rho)^2$$

where $0<\rho\le 1$. On a historical note, Theorem 2.4 is still sharper than Leighton and Wall's classic theorem published in 1936 [5] where

$$|a_{2n-1}|\le\tfrac{1}{4}\quad\text{and}\quad|a_{2n}|\ge\tfrac{25}{4}.$$

THEOREM 2.4. *Let $\{A_n\}$ be a sequence of $n\times n$ matrices satisfying*

$$\|A_{2n+1}\|\le\rho^2\quad\text{and}\quad\|A_{2n}^{-1}\|\le 1/(1+\rho)^2,\qquad n\ge 1,\tag{2.7}$$

where $0<\rho<1$. Then the matrix continued fraction in (2.1) converges.

Proof. The backward recurrence algorithm requires the existence of the inverse of the denominators of arbitrary sections of the continued fraction in (2.1) whose elements satisfy (2.7). Because the even and odd elements satisfy different constraints, let the denominators which start with an even element be denoted by

$$E_{2j;0}=I+A_{2j}\quad\text{and}\quad E_{2j;n}=I+A_{2j}K_{2j+1,2j+n},\qquad n\ge 1.$$

Let the denominators starting with an odd element be denoted by

$$X_{2j+1;0}=I+A_{2j+1} \quad \text{and} \quad X_{2j+1;n}=I+A_{2j+1}K_{2j+1,2j+1+n}, \qquad n\ge 1.$$

An induction argument will show that $E_{2j;n}^{-1}$ and $X_{2j+1;n}^{-1}$ exist for all $j\ge 1$, and $n\ge 0$. Two matrix identities will be useful throughout the remainder of the proof.

$$(I+B)^{-1}=\left[B(B^{-1}+I)\right]^{-1}=(B^{-1}+I)^{-1}B^{-1}=(I+B^{-1})^{-1}B^{-1}$$

and

$$(I+B)^{-1}=\left[(B^{-1}+I)B\right]^{-1}=B^{-1}(B^{-1}+I)^{-1}=B^{-1}(I+B^{-1})^{-1}.$$

$E_{2j;0}^{-1}$ exists since $(I+A_{2j}^{-1})^{-1}A_{2j}^{-1}$ exists by Lemma 2.1. Lemma 2.2 implies that $\|E_{2j;0}^{-1}\|\le 1/(2\rho+\rho^2)=a_2<1/\rho$. Since $\|A_{2j+1}\|\le\rho^2<1$, $X_{2j+1;0}^{-1}$ also exists by Lemma 2.1.

$$E_{2j;1}^{-1}=\left(I+A_{2j}X_{2j+1;0}^{-1}\right)^{-1}=\left(I+X_{2j+1;0}A_{2j}^{-1}\right)^{-1}X_{2j+1;0}A_{2j}^{-1}.$$

Since $\|X_{2j+1;0}A_{2j}^{-1}\|\le(1+\rho^2)/(1+\rho)^2<1$, $E_{2j;1}^{-1}$ exists by Lemma 2.1 and Lemma 2.2. Furthermore,

$$\left\|E_{2j;1}^{-1}\right\|\le\frac{1+\rho^2}{(1+\rho)^2}\Big/1-\frac{1+\rho^2}{(1+\rho)^2}=\frac{1+\rho^2}{2\rho}=a_3<\frac{1}{\rho}.$$

$X_{2j+1;1}^{-1}$ also exists since $\|A_{2j+1}(I+A_{2j+2})^{-1}\|\le\rho^2a_2=\rho/(2+\rho)<1$.

Suppose for $j\ge 1$ and $1\le k\le n$, $\|E_{2j;k-1}^{-1}\|\le a_{k+1}<1/\rho$ and $X_{2j+1;k-1}^{-1}$ exists. Define

$$a_{i+2}=\frac{(1+\rho^2a_i)/(1+\rho)^2}{1-(1+\rho^2a_i)/(1+\rho)^2}, \quad a_0=0, \quad a_1=1, \quad i\ge 0,$$

and note that $a_{i+2}<1/\rho$, $i\ge 0$. An induction argument requires proving that $\|E_{2j;k}^{-1}\|\le a_{k+2}<1/\rho$ and $X_{2j+1;k}^{-1}$ exists for $1\le k\le n, j\ge 1$.

$$\text{(2.8)} \quad E_{2j;k}^{-1}=\left(I+A_{2j}X_{2j+1;k-1}^{-1}\right)^{-1}=\left(I+X_{2j+1;k-1}A_{2j}^{-1}\right)^{-1}X_{2j+1;k-1}A_{2j}^{-1}.$$

$X_{2j+1;k-1}^{-1}$ exists by the induction hypothesis and

$$\left\|X_{2j+1;k-1}A_{2j}^{-1}\right\|=\left\|\left(I+A_{2j+1}E_{2(j+1);k-2}^{-1}\right)A_{2j}^{-1}\right\|\le\frac{1+\rho^2a_k}{(1+\rho)^2}<1.$$

Consequently $E_{2j;k}^{-1}$ exists by the first equality in (2.8) and since $a_k<1/\rho$, $k\ge 2$. Furthermore

$$\left\|E_{2j;k}^{-1}\right\|\le\frac{(1+\rho^2a_k)/(1+\rho)^2}{1-(1+\rho^2a_k)/(1+\rho)^2}=a_{k+2}<\frac{1/(1+\rho)}{1-1/(1+\rho)}=\frac{1}{\rho}.$$

To show the existence of $X_{2j+1;k}^{-1}$, note that $X_{2j+1;k}=I+A_{2j+1}E_{2(j+1);k-1}^{-1}$ and $\|A_{2j+1}E_{2(j+1);k-1}\|<\rho^2(1/\rho)=\rho<1$ and use Lemma 2.1.

To prove that the sequence $\{K_{2,n}\}$ of approximants is a Cauchy sequence, it is necessary to consider $\|K_{2,n}-K_{2,n+k}\|$, $k\ge 1$ when n and $n+k$ are combinations of odd and even approximants. Since all such cases are essentially similar, without loss in generality, let n be even.

As in the proof of Theorem 2.3,

$$K_{2,n}-K_{2,n+k}=(I+A_2K_{3,n})^{-1}A_2(K_{3,n+k}-K_{3,n})(I+A_2K_{3,n+k})^{-1}.$$

Using the identities involving $(I+B)^{-1}$ changes this equation to

$$\begin{aligned}K_{2,n}-K_{2,n+k}&=\left(I+K_{3,n}^{-1}A_2^{-1}\right)^{-1}K_{3,n}^{-1}(K_{3,n+k}-K_{3,n})K_{3,n+k}^{-1}A_2^{-1}\left(I+K_{3,n+k}^{-1}A_2^{-1}\right)^{-1}\\&=\left(I+K_{3,n}^{-1}A_2^{-1}\right)^{-1}\left(K_{3,n}^{-1}-K_{3,n+k}^{-1}\right)A_2^{-1}\left(I+K_{3,n+k}^{-1}A_2^{-1}\right)^{-1}.\end{aligned}$$

But $K_3=(I+A_3K_{4,n})^{-1}$ and $K_{3,n+k}=(I+A_3K_{4,n+k})^{-1}$ imply

$$K_{2,n}-K_{2,n+k}=\left(I+K_{3,n}^{-1}A_2^{-1}\right)^{-1}A_3(K_{4,n}-K_{4,n+k})A_2^{-1}\left(I+K_{3,n+k}^{-1}A_2^{-1}\right)^{-1}.$$

Continuing these matrix manipulations yields

$$\begin{aligned}&(2.9)\quad \|K_{2,n}-K_{2,n+k}\|\\&\qquad\le\left\|(I+A_n)^{-1}-(I+A_nK_{n+1,n+k})^{-1}\right\|\\&\qquad\cdot\prod_{i=1}^{n/2-1}\|A_{2i}^{-1}\|\|A_{2i+1}\|\left\|\left(I+K_{2i+1,n}^{-1}A_{2i}^{-1}\right)^{-1}\right\|\left\|\left(I+K_{2i+1,n+k}^{-1}A_{2i}^{-1}\right)^{-1}\right\|.\end{aligned}$$

To obtain an estimate of the right-hand side of (2.9), use the fact that for $j=n$ or $n+k$,

$$\begin{aligned}(2.10)\qquad \left\|K_{2i+1,j}^{-1}\right\|&=\|I+A_{2i+1}K_{2(i+1),j}\|\le 1+\rho^2\|K_{2(i+1),j}\|\\&=1+\rho^2\left\|E_{2(i+1);\,j-2(i+1)}^{-1}\right\|\le 1+\rho^2a_{j-2i}<1+\rho.\end{aligned}$$

By Lemma 2.2 write for $j=n$ or $n+k$

$$(2.11)\qquad \left\|\left(I+K_{2i+1,j}^{-1}A_{2i}^{-1}\right)^{-1}\right\|<\frac{1/(1+\rho)}{1-1/(1+\rho)}=\frac{1}{\rho}.$$

Combining (2.7), (2.9) and (2.11) yields

$$(2.12)\qquad \|K_{2,n}-K_{2,n+k}\|<\left\|(I+A_n)^{-1}-(I+A_nK_{n+1,n+k})^{-1}\right\|\frac{1}{(1+\rho)^{n-2}}.$$

In (2.12) use $\|(I+A_n)^{-1}\|\le a_2$ and $\|(I+A_nK_{n+1,n+k})^{-1}\|=\|E_{n;\,k}^{-1}\|\le a_{k+2}<1/\rho$ to write

$$(2.13)\qquad \|K_{2,n}-K_{2,n+k}\|<\left(\frac{1}{2\rho+\rho}+\frac{1}{\rho}\right)\frac{1}{(1+\rho)^{n-2}}.$$

Since the right-hand side of (2.13) is independent of k and converges to zero as n increases, $\{K_{2,n}\}$ is a Cauchy sequence and Theorem 2.4 is proved.

Remarks. If n were odd in (2.12), $\|(I+A_n)^{-1}-(I+A_nK_{n+1,n+k})^{-1}\|$ would read $\|I-(I+A_{n+1}K_{n+2,n+k})^{-1}\|$ and the upper limit of the product in (2.9) would be $(n-1)/2$. When n is even, (2.13) may be improved by noting that $a_{2k}=k/\rho(k+1+\rho)$, $k\ge 1$, so that from (2.10) and (2.11), for j even

$$\left\|\left(I+K_{2i+1,j}^{-1}A_{2i}^{-1}\right)^{-1}\right\|\le\frac{\left(1+\rho^2a_{j-2i}\right)/(1+\rho)^2}{1-\left(1+\rho^2a_{j-2i}\right)/(1+\rho)^2}=a_{j-2i+2}<\frac{1}{\rho}\left(\frac{j+2-2i}{j+4-2i}\right).$$

Thus when n is even, this inequality allows the multiplication on the right-hand side of (2.13) by $4/(n+2)$.

In the next theorem the matrix continued fractions correspond to ordinary continued fractions of the form $K(1/b_n)$. For the remainder of this section

$$K_1 \equiv \frac{I}{B_1+} \frac{I}{B_2+} \cdots \tag{2.14}$$

and

$$K_{j,n} \equiv \frac{I}{B_j+} \frac{I}{B_{j+1}+} \cdots \frac{I}{+B_n}, \qquad n \geq j.$$

The proof of the theorem uses induction but matrix operations are different from those used in the previous proofs.

THEOREM 2.5. *Let $\{B_n\}$ be a sequence of $n \times n$ matrices satisfying*

$$\|B_{2n-1}^{-1}\| \leq \alpha \quad \textit{and} \quad \|B_{2n}^{-1}\| \leq 1-\alpha, \quad n \geq 1 \quad \textit{where}\ 0<\alpha<1. \tag{2.15}$$

Then the matrix continued fraction in (2.14) *converges.*

Proof. As in the proofs of Theorems 2.3 and 2.4, the existence of the inverses in the backward recurrence algorithm and that the sequence $\{K_{1,n}\}$ of nth approximants of K_1 is a Cauchy sequence are verified.

Using the identities involving $(I+B)^{-1}$ in the proof of Theorem 2.3, formally write

$$\begin{aligned} K_{i,i+k} &= \frac{I}{B_i+} \cdots \frac{I}{+B_{i+k-2}+\left(B_{i+k-1}+B_{i+k}^{-1}\right)^{-1}} \\ &= \frac{I}{B_i+} \cdots \frac{I}{+B_{i+k-2}+\left(I+B_{i+k-1}^{-1}B_{i+k}^{-1}\right)^{-1}B_{i+k-1}^{-1}} \\ &= \left(I+B_i^{-1}\left(\cdots\left(I+B_{i+k-2}^{-1}\left(I+B_{i+k-1}^{-1}B_{i+k}^{-1}\right)^{-1}B_{i+k-1}^{-1}\right)\cdots\right)^{-1}B_i^{-1}\right. \\ &\equiv (I+X_{i,k})^{-1}B_i^{-1} = \left(I+B_i^{-1}(I+X_{i+1,k-1})^{-1}B_{i+1}^{-1}\right)^{-1}B_i^{-1}. \end{aligned} \tag{2.16}$$

It will be shown by induction that all the inverses in (2.16) exist.

By hypothesis, for all i, $\|X_{i,1}\| = \|B_i^{-1}B_{i+1}^{-1}\| \leq \alpha(1-\alpha) \leq \frac{1}{4}$ implies that the innermost inverse in the nest of inverses in (2.16) exist for all $i \geq 1$. Also for $i \geq 1$, $\|X_{i,2}\| = \|B_i^{-1}(I+B_{i+1}^{-1}B_{i+2}^{-1})^{-1}B_{i+1}^{-1}\| \leq \alpha(1-\alpha)/(1-\alpha(1-\alpha)) < 1$ and the next innermost inverse exists. Now suppose that for all $i \geq 1$, $2 \leq k \leq j$, $(I+X_{i,k-1})^{-1}$ exists and

$$\|X_{i,k-1}\| \leq \frac{\alpha(1-\alpha)}{1-} \frac{\alpha(1-\alpha)}{1-} \cdots \frac{\alpha(1-\alpha)}{1}, \tag{2.17}$$

where the right-hand side is $\alpha(1-\alpha)$ times the $(k-1)$th approximant of the ordinary continued fraction $K(a_n/1)$, $a_1=1$, $a_n=-\alpha(1-\alpha)$, $n \geq 2$. Since the maximum of the function $g(\alpha)=\alpha(1-\alpha)$, $0<\alpha<1$, is $\frac{1}{4}$ when $\alpha=\frac{1}{2}$, Worpitski's theorem for ordinary continued fractions (see [12, p. 42]) is applicable to conclude that all the approximants of $K(a_n/1)$ lie in the disk $S=\{Z\colon \|Z-1\|<1\}$ so that $\|X_{i,k-1}\| \leq \alpha(1-\alpha)2 < \frac{1}{2}$. Therefore

$$\|X_{i,k}\| = \left\|B_i^{-1}(I+X_{i+1,k-1})^{-1}B_{i+1}^{-1}\right\| \leq \alpha(1-\alpha)\frac{1}{1-\|X_{i+1,k-1}\|} < \frac{1}{2},$$

implies that $(I+X_{i,k})^{-1}$ exists by Lemma 2.1 and (2.17) holds for $\|X_{i,k}\|$. In particular for i odd and j even, (2.16) and (2.17) together imply

$$\|K_{i,i+p}\| \leq \frac{\alpha}{1-}\,\frac{\alpha(1-\alpha)}{1-}\cdots\frac{\alpha(1-\alpha)}{1}$$

and

$$\|K_{j,j+p}\| \leq \frac{1-\alpha}{1-}\,\frac{\alpha(1-\alpha)}{1-}\cdots\frac{\alpha(1-\alpha)}{1}, \tag{2.18}$$

where the right-hand sides are respectively α and $(1-\alpha)$ times the $(p+1)$th approximant of the continued fraction $K(a_n/1)$ defined earlier.

To show that $\{K_{1,n}\}$ is a Cauchy sequence, write

$$\begin{aligned} K_{1,n}-K_{1,n+k} &= -K_{1,n}(K_{2,n}-K_{2,n+k})K_{1,n+k} \\ &= (-1)^{n-1}K_{1,n}K_{2,n}\cdots K_{n,n}K_{n+1,n+k}K_{n,n+k}\cdots K_{2,n+k}K_{1,n+k}, \end{aligned}$$

where $K_{n,n}=B_n^{-1}$. It is easily shown that the right-hand sides of (2.18) are bounded respectively by α and $(1-\alpha)$ times the $(p+1)$th the approximant of the ordinary continued fraction

$$\frac{1}{1-}\,\frac{1/4}{1-}\,\frac{1/4}{1-}\,\frac{1/4}{1-}\cdots.$$

The nth approximant of this continued fraction is $h_n=2n/(n+1)$, $n\geq 2$ and $h_1=1$. Without loss in generality let n be even and write

(2.19)

$$\begin{aligned} \|K_{1,n}-K_{1,n+k}\| &\leq [\alpha(1-\alpha)]^n\left(\prod_{i=1}^{n} h_{n+1-i}h_{n+k+1-i}\right)h_k \\ &\leq 4^{n-1}[\alpha(1-\alpha)]^n\frac{n+1}{n+2}\,\frac{n}{n+1}\cdot\ldots\cdot\frac{2}{3}\cdot\frac{n+k+1}{n+k+2}\cdots\frac{k+1}{k+2}\,\frac{2k}{k+1}\leq\frac{4\alpha(1-\alpha)}{n+2}. \end{aligned}$$

Since the right-hand side of (2.19) is independent of k and converges to zero as n increases without bound, $\{K_{1,n}\}$ is a Cauchy sequence and Theorem 2.5 is proved.

3. The matrix exponential. The analysis of continued fraction and Padé table expansions of the matrix exponential was an important portion of Varga's paper [11] on parabolic partial differential equations. In that paper the matrix continued fraction expansion of e^A, A an $n\times n$ Hermitian matrix, was given by

$$e^A=\frac{I}{I-}\,\frac{A}{I+}\,\frac{A}{2I-}\,\frac{A}{3I+}\,\frac{A}{2I-}\,\frac{A}{5A+}\cdots. \tag{3.1}$$

To utilize (2.6) in the proof of Theorem 2.3, write

$$e^A=\frac{I}{I+}\,\frac{-A}{I+}\,\frac{(1/2)A}{I+}\,\frac{(-1/2\cdot 3)A}{I+}\,\frac{(-1/2\cdot 5)A}{I+}\cdots. \tag{3.2}$$

Then for $\|A\|\leq\frac{1}{4}$ (A is not necessarily Hermitian),

$$\|e^A-K_{2,n}\|\leq 2/(n+1),$$

where $K_{2,n}$ is the nth approximant of the matrix continued fraction in (3.2). For matrices with larger norm use the following expansion easily derived from (3.1) (also see [1, p. 70]).

$$e^A=I+\frac{A}{I-}\,\frac{A}{2I+}\,\frac{A}{3I-}\,\frac{A}{2I+}\,\frac{A}{5I+}\cdots.$$

If A^{-1} exists, the above expansion can be written as

$$(3.3)\qquad A^{-1}(e^A-I)=\frac{I}{I+}\,\frac{-(1/2)A}{I+}\,\frac{(1/2\cdot 3)A}{I+}\,\frac{(-1/2\cdot 3)A}{I+}\,\frac{(-1/2\cdot 5)A}{I+}\cdots.$$

Thus if $\|A\|\le\frac{1}{2}$ and A^{-1} exist (2.6) in Theorem 2.3 implies

$$(3.4)\qquad \|A^{-1}(e^A-I)-K_{2,n}\|\le\frac{2}{n+1},$$

where $K_{2,n}$ is the nth approximant of the matrix continued fraction in (3.3). Since $\|A\|\|B\|\ge\|AB\|$, multiplying the inequality in (3.4) by $\|A\|$ yields

$$\|e^A-I-AK_{2,n}\|\le 2\frac{\|A\|}{(n+1)}\le\frac{1}{(n+1)},$$

if A^{-1} exists and $\|A\|<\frac{1}{2}$.

Acknowledgment. I am grateful to the referee for bringing to my attention references [2], [6] and [7], and for many valuable suggestions.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, (eds.), *Handbook of Mathematical Functions With Formulas, Graphs and Mathematical Tables*, Nat. Bur. Standards, Appl. Math. Ser., No. 55, Superintendent of Documents, U. S. Government Printing Office, Washington, D.C., 1966.

[2] G. Copp, *Some convergence regions for a continued fraction*, Dissertation, Univ. of Texas, 1950.

[3] W. Fair, *A convergence theorem for noncommutative continued fractions*, J. Approx. Theory, 5 (1972), pp. 74–76.

[4] T. L. Hayden, *Continued fractions in Banach spaces*, Rocky Mtn. J. Math., 4 (1974), pp. 367–369.

[5] W. Leighton and H. S. Wall, *On the transformation and convergence of continued fractions*, Amer. J. Math., 58 (1936), pp. 267–281.

[6] N. Negoescu, *Sur les fractions continues non commutatives*, Proc. Inst. Math. Iasi, (1974), pp. 137–143.

[7] ______, *Convergence theorems on non-commutative continued fractions*, Rev. Anal. Numér. Théorie Approx., 5 (1977), pp. 165–180.

[8] A. Ralston, *A First Course in Numerical Analysis*, McGraw-Hill, New York, 1965.

[9] C. H. Rasmussen, *Oscillation and asymptotic behavior of systems of ordinary differential equations*, Trans. Amer. Math. Soc., 256 (1979), pp. 1–47.

[10] C. H. Rasmussen and G. H. Ratwischer, *Asymptotic approximations, with error estimates, of the scattering matrix for quantal Coulomb excitation by means of a nonlinear (Riccati) matrix differential equation*, J. Math. Physics, 18 (1977), pp. 395–403.

[11] R. S. Varga, *On higher order stable implicit methods for solving parabolic partial differential equations*, J. Math. Phys., 40, 220–231 (1961), pp. 220–231.

[12] H. S. Wall, *Analytic Theory of Continued Fractions*, Van Nostrand, New York, 1948.

[13] ______, *Partially bounded continued fractions*, Proc. Amer. Math. Soc., 7 (1956), pp. 1090–1093.

[14] P. Wynn, *Continued fractions whose coefficients obey a non-commutative law of multiplication*, Arch. Rat. Mech. Anal., 12 (1963), pp. 273–312.

[15] ______, *On some recent developments in the theory and application of continued fractions*, SIAM J. Numer. Anal., 1 (1964), pp. 177–197.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics
017

SOME USEFUL PROPERTIES OF THE HILBERT TRANSFORM*

M. L. GLASSER†

Dedicated to Professor Walter Kohn

Abstract. The Hilbert transform is shown to be invariant under certain rational transformations of the integration variable. Examples are provided to show how this leads to new transform pairs.

The Hilbert transform of an integrable function F is defined by

$$\mathcal{H}\{F(x);y\}=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{F(x)}{x-y}dx, \tag{1}$$

where y is real. (All integrals in this note denote the Cauchy principal value at poles of the integrand.) These transforms occur in a variety of applications and at least one extensive tabulation is available [1]. In general, Hilbert transforms are difficult to evaluate analytically as well as numerically. In this note I present two apparently new relations satisfied by (1) which are remarkable not only for their generality and power in providing new transform pairs, but also for their elementary character.

In what follows $\{a_1,a_2,\cdots,a_n\}$ will denote a sequence of arbitrary positive numbers, and $\{b_1,\cdots,b_n\}$ an increasing sequence of arbitrary real numbers.

THEOREM 1. *For any function F, whose Hilbert transform is defined,*

$$\mathcal{H}\{F(\phi_1(x));y\}=\mathcal{H}\{F(x);\phi_1(y)\}, \tag{2}$$

where

$$\phi_1(x)=x-\sum_{j=1}^{n-1}\frac{a_j}{x-b_j}. \tag{3}$$

To appreciate the scope of (2) it suffices to examine one or two examples. Thus, from [1, Eq. 15.2(11)] with $\phi_1(x)=x-x^{-1}$, we have immediately ($F(x)=x(x^2+a^2)^{-1}$)

$$\mathcal{H}\left\{\frac{x(x^2-1)}{x^4+(a^2-2)x^2+1};y\right\}=\frac{ay^2}{y^4+(a^2-2)y^2+1}, \tag{4}$$

as can be verified by the tedious process of decomposition into partial fractions. From [1, Eq. 15.2(38)] we find the formula, whose derivation from scratch is less clear,

$$\mathcal{H}\{e^{iax}e^{-ia/x};y\}=ie^{iay}e^{-ia/y},\qquad a>0. \tag{5}$$

An interesting corollary is that

$$\mathcal{H}\{F(\phi_1(x));b_k\}=0,\qquad k=1,2,\cdots,n-1.$$

*Received by the editors July 2, 1982, and in revised form April 1, 1983. This work was supported by the National Science Foundation.

†Department of Mathematics and Computer Science, Clarkson College of Technology, Potsdam, New York 13676.

The proof of Theorem 1 is quite simple. A glance at the graph of $u=\phi_1(x)$ and elementary calculus shows that this equation, which we rewrite

$$G(x)/\prod_{j=1}^{n-1}(x-b_j)=0, \tag{6}$$

where

$$G(x)=u\prod_{j=1}^{n-1}(x-b_j)-f(x)=\prod_{j=1}^{n}(x-x_j) \tag{7}$$

and f is a polynomial, has n distinct roots $x=x_j(u)$ which are differentiable functions of u for $-\infty<u<\infty$. We now write

$$I=\int_{-\infty}^{\infty}\frac{F[\phi_1(x)]}{x-y}dx=\int_{-\infty}^{b_1^-}+\int_{b_1^+}^{b_2^-}+\cdots+\int_{b_{n-1}^+}^{\infty}\frac{F(\phi_1(x))}{x-y}dx. \tag{8}$$

In the range $b_{j-1}<x<b_j$ we set $x=x_j(u)$ and obtain

$$I=\int_{-\infty}^{\infty}F(u)\sum_{j=1}^{n}\frac{x_j'(u)}{x_j-y}du. \tag{9}$$

We have assumed that y does not coincide with any b_j. The case where it does can be recovered by an appropriate limit.However,

$$\sum_{j=1}^{n}\frac{x_j'}{x_i-y}=\frac{d}{du}\log G(y)=\frac{1}{u-\phi_1(y)}, \tag{10}$$

which concludes the proof.

By a similar argument we find

THEOREM 2. *For any function F, whose Hilbert transform is defined,*

$$\mathcal{H}\{F(\phi_2(x));y\}=\mathcal{H}\{F(x);0\}-\mathcal{H}\{F(x);\phi_2(y)\}, \tag{11}$$

where

$$\phi_2(x)=\sum_{j=1}^{n}\frac{a_j}{x-b_j}. \tag{12}$$

For example, from [1, 15.2(38)] with $\phi_2(x)=x^{-1}$ we see that

$$\mathcal{H}\{e^{ia/x};y\}=i(1-e^{ia/y})\qquad(a\geq 0) \tag{13}$$

as can be checked directly. A more interesting case is for (12) to be the Mittag–Leffler representation for a suitable meromorphic function. Thus, noting that

$$\begin{aligned} x-\cot(x^{-1})&=2x\sum_{k=1}^{\infty}\frac{1}{(k^2\pi^2x^2-1)} \\ &=\lim_{N\to\infty}\frac{1}{\pi^2}\sum_{k=1}^{N}\frac{1}{k}\left(\frac{1}{x-(k\pi)^{-1}}+\frac{1}{x+(k\pi)^{-1}}\right), \end{aligned} \tag{14}$$

we find the surprising relation valid for continuous F possessing a Hilbert transform

$$\mathcal{H}\{F(x-\cot(x^{-1}));y\}=\mathcal{H}\{F(x);0\}-\mathcal{H}\{F(x);y-\cot(y^{-1})\}. \tag{15}$$

Using [1, 15.2(6)] this yields

$$(16)\qquad \int_{-\infty}^{\infty}\frac{dx}{(x-y)[x+a-\cot(1/x)]}=i\frac{(y-\cot(1/y))}{a(y+a-\cot(1/y))},\qquad \mathrm{Im}\,a>0,$$

whose direct verification is clearly problematical.

These results can be extended to the Stieltjes transform.

Note added in proof. Prof. M. Klamkin has pointed out (private communication) that these results were described, in somewhat different form, by G. Boole (G. Boole, *On the comparison of transcendents*, Phil. Trans. Part III, 1857). Dr. Klamkin and the author are in the process of modernizing and exploring the consequences of Boole's investigations.

REFERENCES

[1] A. Erdelyi et al., *Tables of Integral Transforms*, Vol. II, McGraw-Hill, New York, 1954, Chapter 15.

SIAM J. MATH. ANAL.
Vol. 15, No. 6, November 1984

© 1984 Society for Industrial and Applied Mathematics
018

SOME REMARKS ON ZEROS OF CYLINDER FUNCTIONS*

M. E. MULDOON[†] AND R. SPIGLER[‡]

Abstract. We prove that the functions $Y_\nu(\nu)/J_\nu(\nu)$ and $Y_\nu'(\nu)/J_\nu'(\nu)$ are monotonic for $\nu>0$. It follows that, for $0\le\theta\le 5\pi/6$, $J_\nu(x)\cos\theta-Y_\nu(x)\sin\theta$ has no x-zero on $0<x\le\nu$ and, for $\pi/6\le\theta\le\pi$, $J_\nu'(x)\cos\theta-Y_\nu'(x)\sin\theta$ has no x-zero on $0<x\le\nu$.

1. Introduction. In some investigations of x-zeros of cylinder functions

$$\mathcal{C}_\nu(x,\theta)\equiv J_\nu(x)\cos\theta-Y_\nu(x)\sin\theta,\qquad \nu>0,$$

and of their x-derivatives $\mathcal{C}_\nu'(x,\theta)$, the question arises whether there are such zeros on the interval $0<x\le\nu$. We show here that there are no such zeros of $\mathcal{C}_\nu(x,\theta)$ in case $0\le\theta\le 5\pi/6$ and no such zeros of $\mathcal{C}_\nu'(x,\theta)$ in case $\pi/6\le\theta\le\pi$. Similar assertions are contained implicitly in [4, p. 707, footnote] and in the assertions in [4] concerning $\sigma(n,t)$, but proofs are not given. It is proved by Sturmian methods in [3, Corollary to Theorem 2] and is also a consequence of the monotonic character of $Y_\nu(x)/J_\nu(x)$ as a function of x, that the x-zeros of $C_\nu(x,\theta)$ decrease as θ increases, $0\le\theta<\pi$. Also the first positive zero of $Y_\nu(x)$ exceeds ν [6, p. 487], so there is a $\theta_0(\nu)$ with $\pi/2<\theta_0(\nu)<\pi$ such that $C_\nu(x,\theta)$ has no x-zeros on $0<x\le\nu$ when $0<\theta<\theta_0(\nu)$. (See the concluding remark below). To obtain a result which is independent of ν we need to study the monotonic character of $Y_\nu(\nu)/J_\nu(\nu)$ and its limit as $\nu\to\infty$.

2. Notation and results. For fixed $\nu\ge 0$ and fixed $x>0$, let $\theta_1(x,\nu)$ and $\theta_2(x,\nu)$ be the unique numbers satisfying $0\le\theta_{1,2}(x,\nu)<\pi$ such that $\mathcal{C}_\nu(x,\theta)$ and $\mathcal{C}_\nu'(x,\theta)$, respectively, vanish for $\theta=\theta_1,\theta_2$. The uniqueness and many of the properties of these functions (with a slightly different notation) have been discussed in [5]. We have

$$\theta_1(x,\nu)=\pi/2-\arctan\left[Y_\nu(x)/J_\nu(x)\right],$$

and

$$\theta_2(x,\nu)=\pi/2-\arctan\left[Y_\nu'(x)/J_\nu'(x)\right].$$

Here we prove the following results.

THEOREM 1. *$\theta_1(\nu,\nu)$ decreases from π to $5\pi/6$ as ν increases on* $(0,\infty)$.

THEOREM 2. *$\theta_2(\nu,\nu)$ increases from* 0 *to $\pi/6$ as ν increases on* $(0,\infty)$.

COROLLARY 1. *If $0\le\theta\le 5\pi/6$, there is no x-zero of $\mathcal{C}_\nu(x,\theta)$ on $0<x\le\nu$.*

COROLLARY 2. *If $\pi/6\le\theta\le\pi$, there is no x-zero of $\mathcal{C}_\nu'(x,\theta)$ on $0<x\le\nu$.*

COROLLARY 1′. $\sqrt{3}J_\nu(\nu)+Y_\nu(\nu)<0$, $0<\nu<\infty$.

COROLLARY 2′. $\sqrt{3}J_\nu'(\nu)-Y_\nu'(\nu)<0$, $0<\nu<\infty$.

3. Remarks. The assertion in Theorem 1 about the decrease of $\theta_1(\nu,\nu)$ is essentially given by Watson [6, p. 515] and the "end-point" values follow from [5, p. 70, (1.9)

*Received by the editors May 25, 1982, and in revised form October 5, 1982. This research was supported partially by the Natural Sciences and Engineering Research Council (Canada) and by the Consiglio Nazionale delle Ricerche (Italy).

[†] Department of Mathematics, York University, Downsview, Ontario M3J 1P3, Canada.

[‡] Courant Institute of Mathematical Sciences, New York University, New York, New York 10012 (on leave from the University of Padua).

and (1.10)]. We are able to shorten Watson's proof slightly. Theorem 2 appears to be new. Corollaries 1 and 2 imply the assertions of Olver [4] mentioned in the introduction.

We observe that $\theta_{1,2}(\nu,\nu)$ represent the values of the parameter θ such that $c_{\nu 1}(\theta)=\nu$, $c'_{\nu 1}(\theta)=\nu$, respectively, $c_{\nu 1}(\theta)$ and $c'_{\nu 1}(\theta)$ being the first positive zeros of $\mathcal{C}_\nu(x,\theta)$ and $\mathcal{C}'_\nu(x,\theta)$. They can be evaluated by using the graphs of $J_\nu(\nu)$, $Y_\nu(\nu)$, $J_\nu(\nu+1)$, $Y_\nu(\nu+1)$ plotted in [1, p. 185]. In order to obtain greater precision, they have been tabulated in [5, pp. 81–82] for $0.1\le\nu\le 10$ with step 0.1.

Corollaries 1 and 2 are important in that some results about the x-zeros of $\mathcal{C}_\nu(x,\theta)$ and, more especially, of $\mathcal{C}'_\nu(x,\theta)$ hold, or are proved, only for those x-zeros which exceed $|\nu|$. This is true, for example, of the result given by Watson [6, p. 488] on stationary values of cylinder functions and of Dixon's theorem on the interlacing of zeros of functions $A\mathcal{C}_\nu(x,\theta)+Bx\mathcal{C}'_\nu(x,\theta)$ and $C\mathcal{C}_\nu(x,\theta)+Dx\mathcal{C}'_\nu(x,\theta)$ where $AD\ne BC$ [6, p. 481]; see also the higher monotonicity results given in [2, Thm. 7.2].

4. Proofs. Watson [6, pp. 444–445, 448, 515] proves the monotonicity assertion in Theorem 1 by showing that $Y_\nu(\nu)/J_\nu(\nu)$ increases for $\nu>0$, i.e. that

$$J_\nu(\nu)\,dY_\nu(\nu)/d\nu-Y_\nu(\nu)\,dJ_\nu(\nu)/d\nu>0,\qquad \nu>0. \tag{1}$$

He proves that the left-hand side in (1) is equal to

$$f(\nu)=2/(\pi\nu)-(4/\pi)\int_0^\infty K_0(2\nu\sinh t)e^{-2\nu t}\,dt$$

and then that $f(\nu)>0$, $\nu>0$. His argument can be shortened a little by using the decreasing character of $K_0(x)$ and the elementary inequality $\sinh t>t$, $t>0$, to get

$$f(\nu)>2/(\pi\nu)-(4/\pi)\int_0^\infty K_0(2\nu t)e^{-2\nu t}\,dt=2/(\pi\nu)-2(\pi\nu)^{-1}\int_0^\infty K_0(u)e^{-u}\,du=0$$

on using [6, p. 388, (9)].

The proof of Theorem 2 is, rather surprisingly, easier than that of Theorem 1. We need to show that

$$J'_\nu(\nu)\,dY'_\nu(\nu)/d\nu-Y'_\nu(\nu)\,dJ'_\nu(\nu)/d\nu<0. \tag{2}$$

Now the left-hand side of (2) is equal to

$$\left[J'_\nu(x)\partial Y'_\nu(x)/\partial\nu-Y'_\nu(x)\partial J'_\nu(x)/\partial\nu\right]_{x=\nu}+\left[J'_\nu(x)Y''_\nu(x)-Y'_\nu(x)J''_\nu(x)\right]_{x=\nu}.$$

The second term here is seen to be 0 on using [6, p. 76, (6)] or the Bessel equation directly. On using [6, p. 445, (3)] we see that the first term is

$$-(4/\pi)\int_0^\infty(\cosh 2t-1)K_0(2\nu\sinh t)e^{-2\nu t}\,dt$$

which is clearly negative. This completes the proof of Theorem 2.

Now suppose that $0<\theta\le 5\pi/6$ and that $\mathcal{C}_\nu(x,\theta)$ has an x-zero at x_0, $0<x_0\le\nu$. From [5], there is at most one $\theta=\theta_1(x_0,\nu)$ for which this can hold. Since $\theta_1(x,\nu)$ is a decreasing function of x, $0<x<\nu$, we have

$$\theta_1(x_0,\nu)\ge\theta_1(\nu,\nu)>5\pi/6,$$

the last inequality coming from Theorem 1. This is a contradiction so Corollary 1 is proved. The case $\theta=0$ is known already [6, p. 485].

Corollary 2 is proved in a similar way using Theorem 2.

Now $\mathcal{C}_\nu(x, 5\pi/6)$ is positive when $x(>0)$ is close to 0. Corollary 1′ is simply a statement of the fact that it is still positive when $x=\nu$. (There is no x-zero on $(0,\nu]$, by Corollary 1.) On the other hand $\mathcal{C}_\nu'(x,\pi/6)$ is negative when $x(>0)$ is close to 0. Corollary 2′ says that it is still negative when $x=\nu$, a fact which follows from Corollary 2.

5. Concluding remark. For a *fixed* ν, $\nu>0$, $5\pi/6$ can be replaced by $\theta_1(\nu,\nu)-\varepsilon$ in Corollary 1 and $\pi/6$ can be replaced by $\theta_2(\nu,\nu)+\varepsilon$ in Corollary 2 where ε is an arbitrarily small positive number. This permits us to complement Corollaries 1 and 2 as follows:

For each fixed ν, $\nu>0$, $\mathcal{C}_\nu(x,\theta)$ *has exactly one x-zero in* $(0,\nu]$ *if and only if* $\theta_1(\nu,\nu)\le\theta<\pi$ *and* $\mathcal{C}_\nu'(x,\theta)$ *has exactly one x-zero in* $(0,\nu]$ *if and only if* $0<\theta\le\theta_2(\nu,\nu)$.

Acknowledgment. The authors wish to thank Professor F. W. J. Olver for his interest and for several helpful comments.

REFERENCES

[1] E. Jahnke, F. Emde and F. Lösch, *Tables of Higher Functions*, McGraw-Hill, New York, 1960.

[2] L. Lorch, M. E. Muldoon and P. Szegö, *Higher monc‘ πicity properties of certain Sturm–Liouville functions*, IV, Canad. J. Math., 24 (1972), pp. 349–368.

[3] L. Lorch and D. J. Newman, *A supplement to the Sturm separation theorem, with applications*, Amer. Math. Monthly 72 (1965), pp. 359–366; 980.

[4] F. W. J. Olver, *A further method for the evaluation of zeros of Bessel functions and some new asymptotic expansions for zeros of functions of large order*, Proc. Cambridge Philos. Soc., 47 (1951), pp. 699–712.

[5] R. Spigler, *Alcuni risultati sugli zeri delle funzioni cilindriche e delle loro derivate*, Rend. Sem. Mat. Univ. Politec. Torino, 38 (1980), pp. 67–85.

[6] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, London, 1944.

© 1984 Society for Industrial and Applied Mathematics
019

A NODAL LINE THEOREM FOR THE SLOSHING PROBLEM*

J. R. KUTTLER†

Abstract. We show that the nodal lines of the *n*th sloshing mode do not divide the region into more than *n* nodal domains. As a corollary, the first nonzero eignevalue is shown to be simple.

The sloshing problem is the eigenvalue problem

$$\begin{aligned} \Delta u &= 0 \quad \text{in } R\ , \\ \frac{\partial u}{\partial n} &= \lambda u \quad \text{on } T\ , \qquad (1) \\ \frac{\partial u}{\partial n} &= 0 \quad \text{on } S\ . \end{aligned}$$

Here u may represent the potential of an incompressible fluid in a container with rigid walls and horizontal free surface [1], [3, Chap. IX]. We consider the two-dimensional problem where R is the cross-section of a uniform tank or canal. We take T to be a finite interval on the x-axis and S the rest of the boundary of R (Fig. 1). We assume that S is smooth and intersects T with nonzero interior angles.

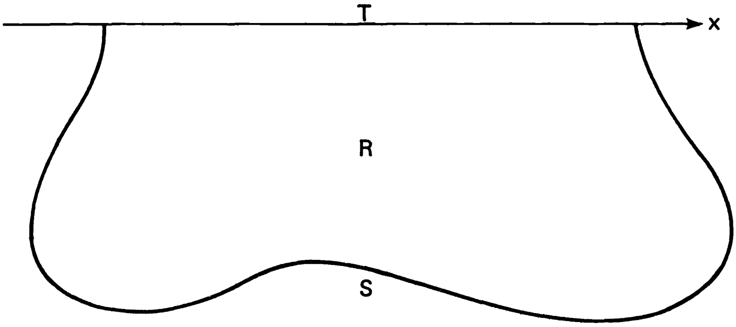


FIG. 1. *The sloshing region.*

If the eigenvalues of (1) are $\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots$ with associated eigenfunctions u_1, u_2, u_3, $\cdots$, they can be characterized by a minimum principle

$$\lambda_n = \min \frac{\int_R (u_x^2 + u_y^2)\, dx\, dy}{\int_T u^2\, dx}\ , \qquad (2)$$

*Received by the editors January 25, 1983, and in revised form June 1, 1983. This research was supported by the Naval Sea Systems Command, U. S. Department of the Navy, under contract N00024-83-C-5301. This paper was typeset at The Johns Hopkins University Applied Physics Laboratory.

†Milton S. Eisenhower Research Center, The Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland 20707.

where the minimum is over all continuous and piecewise differentiable functions on R satisfying

$$\int_T uu_i\,dx = 0\ , \qquad i = 1, 2, \cdots, n-1. \tag{3}$$

A function minimizing (2) subject to (3) is an eigenfunction u_n of (1) associated with λ_n.

The first eigenvalue λ_1 is zero with constant associated eigenfunction. All other eigenvalues are positive, and, because of (3), their eigenfunctions necessarily change sign on T. We are interested in the character of *nodal lines* where $u_n = 0$.

Note that a nodal line of u_n must have one end on T, because a nodal line that is either closed or has both ends on S encloses a subregion on which the function u_n is harmonic and vanishes on part of the boundary while its normal derivative vanishes on the rest of the boundary. Such a function is necessarily zero on the subregion. By the *unique continuation property* of harmonic functions ([2, p. 259]) this would force u_n to vanish on all of R.

LEMMA. *Nodal lines of u_n have one end on T and one end on S.*

Proof. Suppose to the contrary that a nodal line of u_n has both ends on T without intersecting S (Fig. 2). Let R' be the subregion enclosed by this nodal line. Define the function ϕ_1 by

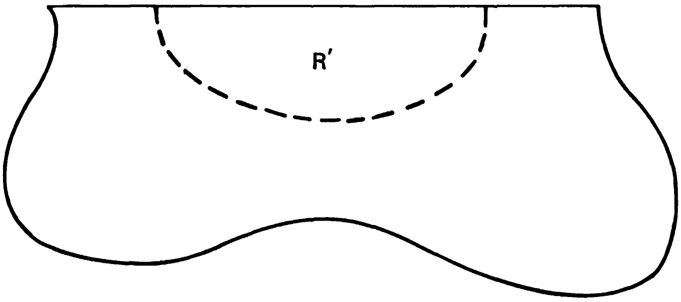


FIG. 2. *A nodal line with both ends on T.*

$$\phi_1 = \begin{cases} u_n & \text{in } R' \\ 0 & \text{outside } R'\ . \end{cases} \tag{a}$$

Let ϕ_i, $i = 2, \cdots, n$, be translates of ϕ in the x-direction by distinct values $t_2, \cdots, t_n$, i.e.,

$$\phi_i(x,y) = \phi_1(x + t_i, y)\ . \tag{b}$$

For sufficiently small t_i, the functions ϕ_i all have their support in R and are continuous and piecewise differentiable. Since they are linearly independent on T, there is a linear combination

$$\Phi = \sum_{i=1}^{n} a_i\phi_i \tag{c}$$

that will satisfy the constraints (3). Since

$$\frac{\partial \Phi}{\partial n} = \lambda_n \Phi \quad \text{on } T \, , \tag{d}$$

it follows that Φ is a function minimizing (2), and is therefore an eigenfunction, hence is harmonic on R. But Φ vanishes on a subregion of R, hence by unique continuation vanishes identically, giving a contradiction.

We now state and prove the main result.

THEOREM. *An eigenfunction u_n associated with λ_n cannot change signs more than $n - 1$ times on T. Its nodal lines divide R into no more than n subdomains.*

Proof. Each point of sign change on T is an endpoint of a nodal line which we have seen must have its other end on S. Moreover, these nodal lines cannot cross, else there is a nodal line with both ends on S. Thus, the nodal lines of u_n divide R into a number of subregions each having a portion of T as part of its boundary (Fig. 3). We must show there are no more than n such subregions. Suppose to the contrary that there are $N > n$. Call them $R_1, \cdots, R_N$ and define

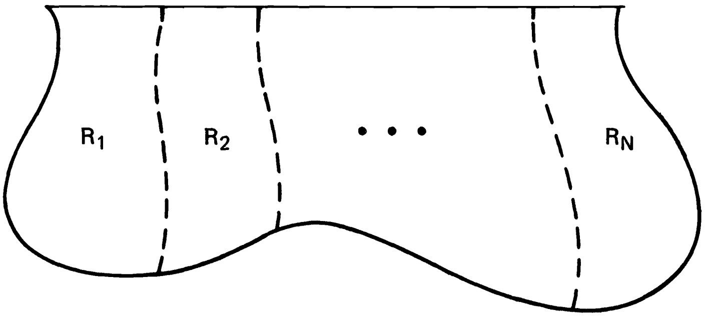


FIG. 3. *Nodal domains of u_n.*

$$\psi_i = \begin{cases} u_n & \text{in } R_i \\ 0 & \text{outside } R_i \, . \end{cases} \tag{e}$$

The ψ_i are continuous and piecewise differentiable on R, and linearly independent on T. There is a linear combination of n of them

$$\Psi = \sum_{i=1}^{n} b_i \psi_i$$

that will satisfy the constraints (3). Since

$$\frac{\partial \Psi}{\partial n} = \lambda_n \Psi \quad \text{on } T \, ,$$

we conclude as in the proof of the lemma that Ψ minimizes (2), is thus an eigenfunction, hence is harmonic on R, but vanishes on a subregion of R, and so vanishes identically, a contradiction.

This proof is modeled on the proof of the famous Courant nodal line theorem for fixed membranes (see, e.g., [4]) \Ve also obtain the following.

COROLLARY. *The first nonzero eigenvalue* λ_2 *is simple.*

Proof. By the theorem, an eigenfunction associated with λ_2 can change signs at most once on T. Since it must change signs at least once by (3), it thus changes sign exactly once. Now, suppose to the contrary that λ_2 is not simple and has two associated eigenfunctions u and v, which are linearly independent on T, so we may assume

$$\int_T uv\,dx = 0 . \tag{4}$$

Now, u and v each have one point on T at which they change sign, and by (4) these points cannot coincide. Thus, without loss of generality, we may suppose that on T

$$\begin{aligned} u(x) &< 0 \quad \text{for } x < x_1 , \\ u(x) &> 0 \quad \text{for } x > x_1 , \\ v(x) &> 0 \quad \text{for } x < x_2 , \\ v(x) &< 0 \quad \text{for } x > x_2 , \end{aligned}$$

and $x_1 < x_2$. Now, consider the linear combination

$$f_t(x) = (1-t)u(x) + tv(x), \qquad 0 \le t \le 1 .$$

Notice that for all such t

$$f_t(x) > 0 \quad \text{for } x_1 < x < x_2 .$$

Now, consider the sets

$$\begin{aligned} T_1 &= \{t : f_t(x) < 0 \ \text{for some } x < x_1\} , \\ T_2 &= \{t : f_t(x) < 0 \ \text{for some } x > x_2\} . \end{aligned}$$

Clearly, T_1 is a half-open interval $[0,t_1)$ and T_2 is a half-open interval $(t_2,1]$. There are two possibilities. If $t_2 < t_1$,. then for $t_2 < t < t_1$, $f_t(x)$ changes sign at least twice on T. If $t_1 \le t_2$, then for $t_1 \le t \le t_2$, $f_t(x) \ge 0$ on T. But as $f_t(x)$ is an eigenfunction associated with λ_2, these are both contradictions.

It is an open question whether or not u_n changes signs exactly $n-1$ times on T and all the eigenvalues are simple.

REFERENCES

[1] D. W. FOX AND J. R. KUTTLER, *Sloshing frequencies,* Z. Angew. Math. Phys., 34 (1983), pp. 668-696.

[2] O. D. KELLOG, *Foundations of Potential Theory,* Dover, New York, 1953.

[3] H. LAMB, *Hydrodynamics,* Dover, New York, 1945.

[4] A. PLEIJEL, *Remarks on Courant's nodal line theorem,* Comm. Pure Appl. Math., 9 (1956), pp. 543-550.